eKNOW 2013

The Fifth International Conference on Information, Process, and Knowledge Management

ISBN: 978-1-61208-254-7

February 24 - March 1, 2013

Nice, France

eKNOW 2013 Editors

Dirk Malzahn, OrgaTech GmbH, Germany

eKNOW 2013

Forward

The fifth edition of the International Conference on Information, Process, and Knowledge Management (eKNOW 2013) was held in Nice, France, February 24 - March 1, 2013. The event was driven by the complexity of the current systems, the diversity of the data, and the challenges for mental representation and understanding of environmental structure and behavior.

Capturing, representing, and manipulating knowledge was and still is a fascinating and extremely useful challenge from both theoretical and practical perspective. Using validated knowledge for information and process management and for decision support mechanisms raised a series of questions the eKNOW 2013 conference was aimed at.

eKNOW 2013 provided a forum where researchers were able to present recent research results and new research problems and directions related to them. The topics covered aspects from knowledge fundamentals to more specialized topics such as process analysis and modeling, management systems, semantics processing and ontology.

We take this opportunity to thank all the members of the eKNOW 2013 Technical Program Committee as well as the numerous reviewers. The creation of such a broad and high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to the eKNOW 2013. We truly believe that, thanks to all these efforts, the final conference program consists of top quality contributions.

This event could also not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the eKNOW 20103 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that eKNOW 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in knowledge management research.

We also hope that Côte d'Azur provided a pleasant environment during the conference and everyone saved some time for exploring the Mediterranean Coast.

eKNOW 2013 Chairs

Dirk Malzahn, OrgaTech GmbH, Germany
Stephen White, University of Huddersfield, UK

eKNOW 2013

Committee

eKNOW 2013 Technical Program Committee

Gil Ad Ariely, California State University (CSU), USA / Interdisciplinary Center (IDC) – Herzliya, Israel
Werner Aigner, Institute for Application Oriented Knowledge Processing – FAW / University of Linz, Austria
Panos Alexopoulos, iSOCO, Spain
Amin Anjomshoaa, Vienna University of Technology, Austria
Zbigniew Banaszak, Warsaw University of Technology, Poland
Ladjel Bellatreche, LISI- ENSMA/ Poitiers University, France
Peter Bellström, Karlstad University, Sweden
Jorge Bernardino, Polytechnic Institute of Coimbra, Portugal
Yaxin Bi, University of Ulster - Jordanstown, UK
Grzegorz Bocewicz, Koszalin University of Technology, Poland
Sabine Bruaux, Picardie Jules Verne University, France
Martine Cadot, University of Nancy1, France
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong
Marco Cococcioni, University of Pisa, Italy
Susan Gauch, University of Arkansas, USA
Olivier Gendreau, École Polytechnique de Montréal, Canada
Conceição Granja, Universidade do Porto, Portugal
Pierre Hadaya, ESG UQAM, Canada
Richard Hussey, University of Reading, UK
Khaled Khelif, EADS- Val de Reuil, France
Daniel Kimmig, Karlsruhe Institute of Technology (KIT), Germany
Marite Kirikova, Riga Technical University, Latvia
Agnes Koschmider, KIT, Germany
Andrew Kusiak, The University of Iowa, USA
Szymon Łazaruk, Poznan University of Economics, Poland
Franz Lehner, University of Passau, Germany
Chee-Peng Lim, Deakin University, Australia
Matthias Loskyll, German Research Center for Artificial Intelligence (DFKI), Germany
Dickson Lukose, MIMOS-Berhad, Malaysia
Hiep Luong, University of Arkansas, USA
Dirk Malzahn, OrgaTech GmbH, Germany
Luis Martínez López, University of Jaén, Spain
Marco Mevius, HTWG Konstanz, Germany
Roy Oberhauser, Aalen University, Germany
Daniel O'Leary, University of Southern California, USA
Andreas Papasalouros, University of the Aegean - Samos, Greece
Tuan D. Pham, The University of Aizu - Aizu-Wakamatsu, Japan

Lukasz Radlinski, University of Szczecin, Poland
Ulrich Reimer, University of Applied Sciences St. Gallen, Switzerland
Erwin Schaumlechner, Tiscover GmbH - Hagenberg, Austria
Jan Sefranek, Comenius University, Bratislava, Slovakia
Pnina Soffer, University of Haifa, Israel
Lubomir Stancev, Indiana University - Purdue University Fort Wayne,USA
Carlo Tasso, Università di Udine, Italy
Jan Martijn van der Werf, Technische Universiteit Eindhoven, The Netherlands
Martin Voigt, Technische Universität Dresden, Germany
Shengli Wu, University of Ulster - Newtownabbey, Northern Ireland, UK
Takahira Yamaguchi, Keio University, Japan

Copyright Information

Table of Contents

Knowledge Management and Business Processes Learning on the Job

A Conceptual Approach and its Prototypical Implementation

Julian Krumeich, Dirk Werth and Peter Loos
Institute for Information Systems (IWi) at the
German Research Center for Artificial Intelligence (DFKI)
Saarbruecken, Germany
{firstname.lastname}@dfki.de

***Abstract*—Business Process Management (BPM) has established itself as an important cross-functional task in many companies. Its primary goal is to optimize the business process design and hence the actual execution of business processes. Since optimizing processes on paper is not sufficient to really boost a company's performance, it is an essential task to optimize the process execution that defines how business processes are actually performed at the end of the day. However, before employees are able to carry out processes, they need a given up-front learning time. Hence, it seems promising to research how business process learning can be realized on-the-job in order to reduce this up-front learning time; thus, being able to work efficiently on processes already from the very beginning. In this paper, we present an approach towards business process learning on-the-job using the concepts of task guidance and process guidance. After introducing the approach, the paper presents a prototypical implementation of it and in doing so proves its feasibility. Afterwards, the paper outlines a promising use case that is going to be supported by our approach after future research.**

Keywords-Business Process Management; Business Process Learning; Action Learning; Process Knowledge; Design Science.

I. Motivation and Contribution

Business Process Management (BPM) has established itself as an important cross-functional task in many companies [1]. Especially in the field of process modeling, a lot of effort is done. The motivation for this is obvious: a strict documentation of business processes fosters the ability to optimize them starting from their modeled as-is state and ending at the optimized to-be state [2]. However, optimizing processes on paper does not really boost a company's performance. Hence, one important facet in the course of BPM is the process execution that defines how business processes are actually performed at the end of the day.

Since most business processes are not performed fully IT-based, human beings do often play a central role in their execution. However, in contrast to IT systems, individuals rely on learning as a basis for knowing how to carry out specific processes. Thus, before persons are able to perform them, there is always a given up-front learning time needed. In addition, not only these high startup costs for being able to perform processes for the first time, but also the risk of wrongly conducting activities are at a high level at the beginning of gaining experience in processes.

During the execution phase of business processes, process guidance has been shown as useful in practice [3]. Hence, the guidance concept could be a way out of the previous described shortcoming. However, in research less effort is put into the question whether process guidance can foster employees in working on business processes that are unfamiliar to them, i.e. without having to learn these processes basically beforehand in a time-consuming and less productive way. Additionally, the same applies when significant changes have been enacted in processes which employees are already familiar with. Since working and business environment has come to ever shorter life cycles, the frequency of changes and hence the need for an efficient change-management including the training and learning becomes more and more important for doing successful business [4]. Consequently, it is often a heavy and time-consuming task for employees to learn unfamiliar processes or adaptations of common ones.

Hence, it seems promising to examine the concept of guidance as an approach to learn business processes on-the-job. Thus, this paper presents an approach demonstrating how task and process guidance can be used to help employees in learning unfamiliar business processes and changes within existing ones with which they are already familiar. To minimize the up-front learning effort, the learning procedure will be implemented into the execution phase of business processes.

The contribution of the paper is two-fold. First of all, the paper presents a conceptual approach that uses the concept of guidance with regard to learn business processes. In addition to that, the paper will further present an implementation of this theoretical concept; hence, following the design science research paradigm [5]. In doing so, the remainder of this paper is structured as follows: Section 2 examines related work in the field of conducting and learning business processes. Afterwards, Section 3 forms the foundation of this paper by introducing the developed concept. Within Section 4, a prototypical implementation of the concept will be presented. In the subsequent Section 5, an additional use case will be outlined that will be supported by the developed approach respectively prototype in the future. Finally, the paper closes with a conclusion and outlook.

ISBN: 978-1-61208-254-7

II. Related Work

One approach heading pretty close towards the direction followed by this paper is the one proposed by Hawryszkiewycz [6]. He aims at integrating reusable *learning components* into business processes in order to allow employees to quickly acquire knowledge in their working context. One of his primary goals is to integrate learning resources within the actual workspace of employees. However, his approach just includes "a link to the learning systems from selected screens in the work process". However, even though employees can start a learning unit by clicking on a button and hence the approach partly integrates business process learning into the actual process execution, employees will not improve or even build up their knowledge *on-the-job* meaning based on actually conducting business processes. In this context, *workplace learning* has evolved in literature describing the significant relationship between working and learning [4]. According to Chen and Kao [7], workplace learning summarizes activities and processes in the workplace by which employees acquire knowledge ranging from basic skills to high qualifications, which they can straightaway use in their job. As stated before, one dimension of workplace learning is on-the-job learning [8].

Apart from the concept of workplace learning, also relevant for the approach to be developed are the so-called learning workflows and the adaptation of workflows, to which lots of work has been done in literature. The goal is to *dynamically and flexibly apply changes into workflow systems*. In doing so, unwanted side effects of complex changes in workflows are avoided since it is very inefficient and often impossible to stop running activities or workflows in order to enact changes. A recent work done by Weber et al. [9] presents a detailed review of challenges and techniques that exist in continuously managing the lifecycle of dynamic processes respectively workflows. As an example, one approach towards this direction is proposed by Dadam et al. [10]. With their ADEPT2 Process Management System, they aim at achieving a quick implementation and deployment of new business processes in order to enable ad-hoc changes of running processes on the fly. To have a broad overview on these aspects, it is also referred to the recent state-of-the-art analysis provided by Burkhart and Loos [11].

While the previous research stream primarily focuses on technical issues regarding how to dynamically enact changes into running business processes, other work explores ways how business processes can be improved by learning from their actual business context. A recent state-of-the-art analysis can be found in Ploesser et al. [12]. They state that *context-awareness* in BPM is a current and future challenges in process management in order to achieve true agility and flexibility. In this context, work dealing with learning how to *improve business processes* can also be regarded [13]. This learning process is also considered as an evolutionary process like BPM and hence must be managed as other business processes are managed in organizations.

III. Business Process Learning on the Job

Within the previous section on related work, some general concepts have been presented and outlined forming the basis of our approach:

- Firstly, business process learning will be intervened with the actual working on processes, i.e. the learning effect results from the process execution. This is basically what is understood as business process learning on-the-job.
- Secondly, the actual learning will be realized by using reusable learning components that are integrated into single process steps. In using the inherent knowledge of these components, users are assisted via task guidance (based on context-oriented knowledge) and process guidance (based on process flow-oriented knowledge).
- Thirdly, this guidance will be realized by recommendations that depend on the underlying business process models and is aware of the context-situation in which the business processes are conducted. This is achieved by continuously monitoring the users' behavior and adapting the process models based on their actual process execution for the individual user and on the other hand for all users within an organization.
- Fourthly, these process model adaptations and optimizations will be dynamically applied into the workflow system. In doing so, the approach aims at overcoming the trade-off between guidance (recommendations) and being flexible and adaptable to support ad-hoc processes (adaptation mechanism) (see Burkhart and Loos [11] for more details).

According to Abecker et al. [14], Allweyer [15] and Lehner [16] *business processes* and *knowledge processes* have to be considered as intervened concepts during their execution (see Figure 1, (1) and (2)). Since the major objective is to learn knowledge linked to the underlying business processes, we need to have a closer look at what process knowledge actually mean. In this regard, Remus [17] distinguishes between two kinds of process knowledge: *process flow-oriented knowledge* and *content-oriented knowledge* in business processes.

Besides combining business processes and knowledge processes there is also the need for including the *training and learning processes* into the overall process design [15] (see Figure 1, (1) and (3)). This means that employees have to learn both types of knowledge in order to know how to perform a business process. To realize this, the approach builds upon two basic concepts of learning that are considered as promising for process learning on-the-job: *task guidance* and *process guidance* (see Figure 1, (b)). While task guidance helps employees in conducting the current active process step (via content-oriented knowledge), process guidance (via process-oriented knowledge) assists them in questions regarding how to proceed in the business process. Based on this guidance during process execution, the concept follows a passive, structured on-the-job learning methodology, which is also called action learning [4].

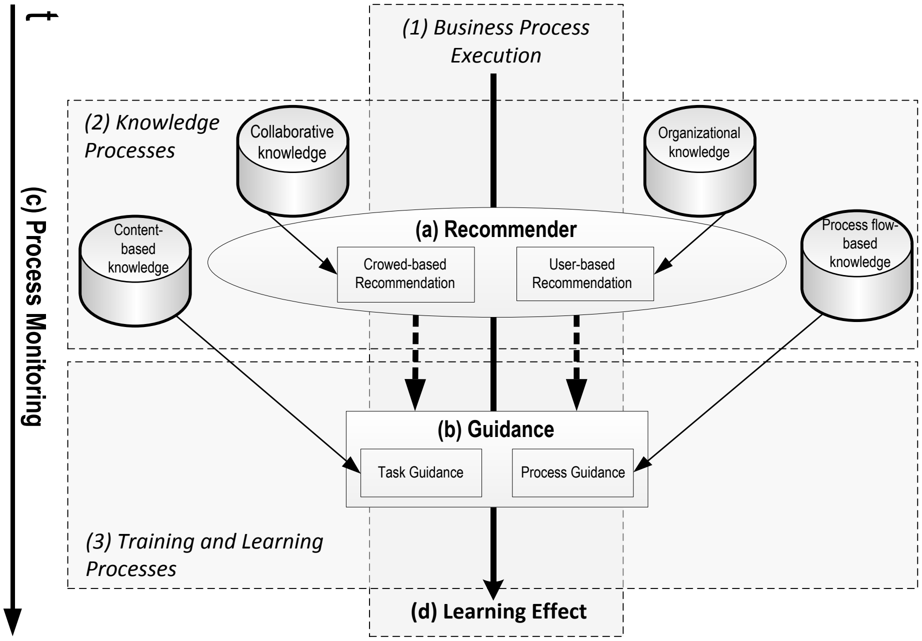


Figure 1. Approach towards Learning Business Processes on the Job

This means "Learning occurs at the actual work setting as a result of using a systems approach, and with limited involvement of a trainer/facilitator". In this regard the adjective "structured" does not describe the underlying business processes (cf. the last bullet point), but outlines the usage of a system that actually helps to learn.

To realized task and process guidance, the concept makes use of recommendations (see Figure 1, (a)) based on the underlying business process models. In this regard, business process models are defined as a combination of single process steps, each of them including task guidance components based on the underlying content-based knowledge, which are considered as learning components. In monitoring users in their work on a continuous manner (see Figure 1, (c)), their behavior can be assigned to a specific process step within a process model. Hence, based on underlying process models, the approach recommends further process steps to the users in order to successfully accomplish the process execution.

However, these process step recommendations are not solely based on the standard underlying process model, but can additionally be crowed-based or user-based. It is distinguished between crowed-based and user-based recommendations since each individual user exhibits a very personalized process that may deviate from the standard business process—as far as it is in line with a company's compliance rules. Of course, user-based recommendations are preferably applied when a user has already acquainted knowledge on the process and not at the beginning when working on a process for the first time. Nonetheless, pure personalized recommendations could reinforce inefficient or even incorrect sequences, such as inadvertently skipping important process steps. Crowd-based recommendations mitigate this shortcoming. In doing so, crowd-based recommendations enrich the set of possible process paths through aggregation of the process experiences from multiple users; hence, it builds upon the knowledge of many. After a given amount of deviations from the standard process model, it will be adapted to the new business situation.

Thus, there is a continuous learning effect resulting in enhanced personal (user learns how processes are conducted) and organizational (business process models adapt to new context situations) knowledge (see Figure 1, (d)).

In doing so, this recommendation methodology (see [18] for more technical details) contributes to the knowledge process that is linked to the business processes. Hence, collaborative knowledge is build up and based on optimizing the process models, the organization knowledge will be enhanced as well. Furthermore, changes can be automatically enacted into (running) business processes as one of the approach's preconditions. Through the task and process guidance, users will be able to learn the process execution on-the-job; hence, they experience a learning effect (see Figure 1, (d)) which reflects the combination of business process execution and the training and learning processes.

IV. Prototypical Implementation of the Concept

Having introduced the theoretical concept in the previous section, this section will show a prototypical implementation illustrating the feasibility of the learning approach in one particular context. This will ease the comprehensibility of the introduced concept and furthermore proves its feasibility to be realized.

In doing so, we put a particular emphasis on email-based processes since email communication has generally become an integral part of daily business activities within companies at any size. On average, employees spend 2.6 hours a day with sending and receiving 33 respectively 72 emails

[19] [20]. Furthermore, not only the time spent with emails as a means of communication, but also the knowledge that is bundled without structure in companies' email repositories is very difficult to manage. This becomes clear, if the number of 75 % is taken into mind representing the percentage of a company's knowledge saved in email messages [21]. As a direct consequence, if employees spend 1/3 of their time with email communication and 3/4 of a company's knowledge is stored in email inboxes, it can be concluded that in various companies a majority of business processes take place via email communication.

Hence, it is promising to ground the approach for process learning in the context of email-based processes. Since emails are a very flexible and ad-hoc means of communication we also address one of our goals, namely supporting flexible and ad-hoc processes.

A. Introducing the Underlying Three-Layer Approach

The developed COPA system, which implements the approach, automatically hooks onto the existing email infrastructure and collaboration systems, e.g., Microsoft Exchange, and assigns incoming emails based on their semantic content to new or running business processes.

This technique allows it to provide users task guidance helping them in conducting the currently active process step within an underlying business process. Furthermore, users will receive process guidance, so that they know how to conduct the following steps within the assigned business process. To explain how the task guidance and process guidance is realized, we initially introduce three layers on which the approach and hence the prototype is based:

- On the level of the system layer, each received email will be intercepted by the system and subsequently be analyzed, archived, decoded and decomposed. Each part of an email, i.e. headers, body or attachments, will be transformed into plain text and merged into a single XML document to allow the other layers to directly access the information for further processing. In addition, the system layer will provide system connectors usable to interface external as well as legacy systems, required to be accessible throughout a task.
- The semantic layer signifies meaningful communication of an enterprise. Outgoing from pattern based information extraction—using e.g., regular expressions—business process and specific process steps within them can be identified and relevant information in this regard will be extracted.
- From a task guidance and process guidance point of view, the process layer is the most important one, since it contributes to the actual process learning on-the-job. The layer is further subdivided into one process build-time (configuration) component and four process run-time components, all of which are described in the following subsection.

B. The Process Layer for Task and Process Guidance

In the following, we take order processes as an example to apply the concept and its implementation to a concrete problem domain. Based on an initial set of business process models stored in an Enterprise Process Repository (EPR), the system can be employed. Therefore, it intercepts the incoming and outgoing email traffic and passes it through the three layers described in the previous section. Figure 2 shows the actual output of a processed email message. Subsequently, the four run-time components are presented and explained by making use of the figure.

Process Detection. The detection component uses the EPR to determine whether an incoming email relates to an already running business process or whether a new process instance has to be initiated. In more detail, based on a semantic analysis performed in the prior semantic layer, the email can either be assigned to an existing process—where it constitutes the next step—or the email is considered as a starting event and triggers a new process. In this case, a new process instance with its specific process ID (see Figure 2, F) will be created outgoing from the corresponding reference model template from the EPR.

Further, the information whether the incoming email is part of an already instantiated process or a completely new one, is being displayed to the user (see Figure 2, F). Future incoming emails concerning this particular business process will be assigned to this process instance henceforth. As mentioned before, the correct assignment of the current process step to the correct template is being realized by an analysis of process characteristics done by the semantic layer. If the detection component assigns an incoming email to a wrong process (step) based on an incorrect semantic analysis, the user still has the possibility to manually reassign the email to another process step (see Figure 2, H). To assist the user, the system provides information about the semantic matching of the email to a process step.

Another feature of the system, which relates to the automatic acquisition of business processes or at least their adaptation to new circumstances, is the following procedure: if the common business process sequence is “A -> B -> C” and after conducting process step A, an incoming email relates to process step C or a user initiates via sending an email process step C and this process adaptation is done a given amount of times then the system automatically adapts the underlying business process in the EPR since this adaption in the process flow might result from a changed business circumstance. In this regard, process models within the EPR can relate to the overall enterprise or just to a subset of employees. This means that such process deviations can on the one hand only affect the personal respectively subset related process template or on the other hand affects the underlying process template of the whole enterprise. How this technique is conceptually realized can be seen in Burkhart et al. [18].

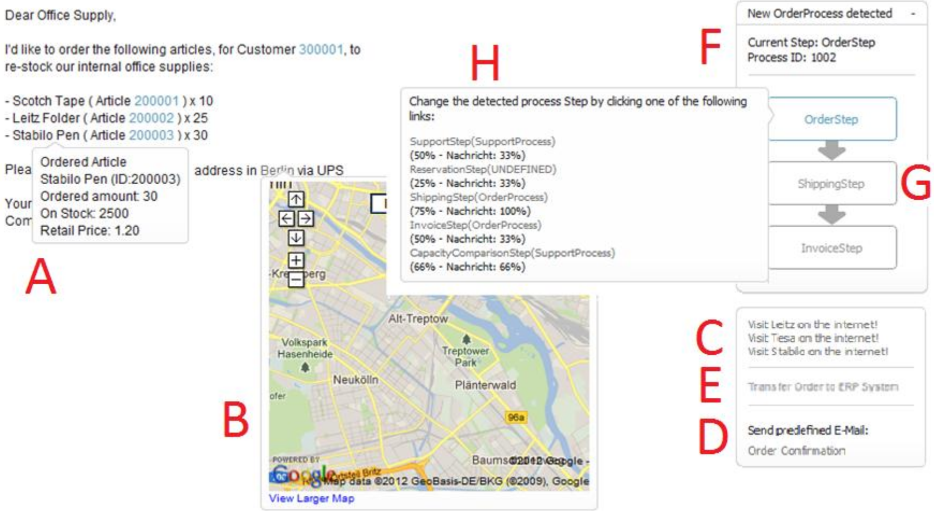
Process Tracking. As the second step along the process layer's execution, the tracking component monitors all incidents occurring within a running process and stores every performed step in context of the related process. This component utilizes the EPR as well as the semantic information gathered from the original incoming email, to track which process is triggered by this email. Additionally, it updates the assigned process instance within the EPR with all important data that can be useful or applicable for future process analysis. Each performed step concerns two occurrences, actions and events. Actions signify human or application triggered activities, whereas events have no active part. In the context of email communication, actions mainly correspond to the activity of sending an email and events to incoming emails. Since every performed step is related to its unique process instance, it can be tracked and on this basis recommendations for further steps can be obtained and provided to the user (see Figure 2, G).

In case the system is applied in a collaborative scenario, it may be possible that the incoming email belongs to an overall process, whose previous steps have been executed by other instances. In this case, the tracking component offers a synchronization functionality, which offers the possibility of synchronizing already executed steps of an overall process throughout several instances. Hereby, the tracking component determines which information has to be gathered from other known instances. Thus, collected information will subsequently be added to the local database and utilized for further enhancement of the generated output. At this point, other beneficial aspects of the tracking component reveal.

The gathered information provides a comprehensible documentation for further disposal. Due to the semantic extraction of process information, e.g., customer information and quantity of ordered goods, the system enables a (mostly) automated build-up of a company unique customer database. The email contains two sets of data informing the user of the present state of the current process, as well as the visualization of the preceding process steps. The first data set contains key information about the email and the process at hand and informs the user about the present status of the process instance the email belongs to (see Figure 2, G). The second data set shows an overview over all preceding steps in the current process including the corresponding emails.

Task Guidance. As the correct process step has already been identified by the detection component and the semantic layer, the task guiding functionality is now deployed in two ways. First, the task guiding component exploits the EPR in order to gather relevant process data. Secondly, the task guiding functionality supplies the user with case-related information about the particular process step. On the one hand, this data consists out of internal information like customer history or article information from an own database (see Figure 2, A). On the other hand, additional external information are offered context-based either in form of a gateway to useful web links (see Figure 2, C) or email-integrated travel details to a location provided by Google Maps (see Figure 2, B). Besides the context-sensitive enrichment of incoming emails with internal and external information, the task guiding component provides the possibility to send email drafts that are context-sensitively selected and recommended to the user (see Figure 2, D). Furthermore, if other software systems are used within the enterprise, e.g., ERP systems, components can be integrated that transfer information out of the email to these systems (see Figure 2, E). Depending on the context, different information can be useful for a particular process. Hence, the type and level of detail of the information to be displayed can be adjusted using a customization tool.

Process Guidance. Due to prior process instances and according user actions, there is already knowledge about the underlying process available, which forms the input for the process guidance functionality. Using the Enterprise Process Repository, the guidance component—as the fourth step in the processing of an incoming email—offers suggestions and recommendations for the further proceedings in a particular process (see Figure 2, G; for detailed information on the recommendation process, it is referred to Burkhart et al. [18]). A second functionality of the process guidance component is to provide advice in actually executing the next process step once the user has chosen one of the provided actions.



Dear Office Supply,

I'd like to order the following articles, for Customer 300001, to
re-stock our internal office supplies:

- Scotch Tape (Article 200001) x 10
- Leitz Folder (Article 200002) x 25
- Stabilo Pen (Article 200003) x 30

Plea address in Berlin via UPS

Your
Com

Ordered Article
Stabilo Pen (ID:200003)
Ordered amount: 30
On Stock: 2500
Retail Price: 1.20

Change the detected process Step by clicking one of the following links:

SupportStep(SupportProcess)
(50% - Nachricht: 33%)
ReservationStep(UNDEFINED)
(25% - Nachricht: 33%)
ShippingStep(OrderProcess)
(75% - Nachricht: 100%)
InvoiceStep(OrderProcess)
(50% - Nachricht: 33%)
CapacityComparisonStep(SupportProcess)
(66% - Nachricht: 66%)

New OrderProcess detected

Current Step: OrderStep
Process ID: 1002

Visit Leitz on the internet!
Visit Tesa on the internet!
Visit Stabilo on the internet!

Transfer Order to ERP System

Send predefined E-Mail:
Order Confirmation

Figure 2. Screenshot of an enriched email message

V. EXPORT—Supporting SMEs in the ATLAS Export Procedure

A. Motivation

Having realized the basic use case regarding a support of order processes, we are currently researching how our approach can be applied to an additional application domain.

A study conducted by the United Nations indicates that inefficient customs processing accounts for 7% of the overall international trading costs. To address this problem, the European Commission finished the eCustoms law initiative in 2003. While this initiative has to be implemented by all member states, the German implementation is called ATLAS (Automatisiertes Tarif und lokales Zollabwicklungssystem). In 2009, this system fully replaced the manual, paper-based processing and became mandatory for the use case of exports. For larger companies, such online-based customs declarations offer the chance to be included into the existing IT-infrastructure and business processes. Thus, they can contribute to process automation. Small and medium-sized enterprises (SME) however face difficulties with the change towards ATLAS [22]. Considering their usually scarce IT landscape that does often not exceed the basic email infrastructure, they often rely on the online platform IAA-Plus as provided by the German customs office or consult an external service provider. Still, these alternatives do not satisfy all special needs and characteristics of SMEs. For instance, they cause additional costs and increase the complexity of process execution. Hence, the main objective of ATLAS, a largely automated handling of border-crossing product exchange resulting in an integrated and predictable supply chain, has not been reached.

B. Use Case Description that will be Supported in Future

A small SME occasionally exports its final products into foreign countries outside the EU. Considering the small size of the company in question, like most other such companies, it does not own an extensive IT landscape with an ERP system. Thus, software solutions that integrate ATLAS with common ERP systems are not applicable. To perform the electronic export declaration nevertheless, the company relies on the web-based platform IAA-Plus. This however turns out to be a time-consuming process because the responsible employee has to gather all relevant data from other colleagues and from Excel sheets (like product tariff numbers). Afterwards, this data has to be put into the web-form manually and, after completion of the export declaration, most of the data has to be transferred again to a logistics service provider. This processing is not optimal, yet it is typical.

EXPORT as an extension of the COPA system will address the mentioned problems. The majority of relevant information for the customs declaration has already been communicated between seller and customer before. For SMEs, this is in most cases done via e-mail. The e-mail infrastructure therefore has valuable information available. The EXPORT tool is simple to integrate into existing e-mail infrastructure. After installation, it extracts the required information from e-mail conversations and generates an ATLAS declaration automatically subsequent to a successful sell. If some piece of information is missing, the tool supports the user during the data input. As an example, it is referred to the product tariff numbers that are a challenge especially for SMEs. Appropriate search mechanisms as well as the automated building of a repository for mapping the own product portfolio to the respective numbers address this problem. As soon as the ATLAS system successfully assigned a unique Movement Reference Number (MRN) on the electronic customs accompanying document (ABD), this number can be forwarded together with the already available information to the cheapest logistics service provider. This happens via an appropriate interface connection and finishes the process.

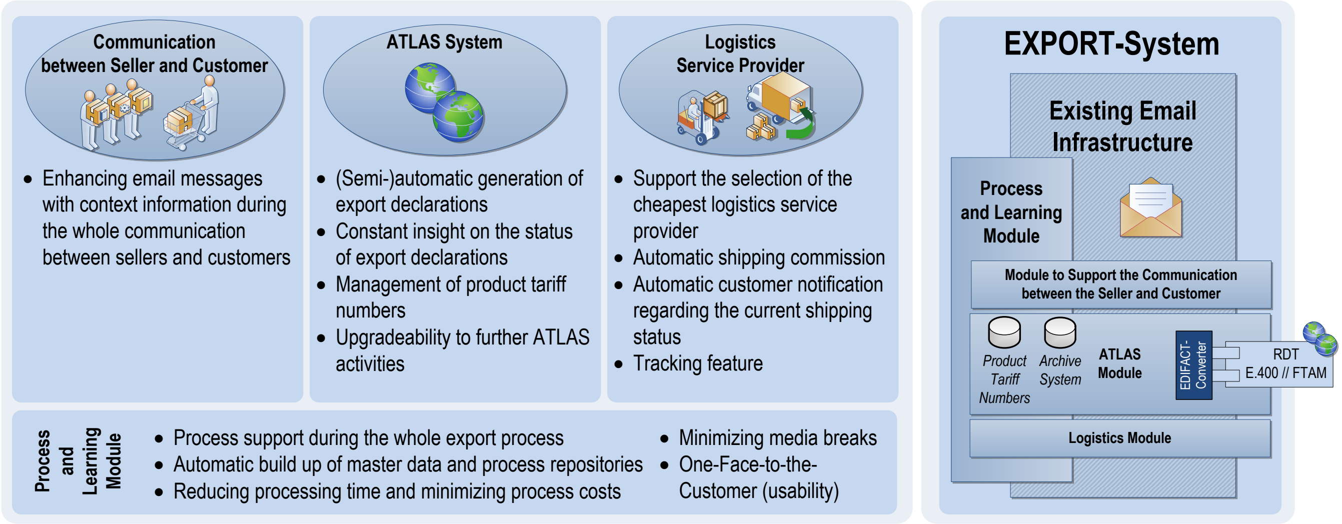

Communication between Seller and Customer
• Enhancing email messages with context information during the whole communication between sellers and customers
ATLAS System
• (Semi-)automatic generation of export declarations
• Constant insight on the status of export declarations
• Management of product tariff numbers
• Upgradeability to further ATLAS activities
Logistics Service Provider
• Support the selection of the cheapest logistics service provider
• Automatic shipping commission
• Automatic customer notification regarding the current shipping status
• Tracking feature
Process and Learning Module
• Process support during the whole export process
• Automatic build up of master data and process repositories
• Reducing processing time and minimizing process costs
• Minimizing media breaks
• One-Face-to-the-Customer (usability)
EXPORT-System
Existing Email Infrastructure
Process and Learning Module
Module to Support the Communication between the Seller and Customer
Product Tariff Numbers
Archive System
ATLAS Module
EDIFACT-Converter
RDT
E.400 // FTAM
Logistics Module

Figure 3. Main functionalities of an EXPORT supporting system and its schematic architecture

A learning component included in EXPORT builds a repository on its own during the tool operation in order to reduce manual data input and maintenance efforts during export processes; hence, ensuring a consequent implicit knowledge management. Additionally, the integrated process component supports the process execution starting from the customer request up to the communication with the logistics service provider and the product deliverance to the foreign customer. Thereby, it also allows for a retrospective view on the executed export processes. In summary, SME are enabled to cover the currently existing disadvantages in this section of the supply chain in an autonomous, straightforward and inexpensive way. Due to the process support that is given during the whole export process (via task and process guidance), users that are not familiar with the overall process can learn it on the job.

The main functionalities of EXPORT as well as a schematic outline of the system are visualized in Figure 3.

VI. Conclusion and Outlook

In this paper, we presented an approach that allows business process learning on-the-job using the concepts of task guidance and process guidance. After introducing the approach, the paper presented a prototypical implementation of the approach and in doing so proved its general feasibility. A first empirical evaluation of the approach and its application has already been conducted and can be found in Burkhart et al. [23]. This evaluation has demonstrated the basic benefits of guidance for carrying out unfamiliar business processes and learning them on-the-job based on real test persons that were involved. As a result, the study proofed that test subjects were able to process an unfamiliar workflow significantly faster by task guidance and process guidance. Furthermore, they experienced the processing as significantly easier and moreover, they were significantly higher satisfied with the result of the conducted workflow.

In the next step, we are going to evaluate our approach using the implementation in a real-world scenario or even in a large-scale field study. Moreover, further highly-important features, which are only testable in a time-consuming way, e.g., the adaptive, flexible and self-learning features, will be evaluated to see how learning is progressing and how organizational-based as well as crowed-based knowledge will increase over time. Furthermore, we are going to implement further features into our prototype to be able to support the promising use case that was presented at the end of this paper. In applying the conceptual approach to this further application domain, we can prove its applicability and feasibility in more general and extended terms.

References

[1] O. Marjanovic and W. Bandara, “The Current State of BPM Eduction in Australia: Teaching and Research Challenges”, Business Process Management Workshops (BPM 2010), pp. 775–789.

[2] E. Kavakli, “Modelling organizational goals: Analysis of current methods”, ACM Symposium on Applied Computing (SAC '04), pp. 1339–1343.

[3] G. Grambow, R. Oberhauser, and M. Reichert, “Contextual Generation of Declarative Workflows and their Application to Software Engineering Processes”, International Journal on Advances in Intelligent Systems, vol. 4, no. 3/4, 2011, pp. 158–179.

[4] R. L. Jacobs and Y. Park, “A Proposed Conceptual Framework of Workplace Learning: Implications for Theory Development and Research in Human Resource Development”, Human Resource Development Review, vol. 8, June 2009, pp. 133–150.

[5] A. R. Hevner, S. T. March, J. Park, and S. Ram, “Design Science in Information Systems Research”, MIS Quarterly, vol. 28, March 2004, pp. 75–105.

[6] I. T. Hawryszkiewycz, “A Framework for Integrating Learning into Business Processes”, South East Asia Regional Computer Science Confederation (SEARCC) Conference 2005, pp. 23–28.

[7] H.-J. Chen and C.-H. Kao, “Empirical validation of the importance of employees' learning motivation for workplace e-learning in Taiwanese organisations”, Australasian Journal of Educational Technology, vol. 28, May 2012, pp. 580–598.

[8] N. Clarke, “Workplace learning environment and its relationship with learning outcomes in healthcare organizations”, Human Resource Development International, vol. 8, March 2005, pp. 185–205.

[9] B. Weber, S. Sadiq, and M. Reichert, “Beyond Rigidity - Dynamic Process Lifecycle Support: A Survey on Dynamic Changes in Process-aware Information Systems”, Computer Science - Research and Development, vol. 23, May 2009, pp. 47–65.

[10] P. Dadam, M. Reichert, S. Rinderle, M. Jurisch, H. Acker, K. Gösner, U. Kreher, and M. Lauer, “Towards Truly Flexible and Adaptive Process-Aware Information Systems“, United Information Systems Conference (UNISCON 2008), pp. 72–83.

[11] T. Burkhart and P. Loos, “Flexible Business Processes - Evaluation of Approaches”, Multikonferenz Wirtschaftsinformatik 2010 (MKWI 2010), pp. 1217–1228.

[12] K. Ploesser, M. Peleg, P. Soffer, M. Rosemann, and J. C. Recker, “Learning from Context to Improve Business Processes”, BPTrends, vol. 6, Jan. 2009, pp. 1–7.

[13] J. Ghattas, P. Soffer, and M. Peleg, “Learning Business Process Models: A Case Study”, International Conference on Business Process Management (BPM'07), pp. 383–394.

[14] A. Abecker, K. Hinkelmann, H. Maus, and H.-J. Müller, “Geschäftsprozessorientiertes Wissensmanagement”, Berlin: Springer-Verlag, 2002.

[15] T. Allweyer, “Geschäftsprozessmanagement: Strategie, Entwurf, Implementierung, Controlling”, Herdecke: W3L-Verlag, 2005.

[16] F. Lehner, “Wissensmanagement: Grundlagen, Methoden und technische Unterstützung”, München: Hanser, 2006.

[17] U. Remus, “Prozessorientiertes Wissensmanagement: Konzept und Modellierung“, Regensburg: University of Regensburg, 2002.

[18] T. Burkhart, D. Werth, and P. Loos, “Flexible process support by automatic aggregation of implicit and explicit user behavior“, Fourth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT 2011), pp. 167–172.

[19] Email Equation, “Email marketing services for small to midsize business”, 2010, retrieved July 20, 2012, from http://www.emailequation.com/emailmetricsroi.html.

[20] The Radicati Group, “Business User Survey”, 2010, retrieved July 20, 2012, from http://www.radicati.com/wp/wp-content/uploads/2010/11/Business-User-Survey-2010-Executive-Summary.pdf.

[21] Messaging Architects, “Policy-Based Email Security and Data Leak Prevention”, 2012, retrieved July 20, 2012, from http://www.messagingarchitects.com/solutions/guardian.html.

[22] J.-B. Delèze and J.-P. Lattion, “Nutzen einer möglichen Beteiligung der Schweiz am E-Zoll-Projekt der EU“, 2011, retrieved July 20, 2012, from http://www.dievolkswirtschaft.ch/editions/201103/Deleze.html

[23] T. Burkhart, J. Krumeich, D. Werth, and P. Loos, “Flexible Support System for Email-based Processes: an Empirical Evaluation”, International Journal of E-Business Development, vol. 2, Aug. 2012, pp. 77–85.

A Framework for Designing Knowledge Management Systems: Aggregating the Existing Approaches

Gaga Ali Thomas, Farhi Marir, Romas Mikusauskas, Preeti Patel
Knowledge Management Research Centre,
Department of Computing
London Metropolitan University, UK
gaga642001@yahoo.com, f.marir@londonmet.ac.uk, r.mikusauskas@londonmet.ac.uk, p.patel@londonmet.ac.uk

***Abstract*—To effectively manage organizational knowledge to promote innovative practice and gain sustainable competitive advantage, there is a need for a system called Knowledge Management System (KMS). This system helps to enhance the organizational process of knowledge creation, storage, retrieval, transfer and application. With the explosive growth of interest in knowledge management, different KMS frameworks have been produced by various researchers for successful implementation of knowledge initiatives. However, it was observed that, the existing frameworks do not provide a complete and generalized framework for designing of knowledge management system by defining the key fundamental attributes of KMS and their inter-relationships. This paper reviews the existing frameworks for designing KMS with the view to improving them by developing a more comprehensive integrated framework from a multi-dimensional approach by incorporating, extending and aggregating attributes that are already available either from academic, theoretical approaches as well as from applied practitioner–like approaches in knowledge management efforts. The development of this framework is based on the analysis of five selected KMS frameworks on which an initial integrated framework for designing KMS is proposed.**

Keywords-Framework; knowledge management system.

I. INTRODUCTION

Today's economy is generally referred to as a knowledge –based economy, where the economy and the wealth has changed from a world where capital is seen to be a physical thing such as plant, machinery and land to a world where the real capital for creating wealth is less tangible.

Knowledge is a multifaceted concept with multilayered meanings, however, knowledge can be defined as "justified true belief" [1]. According to Alavi and Leidner [2], knowledge is information possessed in the mind of individuals. It is personalized information related to facts, procedures, concepts, interpretations, ideas, observations and judgment.

Organizations need to manage knowledge for enabling their employees to learn and develop their competencies efficiently. Knowledge helps employees to be more creative and innovative and managing it efficiently enables individuals, teams and entire organizations to collectively and systematically create, share and apply knowledge to achieve their objectives.

To effectively manage organizational knowledge to promote innovative practice and competitive advantage, there is a need for a system called Knowledge Management System (KMS). This system helps to enhance the organizational processes of knowledge creation, storage, retrieval, transfer and application [1],[2],[4].

Organizational KMSs, usually, require profound cultural renovations, because, traditionally, organizations norms promote knowledge hoarding rather than knowledge sharing. Hence, a major cultural shift is requiring in changing employees' attitudes and behavior so that they willingly and consistently share their knowledge and insights. These attitude and behavior challenges need to be adequately addressed, so as to ensure successful implementation of KMS in an organization. The challenges can be resolved through the development of a comprehensive KMS framework.

A KMS framework is a conceptual model that provides a broad guideline that facilitates effective and efficient implementation of a KM initiative. KMS is not a technology or a set of methodologies; rather, it is a practice or discipline that involves people, processes and technology [1],[2],[3].

This paper is divided into six sections. Sections 1 and 2 are devoted to the introduction and motivation for development of a framework. Section 3 is devoted to presenting previous work, Section 4 analyses a selected number of existing frameworks and Section 5 presents the proposed framework. Sections 6 and 7 are devoted to conclusion and future work.

II. MOTIVATION

With the explosive growth of interest in knowledge management, various Knowledge management system frameworks have been developed by different researchers based on their background and area of interest for successful implementation of a knowledge management initiative.

However, it was observed that most of the existing frameworks do not adequately fulfill the KMS needs of organizations [3]. That is, the existing frameworks do not provide a complete and generalized framework for designing of the knowledge management system by defining the key fundamental attributes of KMS and their inter-relationship.

Hence, there is a need to improve on the existing KMS frameworks so as to ensure that the framework comprehensively consists of key fundamental attributes, in order to reduce the level of failure of Knowledge management (KM) projects and loss in revenue incurred by organizations in implementing KM projects.

Therefore, the purpose of this research is to examine the existing frameworks for designing KMS and to improve upon them by developing a more comprehensive integrated KMS framework from a multi-dimensional approach by incorporating, extending and aggregating attributes that are already available from academic, theoretical approaches as well as from applied practitioners, like approaches in knowledge management efforts.

III. Previous Work

In order to implement KMS successfully, a KMS framework is needed [3]. Rusli et al. [4] define the framework of KMS as the guidelines and directions to set up KMS. In this section, we will review the proposed framework and identify limitations on a number of issues related to KMS such as collaboration, cultural issues, knowledge sharing, methodology for implementing the framework, generalization, leadership, communities of practices, information context, learning elements, usability of KMS, copyright and costs of implementing the framework.

Sajeva [5] acknowledges that the changing business environment, characterized by dynamically discontinuous changing, requires a re-conceptualization of Knowledge Management Systems as they have been understood in information system practice and research. It emphasises that in a dynamically and discontinuously changing business environment, there is a need for a paradigm shift from an information processing view to a sense-making view of Knowledge Management.

Malhotra and Galletta [6] explicitly recognized that knowledge resides in the user and not in the collection of information. It states that, the human aspect of knowledge creation and knowledge renewal cannot be replaced by knowledge management technologies especially in the following areas: imagination and creativity latent in human minds, untapped tacit dimensions of knowledge creation, subjective and meaning making basis of knowledge and constructive aspects of knowledge creation and renewal. The proposed framework offers a combination of flexibility and agility while ensuring efficiencies of the current technology architecture. It allows for continuous re-examination of the assumptions under lying best practices, reinterpretation of this information and efficiencies based on propagation and dissemination of the best practices. Despite the fact that the proposed framework acknowledged the human factor, need for sense-making knowledge in a dynamic business environment, the framework fail to address cultural issues that need to be considered when migrating from the traditional to sense-making approach.

Rusli et al. [7] present a Knowledge Management system framework called Active Design Support (ADS). The frameworks is aimed at providing product designers with critical design knowledge and guide them toward rational design decisions based upon relevant design errors and successful design decisions in the past during product development processes. They considered the design knowledge obtained by individual designers and experts as a valuable asset to an organization for enhancing the competitiveness of products a company’s designs can produce. They stated that insufficient flow of information and shared knowledge in an organization can result in delays, sub-standard product quality and costly errors, due to disregard of previous experiences. Although, the ADS framework enhances and promotes knowledge sharing amongst designers, it is clear that the framework does not stress the importance of collaboration and it undermines the issue of copyright law.

According to Rusli et al. [3], within the general frameworks of KMS, even though accepted, there are some unidentified features that have not been discovered and that the addition of these unidentified features will make the existing framework of KMS more effective. Rusli et al. [7] adopted an earlier KMS framework as a base-line for their research work in investigating the general perception and acceptance of people toward the current KMS implementation in six selected PHLI in Klang Village, Malaysia. From their research, six elements were identified as causes for not successfully implementing KMS in the selected institutions. These elements are: lack of awareness of KMS implementation, unutilized technical component, application and systems, ignorance of advance technology, cost of KMS implementation, lack of incentives and rewards and unaware of KMS audit. Therefore, the KMS framework of Rusli et al. [7] was modified; KMS awareness was defined as individual components for the KMS framework, rather than a part of a component as presented by the earlier work [7] which considered awareness as part of the KMS Psychological component. In addition, the research indicated that KMS Audit gained less attention in KMS implementation. The authors suggested that there should be clear interaction between KMS Awareness and KMS Audit; this will be achieved by implementing the Audit Mechanism as well as feedback mechanism. They also stated that in implementing the KMS framework, the issue of incentives and rewards must be considered, while they neglect the issue of culture as it relates to the individual and the organization.

Roberta [8] presented a new approach to KMSs called Distributed Knowledge Management (DKM) and applied it in a case study with Impres a Pizzarotti & C.S.P.A., a complex Italian building industry. The paper views that the common outcome of the traditional KMS is the creation of Enterprise Knowledge Portal (EKP), a web-based interface which provides a common access point to corporate knowledge. Even if users have different profiling systems, the underlying representation of EKP is typically unique, and is meant to represent a common and shared conceptualization of corporate knowledge that enables communication and knowledge sharing across the entire organization. This approach to KMS is incompatible with the very nature of what is to be managed and consequently are often deserted by users. The author based the concept of DKM on two principles: the principle of autonomy, which grants organizational units a high degree of semantic autonomy in managing their local knowledge and the principle of coordination, which allows each organizational unit to exchange knowledge with other units through processes of double loop learning. According to this approach, complex knowledge-based organizations can be seen as “Constellations” of local organizational units which exhibit some degree of semantic autonomy, with the ability to manage local knowledge and to develop a personal perspective on the world. The resulting KMS aimed at sustaining the creation and management of different conceptual schemes which coexist within a DKM system. The author further explained that within a DKM system, each organizational unit, either formal or informal must be represented and verified by a Knowledge Node (KN), and that each KN should consist of a knowledge owner, a system of artifact and a context. This approach attempts to address why people are led to desert KMS by focusing on the lack of coherence between a privileged, unique and supposedly shared conceptualization of knowledge within KMS and the different ways of thinking of workers, communities, teams and officers that participated in the firm’s activity. While the framework recognized the importance of knowledge nodes in designing KMS, the authors fail to explain the effective way to manage inter KN in knowledge sharing.

Roberta [8] also noted that numerous researchers have proposed several KMS frameworks, many of these frameworks are prescriptive and providing direction on the type of KM procedure without providing specific details on how these procedures should be accomplished. Based on their research work on HLI, they revealed that people mostly concentrate on KMS infrastructure and technology and neglect other very important issues of KMS such as human aspects. Therefore, they proposed a KMS framework that consists of five components. These include: functionality and system architecture as the backbone to support the KMS, psychological and cultural aspects as well as the knowledge strategies and measurement or system auditing. The proposed KMS framework covered both the technological and human aspect of KMS, however key issues like leadership, communities of practices are missing in the framework design, which are very fundamental elements in the success of KMS.

Rusli et al. [9] state that a Learning Organization (LO) still has difficulties in identifying the appropriate KMS architectural framework and KMS technologies for their organizations and that, there is no clear mechanism on how to motivate and encourage a Community of Practice (COP) to share and reuse knowledge, as well as to generate new knowledge in a collaborative environment. The authors proposed a KMS model and architecture for LO that consists of six main components in order to serve a community within a collaborative environment to work together to achieve the desired objectives of an organization. This KMS framework is found to be good for people to share their knowledge in a learning organization, however, it fails to consider the dynamism of the learning environment, information flow and the issue of context of information shared between users of the KMS.

According to Mohd et al. [10], there exist gaps between theory and practice in the current knowledge management framework. The authors used Shell IT International (SITI) Knowledge Management framework as a case study. The authors identified eight activities that are critical in the knowledge management of an organization. The activities are as follows: initiation, production, modelling, repository, distribution and transfer, technology infrastructure, application and retrospect. The authors presented an alternative framework that addresses the entire processes needed for SITI’s internal and external Knowledge Management usage and development. The framework is cyclic in nature, with multiple feedback loops and iteration which means it can provide queries and receive feedbacks from various departments in the organization. The features of the proposed framework are: strategic, model, use, review, and transfer and technology infrastructure. The framework did not provide methodology for implementing the framework, and the research is based on a single entity and cannot be generalized.

Mostafa et al. [11] observed that early KMS concentrated too much on technical issues and hence fails to produce the desired outcome of KMS. The authors presented an integrated KMS framework, which consists of three main layers. The interior layer is the knowledge architecture, which it is considered as the KM backbone. They defined knowledge architecture as a logical set of principles and standards which guide the engineering (high level) design, selection, construction, implementation, support and management of an organization’s Knowledge Management System Infrastructure.

Others factors considered in the interior layer are: Knowledge Strategy, Knowledge Capturing, Knowledge Storage and Knowledge Sharing. The middle layer consists of factors considered as necessary for the successful implementation of a KMS; these factors are business process reengineering, reward and promotion system, pilot, technology, training and education programs. The outer layer includes factors that are classified as general in comparison with the outer factors. These factors are organizational culture, transparency, CEO support and commitment, and trust. The authors explained the methodologies for the adoption of this KMS framework, which takes into account both the technological and human aspects. The framework presents a holistic approach to KMS, but does not mention anything regarding data management and cost effectiveness of the KMS framework.

Chong and Choi [12] reviewed early studies on KMS and noted that many KMS research has taken a narrow view, overlooking important foundations such as law (Knowledge Privacy and Protection), Politics (Knowledge Control and dominance) and marketing (persuasion and knowledge asymmetries). Also that KMS research seldom considered the “dark side” and how it could be used to suppress or distort knowledge to serve a specific agenda. The authors came up with seventeen most desirable capabilities of KMS. The seventeen capabilities were sorted in order of importance as follows: adaptability, cost effectives, first access, ease to use, search and retrieval, security, knowledge creation, content management, quality assurance, collaboration, multimedia, report generation, central repository, push strategy, customizability, metrics and incentive. Their studies focused on the recent changes in the way that organizations view KM and suggested that there should be stronger integration of KMS with the overall technology in organization. More focus should be given to place KM Support in context and integrating KMS with existing technologies, creating integrated knowledge support systems-business technologies enhanced with KM capabilities. The KMS framework presented an approach from a multidimensional perspective; however, the framework fails to consider learning as a key element of KMS.

Weber [13] observed that among the widely discussed categories of KMS are repository-based and expert locations. Repository-based KMS are typically adopted in support of knowledge sharing and leveraging, based on well-maintained databases that store explicit knowledge. Expert locater KMS are systems that link users with experts on the basis of stored experts' skills and competencies. They noted that despite the fact that both the repository-based and expert locator are important to organizations, they are implemented separately by different systems. They proposed a multifunction framework with a single architecture that performs the role of both systems. That is, a multifunctional framework for designing KMS which adopts a single architecture and performs KM functions that originally required multiple architectures. The architecture lies on two databases: structured knowledge artifacts and the experts, where each artifact is associated with the experts. The principles guarding the framework are highlighted as collaboration, transparency, justification, absorbency, technology and verification. The framework focuses more on technical aspects of designing KMS; it does not mention anything regarding easy to use and user friendliness of application. Also, the proposed framework was not subjected to thorough evaluation of the different functionalities.

Hanlie et al. [14] proposed an enhanced framework and methodology for KM system implementation. In developing the proposed framework and methodology, the authors take into consideration recommendations regarding the development of a KM framework presented in Rubenstein-Montano et al. [15]. The proposed framework consists of five phases namely: choosing a strategy, evaluation, development, validation and implementation. Each phase of the framework consists of sub-phases describing the methodology applicable to each phase. The proposed methodology describes the procedure and steps to be followed and is aligned with the proposed framework. The authors claimed that the outcome of the proposed framework was successful. As the proof of concept was carried out on a single organization, hence the generalization and validation of the framework across multiple organizations and sectors of the economy is desirable to ascertain the comprehensiveness of the framework and methodology.

According to Alavi and Leidner [2], many of the past frameworks do not take into account the importance of human aspects in knowledge management. The author suggested a new framework; the emphasis is on the provision of training to the employees, providing incentives and rewards to employees to share tacit knowledge and the importance of information technology. The major constituents of the framework are rewards, technology, culture, training, learning, strategy, structure, system, leadership, personality and attitude. The author claimed that the proposed framework provides a holistic view for KM implementation which earlier frameworks have ignored. Even though, the proposed KMS framework was developed based on practical survey in an Indian organization, there is no evidence of validation of this model in different environments or through case study.

Parag [16] acknowledged that today's global managers are facing unprecedented challenges outside their organizations fueled by environmental forces of changes such as globalization, emerging technologies, emerging best business practices, government regulations, competitive global financial markets, limited knowledge workers and higher worker turnover rates. Also the rapid increases in the development of emerging technologies have forced many managers and executives to reinvent their decision-making methodologies. The author noted that the current KMS may have outlined their usefulness due to the rapid rate of change of technological and economic forces occurring in the global economy. The author suggested that emerging Knowledge Management System will include encryption tools, existing client/server applications, new ultra high speed internet, emerging technologies, mobile devices, government regulations and guidelines, financial information system, accounting information system, best business practices, ethical practices and legal guidelines. The proposed KMS framework will allow the knowledge workers to collaborate remotely on projects via high speed Internet bandwidth and web-based tools and applications. However, the author fails to take into consideration the cost implication of implementing such KMS framework, and the reliability of networks especially in the developing countries.

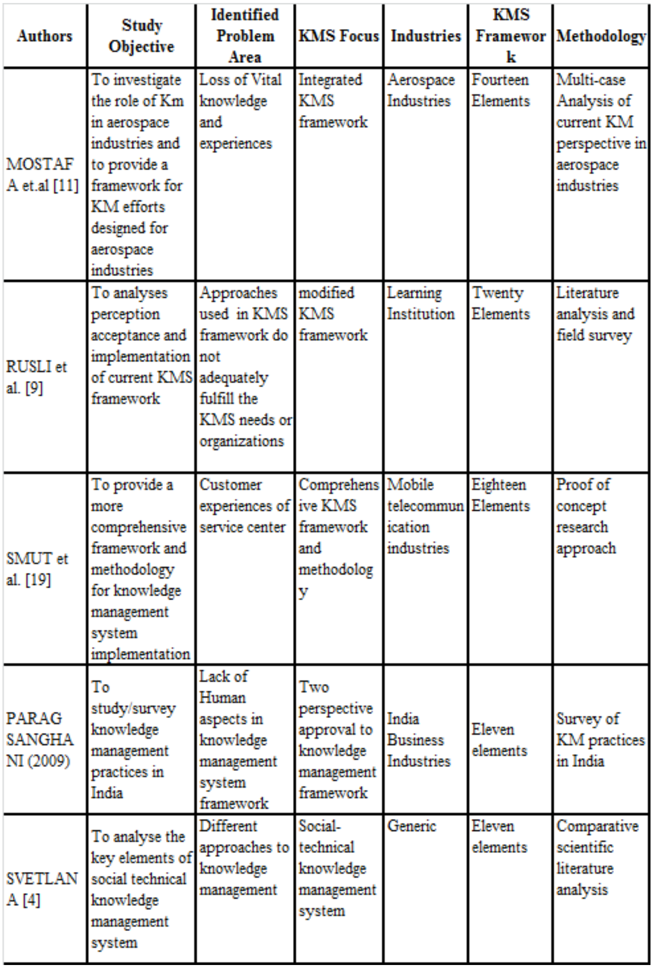
IV. Analysis Of Slected KMS Framework

Since the objective of this research is to develop a comprehensive integrated KMS framework from a multidimensional approach, taking into consideration the key fundamental attributes of KM initiatives, two approaches were adopted: (1) a critical literature review of the existing literature on KMS frameworks. Based on the review, five KMS frameworks (form a social-technical perspective) were selected, as a benchmark for the research; (2) a comparative study of the five selected KMS frameworks was conducted. In a comparative study like most other studies, there are two different approaches: Descriptive and Normative approach.

Since the research is concerned with developing an improved framework, a normative approach was adopted for the study. This is because the normative approach aims at studying, evaluating and improving the present stage of the object of study. Since the normative approach combines empirical observation with normative assessment, it is particularly useful for the analysis of concepts that have both descriptive and evaluation dimension that cannot be disentangled [18].

Based on this study, a more comprehensive integrated KMS framework from a multidimensional approach was developed by aggregating the critical success attributes from the selected framework. The proposed framework was evaluated through a questionnaire to obtain scientific feedback from Developers, Practitioners and Academics in the domain. The aim of the evaluation was to investigate the acceptability of the proposed framework. The analysis of the five selected frameworks is presented in Table 1 and Table 2.

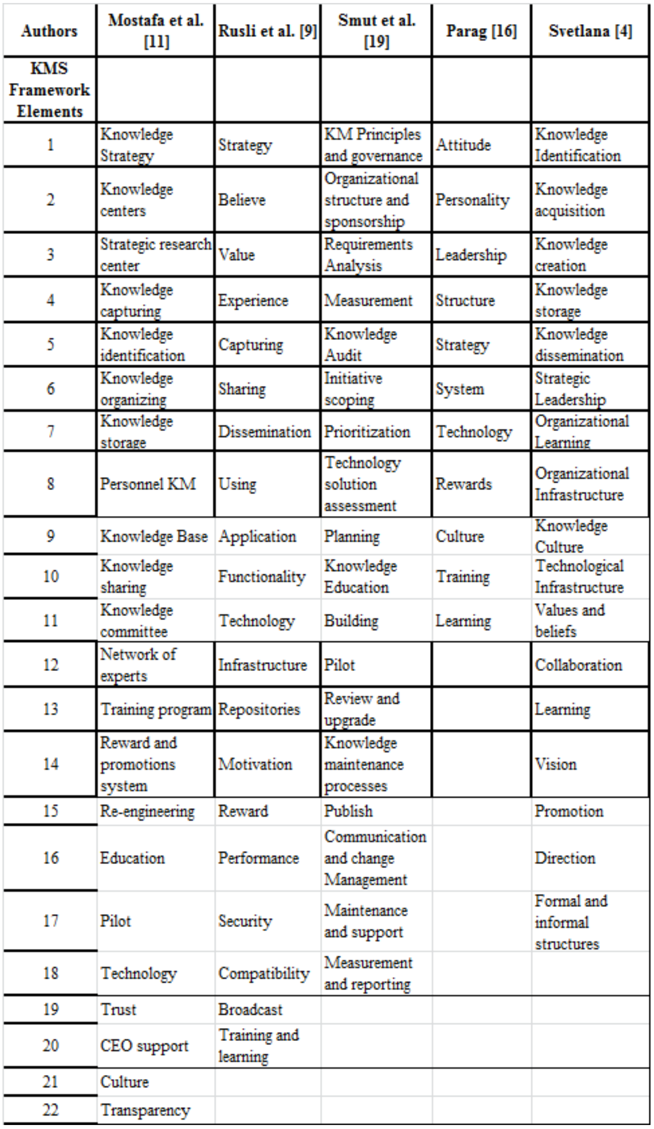
TABLE I. FIVE SELECTED KMS FRAMEWORKS



Authors	Study Objective	Identified Problem Area	KMS Focus	Industries	KMS Framework	Methodology
MOSTAFA et.al [11]	To investigate the role of Km in aerospace industries and to provide a framework for KM efforts designed for aerospace industries	Loss of Vital knowledge and experiences	Integrated KMS framework	Aerospace Industries	Fourteen Elements	Multi-case Analysis of current KM perspective in aerospace industries
RUSLI et al. [9]	To analyses perception acceptance and implementation of current KMS framework	Approaches used in KMS framework do not adequately fulfill the KMS needs or organizations	modified KMS framework	Learning Institution	Twenty Elements	Literature analysis and field survey
SMUT et al. [19]	To provide a more comprehensive framework and methodology for knowledge management system implementation	Customer experiences of service center	Comprehensive KMS framework and methodology	Mobile telecommunication industries	Eighteen Elements	Proof of concept research approach
PARAG SANGHANI (2009)	To study/survey knowledge management practices in India	Lack of Human aspects in knowledge management system framework	Two perspective approval to knowledge management framework	India Business Industries	Eleven elements	Survey of KM practices in India
SVETLANA [4]	To analyse the key elements of social technical knowledge management system	Different approaches to knowledge management	Social-technical knowledge management system	Generic	Eleven elements	Comparative scientific literature analysis

However, none of the selected frameworks presented the whole spectrum of element as depicted in Table 3. Also, each of the KMS framework focused more on one or two sub system(s) than the other, that is some have emphasis on the Human-Social context and Knowledge context than in Technology context [19]. In order to create an effective KMS in an organization, there is need to ensure that all relevant elements are considered in designing and developing the KMS framework. That is relevant elements from the Human-Social, Technology and Knowledge context need to be integrated and harmonized.

TABLE II. ATTRIBUTES OF SELECTED KMS FRAMEWORKS



Authors	Mostafa et al. [11]	Rusli et al. [9]	Smut et al. [19]	Parag [16]	Svetlana [4]
KMS Framework Elements					
1	Knowledge Strategy	Strategy	KM Principles and governance	Attitude	Knowledge Identification
2	Knowledge centers	Believe	Organizational structure and sponsorship	Personality	Knowledge acquisition
3	Strategic research center	Value	Requirements Analysis	Leadership	Knowledge creation
4	Knowledge capturing	Experience	Measurement	Structure	Knowledge storage
5	Knowledge identification	Capturing	Knowledge Audit	Strategy	Knowledge dissemination
6	Knowledge organizing	Sharing	Initiative scoping	System	Strategic Leadership
7	Knowledge storage	Dissemination	Prioritization	Technology	Organizational Learning
8	Personnel KM	Using	Technology solution assessment	Rewards	Organizational Infrastructure
9	Knowledge Base	Application	Planning	Culture	Knowledge Culture
10	Knowledge sharing	Functionality	Knowledge Education	Training	Technological Infrastructure
11	Knowledge committee	Technology	Building	Learning	Values and beliefs
12	Network of experts	Infrastructure	Pilot		Collaboration
13	Training program	Repositories	Review and upgrade		Learning
14	Reward and promotions system	Motivation	Knowledge maintenance processes		Vision
15	Re-engineering	Reward	Publish		Promotion
16	Education	Performance	Communication and change Management		Direction
17	Pilot	Security	Maintenance and support		Formal and informal structures
18	Technology	Compatibility	Measurement and reporting		
19	Trust	Broadcast			
20	CEO support	Training and learning			
21	Culture				
22	Transparency				

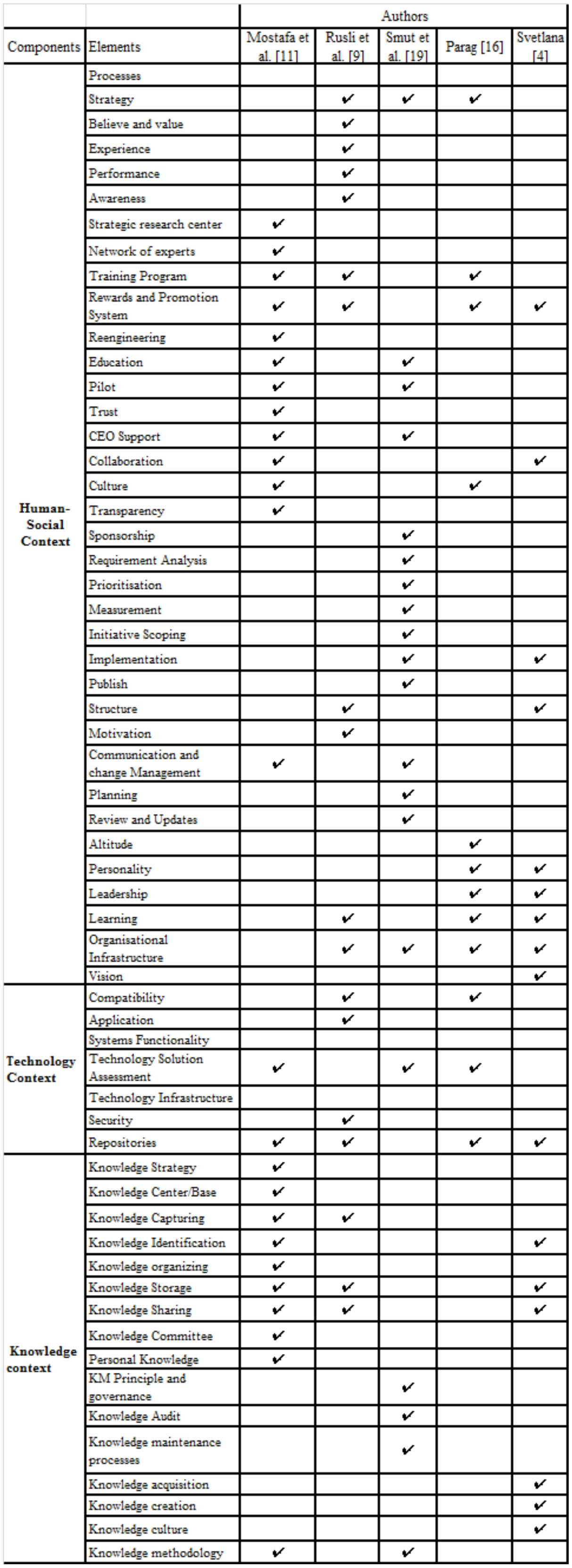
V. PROPOSED FRAMEWORK FOR DEVELOPING KMS

Rubenstein-Montano et al. [15] make the following recommendations as regard to the development of KMS framework:

- A KMS framework should be both prescriptive and descriptive.
- KMS must be directed by learning as feedback loops both single and double.
- The Cultural aspects of the organization must be acknowledged and the practices must be compatible with the culture.
- Planning should take place before any KM activities are conducted.
- The organizational goals and strategies must be linked to KM.
- A KM framework should be consistent with system thinking.

ISBN: 978-1-61208-254-7

TABLE III. ANALYSIS OF SELECTED KMS FRAMEWORKS



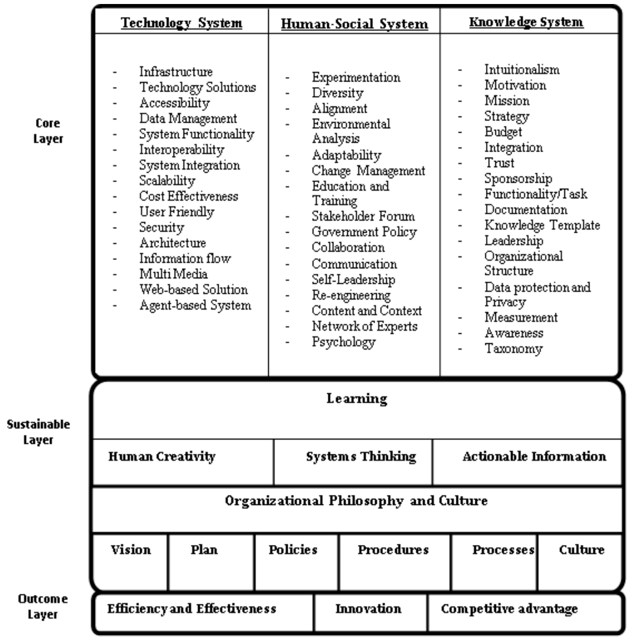
		Authors				
Components	Elements	Mostafa et al. [11]	Rusli et al. [9]	Smut et al. [19]	Parag [16]	Svetlana [4]
Human-Social Context	Processes					
	Strategy		✔	✔	✔	
	Believe and value		✔			
	Experience		✔			
	Performance		✔			
	Awareness		✔			
	Strategic research center	✔				
	Network of experts	✔				
	Training Program	✔	✔		✔	
	Rewards and Promotion System	✔	✔		✔	✔
	Reengineering	✔				
	Education	✔		✔		
	Pilot	✔		✔		
	Trust	✔				
	CEO Support	✔		✔		
	Collaboration	✔				✔
	Culture	✔			✔	
	Transparency	✔				
	Sponsorship			✔		
	Requirement Analysis			✔		
	Prioritisation			✔		
	Measurement			✔		
	Initiative Scoping			✔		
	Implementation			✔		✔
	Publish			✔		
	Structure		✔			✔
	Motivation		✔			
	Communication and change Management	✔		✔		
	Planning			✔		
	Review and Updates			✔		
	Altitude				✔	
	Personality				✔	✔
	Leadership				✔	✔
	Learning		✔		✔	✔
	Organisational Infrastructure		✔	✔	✔	✔
	Vision					✔
Technology Context	Compatibility		✔		✔	
	Application		✔			
	Systems Functionality					
	Technology Solution Assessment	✔		✔	✔	
	Technology Infrastructure					
	Security		✔			
	Repositories	✔	✔		✔	✔
Knowledge context	Knowledge Strategy	✔				
	Knowledge Center/Base	✔				
	Knowledge Capturing	✔	✔			
	Knowledge Identification	✔				✔
	Knowledge organizing	✔				
	Knowledge Storage	✔	✔			✔
	Knowledge Sharing	✔	✔			✔
	Knowledge Committee	✔				
	Personal Knowledge	✔				
	KM Principle and governance			✔		
	Knowledge Audit			✔		
	Knowledge maintenance processes			✔		
	Knowledge acquisition					✔
	Knowledge creation					✔
	Knowledge culture					✔
	Knowledge methodology	✔		✔		

Following the analyses of the selected KMS frameworks and considering what constitutes a KMS framework as described by Rubenstein-Montano et al. [15], a proposed Integrated KMS framework is presented as shown in Table 4.

The proposed framework consists of three layers namely: foundation layer, core layer and outcome layer. The foundation layer is considering being a strategy sustainable layer which consists of two components: the organizational philosophy and learning. Organizational philosophy contains the following attributes: vision, plan, policies, procedures, processes and culture while learning components have system thinking, human creativity and actionable information as attributes. Each of these attributes is considered as necessary critical factors for the successful implementation of knowledge management system.

The core layer consists of three components: technological system, social-human system and knowledge system. The technological system has sixteen attributes namely: Infrastructure, Data Management, Inter-operability, Cost Effectiveness, Technological Solution, System Functionality, System Integration, Scalability, User Friendly, Information Flow, Architecture, Accessibility, Security, Multi-media, Web-based solution and Agent-based system.

TABLE IV. PROPOSED FRAMEWORK



Core Layer

Technology System
- Infrastructure
- Technology Solutions
- Accessibility
- Data Management
- System Functionality
- Interoperability
- System Integration
- Scalability
- Cost Effectiveness
- User Friendly
- Security
- Architecture
- Information flow
- Multi Media
- Web-based Solution
- Agent-based System

Human-Social System
- Experimentation
- Diversity
- Alignment
- Environmental Analysis
- Adaptability
- Change Management
- Education and Training
- Stakeholder Forum
- Government Policy
- Collaboration
- Communication
- Self-Leadership
- Re-engineering
- Content and Context
- Network of Experts
- Psychology

Knowledge System
- Intuitionalism
- Motivation
- Mission
- Strategy
- Budget
- Integration
- Trust
- Sponsorship
- Functionality/Task
- Documentation
- Knowledge Template
- Leadership
- Organizational Structure
- Data protection and Privacy
- Measurement
- Awareness
- Taxonomy

Sustainable Layer

Learning

Human Creativity | **Systems Thinking** | **Actionable Information**

Organizational Philosophy and Culture

Vision | **Plan** | **Policies** | **Procedures** | **Processes** | **Culture**

Outcome Layer

Efficiency and Effectiveness | **Innovation** | **Competitive advantage**

The social-human system has eighteen attributes namely: Psychology, Environmental Analysis, Collaboration, Communication, Re-engineering, Experimentation, Adaptability, Self-Leadership, Education and training, Network of Experts, Alignment, Diversity, Content and Context, Change management, Stakeholder forum, and Government policy .

The knowledge system has also eighteen attributes namely: Mission, Functionality, Strategy, Integration, Institutionalization, Sponsorship, Motivation, Organizational

Structure, Trust, leadership, Budget, Documentation, Knowledge template, Data protection and privacy, Measurement and Awareness.

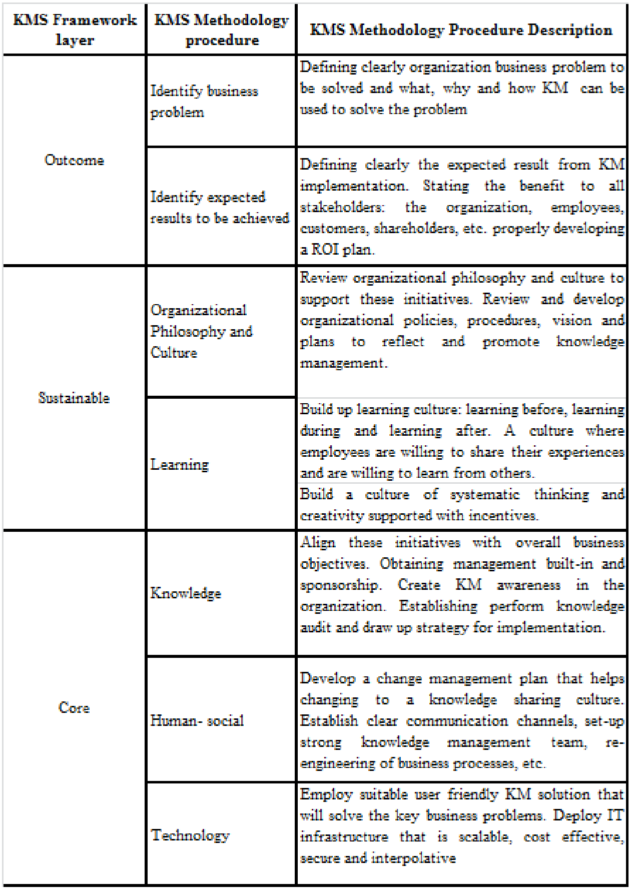
The outcome layer has three attributes namely: efficiency and effectiveness, innovative practice and competitive advantage. The presence of the outcome layer in the framework is to ensure that organization really identify the benefits that they intend to derive from implementing KMS. Without a clear understanding of the benefits of implementing a KMS, it will be difficult to measure the success of KMS.

When an organization has a clear mind set of what they want to achieve from implementing KMS, then Human, cultural and organizational issues need to be addressed to ensure that they support, promote and encourage knowledge management practice. These issues will be addressed in the sustainable layer.

The last phase for implementing KMS is the core layer; here the issue of technological, knowledge and Human-Social system are addressed.

As for implementing this proposed KMS framework, a methodology is proposed in table 5 below. It describes the procedures and steps to be followed in implementing this framework in which detail attributes and activities are contained.

TABLE V. DESCRIPTION OF PROPOSED KMS FRAMEWORK METHODOLOGY



KMS Framework layer	KMS Methodology procedure	KMS Methodology Procedure Description
Outcome	Identify business problem	Defining clearly organization business problem to be solved and what, why and how KM can be used to solve the problem
	Identify expected results to be achieved	Defining clearly the expected result from KM implementation. Stating the benefit to all stakeholders: the organization, employees, customers, shareholders, etc. properly developing a ROI plan.
Sustainable	Organizational Philosophy and Culture	Review organizational philosophy and culture to support these initiatives. Review and develop organizational policies, procedures, vision and plans to reflect and promote knowledge management.
	Learning	Build up learning culture: learning before, learning during and learning after. A culture where employees are willing to share their experiences and are willing to learn from others. Build a culture of systematic thinking and creativity supported with incentives.
Core	Knowledge	Align these initiatives with overall business objectives. Obtaining management built-in and sponsorship. Create KM awareness in the organization. Establishing perform knowledge audit and draw up strategy for implementation.
	Human- social	Develop a change management plan that helps changing to a knowledge sharing culture. Establish clear communication channels, set-up strong knowledge management team, re-engineering of business processes, etc.
	Technology	Employ suitable user friendly KM solution that will solve the key business problems. Deploy IT infrastructure that is scalable, cost effective, secure and interpolative

VI. CONCLUSION AND FUTURE WORK

As earlier stated, KMS is not a technology or a set of methodologies; rather, it is a practice or discipline that involves people, processes and technology [1],[2],[3]. Every organization needs a KMS framework to enable it derive the desired benefits of implementing KM initiative. The proposed framework attempts to build a framework that is comprehensive, integrated and multidimensional in approach into a single framework. Initial analyis of questionnaires and surveys responses on the proposed framework has been very encouraging. The results has showen that most of the issues raised during surevys have been addressed in this proposed framework in particular:

(1) Integration of learning and knowledge management: The need for organizations to become learning organizations requires knowledge management, which in turn is dependent on learning organizations. However, these concepts are addressed separately in most KMS framework. From the research, it is clear that the two concepts are interrelated and dependent. Hence, to enhance organizational creativeness and innovation, the two concepts need to be integrated into a single framework.

(2) Systematic approach to KMS implementation: An Integrated framework that is holistic needs to adopt a number of guiding principles for KM implementation. These principles should include; organizational policies, plan, procedures, philosophy, structure and methods. These principles should present the organizational KM vision and link it to the overall organizational business goals. All these guiding principles are integral of a KMS framework and should be the foundation layer of KM initiative as presented in the proposed KMS framework.

(3) Framework comprehensiveness: The proposed framework presents a fully integrated framework from a multi- dimensional approach by incorporating and aggregating the KMS attributes that are already available from academic, theoretical approaches and as well as from applied practitioner –like approaches in KM efforts.

(4) Human-centric approach in designing KMS: The research work revealed that KM success highly depends on human-social system of KM efforts. That is, Human-centric is the best approach to KM initiative since people are considered as the most critical element in KMS implementation. Hence, frameworks should focus on the importance of people in relation to KMS, and the need to put in place appropriate cultural value that will encourage KMS practice.

(5) Integrating the outcome layer in the KMS framework: A clear understanding of the expected result of implementing KMS by an organization is very critical to the success of the project. Hence organizations need to identify what they want to

achieve in implementing KMS before commencing the implementation.

So, the next stage of this research work will be to get a scientific feedback from the experts and the perceptions on the components, attributes, approach and design of the proposed KMS framework from stakeholders. Practitioners, Academics and Developers are the main stakeholders in the knowledge management domain which could be usefully surveyed with an Internet-based questionnaire.

More questionnaires and surevys will be administered to a breadth of industrial sectors to consolidate the initial results. The findings of the investigation will be interpreted and used to review the proposed KMS framework. This enhanced framework will then be practically tried and tested in a large public-sector organisation like Nigerian Post Office.

References

[1] I. Nonaka, “A Dynamic Theory of organizational knowledge creation” Oganization Science (5:1), pp. 14 – 37, 1994.

[2] M. Alavi and Leidner D. E., “Knowledge Management and Knowledge Management Systems: Conceptual Foundations and research issues” Management Information Systems Quarterly, pp. 107 – 136, 2001.

[3] A. Rusli, I. Hamidah, A. Rodziah, N. Suhaimi, H.S Mohd, H.V. Nurul, and H.H. Sifi “The Development of Bio informatics Knowledge Management System with collaboration Environment” International formal of computer science and Network Security, Vol. 8, No. 2, pp. 309 – 318, February 2008.

[4] S. Svetlana “The analysis of key element of Socio-technical knowledge management system”. Economics and Management, ISBN 1822-6515, pp. 765 – 744, 2010.

[5] P. P. Gandong, H. Sundgo, and P. Sehyung,“A design knowledge management framework for active design support” Proceedings of DETC’ 99: 1999 ASME Design Engineering Technical Conference September 12–15, pp. 22-29, 1999 Las Vages Nevade.

[6] Y. Malhotra and D.F. Galletta, “Role of commitment and motivation in knowledge management system implementation: theory, conceptualization and measurement of antecedents of success”. Proceeding of the 36th Hawaii International Conference on System Sciences, pp. 1 – 10, 6-9 January, Hawai, 2003.

[7] H.A. Rusli, H.S. Mohd, S. Shamsul, and A.A. Rose “A framework for Knowledge Management System Implementation in Collaborative Environment for Higher Learning Institution.” Online Journal of Knowledge Management Practice, March 2005. (Retrieved from http://www.tlainc.com/articl90.htm on January 2013)

[8] R. Cuel “A New methodology for Distributor Knowledge Management Analysis”. 3rd International Conference on Knowledge Management: Proceedings of I-KNOW '03, 2003. Proceedings of: Know-Center, Graz, Austria, pp. 531 – 537, 2nd-4th July 2003.

[9] H.A. Rusli, S. Shamsul, A. A. Rose, and H.S. Mohd “Knowledge Management System Architecture for organization learning with collaboration environment”. International Journal of Computer Science and Network Security, Vol. 6. No: 3A, pp. 237-246, March 2006.

[10] H.S. Mohd, H.A. Rusli, and J.P. Christi “Knowledge Management framework in technology support environment” International Journal of Computer Science and Network Security. Vol. 6, No. 8, pp. 101 – 109, SA August 2006.

[11] J. Mostafa, N. Mehdi, and A. Penman A. “Establishing an Integrated KM System in Iran Aerospace Industries Organization Journal of Knowledge Management, Vol.II, No. 1, 2007, pp. 127-142.

[12] C.S. Chong and Y. S. Choi ”Critical factor in the Successful implementation of knowledge Management” Online Journal of Knowledge Management Practice, June 2005. (Retrieved from http://www.tlainc.com/articl90.htm on January 2013)

[13] R. Weber “Knowledge Management in call Centres”. The Electronic Journal of Knowledge Management Volume 5, Issue 3, 2007, pp. 333 – 346. Available outline at www.ejkm .com (Retrieved Januray 2013)

[14] S. Hanlie, V.D.M Alta, L. Marianne, and K. Paula “ A framework and methodology for knowledge management system implementation” Proceedings of the Conference of the South Africa Institution of Computer Scientists and Information Technologists. ISBN 978-1-60358-6431-4. pp. 70-79, 2009.

[15] N. Rubenstein-Montano, J. Liebowitz, J. Buchwalter, D. McCaw, B.B. Newman, and K. Rebeck “A systems thinking framework for knowledge management”. Decision Support Systems, 2001 a 31 (1): pp. 5 – 16.

[16] S. Parag “Knowledge Management Implementation Holistic Framework based on Indian Study”. Proceedings of Pacific Asia Conference on Information Systems (PACIS), 2009, July 10-12, pp. 60 – 69, India .

[17] S. Hanlie, V.D.M Alta, L. Marianne, and K. Paula “A framework and Methodology for knowledge Management System Implementation” Proceeding of the Conference of the South Africa Institute of Computer Scientist and Information Technologist, pp. 70-77, 11-14th October, 2009, South Africa.

[18] H. Tzyh-Lih, A. Li-min, A. Jen-Her, and A. H. Jang “A framework for Designing Nursing Knowledge Management Systems”. Interdisciplinary journal of Information, knowledge and Management Volume 1, pp. 14 – 22, 2006.

[19] H. Smut, V.D.M Alta, L. Marianne, and K. Paula “A framework and Methodology for knowledge Management Systems Implementation” Proceeding of the Conference of the South Africa Institute of Computer Scientist and information Technologist, pp. 70-77, ISBN 978-1-60558-643-4, 2009, South Africa.

Reuse Cases when Doing Financial Case-Base Reasoning with Respect to Adaptation

Jürgen Hönigl
Institute for Application-Oriented Knowledge Processing
Johannes Kepler University
Linz, Austria
juergen.hoenigl@jku.at

***Abstract*—Case-Based Reasoning (CBR) applies past experience to solve new problems with suitable solutions. This approach presents overloading queries to adapt solutions if necessary. Subpar solutions have to be adapted within a CBR cycle before retaining them to keep a good quality of the case base. Dealing with missing values can be seen as previous step to avoid unnecessary adaptations. Integrate efficient and useful adaptations can be seen as really interesting and challenging task when considering the full CBR methodology. The common CBR principle -similar problems are having similar solutions- can be seen as a rather good point of start when developing an adaptation feature. An adaptation concept and first experience are presented within this paper.**

***Index Terms*—Adaptation, Case-Based Reasoning**

I. Introduction

This paper presents an approach for adaptation of cases. The main goal was achieving a concept, proof the feasibility and get first results which was divided into several sections. Adaptation of cases will be mainly seen as complicated in comparison to retrieve cases because the retrieve step can be clearly divided into different parts such as the case base, a connection between the case base and the CBR system and suitable similarity measures. An adaptation feature depends on the applied domain. For instance, CHEF was using modification rules (change ingredient) and object critics (e.g. cooking time) to modify a cooking recipe namely BEEF-WITH-GREEN-BEANS to BEEF-AND-BROCOLLI. [2] Within the domain regarding this approach, another adaptation process will be used. Similarity measures and queries are suitable for processing different loan applications with numerical and categorial attributes. The similarity value between loan applications can be used for the adaptation when remember the CBR principle that similar problems are related to similar solutions.

Firstly, a brief overview about Case-Based Reasoning will be provided to demonstrate the R^4 model by Aamodt and Plaza. [1] The section Previous Work will briefly introduce associations and similarity measures. Definitions of the case base will be shown which are related to the adaptation of cases. Following two sections are providing the core of this work in progress paper. An adaptation of a case requires consideration of possible missing values which is presented within the next section. Then an adaptation concept will be shown - different queries can achieve different results regarding the precision of relevant and retrieved cases. Then notes regarding first experiments and the concept of evaluation are demonstrated. The manual evaluation by teacher provides a guarantee concerning the quality of the case base. It can be used as a post-condition to the adaptation process. The conclusion and future work are presented at the end.

II. Case-Based Reasoning in a Nutshell

The origin of CBR was given within the research of cognitive science. Schank provides 1982 with his work an approach of Episodic Memory Organization Packets (E-MOPs). [3] CYRUS was a prototype by Kolodner and used meetings and talks by United States of America politician Cyrus Vance to apply E-MOPs to a real scenario. [4] An E-MOP contains a content frame (also known as norm) which stores common information like place, people and subject of a meeting and informations concerning relations to other episodes if necessary. E-MOPs are using a tree-like structure to connect different episodes. In 1994 Agnar Aamodt and Enric Plaza introduced a process model of the CBR cycle which was commonly called the R^4 model. [1] The process involved in this model can be represented by a schematic cycle containing the four R's, namely *Retrieve, Reuse, Revise and Retain*. First, cases are retrieved from the case base which are similar to a new given problem. The old case with a solution will be reused and modified if necessary, an evaluation of the solution will be handled in the Revise step and finally a new case complements the knowledge base in the Retain phase. According to Janet Kolodner a case can be defined as: *"(i) a situation and its goal, (ii) the solution and, sometimes, means of deriving it, (iii) the result of carrying it out, (iv) explanations of results, and (v) lessons that can be learned from the experience."* [5] Anyway, Kolodner also stated that a case can be seen as a *"contextualized piece of knowledge representing an experience that teaches a lesson fundamental to achieving the goals of the reasoner"*. [5]

In CBR, we distinguish three different approaches: conversational, textual and structural. The conversational approach has the intention to provide solutions for many recurring simple problems. Predefined phrases -such as 'Have you tried to turn it off and on again ?' in first instance- will support a user to obtain a solution. These supporting phrases will be shown in the order of their importance for the given new

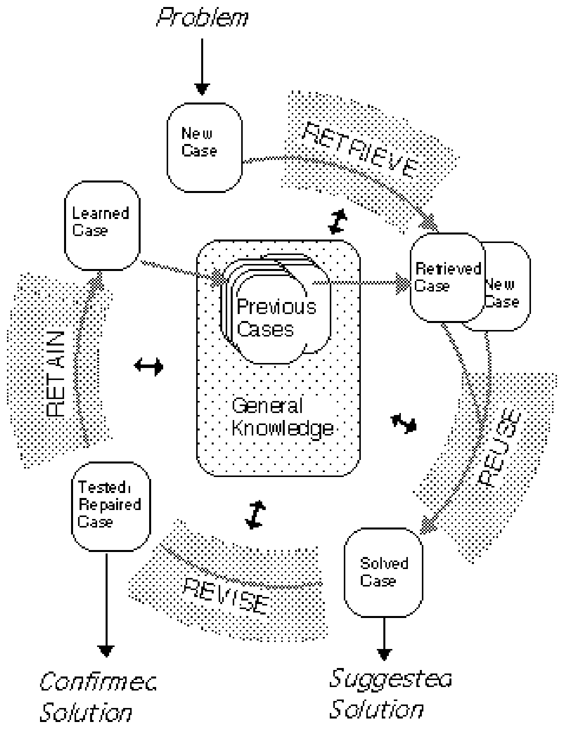


Fig. 1. R^4 model [1]

problem. The case base will be manually organized by the developer, while questions and phrases will be sequentially asked according to a decision tree which must be maintained when adding a new case. The textual CBR approach will be used for many documents which were analyzed concerning their content. The case base should not be greater than a couple of hundred cases, each case containing a short description with three lines. This approach should be aware of synonyms and associations between different terms. The structural CBR approach covers systems which are using a domain model. Therefore, predefined attributes and their representation should be chosen at the beginning of the modeling process. [6]

III. Previous Work

Different issues were researched such as association models and similarity measures. The gained knowledge of the association models were used to model a case base will be regularly used within the retrieve and retain step of the R^4 model. [1] Associations were obtained with the Hotspot algorithm of WEKA (Waikato Environment for Knowledge Analysis). [7] It is a target attribute driven algorithm which clearly presents the associations for a given target in a decision tree like structure. Arguments for this algorithm are the supported segment size of the data, the branching factor (on the top level) and the target attribute. The associations were partially published within [8]. Similarity measures are an ongoing topic and under development. According to the R^4 model they will be mainly used within the retrieve step. [1] Within the following lines they will be briefly described. The distance between two loans regarding the amount can be calculated when using the attribute amount, but to get the nearest cases -within the retrieve step- more attributes has to be considered such as age, purpose and credit history of a customer. Using a similarity measure will be more suitable in comparison to simple distances for these kind of loan cases. Similarity measures which encapsulates support for different attributes were modeled and will be tested due to different aspects such as non-negativity and range, reflexivity and positiveness. Weights will be used in addition if the will improve the functionality of these measures. A few pitfalls were avoided such as the difference between a distance metric and a similarity measure. For instance, distance=0 is equal to similarity=1. Both distance value and similarity value must be greater or equal than zero, but the range of a similarity value ends with 1.

IV. Definition of Problem, Solution and Case

First definitions were made which was a pre-condition for further work regarding the proof of concept. An example for a minor query would be following given problem which contains six attributes which are describing a loan application of a customer.

Problem = {Age, Credit Amount, Credit History, Duration, Income, Purpose}

A query can be extended with an attribute such as guarantors of the debtor. Extending a query towards the prototype will be suitable when the desired data of a customer is available within her or his loan application.

Problem = {Age, Credit Amount, Credit History, Duration, Income, Other Debtors Guarantors,Purpose}

A solution can be abstractly defined with two parts.

Solution = {Cost Factor, Recommendation}

The cost factor can be divided into different elements such as a percentage value of the predicted repayment, an absolute value concerning the amount of an assumed financial loss and a nominal value (e.g. 1 - 5) which describes the cost of this loan. The recommendation can be divided into subparts like a solution quality factor which will be given within the evaluation procedure within the revise step of the R^4 model and a real recommendation regarding the loan query of the customer. In the most efficient representation, the loan recommendation would be a boolean value which will be suitable on the top level for an employee of a financial institute. A case will be a triple of three elements in the minimal form.

Case = {Problem, Solution, Notes}

However, further allocation will be made when developing the prototype. For instance, notes can be used as a relation or as a character large object attribute. These definitions were partially published within [8].

V. Consideration of Missing Values

Unfortunately missing values can affect processing a case within different tasks such as retrieve a case from the case base and reuse a case, for instance. During the work on associations it was obvious that the attribute income was not explicit mentioned within the data definition of the German credit data set. [9] However, income was chosen as a possible attribute for new problems (or queries) which are submitted to the prototype because newer requests can and should provide this information which can be used for the pre-processing and reasoning. The Oracle Database provides a rather good function namely nvl (null value substitution) but for certain cases an implementation concerning a given domain has to be made. Although the income is missing within the German credit data set but this was not a reason to avoid this attribute within the definitions of a small query for a new given problem. Different strategies can be used to minimize the effect of missing values, for instance attribute income.

1) Substitution - Replace the missing value with an estimation: Generating an estimated value for the attribute income can be done with reasoning from other attributes such as the duration of the current employment and the amount of cash on the account of the customer which would be a rough estimation. An estimation function like this can be improved with additional knowledge provided by other attributes like country, job, age and a sub function which returns a range for a given job for a person within a given region of a country.
2) Overload methods to gain other queries - Using internal another query which can be made with using overloading of functions within the code. If the income is missing, then another query will be used with the same attributes except the attribute which refers to income.
3) Using social networks - Retrieving data by application programming interfaces (APIs) from social networks would be another feature but it would somewhat less than perfect. Many social networks are containing fake profiles or orphaned profiles. An additional preprocessing would be necessary to distinguish between fake profiles and real persons. The APIs of social networks are different and another issues like different e-mail-addresses for a person have to be considered if this kind of support really would be used. Extracting data concerning solvency from social networks was an upcoming issue in approx. 2010 and later but using this kind of approach was too buggy and rather subpar concerning many missing values and assumptions made in another approach during the work on this paper. For instance, an assumption was defined as follows: If the address of a house was evaluated as a real address by a Google Maps API call, then the profile of the social network will be classified as a real person which is comprehensible (if a person has not lied regarding the physical address). Another discussed assumption was to check relations to other human beings within the social network and use the gained knowledge to predict solvency for a person which is not suitable. [10]

VI. To Adapt or not to Adapt

Certain pre-conditions were to resolve before developing an adaptation feature which was enumerated within the previous sections. An awareness about the domain was reached with association models. The definitions of case, problem and solution was the basis for the case base. Similarity measures are an essential part of the adaptation procedure. Removing missing values avoids unnecessary adaptations because these could decrease the quality of the case base and increase the runtime cost of the software application. The solution quality factor (SQF) refers to 'TBA - to be announced' because the evaluation will be made within the revise step, but adaptation will be made earlier in the reuse step within the R^4 model. [1] The main idea will be to use a kind of quality factor for a proposed solution which is given by a user. A comparison between the proposed solution and previous solutions can affect and adapt attributes of a new subpar solution. If a similar solved problem exists according to the used similarity measures and the previous solution was marked with a rather good solution factor within the evaluation of a user, then the previous solution can be partially used as a basis for the adaptation of the new solution. Changing the internal queries towards the case base can provide another solutions, as appropriate, which can be used for a comparison with the suggested new subpar solution and adaptation if suitable. Defining a threshold value concerning the similarity measure regarding the new problem and retained cases (especially alternative solutions within these cases) has to be made.

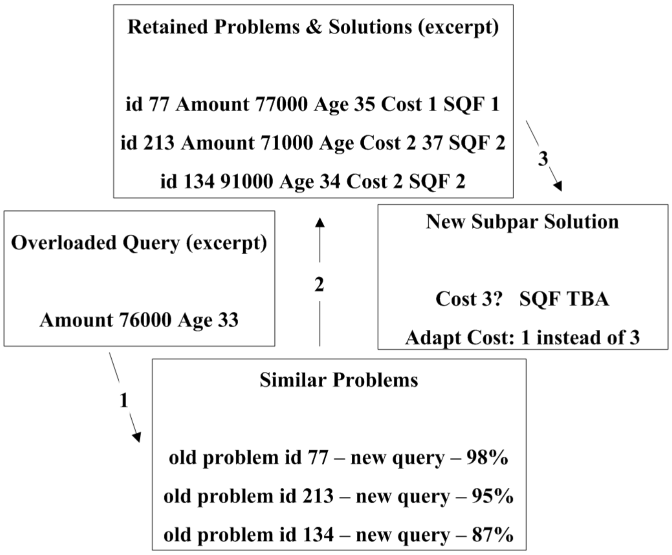

Retained Problems & Solutions (excerpt)
id 77 Amount 77000 Age 35 Cost 1 SQF 1
id 213 Amount 71000 Age Cost 2 37 SQF 2
id 134 91000 Age 34 Cost 2 SQF 2
3
Overloaded Query (excerpt)
Amount 76000 Age 33
2
New Subpar Solution
Cost 3? SQF TBA
Adapt Cost: 1 instead of 3
1
Similar Problems
old problem id 77 – new query – 98%
old problem id 213 – new query – 95%
old problem id 134 – new query – 87%

Fig. 2. Adaptation steps

Firstly, an internal overloaded query has to retrieve another solutions if available. Secondly, a similarity measure, according to the arguments of a modified query, can order the alternative solutions. Thirdly, these retrieved solutions will be used to modify attributes of a subpar solution. Within the reuse step, testing a modified solution can be made again with a similarity measure.

VII. Experiments

When comparing similarity values between a new searched solution and an initial query (also known as problem), it was clear that a similarity measure with less attributes -in comparison to a former used similarity measure- could impair the quality of the case base. Therefore, a similarity measure should use attributes comparable to the initial given query. Otherwise the evaluation feature of a CBR system will be required. For instance, a similarity measure with only two attributes (age and credit amount) would deliver a subpar similarity result (52 per cent) when comparing a retained case (age 58, credit amount 6143) with a random query (age 27, credit amount 10467). However, interesting alternatives can be found when using a simpler similarity measure. Using weighting of attributes must be carefully considered to keep a good precision of search results within retained cases.

VIII. Evaluation by Teacher

According to Aamodt and Plaza, the assessment of a new solution by a user was defined as evaluation by teacher within the revise step of the R^4 cycle. [1] This concept clearly provides an advantage that probably wrong data or subpar solutions can be modified. The solution can be marked as helpful with a degree from A - Excellent to E - Not helpful. Evaluation by teacher can be seen as improving a CBR software application, but it can not circumvent the adaptation procedure of the reuse step. A manual repair step made by a user would be possible according to the R^4 model by Aamodt and Plaza, but this would not be suitable for every single case if a big volume, velocity and variety of data will occur. Current tendencies such as big data within the future can not be precluded.

IX. Conclusion

Adaptation of cases was the core of this work in progress paper. The adaptation was achieved when overloading queries. There exists a significant difference between different queries regarding the precision of the result which leads to different solutions.

1) The Good - additional data -if not redundant- can be an enrichment for the case base if used in a proper way.
2) the Bad - missing values can hide the actual nearest case.
3) and the Ugly - neglect both adaptation and evaluation would be subpar concerning the final solutions of a case. [11]

Testing and evaluation of new and adapted solutions should be made within the reuse and revise step of the CBR cycle to keep the quality of a case base. Automatically testing of an adapted solution fits to the reuse step, a manual evaluation of an adapted solution fits to the revise step within the CBR cycle.

X. Future Work

Many issues are open like finishing the work on similarity measures, adapt cases to improve solutions, develop an evaluation by teacher component and integrate all of these parts within one prototype. Another interesting point to research will be a deletion strategy to avoid inflating the case base with many (too) similar cases which affects the efficiency of a reasoning process.

The following real world example clearly shows a motivation to model and implement a deletion strategy for both too similar and redundant cases. Boeing has obtained more than eighty million flight hours after ten years which resulted in 23000 troubleshooting reports submitted by SNECMA Services. The maintenance of their engines was supported by a CBR system. At a certain point, they have retained too many similar and redundant cases because a deletion strategy was missing at the begin within their software application. An employee was used to check and remove, if necessary, manual redundant cases *at the rate of 15 cases per hour*. At the end, their system contained 1500 *"clean"* cases. [12]

Acknowledgment

Grateful acknowledgement for proofreading and providing hints to improve the paper go to Yuliya Nebylovych.

References

[1] A. Aamodt and E. Plaza, "Case-Based Reasoning: Foundational Issues, Methodological Variations and System Approaches," *AI Commun.*, vol. 7, no. 1, pp. 39–59, 1994.

[2] K. Hammond, "Case-Based Planning: A Framework for Planning from Experience." *Cognitive Science*, vol. 14, pp. 385–443, 1990.

[3] R. Schank, "Dynamic Memory: A Theory of Learning in Computers and People," *New York, Cambridge University Press*, 1982.

[4] J. L. Kolodner, "Reconstructive Memory: A Computer Model," *Cognitive Science*, vol. 7, 1983.

[5] R. Bergmann, J. L. Kolodner, and E. Plaza, "Representation in Case-Based Reasoning," *Knowledge Eng. Review*, vol. 20, no. 3, pp. 209–213, 2005.

[6] J. Hönigl, H. Kosorus, and J. Küng, "On Reasoning within Different Domains in the Past, Present and Future," in *23rd Database and Expert Systems Applications (DEXA), 2012. 2nd International Workshop on Information Systems for Situation Awareness and Situation Management - ISSASiM'12*, September 2012.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The Weka Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, Nov. 2009. [Online]. Available: http://doi.acm.org/10.1145/1656274.1656278

[8] J. Hönigl and Y. Nebylovych, "Building a Financial Case-Based Reasoning Prototype from Scratch with Respect to Credit Lending and Association Models Driven by Knowledge Discovery," *Central & Eastern European Software Engineering Conference in Russia*, November 2012.

[9] A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml

[10] M. A. Stetco, "Creditworthiness Analysis using Data Gathered from Social Network Sites using a Supervised Learning Approach," 2012, Master Thesis, Johannes Kepler University, Linz, Austria.

[11] S. Leone, "The Good, the Bad and the Ugly. il buono, il brutto, il cattivo. (original title)," 1966.

[12] R. Bergmann, K. D. Althoff, S. Breen, M. Göker, M. Manago, and S. Wess, *Developing Industrial Case-Based Reasoning Applications: The INRECA Methodology*. Springer Verlag, 2003, vol. Lecture Notes in Artificial Intelligence Berlin, LNAI 1612, Berlin.

Shape Feature Extraction for On-line Signature Evaluation

Jungpil Shin
School of Computer Science and Engineering
The University of Aizu
Fukushima, Japan
e-mail: jpshin@u-aizu.ac.jp

Weichen Lin
School of Computer Science and Engineering
The University of Aizu
Fukushima, Japan

***Abstract*—In the past few years, banks and companies have increased security by switching from simple static passwords to more dynamic security measures that offer greater protection for users of mobile and web commerce. The most personal method for authentication is analysis of handwritten signatures. Signature evaluation determines whether an individual's signature is considered "good" or "bad." A good signature is more complex and difficult to impersonate, whereas a bad signature is simple and easy to impersonate. A signature typically contains many angles, whether big or small. A signature with a higher number of angles is more complex and is considered "good"; therefore, the number of internal angles was calculated to determine the quality of the signature. In this paper, geometry was used to decide the kinds of angles to analyze. After the evaluation, verification was performed to analyze the EER (Equal Error Rate), and it has been concluded that it is best to use the Neighbor method to create the best signature evaluation system.**

Keywords-shape feature; signature verification; human–computer interaction

I. Introduction

Handwriting is the most natural way to enter text, and it allows users to replace the keyboard or mouse for input. Many kinds of input devices are available for handwriting, including the TouchPad, the pressure-sensitive tablet, the PDA touch panel, or other input panels. Online handwriting verification systems analyze the signature as a series of coordinate points, which are based on the writing movements made with a pen, in relation to the trajectory of the coordinates on the XY axis, including where the nib started to write, as well as the number of writing strokes [1]. The series of coordinates will first undergo pre-processing to remove noise in the handwriting and to reduce the variability of handwriting, including repetition of the coordinates of the point, smoothness, and size. Then, features are extracted around the contours of characters to identify illegible ligatures, with allowances for different strokes due to different writing habits of different users. Offline handwriting recognition systems allow users to write on paper, from which the image is scanned using a scanner or a camera. However, offline systems are unable to obtain any dynamic feature; therefore, it is more difficult to recognize the signature. The typical flow of the signature evaluation system includes image input, image recognition, and result output. Using images from devices such as digital cameras and digital scanners, image processing is performed in three steps: image pre-processing, feature extraction, and recognition. The available methods for offline signature recognition are based on a wide range of concepts. The research can be categorized according to the way that it handles the problem, as methods based on holistic, regional, and local properties.

Character feature extraction is the basis of character recognition. It is one of the most popular research topics in pattern recognition and is widely used in many areas, such as edge extraction, character learning, automatic letter sorting, and automatic license plate recognition. In license plate character recognition, some character feature extraction methods are used, including outline feature extraction and coarse grid feature extraction.

Signature verification is categorized into two main types: static and dynamic [2]. The static type, which is also known as offline verification, uses a scanner or a digital camera to obtain the image of the signature to be verified. The dynamic type, which is also known as online verification, verifies a signature that is entered using a digital pen and a tablet PC. The difficulty in signature recognition is that handwriting is affected by complex personal factors; therefore, the same character could have slightly different shapes.

Signature evaluation systems are used to analyze human signatures whether they are categorized as good or bad. If a signature is too simple or easy to replicate, it is considered a bad signature. A good signature must contain several angles or many strokes. In this paper, an evaluation system was used to determine if a signature is good or bad.

Online signature system can extract additional human-writing parametrics based on a time function (e.g., position trajectory, velocity, acceleration, pressure, direction of pen movement and azimuth); whereas, offline systems evaluate signatures using only scanned images. For this reason, online signature systems are much better than offline signature systems.

Most of previous research on signature evaluation focused on the development and the implementation of new algorithms. This paper discusses the analysis and evaluation of Chinese kanji from the "One Hundred Family Names" (百家姓), a collection of characters that are often used in Chinese family names.

If internal angles are used when evaluating good and bad signatures, a character could contain too many angles between 0° and 180°. In order to define the optimal angles, a method was used to determine how many angles usually appear in a character; therefore, 100 characters were selected and all of the angles that typically appeared in them were calculated.

II. RELATED WORK

Each person has a different signature, which could be slightly different in shape in different situations. To process these signatures, they need to be standardized based on some method.

Previous research usually focused on signature verification by using matching techniques based on dynamic time warping (DTW) [3], hidden Markov model (HMM) [4], and support vector machine (SVM) [5]. These techniques are useful for static signature verification, as well as dynamic verification when the signature image is combined with several features based on average speed XY, signature width, signature height, pen direction, number of strokes, pen azimuth [6], etc.

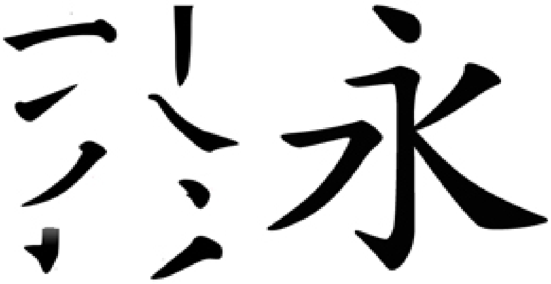


Figure 1. Eight principles of Yong

III. SIGNATURE EVALUATION METHODOLOGY

A. Eight Principles of Yong

The proposed kanji signature evaluation process is used to recognize an individual's handwritten signature whether it is considered good or bad, by applying the method only to internal angles. Twenty characters from the "One Hundred Family Names" collection were selected to calculate the averages on which to base the evaluations. Each character contains various angles, which were documented and compared. In the past, all characters were created with eight strokes; for example, this kanji contains eight strokes"永": 點, 横, 豎, 勾, 彎, 提, 撇, and 捺, where each stroke has different characteristics. Statistics on every stroke's angle can be obtained, as well as statistics on the angles between the lines (Figure 1).

B. "One Hundred Family Names" (百家姓)

The "One Hundred Family Names" (Chinese: 百家姓; pinyin: Bai jia xing) [7] is a classic Chinese book composed of common surnames in ancient China. Based on these family names, the average angles of the characters could be calculated.

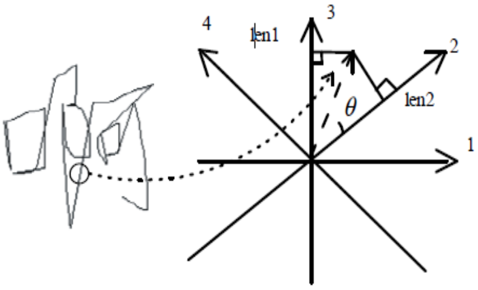


Figure 2. Internal angle

C. Internal Angle

An interior angle (or internal angle) of a kanji character is an angle formed by two connected lines [8] of a polygon stroke (Figure 3). Some people have a habit of writing kanji where the closed polygon is less than 180°; in that case, the polygon is called "convex." More complex signatures have more varied angles that were used for the calculations. The following are the different types of internal angles typically found in kanji:

- Equiangular: All angles are equal.
- Cyclic: All corners lie in a circular format.
- Vertex-transitive: All corners lie within the same symmetry orbit. The polygon is also cyclic and equiangular.
- Edge-transitive: All sides lie within the same symmetry orbit. The polygon is also equilateral.
- Tangential: All sides are tangential to an inscribed circle.
- Regular: The polygon is both cyclic and equilateral.

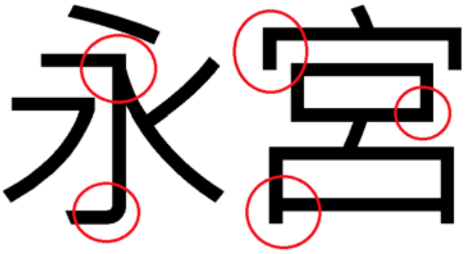


Figure 3. Internal angles in kanji

With simple characters and signatures, angles that were deemed important were manually searched for and selected. With complex signatures, a polygon algorithm was used to calculate the complex angles (Figure 2).

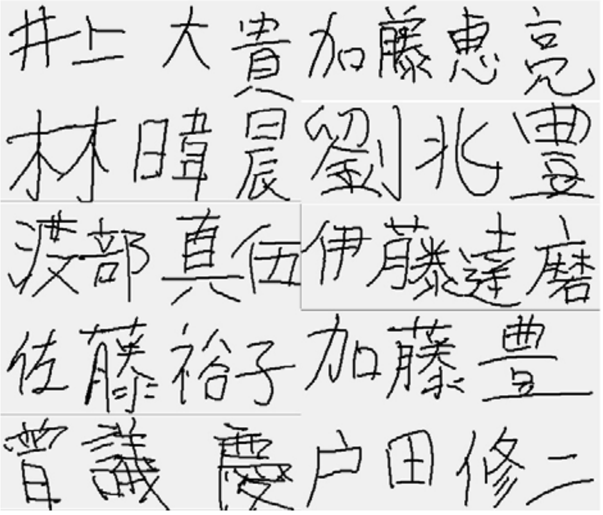


Figure 4. Example of signatures of twenty people

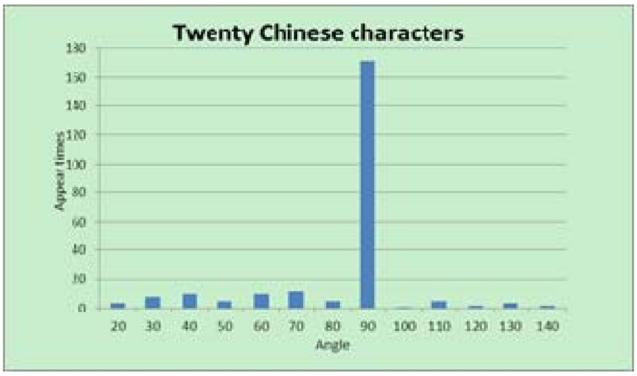


Figure 5. Analysis of internal angles occurring in twenty Chinese characters from “One Hundred Family Names”

D. Geometry

Geometric features [9][10] are based on multi-stroke recognizer techniques and are extracted to collect the internal features of the characters. Angles whose sum is a right angle (90°) are called “complementary.” Complementary angles are formed when one or more rays share the same vertex and are pointed in a direction between the two original rays that form the right angle. The number of rays between the two original rays is infinite.

Angles whose sum is a straight angle (180°) are called “supplementary.” Supplementary angles are formed when one or more rays share the same vertex and are pointed in a direction between the two original rays that form the straight angle. The number of rays between two original rays is also infinite.

Although the geometry method is a useful algorithm, the computer is unable to find some of the angles in a kanji character. In the end, all useful angles were manually collected.

E. Measuring Angle

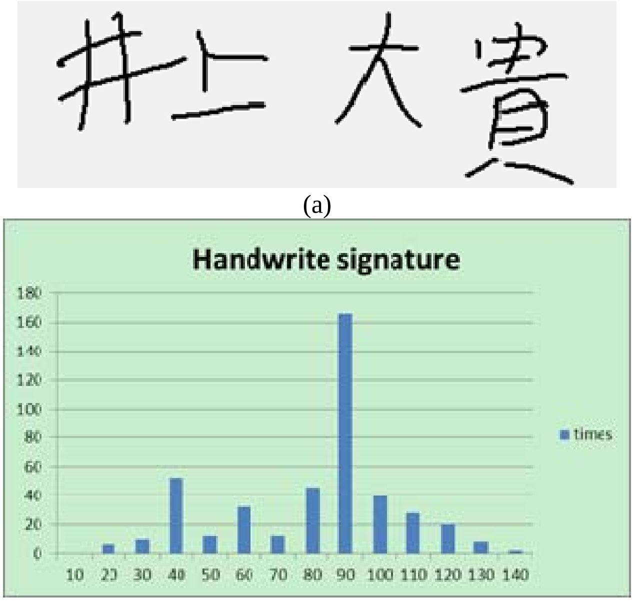
The following data were used in the analysis:

1) Eight Principles of Yong: This basic character set was used as a reference to obtain statistics on all the angles that usually appear in kanji characters.

2) “One Hundred Family Names” (百家姓): Twenty characters were randomly selected, and statistics were generated on the most common angles found in them.

3) Own Individual Signatures: Twenty test subjects wrote their own signatures, and statistics were also generated on the most common angles found in them. In addition, the analyses on handwritten and printed characters were compared.

4) Unfamiliar Signatures: The same twenty test subjects were asked to handwrite newly designed signatures that were unfamiliar to them and to handprint other characters. This set of experimental data was used to determine whether copying an unfamiliar signature would create different internal angles.



(a)

(b)

Figure 6. (a) Handwritten samples of test subjects’ own signatures with simple patterns (b) The analysis of internal angles occurring in those signatures

III. EXPERIMENTAL RESULTS

Figure 4 shows an example from our database of signatures of people who handwrote their own names.

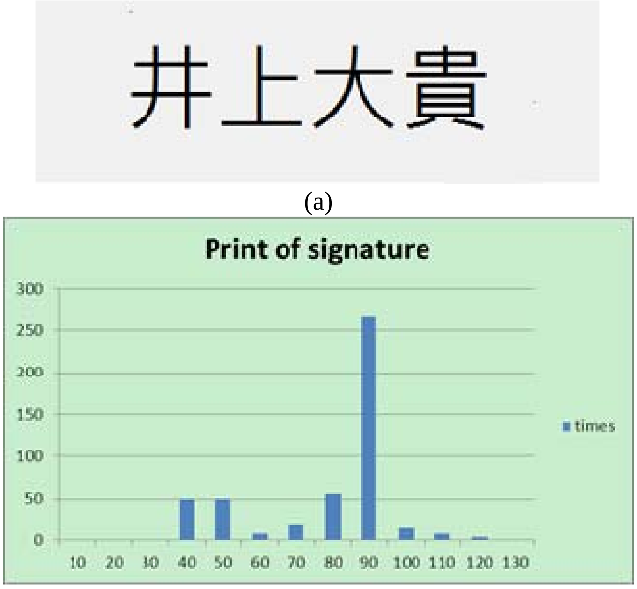
A. Analysis of Twenty Printed Chinese Characters

Twenty Chinese characters obtained from the “One

Hundred Family Names" collection were analyzed, as shown in Figure 5.

B. Analysis of Own Handwritten Signatures

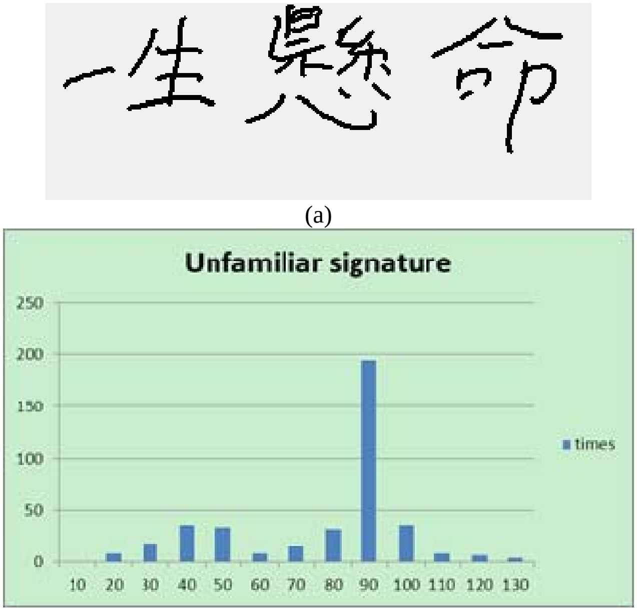
The analysis of the handwritten signatures occurred in two steps. First, twenty people handwrote their own signatures and those were analyzed (Figure 6). Then the results were compared with the analysis of the computer-printed signatures (Figure 7).



(a)

(b)

Figure 7. (a) Computer-printed samples of test subjects' own signatures with simple patterns (b) The analysis of internal angles occurring in those signatures



(a)

(b)

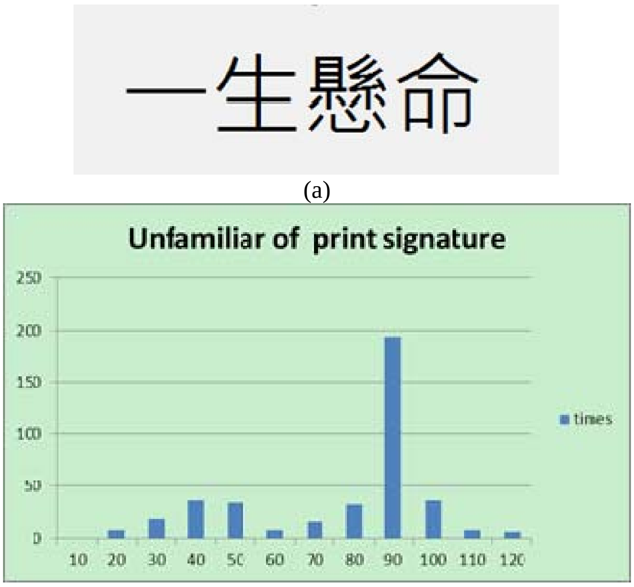
Figure 8. (a) Hand-copied samples of unfamiliar signatures (b) The analysis of internal angles occurring in those signatures

C. Analysis of Unfamiliar Signatures

Analyzing people's own signatures is not enough; therefore, in order to collect additional data about character angles, the research included an analysis of signatures that were unfamiliar to people who were writing them (Figures 8 and 9).

D. Chinese and English Character Angles

For this research to be more complete, the same analysis was performed not only on Chinese character signatures, but also on English characters from A to Z and English signatures, as shown in Tables 1 and 2.



(a)

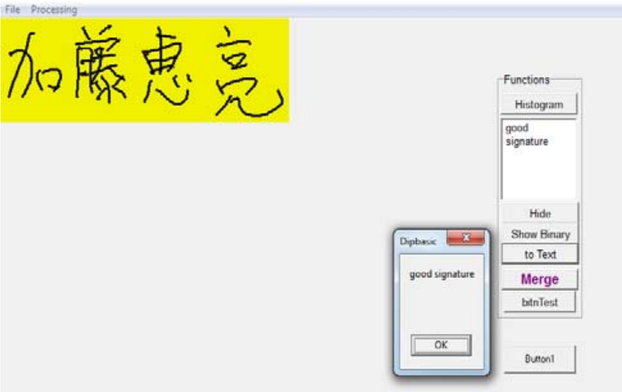
Figure 9. (a) Computer-printed samples of unfamiliar signature with simple patterns (b) The analysis of internal angles occurring in those signatures

TABLE I. ANALYSIS OF ALL ANGLES FOUND IN COMMON CHINESE CHARACTERS

	40°	50°	60°	70°	80°	90°	100°
Handwritten Sig.	52	12	32	12	46	166	40
Twenty Chinese	16	10	16	18	4	160	2
Print of Sig.	50	50	8	18	56	266	14
Unfamiliar Sig.	36	34	8	16	32	194	36
Print of Unfamiliar	2	5	0	0	1	12	0
Total	156	111	64	64	139	792	92

TABLE II. ANALYSIS OF ALL ANGLES FOUND IN COMMON ENGLISH CHARACTERS

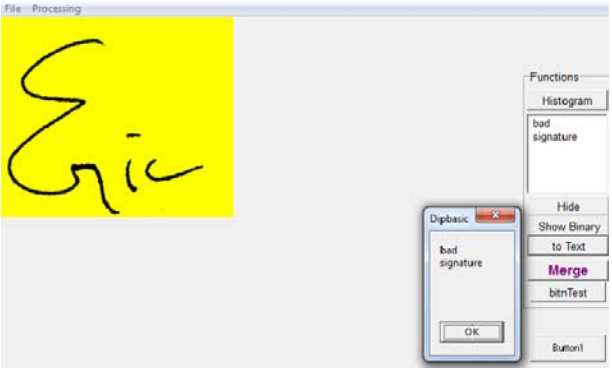
	10°	20°	30°	40°	50°	90°	120°
English A–Z	80	46	12	54	80	120	50
English Sig.	30	43	50	20	18	200	20
Total	110	89	62	74	98	320	70



good signature

good signature

Figure 10. An example of a good signature



bad signature

bad signature

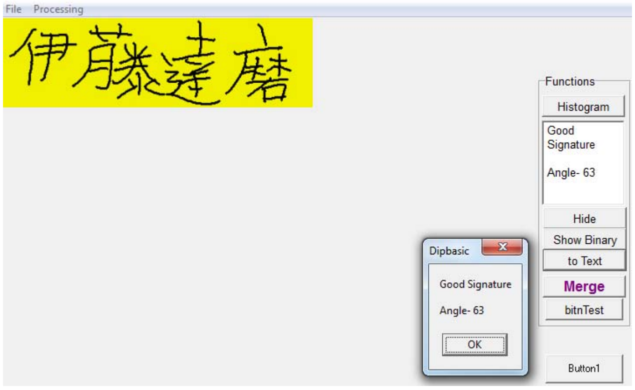
Figure 11. An example of a bad signature

V. EVALUATION SYSTEM

In the proposed evaluation system, the signature could be considered good and bad, as shown in Figures 10 and 11.

A. Comparison with Nearest Neighbor method and BPN

The analysis of the signatures was performed using the Neighbor method and the Neural Network BPN method. Although either of these methods would be effective, the two methods were still compared to find the best method. The results of the comparison are shown in Figure 12, Figure 13, and Table 3. We compared the Neighbor method and the Neural Network BPN method by analyzing Chinese basic characters, handwritten Chinese signatures, computer–printed signatures, unfamiliar Chinese signatures, and English signatures for analysis.



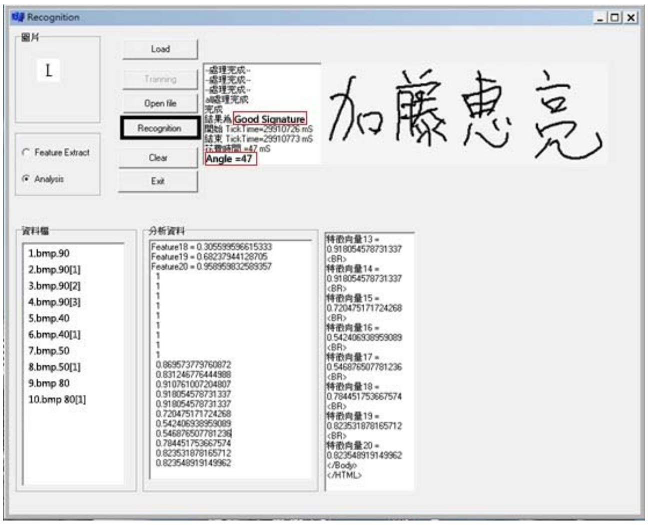
Good Signature

Angle- 63

Good Signature

Angle- 63

Figure 12. Angle of Neighbor method



```
--處理完成--
--處理完成--
--處理完成--
a處理完成
完成
結果為 Good Signature
開始 TickTime=29910726 mS
結束 TickTime=29910773 mS
花費時間 =47 mS
Angle =47
```

```
1.bmp.90
2.bmp.90[1]
3.bmp.90[2]
4.bmp.90[3]
5.bmp.40
6.bmp.40[1]
7.bmp.50
8.bmp.50[1]
9.bmp 80
10.bmp 80[1]
```

```
Feature18 = 0.305599596615333
Feature19 = 0.68237944128705
Feature20 = 0.958959832589357
 1
 1
 1
 1
 1
 1
 1
 1
 1
 0.869573779760872
 0.831246776444988
 0.910761007204807
 0.918054578731337
 0.918054578731337
 0.720475171724268
 0.542406938959089
 0.546876507781236
 0.784451753667574
 0.823531878165712
 0.823548919149962
```

```
特徵向量13 =
0.918054578731337
<BR>
特徵向量14 =
0.918054578731337
<BR>
特徵向量15 =
0.720475171724268
<BR>
特徵向量16 =
0.542406938959089
<BR>
特徵向量17 =
0.546876507781236
<BR>
特徵向量18 =
0.784451753667574
<BR>
特徵向量19 =
0.823531878165712
<BR>
特徵向量20 =
0.823548919149962
</Body>
</HTML>
```

Figure 13. Angle of neural network BPN method

Table 3 illustrates that the Neighbor method can find more angles than the Neural Network BPN method, because the Neighbor method scans all pixels; therefore, it can be more accurate in extracting angles.

On the other hand, the Neural Network BPN method uses pattern recognition. In this research, the neural network was taught a total of 20 pattern designs, covering all angles (40°, 60°, 70°, and 90°). Because of the limited number of learned patterns, the neural network is unable to find all angles in an image. However, given additional patterns to learn, the Neural Network BPN system would require more processing time.

In a set of twenty randomly selected characters from “One Hundred Family Names,” the most common angles are 40°, 60°, 70°, and 90°, as shown in Figure 5. In a set of twenty handwritten signatures, the most common angles are 40°, 60°, 80°, 90°, and 100°, as shown in Figure 6 (b). This set was also compared with the computer–printed versions of these characters to check whether these two categories would have similar analyses, as shown in Figure 7 (b).

TABLE III. COMPARISON RESULTS OF NUMBER OF ANGLES WITH NEIGHBOR AND BPN METHOD (A) CHINESE SIGNATURE (B) CHINESE CHAR. (C) COMPUTER PRINT (D) UNFAMILIAR AND (E) ENGLISH SIGNATURE

	(a)	**(b)**	**(c)**	**(d)**	**(e)**	**Total**
Neighbor	80	56	103	74	35	348
BPN	45	53	60	95	37	290

In order to increase the accuracy of the system, the same twenty people handwrote unfamiliar signatures that were not their own. Then those signatures were analyzed, as well as the computer-printed versions of the same characters, as shown in Figure 8 (b) and Figure 9 (b). In handwritten characters, the most common angles are 40°, 50°, 80°, 90°, and 100°.

To reduce the number of angles between 0° and 180° that the system would have to analyze, the analysis was focused on the five most common angles in Table 1: 40°, 50°, 80°, 90°, and 100°. In future work, the automatic signature verification system could use this research data to decide the best angles to analyze.

In addition to collecting and analyzing Chinese characters, computer-printed English characters (uppercase A to Z and lowercase a to z) and twenty handwritten English signatures were selected, and the most common angles in those samples are 10°, 20°, 30°, 40°, 50°, 90°, and 120°. Based on this result, the evaluation system could be designed to be more accurate, based on these angles and on the kind of characters that need to be evaluated.

TABLE IV. FRR AND FAR FOR RANDOMLY CHOSEN SIGNATURES

Total/Signature		**Good Signature**	**Bad Signature**
60 Angles	FAR	2/20 10%	8/20 40%
	FRR	2/20 10%	6/20 30%
80 Angles	FAR	1/15 6%	3/25 12%
	FRR	2/15 13.3%	5/25 20%
100 Angles	FAR	0/8 0%	2/32 6.25%
	FRR	0/8 0%	2/32 6.25%
Average total FAR		5.33%	19.4%
Average total FRR		7.76%	18.75%

B. Calculate FRR and FAR from Nearest Neighbor

To determine whether the system was optimal, random signatures were evaluated and categorized as good or bad.

In Table 4, we determined the false acceptance rate (FAR) and false rejection rate (FRR) values for randomly selected signatures. FRR is affected when users write their own individual signature, and FAR is affected when people forge someone else's signature. With bad signatures, FAR and FRR are expected to be high. If the verification of a good signature reduces FRR and increases FAR, it means that the system performs good evaluations. If the verification of a bad signature raises the FAR higher than the FRR, the bad signature is easy to forge. However, a low FRR value with a bad signature is not sufficient for evaluation. In this system, the research focused on evaluating good signatures, whether they are easy to forge or not. Sometimes, signatures can be successfully copied; therefore, this system is not foolproof.

VI. CONCLUSION

This paper focused on Chinese characters and signatures for evaluation and discussed the analysis of the handwriting of twenty people who wrote familiar characters and unfamiliar characters. These two categories of samples were compared with characters that appear in family names to calculate which characters appear more often. After the evaluation, a verification system analyzed the EER, and it is concluded that the Neighbor method creates the best signature evaluation system.

REFERENCES

[1] D. Y. Yeung, H. Chang, Y. Xiong, S. George, R. Kashi, T. Matsumoto, and G. Rigoll, "SVC2004: First international signature verification competition," in: Proceedings of the International Conference on Biometric Authentication (ICBA), vol. 3072, Springer LNCS, China, pp. 16-22, 2004.

[2] D. S. Guru and H. N. Prakash, "Online signature verification and recognition: An approach based on symbolic representation," IEEE Transaction on Pattern Analysis and Intelligence, vol. 31, no. 6, pp. 1059–1073.

[3] M. I. Khalil, M. N. Moustafa, and H. Abbas, "Enhanced DTW-based on-line signature verification," in: Proceedings of ICIP 2009, 2009, pp. 2713–2716.

[4] S. M. A. Ahmad, A. Shakil, and R. M. Anwar, "Stability and repeatability of HMM based probability outputs across dynamic handwritten signature features," IEEE 2008 International Symposium on ITSim Information Technology,

[5] N. Abbas, "Combination of off-line and on-line signature verification systems based on SVM and DST," IEEE International Conference on Intelligent Systems Design and Applications, 2011, pp. 855–860.

[6] G. Taherzadeh, "Evaluation of online signature verification features," International Conference on 2011 Advanced Communication Technology (ICACT), pp. 772–777.

[7] Y. Nakamura and M. Kidoe, "Extraction of individual handwriting characteristics from handwritten Chinese characters based on knowledge of handwriting analysis," Technical Report of IEICE PRMU2004-23,MI2004-23,WIT-23 (2005405).

[8] S. Huang, L. Jin, and J. Lv, "A novel approach for rotation free online handwritten Chinese character recognition," International Conference on Document Analysis and Recognition, 2009. ICDAR, pp. 1136–1140.

[9] K. Franke, "Analysis of Authentic Signature and Forgeries," in: Proc. IWCF, 2009, pp. 150–164.

[10] B. Fang, C. H. Leung, Y. Y. Tang, K. W. Tse, P. C. K. Kwok, and Y. K. Wong, "Off-line signature verification by the tracking of feature and stroke positions," Pattern Recognition, 2003, pp. 91–101.

An Expert-Driven Bayesian Network Model for Simulating and Predicting Software Quality

Lukasz Radlinski
Institute of Information Technology in Management
University of Szczecin
Szczecin, Poland
lukasz@radlinski.edu.pl

***Abstract*— The main goal of this work is to build an expert-driven Bayesian network model for simulating and predicting software quality. In contrast with earlier models, this model represents software quality as a hierarchy of features and their sub-features where the features are interrelated with other. It contains a range of project and process factors that influence particular quality features. It has been pre-calibrated using results from the questionnaire survey performed among software engineers and managers in various software organizations. Managers in software projects can use such model to simulate and predict various aspects of software quality, typically at the early stage of project lifecycle. Proposed may become a central part of the future decision support system aimed to analyze, understand, manage and optimize a software development process.**

Keywords- Bayesian network, modeling, software quality, expert knowledge, simulation, prediction

I. Introduction

Software quality prediction is an extensively covered area of software engineering. Various models have been developed to predict different features of software quality. These models typically focus on a single aspect of software quality, for example on number of defects [4], defect proneness [2], maintainability [15] or reliability [7]. However, software quality is a combination of various features that are interrelated with each other and influenced by other factors. Unfortunately, very few predictive models, discussed in Section 2, integrate multiple aspects of software quality.

Based on the review of existing models we decided to develop a new model that would overcome their limitations. The main requirements for of the new model are the following:

- Integration of variety of quality features along with their sub-features and measures;
- Integration of project and process factors that influence quality features;
- Incorporation of expert knowledge and empirical data;
- Ability to perform various ‘what-if’ and ‘goal-seeking’ as well as advanced simulations;
- Ability to run with missing data;
- Ability to adjust the model based on new knowledge or data by the end user.

Based on an earlier analysis of different modeling techniques [13], we decided to use a Bayesian network as a formal representation of the model. With Bayesian network it is possible to satisfy all of the above requirements.

The main goal of this paper is to present selected details of the new Bayesian network for integrated software quality prediction and simulation. The model can be used in numerous analyses by answering questions such as:

- How levels of effort in various development activities influence specific quality features?
- Given a typical distribution of effort, how do environmental project factors influence software quality?
- In a project with specific project factors, how much effort should we allocate to achieve some target levels of software quality?

Earlier work on this model has been already published in [9][10][12][13]. Since the model has been evolving for about two years this paper focuses on the most recent version that satisfies all requirements stated earlier in this section. Due to limited space, this paper focuses on new results and does not cover detailed background discussion and justification that have been published in earlier papers. This paper makes the following new contributions by providing:

1. A discussion of the preferences for the expected contents of the model according to the opinions of the respondents provided during questionnaire survey.
2. The details of the most recent structure of the model defined after the questionnaire survey and based on its results.
3. The behavior of this edition of the model by discussing the results of the validation process.

This paper is organized as follows: Section 2 briefly discusses the hierarchy of software quality and revisits earlier work. Section 3 investigates the respondents preferences on the expected scope of a new model. Section 4 explains the structure of the new Bayesian network model. Section 5 discusses the behavior of this model based on the results of the validation process. Section 6 covers limitations and threats to validity of obtained results. Section 7 draws conclusions and ideas for future work.

ISBN: 978-1-61208-254-7

II. Background

A. Software Quality

Software quality is a combination of various features. These features are often organized in a hierarchy. A variety of such hierarchies, known as software quality models, have been proposed in software engineering literature, starting from early work by Boehm and McCall at the end of 1970's. We use a hierarchy proposed in an ISO 25010 standard [6]. We chose it due to its popularity among researchers and in industry and because it has been published very recently but is based on an earlier 9126 standard – thus it can be considered as both mature and contemporary.

TABLE I. Hierarchy of software quality

Features	Sub-features
Functional suitability	Functional completeness Functional correctness Functional appropriateness
Performance efficiency	Time behaviour Resource utilisation Capacity
Compatibility	Co-existence Interoperability
Usability	Appropriateness recognizability Learnability Operability User error protection User interface aesthetics Accessibility
Reliability	Maturity Availability Fault tolerance Recoverability
Security	Confidentiality Integrity Non-repudiation Accountability Authenticity
Maintainability	Modularity Reusability Analyzability Modifiability Testability
Transferability	Adaptability Installability Replaceability
Effectiveness	Effectiveness
Efficiency	Efficiency
Satisfaction	Usefulness Trust Pleasure Comfort
Freedom from risk	Economic risk mitigation Health and safety risk mitigation Environmental risk mitigation
Context coverage	Context completeness Flexibility

This hierarchy assumes three levels of quality – characteristics, sub-characteristics, and measures. Table I lists the first two groups that we call features and sub-features in our study – by changing in these names we stress that, although our predictive model is based on the ISO 25010 hierarchy of software quality, it can be relatively easy adapted to a another hierarchy, i.e., taken from a different quality model.

B. Related Work

Very few predictive models integrate multiple software quality features and enable comprehensive quality prediction together with the ability to perform advanced simulations. Each of these models, besides important benefits, also has some disadvantages. Wagner [16] proposed a set of models – each for a separate quality feature. Thus, this approach does not provide an integrated model with relationships between quality features. Beaver [1] proposed a model which contains a variety of links between quality features. However, this model was developed using data only from very small student projects and thus does not generalize to larger industry-scale projects. Fenton et al. [3] developed a model that incorporates empirical data and expert knowledge from industrial projects and in which quality features are linked together. However, that model contains only two quality features.

Various authors [8][17] proposed approaches or frameworks to integrated quality modeling. They do not propose a working predictive model but rather a meta-model that integrates various concepts of software quality. It can be used to support the process of building a predictive model. Such approaches may seem to be useful for developing a larger knowledge base for populating predictive models from it. However, the process of building them is time consuming. Thus, in our work we develop a predictive model directly, i.e. without an overhead of such type of framework.

III. Analysis of the preferences on the expected scope of the model

To gather data required for calibrating a new model we performed a questionnaire survey among experienced software architects and project managers. Results from the main part of that survey have been discussed in [14]. Before that main part, we asked respondents to rate five predefined versions of the model with different structures.

The main differences between model versions were related to the model complexity and the number of variables. Table II summarizes these differences. Model A was the simplest and model E was the most complex. Models B, C and D were between models A and E in terms of their complexity.

TABLE II. Overview of model versions

Characteristic \ Model	A	B	C	D	E
# process activities	3	3	1	1	3
# of process factors per activity	3	3	18	18	18
# of project factors	0	3	6	3	6
# of quality features	8	13	8	13	13
# of levels in quality hierarchy	1	2	2	3	3
reflects software composition	no	no	no	no	yes

Table III summarizes ratings for different versions of the model. We investigated six criteria: clarity, complexity, coverage, adequacy, adaptability, and usefulness. The scale available for these criteria was a range of integer numbers from ‘1’ to ‘5’, i.e. from low to high level of intensity of a given criterion. For all criteria, except complexity, the most desirable value was ‘5’, i.e. that the model is clear, covers all important aspects, is adequate for a given environment, can be adapted relatively easy, and is useful. For complexity the meaning of the scale was slightly different – with a value ‘3’ being the most desirable, and the values above ’3’ indicating too high level of model complexity.

We aggregated the ratings provided by respondents by calculating a weighted mean for each model and each criterion (with the necessary adjustment for the complexity). We arbitrarily defined the weights based on respondent’s experience and motivation to participate in the survey. Table III shows the values of these weighted means, with the values closest to the most desirable value marked in bold.

TABLE III. RATINGS FOR MODEL VERSIONS

Model Criterion	Weight for criterion	A	B	C	D	E
Clarity	3	**4.4**	4.0	4.1	3.5	3.0
Complexity	2	1.9	2.5	**2.8**	3.5	4.2
Coverage	1	2.3	2.9	3.3	3.8	**4.4**
Adequacy	1	3.0	2.9	**3.2**	**3.2**	2.9
Adaptability	1	3.2	2.9	**3.3**	2.9	2.3
Usefulness	5	2.6	2.4	3.1	**3.3**	2.8
SCORE	–	2.67	2.72	**3.12**	3.01	2.55

By analyzing these ratings, we can conclude that none of the model versions won in all categories. In fact, all versions except ‘B’, won in at least one criterion. Model A was rated as very clear but too simple for most respondents. On the other hand, model E was rated as moderately clear but too complex. The overall rating has been calculated as the weighted mean of ratings for each model. Based on this value, we can conclude that the model C was rated as the best overall, while model E as the worst overall.

IV. STRUCTURE OF THE NEW MODEL

The structure of the model presented in this section is a slightly adjusted version of model ‘C’ that was rated as the best overall by the respondents. Since this model is a proof-of-concept, we decided to enhance model ‘C’ by extending the number of process activities to three and to use all 13 quality features from the quality model proposed in the ISO 25010 standard [6]. These enhancements not only provide more functionality of the model but also gave us an opportunity to investigate the model complexity in terms of the calculation time.

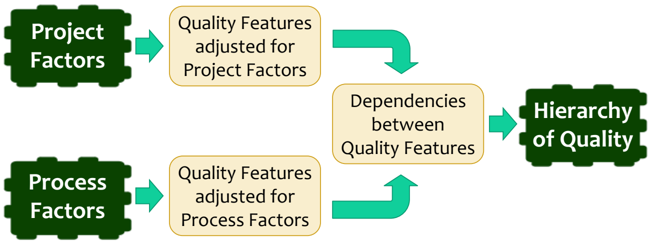

Project Factors
Quality Features adjusted for Project Factors
Dependencies between Quality Features
Hierarchy of Quality
Process Factors
Quality Features adjusted for Process Factors

Figure 1. Schematic of the proposed model

The high level schematic of this model has been illustrated in Figure 1. The core of the model consists of a set of quality features organized in a hierarchy of features, sub-features and measures, with some explicit links between main level features. These features are influenced by two groups of factors, i.e., project factors and process factors. The complete model structure and a ready-to-use model is available on-line [11].

Figure 2. illustrates the structure of the sub-network with project factors. The model contains seven project factors that describe the nature of the project. Project factors define the priors of quality features, i.e. default distributions, according to the information provided by the respondents during a questionnaire survey. A set of five ‘quality in use’ features (bottom of Figure 2) is much less influenced by project factors than remaining internal and external quality features. This is caused by the fact that quality in use strongly depends on the specific context/environment of use rather than on those project factors.

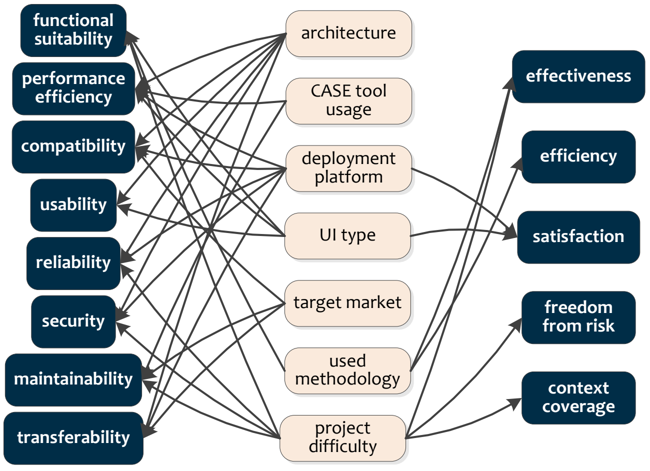


Figure 2. Structure of the sub-network with project factors and priors of quality features

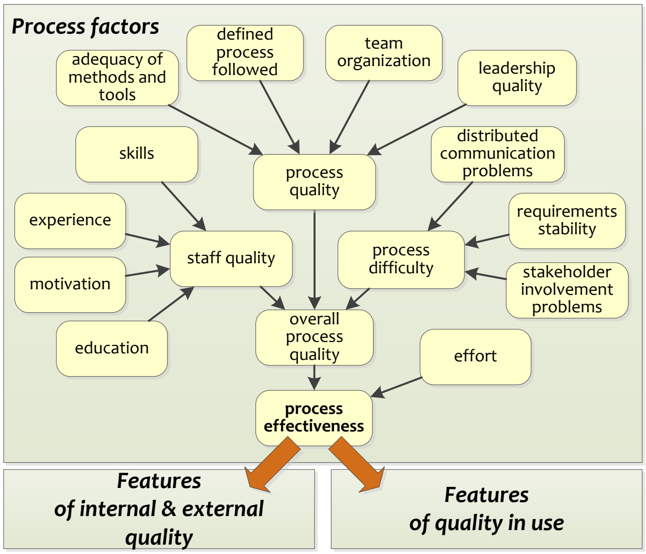


Figure 3. Structure of the sub-network with process factors

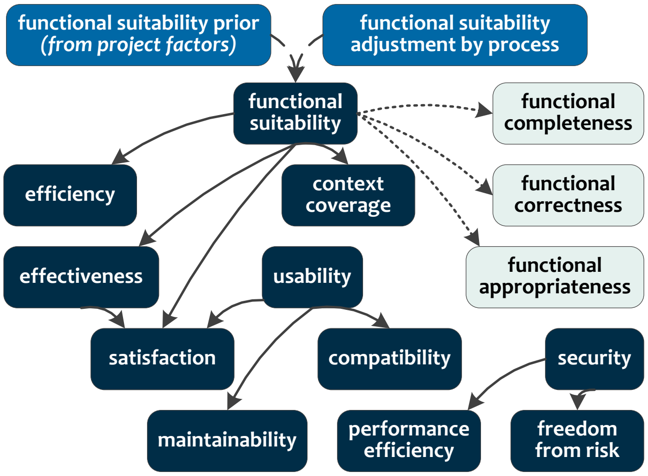


Figure 4. Links between quality features, influence of project and process factors, and a part of the quality hierarchy

Figure 3. illustrates the structure of the sub-network with the process factors. This structure has been strongly adjusted since a version used in the questionnaire survey [13]. The new structure of a converging star is clearer and more user-friendly. It enables easier adjustment directly by users – the variables that aggregate their parents are defined using expressions, most often as a 'weighted min' [5], thus adding or removing a parent variable requires only adjustment of that expression rather than manually rebuilding the whole probability table for the aggregate variable. The model contains three process sub-networks, one for each main activity, i.e. specification, development and testing.

Figure 4. illustrates the links between quality features, influence of project and process factors, and a part of the quality hierarchy. Each quality feature has its own hierarchy, i.e. a set of sub-features and measures, and is defined individually by project and process factors. Figure 4. shows all existing links between quality features but, due to a limited space, sub-features and influences from project and process factors only for an example feature – *functional suitability*. However, each quality feature is defined in a similar way with its own set of sub-features and links from sub-networks with project and process factors. Two quality features, reliability and transferability, are not directly linked with any other quality feature. It does not mean that these two features are not related with any other quality features but rather that there are no direct relationships.

V. Model Validation

To validate the developed model, we performed a variety of analyses of results provided by the model. In this paper, we discuss three of such analyses. Each of them investigates how the model behaves when an observation is entered to a single variable, i.e., what are the predictions for the other variables. Since the model is a Bayesian network, the predictions are provided not as point numeric values but as probability distributions. In our analyses, we investigated the whole probability distributions but to keep the paper concise we report the median values from predicted distributions.

All variables involved in this analysis are expressed on a 5-point ranked scale, typically from 'very low' to 'very high' but for some variables a reverse order of states is used. This ranked scale is internally transformed to a continuous scale where a state 'very low' represents a range [0, 0.2], 'low' a range [0.2, 0.4], etc. until the last state 'very high' represents a range [0.8, 1]. With such transformation it is possible to calculate statistical measures describing a probability distribution, including a median that we used in this paper.

In the first analysis, we investigated how a change of one quality feature influenced remaining quality features. First, we set an observation 'very low' to one quality feature and calculated the model. Second, we set an observation 'very high' for the same variable and calculated the model. Then, for each predicted variable we calculated the difference between median values from these two predictions (calculations) as shown in Equation 1.

$$\text{Difference}(\mathit{feature\_i}) = \mathit{Median}(\mathit{feature\_i}_{\text{prediction\_1}}) - \mathit{Median}(\mathit{feature\_i}_{\text{prediction\_2}}) \quad (1)$$

TABLE IV. Predicted relationships between quality features

Quality features	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) functional suitability	–	0.06	0.20	0.20	0.12	0.14	0.17	0.11	0.55	0.53	0.38	0.21	0.46
(2) performance efficiency	0.07	–	0.11	0.07	0.09	-0.29	0.10	0.05	0.13	0.09	0.10	-0.05	0.11
(3) compatibility	0.12	0.06	–	0.36	0.11	0.10	0.22	0.11	0.16	0.14	0.18	0.12	0.15
(4) usability	0.18	0.07	0.49	–	0.10	0.10	0.44	0.09	0.19	0.18	0.28	0.16	0.18
(5) reliability	0.12	0.08	0.17	0.11	–	0.20	0.22	0.17	0.18	0.18	0.17	0.15	0.19
(6) security	0.13	-0.26	0.15	0.11	0.18	–	0.20	0.16	0.15	0.17	0.15	0.42	0.18
(7) maintainability	0.10	0.06	0.22	0.32	0.13	0.13	–	0.11	0.16	0.13	0.16	0.14	0.18
(8) transferability	0.10	0.04	0.16	0.09	0.16	0.16	0.17	–	0.15	0.15	0.15	0.10	0.15
(9) effectiveness	0.38	0.09	0.17	0.14	0.12	0.11	0.17	0.11	–	0.37	0.29	0.17	0.29
(10) efficiency	0.34	0.05	0.14	0.12	0.11	0.11	0.13	0.10	0.35	–	0.21	0.14	0.25
(11) satisfaction	0.23	0.06	0.17	0.19	0.10	0.10	0.16	0.09	0.28	0.20	–	0.12	0.21
(12) freedom from risk	0.11	-0.03	0.11	0.09	0.08	0.25	0.12	0.06	0.14	0.12	0.11	–	0.16
(13) context coverage	0.32	0.07	0.15	0.12	0.11	0.11	0.17	0.09	0.29	0.25	0.21	0.19	–

TABLE V. PREDICTIONS FOR QUALITY FEATURES DEPENDING ON OBSERVATIONS FOR PROCESS FACTORS

Process factors \ Quality features	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**
spec. overall process quality	0.21	0.13	0.19	0.22	-0.01	-0.02	0.09	-0.01	0.30	0.18	0.20	0.27	0.27
spec. effort	0.22	0.14	0.27	0.23	0.18	0.18	0.24	0.19	0.33	0.26	0.27	0.25	0.27
spec. process effectiveness	0.47	0.31	0.48	0.49	0.14	0.12	0.31	0.14	0.57	0.47	0.49	0.53	0.59
dev. overall process quality	-0.01	0.09	0.10	-0.01	0.21	0.17	0.30	0.20	0.17	0.09	0.12	0.06	0.17
dev. effort	0.17	0.13	0.24	0.17	0.23	0.21	0.28	0.22	0.30	0.24	0.25	0.21	0.25
dev. process effectiveness	0.12	0.23	0.34	0.13	0.47	0.40	0.59	0.45	0.47	0.32	0.37	0.26	0.43
test. overall process quality	0.17	0.00	0.19	0.16	0.19	0.22	0.07	0.20	0.07	0.20	0.16	0.08	0.07
test. effort	0.21	0.11	0.27	0.22	0.22	0.22	0.23	0.23	0.28	0.27	0.26	0.22	0.25
test. process effectiveness	0.41	0.09	0.48	0.40	0.43	0.49	0.28	0.45	0.33	0.50	0.43	0.30	0.30

TABLE VI. PREDICTIONS FOR QUALITY FEATURES DEPENDING ON OBSERVATIONS FOR PROJECT FACTORS

Project factors \ Quality features	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**	**(8)**	**(9)**	**(10)**	**(11)**	**(12)**	**(13)**
architecture	0.00	0.06	0.06	0.06	0.01	0.07	0.03	0.06	0.00	0.00	0.01	0.03	0.00
case tool usage	0.00	0.03	0.00	0.00	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
deployment platform	0.00	0.17	0.11	0.00	0.11	0.02	0.00	0.04	0.00	0.00	0.02	0.01	0.00
UI type	0.12	0.05	0.13	0.29	0.00	0.00	0.12	0.00	0.05	0.05	0.17	0.00	0.04
target market	0.00	0.00	0.03	0.00	0.00	0.00	0.11	0.04	0.00	0.00	0.00	0.00	0.00
used methodology	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.20	0.04	0.00	0.02
project difficulty	0.12	0.07	0.01	0.01	0.10	0.14	0.19	0.01	0.04	0.09	0.02	0.23	0.22

Table IV provides those differences from this first analysis. Each row represents results for quality features given a change in quality feature shown in the first column. For example, a row marked as *(1) functional suitability* contains differences in predictions for remaining quality features than occurred as a result of setting *functional suitability* to ‘very low’ and ‘very high’. A value higher than ’0.2’ can be considered as representing a positive relationship between a pair of quality features, and a value lower than ‘-0.2’ as representing a negative relationship. These results show that the model usually properly incorporates relationships identified during a questionnaire survey and discussed in [14].

However, each quality feature is at least slightly related with other quality features although often no direct relationships exist in the model. This happens because and observation in one quality feature causes revised predictions for its parents, and some of these parents then influence other quality features (i.e. there are common causes for quality features).

In the second analysis, we investigated how observations set to selected process variables influence quality features. Table V provides results for this analysis and, similarly as in the first analysis, contains the differences between the median values for quality features depending on setting observations ‘very low’ and ‘very high’ to process factors. The numbers in the first row refer to quality features according to the numbering as in Table IV. Higher values confirm that the model incorporates a relationship between particular process factor and a quality feature. These relationships are consistent with those identified and discussed in [14].

In the third analysis, we investigated how project factors influence the quality features. Table VI reports the results in the same way as in two previous analyses. However, some project factors are not expressed on a ranked scale but have labeled states. For these factors we calculated predictions by setting an observation for all possible states of a project factor (one at the time). The values of these differences also confirm that the model properly incorporates the influence of project factors on quality features as identified in the questionnaire survey [14].

VI. LIMITATIONS AND THREATS TO VALIDITY

During the work on this model and its validation, we noticed several limitations and threats to validity of obtained results. First, the relationships between quality features, illustrated in Table IV, are not always symmetrical. During the questionnaire survey, we asked respondents about such relationships without investigating the direction of the link. When building a Bayesian network, which is a directed graph, we defined directions of such link typically according to the cause-effect relationship. However, this relationship is of stochastic nature and together with other links in the model it is not possible to define links between quality features that would be symmetrical.

Second, during a questionnaire survey, respondents identified relationships between specific pairs of variables. However, very often respondents did not provide information on the details of such relationship. Even further, there were cases when one respondent provided information on the strong positive relationship between two variables, whereas according to another respondent this relationship was negative. Thus, the model cannot incorporate all information gained because of these contradictory answers.

Furthermore, in this paper, we focus on model validation that involved a change of one variable at a time. We did not report results of analyses of scenarios where multiple variables were set with observations.

Finally, model applicability is limited to software projects which follow the rationale for this model. Specifically, this includes large and long-lasting projects with a full development lifecycle. The model can be applied to other projects but after significant adjustments which may cost-ineffective.

VII. Conclusions and Future Work

The developed Bayesian network model, discussed in this paper, is an extended and improved version of the model discussed in earlier papers. It has a simpler and clearer structure and still offers higher functionality due added useful variables. This model properly incorporates expert most knowledge gathered during the questionnaire survey as we confirmed it during the validation stage. A predefined Bayesian network model well fits the user expectations in terms of its scope, complexity and usefulness.

Although the model has been pre-calibrated, it may and should be recalibrated in the target environment. Depending on the user needs, this may involve adding or removing variables, adding or removing links, and changing the quantitative definitions of variables (e.g. the sensitivity of the changes between different variables). We believe that because of the modular structure and the usage of expressions [5] such adjustments are fairly simple.

In the future, we plan to extend the model by using detailed software measures, i.e., metrics. During the questionnaire survey we were aware that it would be difficult to obtain real data on them. We hope that, after presenting the results from the proof-of-concept model, the companies would be willing to cooperate tighter to calibrate the model to their own needs. We also plan to work on the tool support for the model, so that the model will be a part of a lager expert-based decision support system aimed to analyze, understand, manage and optimize a software development process.

Acknowledgments

I would like to thank all participants of the questionnaire survey for providing useful information and additional thoughts inspiring this and further research. I would also like to thank Magne Jørgensen and Richard Torkar from Simula Research Laboratory in Norway and Norman Fenton from Queen Mary, University of London for valuable comments on building a Bayesian network model discussed here. This work has been supported by research funds from the Ministry of Science and Higher Education / National Science Centre in Poland as a research grant no. N N111 291738 for years 2010–2012.

References

[1] Beaver JM. A life cycle software quality model using bayesian belief networks. Ph.D. Dissertation. University of Central Florida. Orlando; 2006.

[2] Catal C, Diri B. A systematic review of software fault prediction studies. Exp Syst Appl 2009, 36(4): 7346–7354.

[3] Fenton N, Marsh W, Neil M, Cates P, Forey S, Tailor M. Making Resource Decisions for Software Projects. In: Proceedings of the 26th International Conference on Software Engineering. Washington, DC: IEEE Computer Society; 2004, p. 397–406.

[4] Fenton NE, Neil M. A critique of software defect prediction models. IEEE Trans Softw Eng 1999; 25(5): 675–689.

[5] Fenton NE, Neil M, Caballero JG. Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks. IEEE Trans Knowl Data Eng 2007; 19(10): 1420–1432.

[6] ISO/IEC 25010:2011(E), Software engineering – Software product Quality Requirements and Evaluation (SQuaRE) – System and software quality models, 2011.

[7] Musa JD. Software Reliability Engineering: More Reliable Software Faster and Cheaper. Second Edition, Authorhouse; 2004.

[8] Nelson HJ, Poels G, Genero M, Piattini M. A conceptual modeling quality framework. Softw Qual J 2011; 20(1): 201–228.

[9] Radliński Ł. A conceptual Bayesian net model for integrated software quality prediction. Annales UMCS, Informatica 2011; 11(4): 49–60.

[10] Radliński Ł, A Framework for Integrated Software Quality Prediction using Bayesian Nets. In: Murgante B, Gervasi O, Iglesias A, Taniar D, Apduhan B, editors. Computational Science and Its Applications - ICCSA 2011. Lecture Notes in Computer Science; 6786, Berlin / Heidelberg: Springer; 2011, p. 310–325.

[11] Radliński Ł. Bayesian Network Model for Integrated Software Quality Prediction. 2012, http://lukrad.univ.szczecin.pl/projects/banisoq/.

[12] Radliński Ł. Empirical Analysis of the Impact of Requirements Engineering on Software Quality. In: Regnell B, Damian D, editors. Requirements Engineering: Foundation for Software Quality. Lecture Notes in Computer Science; 7195, Berlin / Heidelberg: Springer; 2012 p. 232–238.

[13] Radliński Ł. Enhancing Bayesian Network Model for Integrated Software Quality Prediction. In: Mauri JL, Lorenz P, editors. Proc. Fourth International Conference on Information, Process, and Knowledge Management, Valencia: IARIA; 2012, p. 144–149.

[14] Radliński Ł. Towards expert-based modeling of integrated software quality. J Theor Appl Comp Sci 2012 (under review).

[15] Riaz M, Mendes E, Tempero E. A systematic review of software maintainability prediction and metrics. In: Empirical Software Engineering and Measurement, Washington, DC: IEEE Computer Society; 2009, p. 367–377.

[16] Wagner S. A Bayesian network approach to assess and predict software quality using activity-based quality models. Inf Softw Technol 2010; 52(11): 1230-1241.

[17] Wagner S, Deissenboeck F. An Integrated Approach to Quality Modelling. In: Fifth International Workshop on Software Quality (WoSQ'07: ICSE Workshops 2007), Washington, DC: IEEE; 2007, p. 1.

Extraction of Semantic Relationships from Academic Papers using Syntactic Patterns

Akihiro Kameda, Kiyoko Uchiyama, Hideaki Takeda, Akiko Aizawa
National Institute of Informatics
Tokyo, Japan.
{*kameda, kiyoko, takeda, aizawa*}*@nii.ac.jp*

***Abstract*—Integrating concept and citation networks on a specific research subject can help researchers focus their own work or use methods described in prior works. In this paper, we propose a method to extract semantic relations from concepts and citation in the descriptions of related work. Specifically, we examined (i) topic-paper relations between research topics and reference papers and (ii) method-purpose relations between research topics. We also defined 15 lexico-syntactic patterns for the relation extraction. Results of experiments using a manually annotated dataset of 15 papers demonstrated the effectiveness of using the proposed lexico-syntactic patterns.**

Keywords*-*Relation extraction; Citation context; Knowledge extraction.

I. INTRODUCTION

Most researchers locate, read, and analyze relevant papers to investigate prior studies related to their research fields when they are about to narrow their subject of research or publish their findings in a paper. The objective of such prior study searches is to position their own work among problems in related works or their surroundings and to clarify their own predominance. This type of background work is crucial in terms of qualifying a paper for publication.

In this paper, we examine related work descriptions in academic papers to support such research activities. Our approach is based on a knowledge representation that integrates the citation structure and research concept network of related topics. The proposed representation contains two kinds of node – a Paper and a Concept nodes – and links that connect these nodes. In order to extract comprehensible mutual relations, we define a concept as a text span that is associated with a specific paper in the citation context. These text spans may include different linguistic units such as technical terms, verb phrases, and clauses. For example, to a research the theme “modeling text and citation together in a similar corpus”, technical terms such as “PHITS” and “PLSA”, noun phrases such as “influence propagation model”, and verb phrases such as “explain various phenomena related to linked structure of the corpus” are recognized as Concept Nodes (Fig. 1).

Related work descriptions include information such as methods used in prior studies, other themes addressing the use of these methods, and the original papers in which the methods were proposed. As such, the generated network of preceding studies can help researchers position their own studies properly into surrounding fields and related works. In the following, we propose our method of constructing a paper knowledge network and report the results of experiments conducted with a dataset of academic papers written in English.

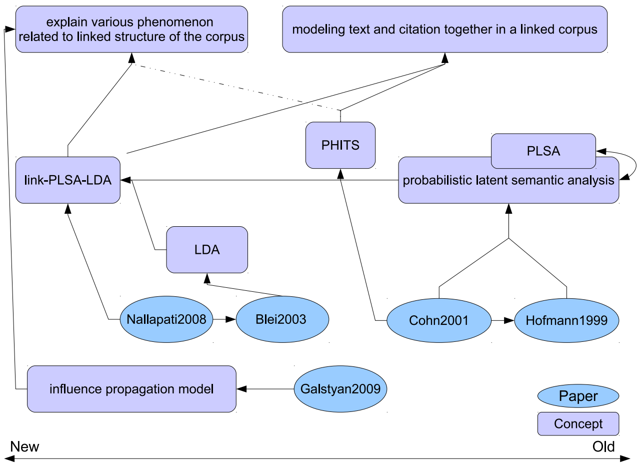

explain various phenomenon related to linked structure of the corpus
modeling text and citation together in a linked corpus
PHITS
PLSA
link-PLSA-LDA
probabilistic latent semantic analysis
LDA
Nallapati2008
Blei2003
Cohn2001
Hofmann1999
influence propagation model
Galstyan2009
Paper
Concept
New
Old

Figure 1. Example of a network representation of papers and concepts.

II. RELATED WORK

Zhang et al. [1] proposed a method to extract the relations of the key concepts of academic papers based on the clustering results. In their method, papers were first clustered and then keywords representative to each cluster were extracted. Then, a hierarchical structure of the keywords was constructed. While their study considered only the existence of a keyword in a target paper, our study looks into more detailed relations, such as whether a certain keyword is mentioned as a method or a research target.

Regarding the refinement of relations among papers, Nanba et al. [2] analyzed sentences that contain references to other papers and classified them into three categories: those citing other papers as their basis, those pointing out relative differences, and otherwise.

There is another similar framework in a study by Teufel et al.[3], where relations among papers were extracted based on the classification of the citation context. Dunne [4] also proposed a method to generate a targeted field overview by clustering the citation network. A devised interface that displays the summary of a citation context simultaneously enables users to obtain a detailed understanding of the topic. However, none of these studies investigated the semantic relations among different citation contexts. It remains unclear which portion of the sentence corresponds specifically to the concept such as a name of method or a expression of purpose appearing in each sentence.

III. Network Representation

A. Paper and Concept Nodes

We prepared two types of node to represent a network structure: *Paper* and *Concept* nodes. Concept nodes are labeled by character strings such as technical terms and noun and verb phrases, while paper nodes are identified by their own URIs or DOIs in a digital library. These nodes are extracted on the basis of lexico-syntactic patterns explained in section IV-B. Both types of node can connect with each other and represent various relations, as follows.

B. Method-Purpose Relation

We examined the relation between a method and its purpose and application in building a knowledge network from papers, hereafter referred to as *the Method-Purpose relation*. This relation consists of Concept nodes. An example sentence is

> *Similar observations have also been made in [38] where Probabilistic Latent Semantic Indexing (PLSI) was used to learn a lower dimension representation of text in terms of probabilistic topics.*

The text provides a relation in that "Probabilistic Latent Semantic Indexing" contributes "to learn a lower dimension representation of text in terms of probabilistic topics."

C. Paper-Topic Relation

A relation to connect a concept with the paper in which it is mentioned is referred to as *the Paper-Topic relation*. For instance, in the above example sentence, the relation with which the concept of "Similar observations" is addressed in reference "[38]" is extracted as the Topic-Paper relation. In this case, coreference resolution should be done for extracting truly semantic relation because the word "Similar observations" by itself is not sufficient for semantics. However, that relation is regarded as correct relation in this paper and we would like to consider coreference resolution for our future work.

D. Other Relation Types

In addition to the two types of relations described thus far, others that play important roles include *the ¬ Method-Purpose relation*, that is, the negation of the Method-Purpose relation, *the Citing-Cited relation* between a citing paper and a cited paper, *the Same-As relation* representing a synonym, and *the Super-Sub relation* expressing a hierarchical or whole-part relation.

IV. Relation Extraction Method

A. Definition and Notation of Extraction Patterns

In the proposed method, we use lexico-syntactic patterns to extract relations. First, we split related work sections into sentences, and then, we apply a syntactic parser to obtain a syntax tree. Syntactic tags such as "Noun Phrase (NP)" or "Verb Phrase (VP)" are labeled for each span and our system can exploit them. For example, in the sentence "*DRAGO [10] specifically examines a distributed reasoning based on the P2P-like architecture*" , while the expression "based on" acts as a key to extract the Method-Purpose relation that "the P2P-like architecture" is applicable to "a distributed reasoning", excessive spans such as "DRAGO [10] ..." cannot be excluded with the simple regular expression "*-based on-*". Syntactic tags can help system determine the boundaries of such spans for the targeted concepts.

The system uses the following extraction rule as a notation: "({NP}) based on ({NP}) = <mp> /2 /1", where <mp> represents the Method-Purpose relation and "/2 /1" denotes that a noun phrase appearing first expresses a purpose and the one appearing next is a method. A general regular expression can also be used in a rule.

B. Corpus Analysis for Extraction Patterns

In our study, we focused on two relations – the Topic-Paper (hereafter designated as <tp>) relation and the Method-Purpose (hereinafter designated as <mp>) relation – and analyzed the related work sections to identify frequent lexico-syntactic patterns.

In our analysis, we first chose 18 papers from the proceedings of the Association for the Advancement of Artificial Intelligence (AAAI2010). We manually annotated all the relations in the sentences of related work section and established 14 lexico-syntactic patterns for automatic relation extraction of <mp>. Additionaly, only the pattern appearing most frequently with a noun phrase immediately before a cited reference sign was used as lexico-syntactic patterns of the <tp> relation. The lexico-syntactic patterns obtained by the analysis are listed in the results section (Table I, Table II).

C. Rule Application and Extraction

Given one sentence in related work section and an extraction pattern described above, we apply the pattern to the sentence in following method:

1) Delete symbols that a parser cannot process properly: About parentheses and blackets, we delete them and words inside them. If there is any citation mark, their positions are recorded for later use. Words between quotations are concatenated with hyphens and quotation marks are deleted.
2) Berkeley Parser http://code.google.com/p/berkeleyparser/ is used for analysing syntactic tree.
3) Syntactic tag such as "{NP}" or "{VP}" is replaced by wildcard of regular expression ".$*$" and sentences which matches that expression are extracted as candidates.
4) For each candidate sentence, the span of words corresponding to wildcard are examined using syntactic tag of the rule and the parsed tree. If those syntactic information matches, the span is extracted as a Concept node.

V. Experiment

We performed two experiments. In the first experiment, we checked the recall and precision of our method in a small but clean data set. Error analysis was also performed. In the second experiment, we evaluated the precision on a knowledge network extracted from a large data set. Owing to the large amount of data, many meaningful knowledge relationships were extracted. We describe some examples and discuss their implications in the next section.

A. Experiment 1

As we mentioned in Section IV, we need sentences in the "Related Work" section to use as an input. This means that various pre-processing steps are needed to extract a desirable format of the input.

1) PDFs of papers were converted to texts using a conversion tool (pdftotext http://www.foolabs.com/xpdf/).
2) Related work chapters were extracted from papers.
3) Reference sections were extracted from papers and divided for each paper.
4) Cited reference signs were extracted from the related work descriptions and matched with those in 3.
5) Related works descriptions were divided into sentences with a sentence division tool (GENIA Sentence Splitter http://www-tsujii.is.s.u-tokyo.ac.jp/y-matsu/geniass/).

Assuming that a set of co-citing papers contain closely related concepts, we selected 15 papers that cited either of the two papers: "Probabilistic latent semantic analysis"[5] or "Probabilistic latent semantic indexing"[6]. Because the test data are in the same research field as the data for lexico-syntactic pattern generation, their writing styles are expected to be similar.

These papers were processed in the same manner described in the previous section, and the correct relation was annotated using an annotation tool called brat http://brat.nlplab.org/.

Comparing those, recall and precision were calculated.

B. Experiment 2

In experiment 1, the data set was quite clean and small and therefore appropriate for basic statistics. However, we could not extract many meaningful knowledge networks because of the smallness, and it would not be feasible to increase the size of the dataset because some processes rely heavily on manual effort.

In experiment 2, we used a large dataset from Microsoft Academic Search http://academic.research.microsoft.com/ to evaluate the precision of each rule and extract knowledge networks.

The data set consisted of 906,788 cited papers and 8,388,909 citation context sentences. The domain was confined to computer science. From those, the number of papers whose citation count was no less than 100 was 9,252, and their citation context sentences numbered 1,952,112 in total. We used 18 paper of them and its 3099 citation context sentences.

VI. Results and Discussion

Table.I and Table.II show a part of the result of Experiment 1 and 2. Each row corresponds to the evaluation result extracted by each rule. Our method obtained an overall accuracy of 76.9% for Experiment 1 and 71.7% for Experiment 2.

Some rules lowered the accuracy and others were very accurate. The errors resulting from failure in syntactic analysis are unavoidable since the reported accuracy of the parser is about 90% without domain dependency problem [7]. Accuracy of parsing is low when the sentence contains a present or past participle because of intrinsic ambiguity. However, the influence of this type of errors (e.g. "close sense clusters" is extracted for correct span "finding close sense clusters") on the knowledge network may be limited if these extracted Concept nodes can be properly unified.

On the other hand, recall of Experiment 1 is 12.2%. This is quite low and thus we need to construct meta heuristics of making more rules. Besides, our goal is not to extract all the relationship pieces from each paper, but to describe whole image with enough semantic details. So, in the future, we plan to examine intra-paper and inter-paper redundancy of relationships and comprehensibility of result knowledge network representation.

From the result of Experiment 2, we were able to construct meaningful knowledge graph. For example, sentences and extracted relations from it are shown as follows.

Original Sentences

The Gen 2 MAC protocol is based on Framed Slotted Aloha [19].

At the MAC layer, readers and tags use a variation on slotted Aloha [14] to solve the multi-access problem in a setting where readers can hear tags but tags cannot hear each other.

Extracted Relations

Table I

Total No. of <mp> relationships extracted and accuracy about each rule – Experiment 1

No.	Rule	Total	Accuracy(%)
1.	= ({NP}) {be} based on ({NP}) = <mp> \2 \1	6	100.0
2.	= ({NP}) based on ({NP}) = <mp> \2 \1	15	73.3
3.	= ({VP}) using ({NP}) = <mp> \2 \1	16	50.0
8.	= use(?:s|d)? ({NP}) to ({VP}) = <mp> \1 \2	5	100.0

Table II

Total No. of <mp> relationships extracted and accuracy about each rule – Experiment 2

No.	Rule	Total	Accuracy(%)
1	= ({NP}) {be} based on ({NP}) = <mp> \2 \1	48	93.8
2	= ({NP}) based on ({NP}) = <mp> \2 \1	106	65.1
3	= ({VP}) using ({NP}) = <mp> \2 \1	154	52.6
8	= use(?:s|d)? ({NP}) to ({VP}) = <mp> \1 \2	97	95.9
13	= ({NP}) {be} used to ({VP}) = <mp> \1 \2	29	93.1
14	= ({NP}) {be} proposed to ({VP}) = <mp> \1 \2	14	100.0

- <tp> relation of "Framed Slotted Aloha" and "[19]"
- <mp> relation of "Framed Slotted Aloha" and "The Gen 2 MAC protocol"
- <tp> relation of "a variation on slotted Aloha" and "[14]"
- <mp> relation of "a variation on slotted Aloha" and "solve the multi-access problem in a setting where readers can hear tags but tags cannot hear each other"

The paper represented as "[19]" or "[14]" is identified by URL http://academic.research.microsoft.com/Publication/1242802/ and its title is "ALOHA packet system with and without slots and capture".

Two types representation of "Framed Slotted Aloha" contribution – descriptive explanation and the name of succession protocol – helps us understand the position of that research.

VII. Conclusion and Future Work

As described in this paper, we have proposed an approach for extracting relations among papers and concepts to construct a paper knowledge network. A sentence citing another paper is extracted from a related work chapter, and a lexico-syntactic pattern is established for extracting semantic relations between the papers from the quoted sentence. We then performed extraction experiments using academic papers written in English. Analysis of failure examples in the experiment revealed that analytical failures can be attributed to the parser that was used and to a limited number of ambiguous lexico-syntactic patterns. We expect to improve the accuracy by performing post-processing for the acquired set of relations and the addition of lexico-syntactic patterns. In the future, we hope to express a paper knowledge network for a whole field on the basis of relations among papers and concepts by addressing the integration of knowledge extracted from multiple papers.

References

[1] C. Zhang and D. Wu, "Concept extraction and clustering for topic digital library construction," *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3, pp. 299–302, 2008.

[2] H. Nanba and M. Okumura, "Towards multi-paper summarization reference information," in *Proceedings of the 16th international joint conference on Artificial intelligence - Volume 2*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 926–931. [Online]. Available: http://portal.acm.org/citation.cfm?id=1624312.1624351

[3] S. Teufel, A. Siddharthan, and D. Tidhar, "Automatic classification of citation function," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 103–110. [Online]. Available: http://dl.acm.org/citation.cfm?id=1610075.1610091

[4] C. Dunne, B. Shneiderman, R. Gove, J. Klavans, and B. Dorr, "Rapid understanding of scientific paper collections: Integrating statistics, text analytics, and visualization," *JASIST: Journal of the American Society for Information Science and Technology*, 2012. [Online]. Available: http://www.cs.umd.edu/localphp/hcil/tech-reports-search.php?number=2011-16

[5] T. Hofmann, "Probabilistic Latent Semantic Analysis," in *Uncertainty in Artificial Intelligence*, 1999, pp. 289–296.

[6] T. Hofimann, "Probabilistic latent semantic indexing," in *Research and Development in Information Retrieval*, 1999.

[7] S. Petrov and D. Klein, "Improved inference for unlexicalized parsing," in *HLT-NAACL*, C. L. Sidner, T. Schultz, M. Stone, and C. Zhai, Eds. The Association for Computational Linguistics, 2007, pp. 404–411. [Online]. Available: http://dblp.uni-trier.de/db/conf/naacl/naacl2007.html#PetrovK07

A Text Mining Approach to Studying Matsushita's Management Thought

Xiaojun Ding
Graduate School of Management
Kyoto University
Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan
xjding@gsm.kyoto-u.ac.jp

***Abstract*— The development of technology in text data collection and handling and the increasing use of textual knowledge applications caused a boost in text mining research of various fields like sociology, business science, and so on. In recent years, the research approach has also been applicable to the field of management philosophy and management principle. In this study, with the use of text mining approach, we analyze text data in Collected Sayings of Konosuke Matsushita, which involves more than 100 lectures and remarks delivered by Konosuke Matsushita, a distinguished and influential business executive in Japan, between 1940 and 1987. The text data were compiled and edited by PHP Research Institute, Inc. In this paper, we categorized these lectures and remarks into two groups: lectures delivered for external business people and remarks delivered for employees of the Matsushita Electric Group (Later known as Panasonic Corporation). By doing so, the study aims to condense the content of lectures and remarks delivered by Konosuke Matsushita over the past few decades, to determine the recognizable patterns of the keywords spoken on these lectures or remarks, and to understand the possible characteristics and changes in his management thought as indicated by the text data. This study is an interdisciplinary research attempt between the field of management philosophy research and the field of text mining, textual knowledge applications, and knowledge discovery. It systematically integrates elements of science, such as statistical mathematics and the ability to analyze text data, and humanities, such as management thought, perceptiveness to understand concepts, and ability to comprehend. The establishment of scientific and objective methodologies in the field of management philosophy has facilitated the effective utilization of text data and information.**

Keywords—Konosuke Matsushita; management thought; management philosophy; text mining

I. INTRODUCTION

Recent years have seen various efforts to identify the fundamentals of business management. This has directed attention to studies on management philosophy and management thought of business executives not only in Japan, but also from more global perspectives. The number of researchers and business managers interested in this study field is also increasing.

The field of management philosophy research involves various qualitative and subjective factors, such as comprehension, reasoning, and explanation. There has not been much quantitative and objective analysis on management thought and philosophy done in the past owing to the difficulty in analyzing a large amount of accumulated text data. Meanwhile, in today's era of a knowledge-based society of the new century, the progress of information digitization has facilitated the utilization of vast amounts of text data.

The technique of analyzing text data with various quantitative methods, mining the patterns from natural language rather than from structured database of facts, and discovering new and useful knowledge and information from textual document repositories is called text mining. It is a process that employs a set of algorithms for converting unstructured text data into structured data conveying the insightful information [1]. On the basis of the idea this study aim to discover useful information related to management philosophy or management thought hidden in textual records of some famous business executives.

Post World War II, Japan has experienced periods of reconstruction, high economic growth, stable growth, and the bubble economy, followed by the collapse of the bubble economy. In addition, the twenty-first century is said to be the era for creating a mechanism for the establishment of a sustainable society. In such an era, corporations are under pressure to demonstrate sustainability. It seems necessary to consider and analyze management thought backed by the experience of distinguished Japanese entrepreneurs and executives, as well as the impact of such thought in order to clarify best practices and attitudes that can be adopted by the next generation of executives and applied to their activities in the future.

The text data in Collected Sayings of Konosuke Matsushita between 1940 and 1987 are the object of this study. Konosuke Matsushita (1894–1989) was one of the most famous Japanese industrialists. He is the founder of Panasonic Corporation, a Japanese multinational electronics corporation, and a global leader in the development and manufacture of consumer, professional and industrial electronics. Since his name has never been prominently displayed, Konosuke Matsushita is not as well known as Henry Ford or Honda or any of the other business giants who used their names on their products. But his company generated more revenue during his lifetime than any of the others. For many Japanese, he is known as "the God of Management". His business books have sold more than 18 million copies in Japan alone. Konosuke Matsushita's management philosophy is widely adopted among Japanese enterprises and laid the foundation of economic growth in Japan.

The organization of this study is as follows. In the next section, we describe the research method and approach of our study for analyzing Konosuke Matsushita's management thought. The third section brings in the discussion and consideration about the results of our analysis. Finally, the fourth section closes with a midterm summary and an outlook for the next step.

II. Research Method and Approach

Text data are data with no numeric values. In the past, such data were often analyzed using a method in which the analyst quoted parts of data and added interpretation and insights. In contrast, text mining approach involves a quantitative analysis of the text data through a numerical conversion process, which includes text preprocessing, feature generation and selection, pattern extraction, and result analyzing [2].

The main purpose of performing a quantitative text analysis is twofold: to explore the data and to improve objectivity. For instance, one can read the content of presentations and remarks in order to grasp the overall impression of the message, but how can this impression be objectively conveyed to a third party? What if the sheer amount of text data poses a difficulty? Furthermore, how can one examine possible changes in lecture content that are influenced by a variety of factors, including different years, different durations of lectures, and different audiences? By taking advantage of the quantitative analysis of text mining, you can deal with these issues.

There are three very important components in actually utilizing text mining: the first is extraction of information, the second is analysis of the extracted information, and the third is visualization of the results. In other words, three points must be considered: the first is how do you collect only the necessary information by minimizing the noise? The second is what analytic technique do you use in order to correctly examine and understand the information collected? The third is what visualization do you use in order to facilitate examination and understanding of the analysis results?

In order to accomplish the research objective of scientifically and objectively analyzing the overall management thought of Konosuke Matsushita and examining the characteristics and their changes over different time periods, this study goes through the procedure in the following key steps: data preparation, data analysis, and result discussion.

- Data preparation. A research platform is constructed. More specifically, records of lectures delivered by Konosuke Matsushita to external business managers as well as his talks within the Matsushita Electric Group between 1940 and 1987 were used to compile a database [3]. Besides, the text data were categorized into two groups: 52 lectures delivered for external business people between 1953 and 1983 and 54 talks or remarks delivered for employees of the Matsushita Electric Group between 1940 and 1987. With the help of KH Coder, which is a text mining software package for quantitative content analysis or text mining [4], a library of relevant keywords was created from the database.
- Data analysis. On the basis of the research platform, a variety of analyses are performed from both an overall view of the keywords and the perspective of characteristic keywords.
- Result discussion. The discussion and consideration about the results are summarized, furthermore, scope for further research is also considered.

III. Data Analysis and Discussion

Analysis of the text data can be carried out from various angles. In this section we make some analysis mainly from both an overall view of the keywords (such as frequently-appearing-keyword analysis, network analysis, etc.) and the perspective of characteristic keywords (such as trend analysis of characteristic keywords, analysis of related keywords, etc.).

A. *An Overall View of the Keywords*

1) Morphological analysis and frequently-appearing keywords: Text mining is an approach that assigns numeric values to text data for analysis. It does not refer to a specific analytic method or process flow. Therefore, there are various types of analytic methods, such as extraction of frequently-appearing keywords from the text data, categorization of observed data, and analysis of the emerging tendency of specific keywords in the text data, and so on. These are all based on morphological analysis. A morpheme is the smallest semantically meaningful unit of language; breaking it down further will make the unit meaningless. By using morphological analysis, we can divide a sentence into morphemes and obtain basic statistics as to what type of words, or more accurately, morphemes, are being used.

In this study, we first conducted morphological analysis on all lectures and remarks selected for the study, and performed noise reduction (eliminating irrelevant and less frequently appearing words). We then extracted the most frequently appearing keywords; summarized aspects such as the number of involved lectures that means the number of lectures and remarks in which those keywords appeared, and the total appearance frequency of the keywords. We also conducted analyses such as comparison between the lectures for external business managers and the remarks within the Matsushita Electric Group as well.

The differences and similarities were summarized as follows. Firstly, the results show that there is relativity between the total appearance frequency of the keywords and the number of involved lectures. In addition, frequently appearing keywords were diverse, including “management”, “politics”, “region”, “business”, “labor and employment”, and so on. However, it is observed that the characteristic keywords used by Konosuke Matsushita do not appear frequently. Meanwhile, the results of the frequently appearing keywords indicate that external lectures are largely based on macro perspectives, while internal remarks are often based on micro perspectives such as employees, staff, and individuals.

2) Trend analysis of keyword groups by year: The extracted keywords were categorized into several different categories, such as “business”, “manufacturing”, “religion”, “moral and value”, “management”, and “labor and employment”. Then the trends and changes in each keyword category by year were summarized separately for internal remarks and external lectures.

The results show that the appearance frequencies of keywords falling under the categories of “business”, “manufacturing”, “moral and value”, and “management” are not a function of year and time period, while the appearance frequencies of keywords falling under the “religion” and “labor and employment” categories varied depending on the time period. The trends and changes in the “religion” and “labor and employment” categories are considered to be related to Konosuke Matsushita’s religious background and the history of the labor movement in Japan.

3) Analysis of co-occurrence networks for keywords in different time periods: Co-occurrence relations of keywords can be described by a network graph, where the lines link the keywords with a high degree of co-occurrence together. In addition, co-occurrence relations between the keywords and time periods can be connected as well. In other words, it is possible to examine how much attention is being paid to an extracted keyword in different time periods and how these keywords are linked each other.

On the basis of the next two different perspectives, the perspective of the career of Konosuke Matsushita and the perspective of the economic fluctuation in Japan, the lectures and remarks that were delivered over a period of more than 40 years were divided into the different time periods for analysis as follows:

- The perspective of the career of Konosuke Matsushita. First, the 1940s to the 1950s, when Konosuke Matsushita was the President of Panasonic Corporation. Second, the 1960s to the early 1970s, when he was the Chairman of Panasonic Corporation. Third, the late 1970s to the 1980s, when he was the executive adviser of Panasonic Corporation.
- The perspective of the economic fluctuation in Japan. Economic fluctuation includes the period marking an increase in the activities of economic society (expansion) and the period marking the stagnation of the economic society (recession). Because economic fluctuation occurs as repeated cycles over several years alternating expansion and recession, it is also called the economic cycle. First, the economic boom of the mid-1950s. Second, the Inventory recession of 1957-1958. Third, the economic boom of 1958-1961. Fourth, the recession of 1961-1962. Fifth, the economic boom of 1962-1964 created by the Tokyo Olympic Games. Sixth, the Securities depression of 1964-1965. Seventh, the economic boom of 1965-1973. Last, the recession of 1973-1986.

The analysis results indicated that although there were some common words across time periods, Konosuke Matsushita focused on different keyword groups during the lectures and remarks in different time periods. In particular, the text mining analysis made it clear that the economic fluctuation in Japan strongly influenced changes in the keywords used.

B. The Perspective of Characteristic Keywords

1) Trend analysis of characteristic keywords by year: The popularity of Konosuke Matsushita’s management philosophy stems from his practical thoughts and ideas on real-life management practices, assuming a form of aphorism in his lectures and writing. In this section, we mainly focus on the following characteristic keywords of Konosuke Matsushita’s management philosophy [5] and verify in detail the change in usage of these keywords in internal remarks and external lectures.

- Management through collective wisdom. A management approach that takes advantage of collective wisdom obtained from as many employees and persons as possible.
- Coexistence and mutual prosperity.
- The right person in the right place.
- Appropriate management. Accurately understanding the company’s comprehensive abilities, such as technical capabilities, financial strength, and management skills, and striving to manage the business realistically within one’s own capability.
- Autonomous responsible management.

The time periods when the characteristic keyword attracted attention and the relevant background information are analyzed and made into a line graph.

2) Analysis of related keywords: Some specific topic-related keywords (i.e., related keywords of some intriguing topics) are the focus in this section. This relation is determined on basis of calculating conditional probability, that is to say, the keywords that are highly likely to appear in lecture texts involving a specific topic are so-called the topic-related keywords. For example, the results showed that in the case of the analysis of keywords related to the topic “business manager”, the “business manager”-related keywords turned out to be “productivity”, “industry”, “business world”, “business people”, “social conditions”, “employees”, “labor union”, “politician”, and so on. Examining such results from various perspectives based on concept of business manager, roles of business manager in economic society, etc., it became clear that for Konosuke Matsushita, important matters strongly associated with business managers, or leaders in business, are pluralistic and multilayered, including factors such as industry, production, economy, society, employees, politics, etc..

In addition to creating a list of keywords highly associated with a particular topic, the analysis of topic-related keywords can help draw a co-occurrence network diagram of a specific topic. Such a diagram places the specific topic selected for discussion in a double-border square and links the topic-related keywords with lines.

Furthermore, in order to better illustrate the characteristics, we can use a thicker line to indicate a stronger co-occurrence relationship, or place keywords that appear more often in larger circles, and so on.

IV. SUMMARY AND FUTURE DIRECTION

This study aims to show that text mining is a new and effective research approach to the field of management philosophy and management principle. The present results confirmed that the keywords in Konosuke Matsushita’s lectures or remarks changed over the course of time and also depending on the person he was talking about. By analyzing the accumulated text data using text mining, some existing interpretation can be verified in a more objective manner.

For the next step, there are a lot of open problems and research directions for further developing this study. First, we plan to focus on adding discussions about other keyword categories and notable characteristic keywords. Moreover, we hope to concentrate on some factors such as the depth of keywords, inter-area analysis, etc. as well. In the future, we will attempt to make analysis from the macro-level and global perspectives. In addition, we are going to work on a comparative study of Konosuke Matsushita’s management philosophy and management philosophy of other influential business leaders, such as Kazuo Inamori, who ranks alongside Sony's co-founder, and Soichiro Honda as one of the great Japanese postwar entrepreneurs.

ACKNOWLEDGMENT

This work is supported by Kyocera Chair of Management Philosophy and financially supported by the Young Researchers Startup Grant from Graduate School of Management, Kyoto University. The author also would like to thank PHP Research Institute for its support of providing the original materials and data.

REFERENCES

[1] R. S. Segall, Q. Zhang, and M. Cao, “Web-Based Text Mining of Hotel Customer Comments Using SAS Text Miner and Megaputer Polyanalyst,” Proc. SWDSI 2009 Annual Conference, Oklahoma City, Feb. 2009.

[2] J. W. Liang, “Introduction to Text and Web Mining,” Seminar at North Carolina Technical University, http://www.database.cis.nctu.edu.tw/seminars/2003F/TWM/slides/p.ppt, 2003.

[3] PHP Research Institute Publication: Collected Sayings of Konosuke Matsushita (in Japanese), 1996.

[4] http://khc.sourceforge.net/.

[5] PHP Research Institute Publication: A Dictionary of Matsushita’s Life and Thought (in Japanese), 1999.

Exploring your Business Leaders through Virtual Characters

- Persona Design and Small Experiments -

Yasuhiro SASAKI
Tokyo Institute of Technology
Mitsubishi Research Institute, Inc.
Tokyo, JAPAN
sasaki.yasuhiro@gmail.com

Hikaru UCHIDA
Masaaki KUNIGAMI
Atsushi YOSHIKAWA
Takao TERANO
Tokyo Institute of Technology

***Abstract*— We propose a novel method to identify desirable leader images in each situation. Leader's characteristics have been researched. However, many researchers could not find unique his/her characteristics. Therefore, desirable characteristics are different in each situation which he/her meets. In order to capture the desirable characters in each situation, we apply the persona technique, which was firstly introduced by Cooper and then has been spread in the field of product development. The proposed method is characterized by the procedure: 1) set persona characters with various attributes, 2) assign the attribute values by orthogonal design techniques, 3) set virtual business situations, 4) based on the business situations, get questionnaire data on the personas from subjects, and 5) evaluate the data to get the leader images of the corresponding business situations. As results of our method, we detected the different in the sense of designed characters' attributes in each situation.**

Keywords- Personas; Competency; Leadership; Personal requirements; Human developments.

I. INTRODUCTION

A. Necessity and difficulty of the research on leader images

According to literature[1], leaders are made, not born, and made more by themselves. We also have the position that leaders can raise later.

In business organizations, leaders' roles are very important. Leaders' capability influences the rise and fall of their organization. Leaders have the role which guides and raises subordinates. If leaders are excellent, the organization can gain a competitive advantage. Therefore, it is the most important proposition to raise excellent leaders in business organizations.[2]

However, the research on training of leaders is incomplete. Leaders' required capability is not uniform. Leaders need to be well versed in the operation of their organization. Moreover, they are urged to carry out business smoothly with bosses and subordinates. Furthermore, they need to learn the conceptualization capability to understand and draw enterprises. Since business conditions differ for every organization, you cannot define leaders' capability uniformly. This is one of the difficult reasons for the research on leader images.

B. What is "personas" − From Cooper's Idea

In marketing and user-centered design, personas are virtual characters created to represent the different user types within a targeted demographic, attitude or behavior set that might use a site, brand or product in a similar way. [4] Personas are useful in considering the goals, desires, and limitations of brand buyers and users in order to help to guide decisions about a service, product or interaction space such as features, interactions, and visual design of a website. Personas may also be used as part of a user-centered design process for designing software and are also considered a part of interaction design, having been used in industrial design and more recently for online marketing purposes. A user persona is a representation of the goals and behavior of a hypothesized group of users. In most cases, personas are synthesized from data collected from interviews with users. They are captured in 1 page descriptions that include behavior patterns, goals, skills, attitudes, and environment, with a few fictional personal details to make the persona a realistic character.

Alan Cooper, a noted pioneer software developer, developed the concept, which he named personas. From 1995 he became engaged with how a specific rather than generalized user would use and interface with software. The technique was popularized for the online business and technology community in his 1999 book “The Inmates are running the Asylum”.[3] In this book, Cooper outlines the general characteristics, uses and best practices for creating personas, recommending that software be designed for single archetypal users.

In this paper, we applied the persona technique to drawing the leader images of companies. Usually, in the case of "persona" creation, minute fixed-quantity investigation and close qualitative investigation are required. [5] Actually, neither expense nor time and effort can be applied to "persona" creation. So, in this research, we devised the simple technique which collateralizes fixed rationality and probability.

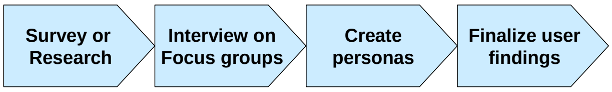


Figure 1. Usual persona generation and measure examination process

II. Purpose of the Research

In knowledge management literature, it is critical to identify desirable leader characteristics. Different business environments, however, require different types of business leaders. While business environment changes early, it is difficult to take agreement about the desirable leader characteristic. In this paper, we propose a novel method to identify desirable leader images using virtual characters. We take in personas to research of the leader characteristic.[8]

By application of an experimental design, we aim at making the number of choices of questionnaire items into the minimum. We aim at obtaining high-precision results with few samples by utilizing conjoint analysis.[7]

The outline of the proposed method is summarized as follows: 1) set persona characters with various attributes, 2) assign the attribute values by orthogonal design techniques, 3) set virtual business situations, 4) based on the business situations, get questionnaire data on the personas from subjects, and 5) evaluate the data to get the leader images of the corresponding business situations.

III. Brief Description on the Proposed Method

A. From Leaders' Requirements to Their Capability Assignment

In the research [6] on the leader image of ICT organizations, 25 requirements were probed as capability required for the leaders. In the paper, nine core competencies are drawn by systematic examination as most important requirements for capability, which is summarized in TABLE I.

TABLE I. Main requirements for capability which business leader is expected

No.	Competency	The example of action
1	Practical Skill (ICT skills and Industry Knowledge)	The thing required for mind and logic composition for which it has deep knowledge broadly and moderately
2	Achievement volition and positivism	The subject of a high level is set up actively and the best is always concentrated towards achievement.
3	Judgment	It can judge [whether while there is an uncertain element, a project can be promoted and it can lead to a success, and] by itself.
4	Cultivating Human Resource	Advice according to the feature and characteristic of a place of work and the member of a project is performed.
5	Vision/ Imagination	Business planning, such as a development project which can promote two or more proposal affair efficiently, is drawn up.
6	Communication	A confidential relation with a partner can be built through "hearing it" and "talking."
7	Negotiation/ Adjustment power	The merit to the company by developing a system is recognized, and in-company adjustment can be carried out.

Since we used the orthogonal array (L8) this time, we needed to extract the item from nine to seven. Then, we removed "management control and the initiative", and "theory and policy." We take up "Practical Skill", "achievement volition and positivism", "judgment", "subordinate training", "the imaginative power and a vision", "communication", and "negotiation / adjustment power."

Of course, the person who has full capacity is excellent. However, there are few such business leaders and it is difficult to expect young leaders' full capacity. Then, we created the orthogonal array as shown in TABLE II, and it was made to get an experiment candidate to choose talented people suitable as a business leader. By using the orthogonal array as shown in TABLE II, we did the work which can perform many comparisons in spite of few selections.

TABLE II. Combination of business leader's capability

	Practical Skill	Posit -iveness	Judgment	Cultivating HR	Vision	Commu -nication	Negotiation
typeA	-	-	-	-	-	-	-
typeB	-	-	-	○	○	○	○
typeC	-	○	○	-	-	○	○
typeD	-	○	○	○	○	-	-
typeE	○	-	○	-	○	-	○
typeF	○	-	○	○	-	○	-
typeG	○	○	-	-	○	○	-
typeH	○	○	-	○	-	-	○

(L8 array)

In this array, Type A means holding all the capability on the average. Type B to H means excelling in any 4 capability.

B. Setups of of the Experiment

We have designed the experiment as follows. We prepared the questionnaire vote. Respondents should answer individually about the following three cases. From the viewpoint of the conformity to the case, we requested respondents to attach leader type ranking.

TABLE III. Setup of each case and others

	The situation of organizations
Case 1	The achievements of this organization tend to descend. If the leader does not take bold measures, this organization does not have the future.
Case 2	The achievements of this organization are upward. The atmosphere of this organization is bright.
Case 3	The achievements of this organization are safe for the time being. However, the future of this organization is slightly opaque.

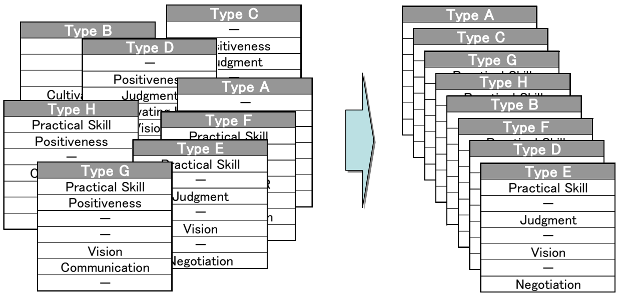


Figure 2. Rearrangement of cards by participants (image)

We printed eight personas (from Type A to Type H) on small papers, and arranged on the desk. We requested respondents to rearrange those papers from the viewpoint of the conformity to the case (Figure 2).

At the following three places, we used the same questionnaire and collected the reply data of 25 votes.

1) The seminar of the Tokyo Institute of Technology graduate school : 13 votes
2) The study group of the Mitsubishi group : 5 votes
3) The seminar hall of a certain society : 7 votes

IV. Results and Discussion

A. Results of Simple Totals

1) Case 1 (the achievements of this organization tend to descend)

At the very top was Type G (Practical Skill, achievement volition, imaginative power, communication) among the types chosen as the 1st place. The next was Type D and Type E.

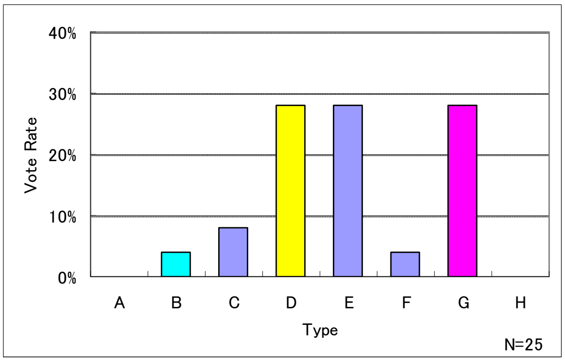


Figure 3. (Case 1) Type selected as the 1st place

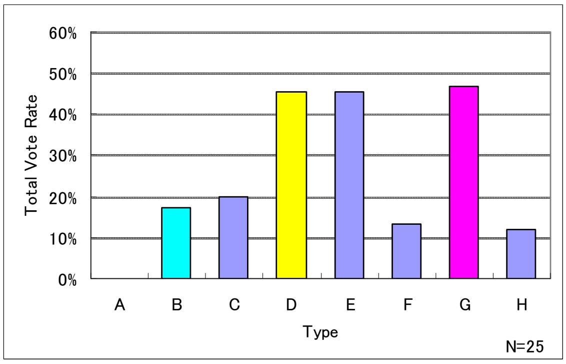


Figure 4. (Case 1) Assigned total score

We assigned 8 points to the type chosen as the 1st place. Similarly, we assigned 7 points to the 2nd place and ..., assigned 1 point to the 8th place, and we totaled the whole. As a result, at the very top was Type G among, and the next was Type E and Type D.

2) Case 2 (the achievements of this organization are upward)

At the very top was Type B (Cultivating HR, the imaginative power, communication, and negotiation power) among the personas chosen as the 1st place. Although the next was Type F and H, there was no difference not much. (Fig. 5)

The assigned sum total vote was also the same result.

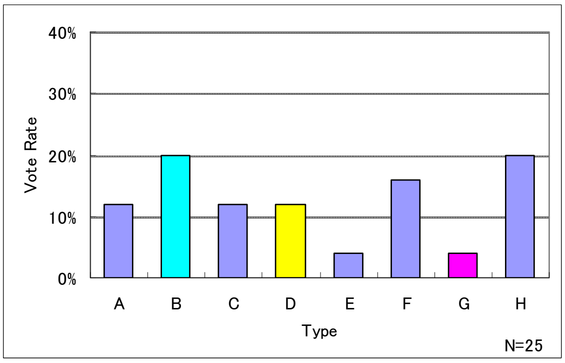


Figure 5. (Case 2) Type selected as the 1st place

3) Case 3 (the achievements of this organization are safe for the time being)

At the very top was Type D (achievement volition, judgment, Cultivating H R, imaginative power) among the types chosen as the 1st place. The next was Type B. Other types are seldom chosen. (Fig. 6)

The assigned total score have also shown also the similar results.

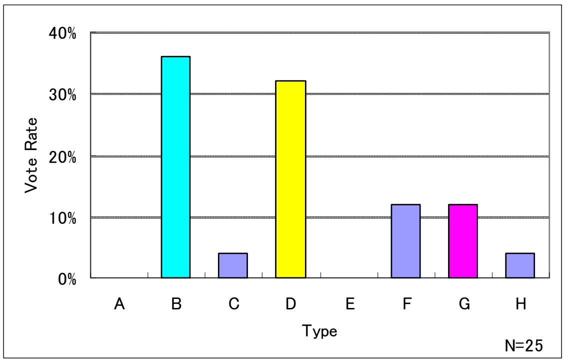


Figure 6. (case 3) Type selected as the 1st place

ISBN: 978-1-61208-254-7

B. *Results of Requirements for Capabilities*

Below, we analyzed questionnaire results taking advantage of this orthogonal array. The seven competencies were allocated to each persona by the L-8 orthogonal array. We analyzed which the competencies for the leader were thought as important, when the respondents rank (the best: 1st - the worst: 8th) the leader personas. Then as well as the conjoint-analysis, we applied the multiple regression analysis for the sensitivity of the respondents' evaluation. In the multiple regression analysis, the response variable is "score (= 8 - rank, the best: 7 - the worst: 0)" and the explanatory variables are the existence of the seven competencies (0-1 data, absence: 0, existence: 1). The multiple regression analysis under the three cases of organizational situation illustrates how the seven competencies contribute to the score.

1) Case 2 (the achievements of this organization are upward)

The requirements for the business leader needed a little for this organization are "subordinate training" capability. There was no significant difference between the seven competencies, the effect-size: r^2 (coefficient of determination) = 0.08.

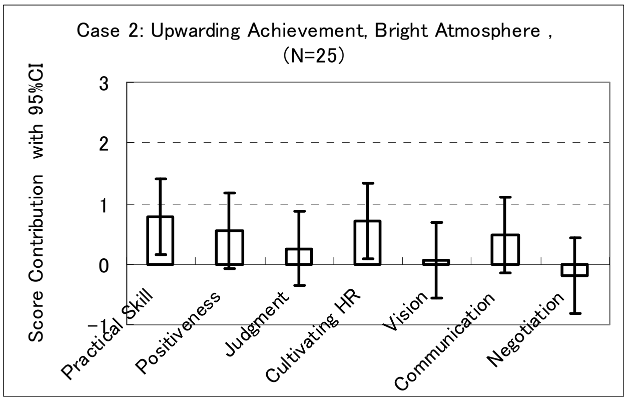


Figure 7. (Case 2) Requirements for capability in high performance organization

2) Case 3 (the achievements of this organization are safe for the time being)

The requirements for the business leader needed for this organization are "the imaginative power and a vision", and "subordinate training." Clearly different from the Case 2, the outcome shows the relatively middle – long range (Vision, Cultivating HR) competencies are clearly significant. The effect-size: r^2 (coefficient of determination) = 0.29.

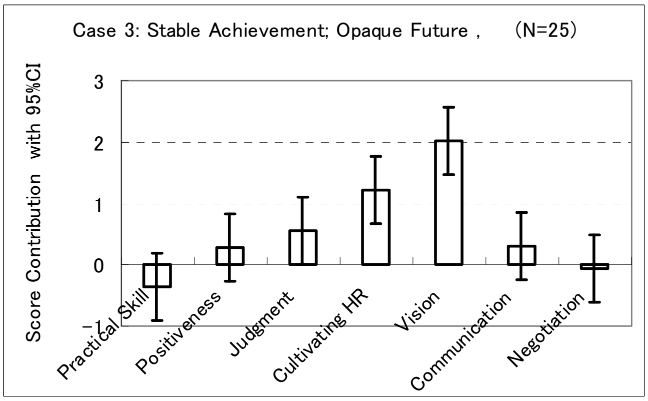


Figure 8. (Case 3) Requirements for capability in dull organization

3) Case 1 (the achievements of this organization tend to descend)

The requirements for the business leader needed for this organization are "the imaginative power and a vision." Subsequently, "judgment" and "achievement volition and positivism" are needed. Clearly different from both of the Case 2&3, the outcome shows the relatively short range competencies become clearly significant instead of the relatively long range one (Cultivating HR). The effect-size: r^2 (coefficient of determination) = 0.35.

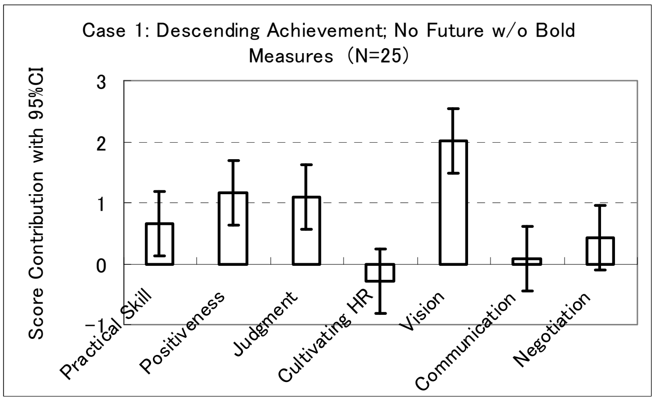


Figure 9. (Case 1) Requirements for capability in low performance organization

V. CONCLUSION AND FUTURE WORK

In the literature of competency research so far, leaders' requirements for capability were explored by decomposing leaders' behavioral trait. Therefore, it was difficult to discern the strength between capability elements. We were able to draw the requirements for business leaders by the easy questionnaire.

Compared with such conventional research methods, the proposed technique is based on the idea of questionnaire design using an orthogonal array. Rearrangement of 8 types is equal to having performed the paired comparison 28 times. It has succeeded in eliminating principle and pulling out respondents' potential opinion with this technique.

Another unique feature of the method comes from the concept of a "disposable persona." People have concern strong against people. Then, we set up the personas as items of the questionnaire.

We have observed the statistical difference by easy situation setup. The results have suggested that, according to the business situations, the desired leader images are clearly different in the sense of designed characters' attributes.

In our future work, we have a plan to have interview sessions with excellent leaders of the cooperation companies. Then we will extract the parameters of the rectangular tables. Such future work includes the following points:

- How are appropriate items incorporated in an orthogonal array?
- What is the suitable method of analyzing the collected data?
- Is the simple persona of this technique appropriate compared with an originator persona?

Then, we will expand the research topics in the following points:

- We will conceive of the new human resource management technique by the experiment base.
- We are going to collect data in ICT leaders' training domain. We will build the practical simulation base.

REFERENCES

[1] Warren Bennis: “On Becoming a Leader”, Basic Books, 1989.

[2] Noel M. Tichy: “The Leadership Engine: How Winning Companies Build Leaders at Every Level ”, Harper Business, 1997.

[3] Alan Cooper: “The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity”, Macmillan, 1999.

[4] Grudin, J. & Pruitt, J.: “Personas, participatory design, and product development: An infrastructure for engagement”. Proc. PDC 2002, 144−161.

[5] John Pruitt, Tamara Adlin “The Persona Lifecycle: Keeping People in Mind Throughout Product Design (Interactive Technologies)”, Morgan Kaufmann, 2006.

[6] Yasuhiro Sasaki, Atsushi Yoshikawa and Takao Terano , “Identifying and Evaluating Next ICT Leaders of a Company - A Competency Oriented Approach -”, Advances in education research, 2012 2nd International Conference on Education and Education Management(EEM2012), Information Engineering Research Institute, pp282-pp288, 2012.

[7] Masaaki Kunigami & Takao Terano, Experiments Based Management and Administrative Science - A Manifesto -, Proc. of General conference on Emerging Arts of Research on management and administration (GEAR2012), 2012.

[8] Hikaru Uchida, Akiko Orita, Masaaki Kunigami, Atsushi Yoshikawa, Takao Terano. “Persona Conjoint Method: Measuring Learners' Latent Understandings and the Effect of Stereotypes in Complex Business Situations”, Grace Hopper Celebration of Women in Computing (GHC2012), The Anita Borg Institute For Women And Technology, 2012.

Profiling of Patrons' Interest Areas from Library's Circulation Records
– An Approach to Knowledge Management for University Students –

Toshiro Minami
Kyushu Institute of Information Sciences,
Dazaifu, Fukuoka, Japan, and
Kyushu University Library,
Fukuoka, Japan
Email: minami@kiis.ac.jp, minami@lib.kyushu-u.ac.jp

***Abstract*—The concept of knowledge management is important not only in industries but also in educational organizations like universities. Considering the importance of this concept, it is not surprising that many universities have introduced the database system for saving the profiles and history information of students and utilize them in order to improve their educational abilities. In order to make the information more effective in education, it is preferable to collect not only the information in raw but also the knowledge that is found by the data with data analysis and data mining. In this paper, as a new approach to knowledge mining in education as a part of educational knowledge management, we deal with the circulation records of a university library as the target data. The library's circulation records show the relationship between the patrons and the books, which are usable to know the patrons about their fields of interest, knowledge levels, and other information. In this paper, we put special emphasis on the investigation of the profiling of students as a knowledge management. As a part of this, we deal with the interest area of a student and explore the measuring methods for the profiling of the student patrons.**

***Keywords*-knowledge management; knowledge discovery; library marketing; data analysis; data mining;**

I. Introduction

The most important mission of a university as an educational organization is to provide its students with good learning environment. The concept of knowledge management is important not only in industries but also in such an educational organization so that it can manage the data concerning its students' learning ability, willingness to study, and other aspects that relate to learning and studying.

Considering the importance of this issue, it is getting to be popular and many universities have introduced the database systems for saving the profile and history information of students and utilize them in order to improve their educational abilities. Such a database is sometimes called a learning portfolio, student portfolio, or digital portfolio, etc.

In order to make the information more effective in education, it is preferable to collect not only the information recorded by teachers and other staff as an original data but also the knowledge that is found by the data with data analysis and data mining.

In this paper, as a new approach to knowledge mining in education, we deal with library data analysis, especially the circulation records. The library's circulation records have an advantage because every university has a library network and every library must have the circulation records as necessary data in their services. In this paper, we take the circulation data of a university library as the target for analysis and show how to extract useful information out of them. The library's circulation records deal with the relationship between the patrons and the books, which can be used to know the patrons about their fields of interest, knowledge levels [6], and other information.

So quite a lot of researches have been conducted so far. For example circulation records are used for evaluation of collections of library in [2]. They are usually analyzed with various kinds of statistical methods, which are very useful to efficiently recognize the representative image of the total data. The system WorldCat Collection Analysis [11], for example, provides an easy-to-use and easy-to-recognize analysis environment to librarians, based on the standard statistical methods. A research on circulation record analysis for evaluating the usage of e-books is reported in [2]. Yamada analyzed the circulation records of a university library with considering the material age of the circulated books [12]. In addition to these research based on the statistical methods, investigation of the association rules in classification category of books using a data mining method is reported in [1].

Our approach to circulation record analysis is different from such standard methods. We take the analysis methods combining two ways; one is the statistical one for surveying the general tendencies of the patrons, and another one is the new way trying to find the more realistic patron's behavioral model and to understand the patrons' behavior in reading, studying, and using of libraries more precisely, including their underlying needs, preferences etc. In this paper we show some example data analysis experiences as a case study using the circulation records of the Central Library of Kyushu University, Japan (KUL) for the academic year 2007. The results shown in this paper are extensions to the results reported in [3], [4], [5], and [8]. Our aim in this paper is to propose new methods for getting more precise patron profiling as a whole and a patron's preference, knowledge level, eagerness to learning, etc. that will be helpful for personalized services in learning assistance.

In the paper [6], we proposed the concept of p-rank for measuring expertise level of a book and of a patron. In this paper we propose two new concepts for measuring interest range size and earnestness in learning. We compare faculties in their features by applying these measures.

The rest of this paper is organized as follows: In Section II we review our previous research results and show some case studies that inspire the research presented in this paper. In Section III, we propose a concept of the profile of a patron that indicate the patron's interest. Then we define two concepts for measuring interest range size and strength of earnestness for studying from the patron's profile. We then investigate how to capture the patrons behavior through these measures. We also define the similar concepts for a group of patrons and apply them to the comparative study of faculties. Finally in Section IV, we summarize what we have done in this paper and prospect possible future works.

II. Profiling of Patron with Data Analysis from Library Data

This section describes some of our case studies on data analysis of library data. The study in the next section is an extension to these analysis experiences.

A. Target Data for Profiling with Data Analysis

In this paper, we use the circulation records obtained in the Central Library of Kyushu University, Japan, in the academic year 2007; i.e. from April 2007 to March 2008. The whole data contain 67,304 circulation records. A record item consists of the book ID, book's classification number, call number, borrower's patron ID (renumbered one so that the record does not link to the real patron ID), borrower's affiliation, borrower's type (undergraduate student, masters student, Ph.D student, professor, staff, others), and the timestamps for borrowing and returning, etc.

The number of patrons, who borrowed at least one book during this year period, is 6,118 in all and the average number of borrowed books per patron is about 11.

A circulation record has 10 patron types: undergraduate student (Bachelors-1 to 6, or B1 to B6), masters student (M), Ph.D students (D), academic staff (Professors, P), and others (O). About 45% of books are borrowed by undergraduate students and 24% by masters and 15% by Ph.D students. Thus about 80% of books are found to be borrowed by students; which supports based-on the objective data that the frequently-told saying that most important patrons of university libraries are students.

B. Preprocessing of Circulation Records

As a preprocessing, we eliminate the records that have inappropriate values and no data for the inevitable properties (items) that are necessary to deal with in the analysis in this paper. For example 244 records have NDC (Nippon Decimal Classification) numbers that are greater than 1000 and 7,260 records have the non-numeric values for this item and thus have eliminated from the original records. After elimination, 53,182 records are left as those for analysis.

C. Case Study: Expertise Level as a Profile for Library Patrons and Library Books

The concept of the expertise level of a patron is useful in various purposes in such cases as to recommend books to read, to form a study group, to estimate the period of times to need for the patron to study some specific subject, etc. We defined an expertise level measure of a book and a patron, which we call p-rank in both cases [6].

We defined the expertise level of a book as the average value of its borrowers' initial expertise levels; where the initial expertise levels of B1 to B6 are set to 1 to 6, respectively, 8 for M, 9 for D, and 10 for P. We do not count the patrons of the type (O). Then we define the expertise level of a patron as the average expertise levels of the books the patron borrows in the circulation records. See the paper [6] for more detail about p-rank, and c-rank, which is another definition for expertise level of a book.

D. Definition of a-value as Another Measure for Expertise Level

As another idea for defining expertise level of books with assuming that if a book is borrowed by a limited number of patrons then its expertise level is high. In other words, if a book is borrowed by a wide range of patrons, its expertise level is low.

Based on this assumption we define the a-value (affiliation based expertise level) of a book [5]. Firstly we have to choose the faculties as the representatives of expertise fields. Affiliations of Kyushu University consist of not only the faculties for undergraduate students but also of some number of research centers, library, communications center, and others. We will take 12 faculties together with the graduate schools for graduate students relating fields and research organizations for professors; precisely, SC for (Faculty of) Sciences, AG for Agriculture, TE for Engineering, MD for Medicine, DD for Dental, PS for Pharmaceutical, LA for Law, LT for Letter, EC for Economy, ED for Education, DS for Design, and 21 for the special faculty of Kyushu University called 21st century program, which was founded for the students who are willing to study from a wide variety of learning fields.

So there are 12 groups based on the faculties. Let m be 12 as the number of categories and let F_i $(i = 1, 2, \ldots, m)$ be the i-th faculty. The a-value of a book is calculated as

ISBN: 978-1-61208-254-7

follows. Let CR be the set of circulation records; $CR = \{r =< BookID, NDC\ Number, Borrower, Borrowed\ Day\ and\ Time, Returned\ Day\ and\ Time, \ldots >\}$. We use $BI(r)$ for the book ID, $Cls(r)$ for the NDC number, $B(r)$ for the borrower, $Bd(r)$ for the borrowed day and time, and $Rd(r)$ for the returned day and time, of r. For a given book b, let us define the number $s_i\ (i = 1, 2, \ldots, m)$ of the book for the faculty F_i by $s_i = \#\{r \in CR | BI(r) = b, B(r) \in F_i\}$, where $\#$ is the number of elements. We put 0 for a-value if all the s_i's are zero; i.e. that the book is borrowed by the patrons who do not belong to these faculty related affiliations. Let us set $s = \Sigma_{i=1}^{m} s_i$, i.e. total number of circulations borrowed by the patrons belonging to either one of the nominated faculties. Then we define the a-value of the book b as $10 \times \Sigma_{i=1}^{m} (s_i/s)^2$, where multiplication value of 10 is used in order to make the maximum value to 10 so that it becomes easier to compare it with other values.

III. Interest Area Analysis from Circulation Records

A. Profiling Interest Area of Patrons

The eventual goal of the study in this paper is to provide library patrons with good learning environment. Mostly the services provided by libraries are intending to be universal; to every patron in a uniform way and thus in a uniform level. However toward the future, personalized services are expected to be more and more important for libraries. With personalized services, patrons are able to get better assistance that matches more to the patrons' needs and will have better effects in learning. In order to provide with unique services we would like to investigate in developing methods of analyzing the library data and to obtain knowledge about the profiles of patrons.

In this paper, we deal with profiling the patrons' interest areas by analyzing circulation records. The concept of interest areas of a patron may be considered good for characterizing the patron's attitude to learning. We would be able to extend the profile on interest areas to other properties of patron that relate more on knowledge level, learning abilities, learning styles, etc.

We use the classification field of book using the NDC number of the book. NDC is a decimal classification system like DDC (Dewey Decimal Classification) localized to Japan. The top level categories consist of the following 10 topics; 000 for General Works, 100 for Philosophy and Religion, 200 for History and Geography, 300 for Social Sciences, 400 for Natural Sciences, 500 Technology (Engineering), 600 for Industry and Commerce, 700 for Arts, 800 for Language, and 900 for Literature. Note that NDC classification items are different from those of DDC.

For a patron p, we define the profile $Prof(p)$ of p as the vector of frequencies of the books borrowed by the patron p according with the books' 10 classification numbers from 000 to 900 in NDC.

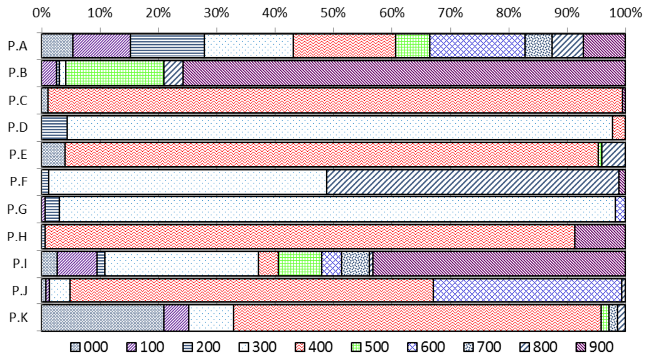


Figure 1. Profiles of the Top 11 Patrons in the Numbers of Borrowed Books, or Items

For a patron p, we define $Prof(p) = \{< bt(c) >_{c \in NDC} \mid bt(c) = \#\{r \in CR \mid B(r) = p, \text{ and } Cls(r) = c\}\}$.

We apply this definition of $Prof(p)$ to the 53,182 circulation records that are described in Section II-B. Figure 1 shows the profile patterns in 100% stacked column chart of top 11 patrons in terms of the number of borrowed books. We will call them from Patron A (P.A) to Patron K (P.K) in the order of the numbers of borrowed books; from 388 by P.A, 268 by P.B to 143 by P.J and P.K.

It is easy to see that the ratios of books according to the classification number, or topic area, vary from patron to patron. For example, P.A borrows quite a wide area of books with NDC number from 0 to 9. On the other hand, P.C borrows mostly with the classification number 400 (Natural Science). Such difference about the range of topic areas indicates a character of the patron in his or her interest range, or curiosity range. Together with the number of the borrowed books, this range can be good measures for characteristic features of a patron, which will be discussed in more detail in the next section.

B. Interest Range and Interest Strength Measures

In order to analyze deeper about the interest profiles of patrons, we define 2 new measures; strength and range of interest of a patron.

The interest strength is a concept intending to measure the willingness to learn and to know, or earnestness to obtain new knowledge. We use the number of books, or items, that the patron borrows as the measure for interest strength; $Str(p) = \#\{r | r \in CR, B(r) = p\}$ for a patron p.

Interest range is also a very important measure to describe about the willingness of learning of a patron. As we see Figure 1, we can easily recognize that P.A is interested in quite a wide areas of topics, whereas P.C is mostly interested in one subject only. In order to compare such difference of the patterns of interest we propose a new measure for the amount of the width of interest of patron. We use the concept of entropy, or the amount of information, for the interest range of a patron. Let p be a patron. We define the

Table I
COMPARISON OF PATRONS IN THEIR PROPERTIES

Patron	Range	Strength	Affiliation	Type
P.A	0.95	388	O	O
P.B	0.34	268	LT	D
P.C	0.04	185	SC	B4
P.D	0.12	183	LA	D
P.E	0.16	173	SC	B3
P.F	0.35	168	LA	D
P.G	0.10	167	LA	D
P.H	0.15	150	SC	B4
P.I	0.72	148	O	M
P.J	0.38	143	AG	B3
P.K	0.49	143	SC	M

(interest) range of p as follows:
$Range(p) = \sum_{Prof(p)_c \in NDC} (\frac{Prof(p)_c}{Str}) log(\frac{Prof(p)_c}{Str})$
where $Str = Str(p)$ and $Prof(p)_c$ is the number of books borrowed by the patron p having the NDC number c among $NDC = \{000, 100, 200, \ldots, 900\}$. We take 10 for the base of the logarithm so that the maximum value of the range becomes 1 because the number of the categories, i.e. number of the NDC values, is 10.

Table I shows the range, strength, affiliation, and type of the 11 patrons from P.A to P.K. As has been predicted the range of P.A (0.952) is quite high; the highest among 11 patrons. On the other hand P.C has the minimum range value (0.04), who's affiliation is SC and the year 4 undergraduate student (B4).

To have a closer look at the table, there are 4 students with affiliation of SC (Sciences) and 2 of them are B4 (P.C and P.H) and 1 (P.E) is B3 and another one (P.K) is M (Masters). The 3 undergraduate students have very low range values from 0.04 to 0.16. They are very concentrated in learning just like P.C. It is interesting to see that the remaining masters student (P,K) has relatively bigger range value 0.49. He or she borrows the books not only in the natural science field (with NDC 400), but also the books in general topics (with NDC 000), social sciences (with NDC 300) and others as well.

There are 3 Ph.D students with affiliation LA (Law); P.D, P.F, and P.G. The patrons P.D and P.G have similar range values 0.12 and 0.10, whereas P.F has bigger value 0.35. The former 2 students borrow the books with NDC 300 (Social Sciences) mostly, whereas the latter student borrows not only the books of social sciences but also the books with NDC 800 (Language) as many as of 300.

Figure 2 shows the correlation between the range size (x-axis) and the strength (y-axis) for all patrons. The range value 0 means that the patron borrows only one book. The range value is 1 if the patron borrows the books with all the NDC numbers, i.e. from 000 to 900, exactly the same number from each category. The location of the numbers parenthesized with [n] indicate that it is the range value, i.e. entropy, for the case that n categories have equal numbers of books, or $log_{10}n$, which is the maximum value for having values in n categories. The location of the numbers parenthesized with (n) indicate that it is the range value for 2 categories in which one category has the possibility of n and the other has the possibility of $1 - n$. In this case, the maximum range value is $log_{10}2 = 0.30$ when n=0.5, i.e. half and half.

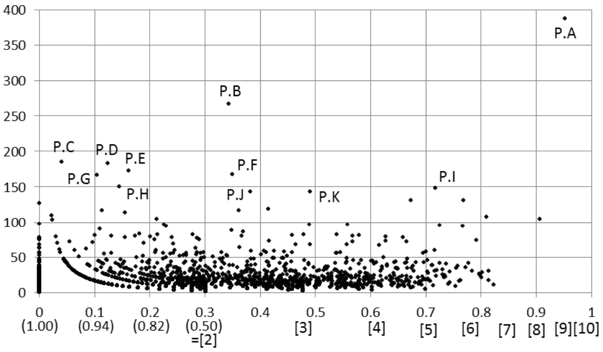


Figure 2. Correlation between the Range (x-axis) and the Strength (y-axis) of All Patrons

The patrons from P.A to P.K are named according to the order of the strength, or the number of borrowed books, so they are located in the upper part of the graph. Patron A (P.A) is located to the right-most and top-most place, which means he or she borrows the books from all the NDC categories with borrowing almost the same number of books each. Furthermore P.A borrows nearly 400 books, which is over 100 books more than the second one, i.e. P.B, who borrows more than 250 books.

The patrons C, D, E, G, and H are located in the left-most part of the graph having the value less than 0.2, which means they borrow books with one category more than 80% of times and other ones less than 20%. Thus they have very limited range of interest.

The patrons B, F, J, and K are located in the range with the range value from 0.3 to 0.5, which means, roughly speaking, they mainly borrow books with 2 or 3 categories.

C. Interest Profile for a Group of Patrons

The definition of profile of a patron is naturally extendable to a group of patrons. Let P be a group of patrons, then the profile of the group P is defined as follows:
$Prof(P) = \{< bt(c) >_{c \in NDC} \mid bt(c) = \#\{r \in CR \mid B(r) \in P$, and $Cls(r) = c\}\}$. In other words, the c-th component of $Prof(P)$ is the sum of the c-th component of the members of the group P; $Prof(P)_c = \sum_{p \in P} Prof(p)_c$.

Figure 3 shows the profiles of the affiliations of patrons. The names in the figure come after the faculty names; the graduate students and academic staff's affiliation names are assigned to the faculty names that are mostly closed to the patrons' affiliations. Patrons who have no appropriate relationship to a faculty is assigned to other (O).

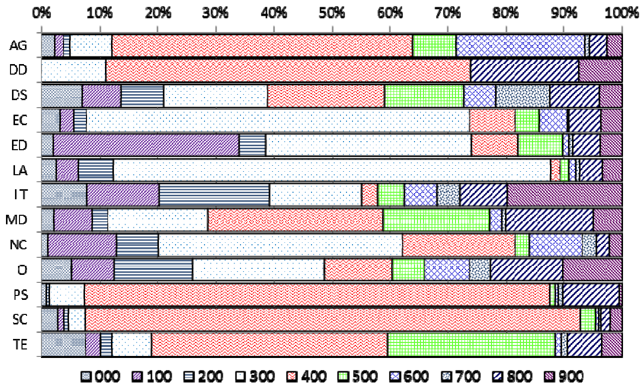


Figure 3. Profiles of Patrons’ Affiliations (or “Faculties”)

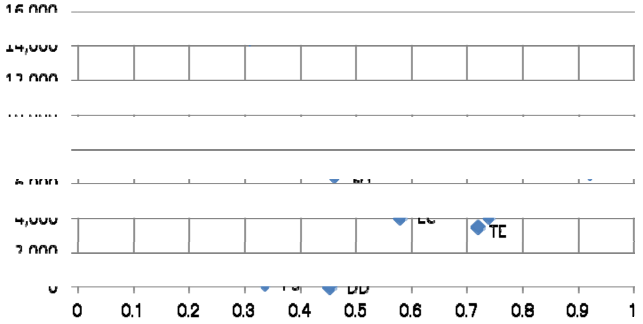


Figure 4. Correlation between Region Size (x-axis) and Strength (y-axis) of Faculties

Figure 4 shows the correlation between region size and strength of faculties. SC (Sciences) is far away from other faculties in both axes. It has the lowest value in region size and the highest in strength, which mean that patrons in SC borrow the books in natural sciences (NDC 400) mostly and the number of the borrowed books are quite high, which probably because that their places locate very close to the library and thus it is quite easy for them to visit the library and borrow many books.

PS (Pharmaceutical), DD (Dental), and LA (Law) are located in the left part from the line with the range size 0.5, which means that their patrons also borrows books of their expertise area mainly than other faculties. The reason why the strengths, or the numbers of borrowed books, of PS and DD is that their faculties locate in a different campus from where the library locates. Thus the patrons in PS and DD visit the library in order to get the books they could not find in the libraries in their own campus. LA is, on the other hand, located in the same campus as the library and also the number of the members is larger than that of PS and DD.

It is interesting to see that DS (Design) and MD (Medical) are located in the lower right part where their range size is relatively large. Even though MD locates in the same campus as PS and DD, its range size is far bigger than these two.

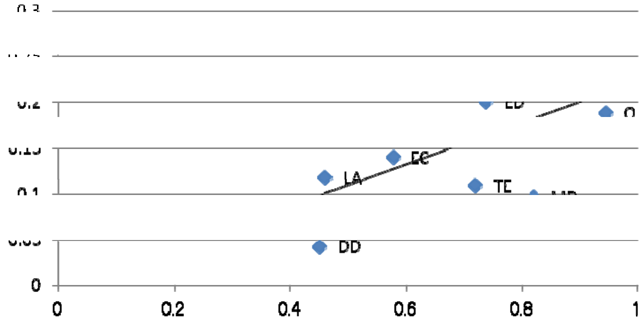


Figure 5. Comparison between the Range Size (x-axis) and the Average Range Size of the Members (y-axis) of Faculties

In order to find the reason of this fact, we investigate more on the patrons’ behavior. Anyway in some reason the members of MD visit the library in a different campus in order not to find the books relating to their study in their expertise field but to find books in a wide variety of books.

DS locates in a campus of it own, i.e. different campus from that of library and even farther than that of MD, PS, and DD. The strength, i.e. the number of borrowed books, is small probably because of this reason. DS is a faculty that relates both to engineering and design, and thus it is easy to guess that their interest range as a whole is wide. However it is still a surprising fact that its range size is larger than any other faculties including O (Other, or unclassified).

Another surprise is that LT (Letter) has high range size. LT patrons borrow books not only of literature (NDC 900), but also of those in other areas nearly as many as of literature.

We are able to define the concept of interest range of a group, or faculty in the current situation, in another point of view; the average range size of the members of the group. Formally, $AverageRange(P) = average\{Range(p) \mid p \in P\}$ for a group P of patrons. Figure 5 shows the difference of the range size and the average range size of faculties. The line in the graph is the linear approximation line.

From the definition, we can see that even if each member have low range size, that as a group is much wider if each member’s interest area is different. In other words, if all the members of a group have exactly the same profile and thus have the same range size, the average range size is the same value as of members. So we can say roughly that the difference between the range size and the average range size indicates the varieties of the interest profiles of the members.

From this view, or interpretation, quite many faculties are close to the approximation line and thus they have average interest varieties of the members. The faculties LT and NC have larger average range size than the one on the line. So we can say that these faculties have a wider variety of members in terms of interest ranges.

On the other hand DD, MD, and TE locate in the lower area of the line. So we can say that the members in these faculties have a narrower variety of interest profiles than other faculties; i.e. the members have somewhat similar interest ranges.

IV. Concluding Remarks

We proposed a new concept of interest area profile of a patron using a set of circulation records of a university library. Considering that one of the most important missions of a library, especially of a university library, is to help its patrons with learning more effectively and more efficiently in the comfortable and enjoyable environment. In order to achieve this goal, capturing the profiles of patrons in terms of their learning styles, their learning histories, their knowledge levels, their interests, their preferences, and so on, is important.

In order to compare the profiles of patrons in more practical ways, we additionally proposed new concepts of strength and range (size) of the profile. The strength intends to represent the eagerness or diligence to learning of the patron and we take the number of the borrowed books, or items, of the patron in this paper. The range size intends to represent the amount of eagerness or earnestness of the patron in terms of the width of the topic areas. We took the information entropy for defining this concept in this paper. We see the patrons' characters not only with profiles but also with these two values.

The concept of profile was extended to a group of patrons, especially to faculties. The concepts of strength and range were also extended to groups. We compared and characterized the faculties by these concepts and analyzed the characteristic features of faculties.

Our approach to library data analysis is quite new and there are no other such studies to our knowledge. Even though our current analysis methods are still in a primitive level, we are convinced from our experience that our methods have high potential as a tool for library marketing, and thus it will become an essential tool in the future.

The research directions of this paper include the topics:

- Investigation of more appropriate definitions of the concepts of the amount of eagerness to learning and the interest range size
- Exploration of defining other concepts such as style of learning, learning pace, preference in learning, etc. of a learner
- Utilization of circulation records and other data that are obtainable by libraries
- Usage of other data from different sources; for example usage of lecture data
- Systematizing the analysis methods and developing a learning support system and/or knowledge management system

This research was partly supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Scientific Research (C), 24500318, 2012.

References

[1] S. J. Cunningham and E. Frank, Market basket analysis of library circulation data, Proceedings of 6th International Conference on Neural Information Processing, 825-830. IEEE Computer Society, Perth, WA, Australia, 1999.

[2] J. Littman and L. S. Connaway, A Circulation Analysis of Print Books and e-Books in an Academic Research Library. Library Resources & Technical Services, 48(4), 256-262, 2004.

[3] T. Minami and E. Kim, Data Analysis Methods for Library Marketing, The 2009 International Conference on Database Theory and Application (DTA 2009), LNCS, 5899, 26-33, Springer, Heidelberg, 2009.

[4] T. Minami, Challenge toward Patron Understanding - A Search for Patron's Profile through Circulation Data of Library –, Kyushu University Library: Annual Report 2010/2011, 9-18, 2011. (in Japanese)

[5] T. Minami, Book Profiling from Circulation Records for Library Marketing – Beginning from Manual Analysis toward Systematization –. International Conference on Applied and Theoretical Information Systems Research (ATISR 2012), pp.15, 2012.

[6] T. Minami, Expertise Level Estimation of Library Books by Patron-Book Heterogeneous Information Network Analysis – Concept and Applications to Library's Learning Assistant Service –. The 8th International Symposium on Frontiers of Information Systems and Network Applications (FINA 2012), DOI 19.1109/WAINA.2012.184, pp.357-362, 2012.

[7] T. Minami and Y. Ohura, Toward Learning Support for Decision Making – Utilization of Library and Lecture Data –, 4th International Conference on Intelligent Decision Technologies (KES-IDT'2012), Springer Smart Innovation, Systems and Technologies 16, pp.137-147, 2012.

[8] T. Minami and K. Baba, Investigation of Interest Range and Earnestness of Library Patrons from Circulation Records, International Conference on e-Services and Knowledge Management (ESKM 2012), as a part of the 1st IIAI International Conference on Advanced Applied Informatics (IIAI-AAI 2012), IEEE CPS, DOI 10.1109/IIAI-AAI2012.15, pp.25-29, 2012.

[9] T. Minami and Y. Ohura, An Attempt on Effort-Achievement Analysis of Lecture Data for Effective Teaching, The 2012 International Conference on Database Theory and Application (DTA 2012), T.-h. Kim et al. (Eds.): EL/DTA/UNESST 2012, CCIS 352, pp. 50–57. Springer, 2012.

[10] T. Minami and Y. Ohura, Towards Development of Lecture Data Analysis Method and its Application to Improvement of Teaching, 2nd International Conference on Applied and Theoretical Information Systems Research (2ndATISR 2012), 2012. (to appear)

[11] Online Computer Library Center, Inc. (OCLC), WorldCat Collection Analysis. http://www.oclc.org/collectionanalysis/

[12] S. Yamada, Analysis of Library Book Circulation Data: Turnover of Open-shelf Books, Journal of College and University Libraries 69, 27-33, 2003. (in Japanese)

Generalised Atanassov Intuitionistic Fuzzy Sets

Ioan Despi
School of Science and Technology
University of New England
Armidale-2351, NSW, Australia
Email: despi@turing.une.edu.au

Dumitru Opriş
Faculty of Mathematics
West University of Timişoara
Timişoara, Romania
Email: opris@math.uvt.ro

Erkan Yalcin
Business School
University of New England
Armidale-2351, NSW, Australia
Email: eyalcin@une.edu.au

***Abstract*—When Atanssov created Intuitionistic Fuzzy Sets, he imposed the condition that the sum of membership and non-membership values for each point in the universe of discourse should be less than or equal to one. We challenge this constraint and define some new types of Intuitionistic Fuzzy Sets such that, for any point in the universe of discourse, the sum of membership and non-membership values can be greater than one, or their difference can be negative or positive, while one value is greater than the other, or the sum of their squares is less than or equal to one.**

***Keywords*-Intuitionistic fuzzy set, Interval-valued fuzzy sets.**

I. Introduction

`Fuzzy Sets` concept was introduced by Zadeh [1] in 1965. Given an non-empty universe of discourse X, one can define a fuzzy set A based on its membership function $\mu_A : X \to [0,1]$, that is A is a set with 'vague boundary' when compared with crisp sets, where $\mu_A : X \to \{0,1\}$. Of course, the bigger the value of $\mu_A(x)$ is, the greater the degree of membership of x to A is, so $\mu_A(x) = 1$ represents the full membership of x to A.

In 1983, Atanassov generalized the concept of fuzzy set by using two membership functions for the elements of the universe of discourse. The English version appeared in 1986 [2].

Definition 1. *Let X be an non-empty universe of discourse. An* `(Atanassov's) Initutionistic Fuzzy Set` *(AIFS or IFS) is described by:*

$$A = \{(x, \mu_A(x), \nu_A(x)) \,|\, x \in X\} \tag{1}$$

where μ_A is used to define the degree of membership (membership function) and ν_A is used to define the degree of non-membership (non-membership function) of x to A

$$\mu_A : X \to [0,1] \qquad \nu_A : X \to [0,1] \tag{2}$$

satisfying the condition

$$0 \le \mu_A(x) + \nu_A(x) \le 1, \, \forall x \in X \tag{3}$$

The word `intuitionistic` was added to suggest that the principle of excluded middle does not hold, so to say $\mu_A(x) + \nu_A(x) = 1$ is not true for all $x \in X$ if one interprets ν as a sort of negation of μ. Some operations on IFSs have been also introduced in [2]:

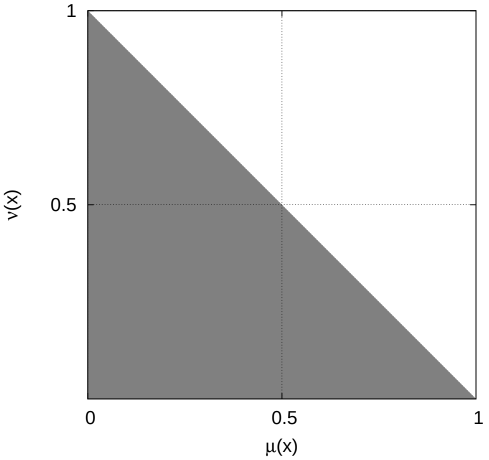


Fig. 1. Atanassov Intuitionistic Fuzzy Set, $\mu_A(x) + \nu_A(x) \le 1$, $\forall x \in X$

Definition 2. *Given two IFSs A and B over an universe of discourse X, one can define the following relations:*
$A \subset B$ *iff* $\forall x \in X \; \mu_A(x) \le \mu_B(x)$ *and* $\nu_A(x) \ge \nu_B(x)$
$A = B$ *iff* $A \subset B$ *and* $B \subset A$
as well as the following operations [2]:
$\bar{A} = \{(x, \nu_A(x), \mu_A(x) \,|\, x \in X\}$
$A \cap B = \{(x, \mu_{A\cap B}(x), \nu_{A\cap B}(x)) \,|\, x \in X\}$, *where*
$\mu_{A\cap B}(x) = \min\{\mu_A(x), \mu_B(x)\}$ *and*
$\nu_{A\cap B}(x) = \max\{\nu_A(x), \nu_B(x)\}$
$A \cup B = \{(x, \mu_{A\cup B}(x), \nu_{A\cup B}(x)) \,|\, x \in X\}$, *where*
$\mu_{A\cup B}(x) = \max\{\mu_A(x), \mu_B(x)\}$, *and*
$\nu_{A\cup B}(x) = \min\{\nu_A(x), \nu_B(x)\}$
$A + B = \{(x, \mu_{A+B}(x), \nu_{A+B}(x)) \,|\, x \in X\}$, *where*
$\mu_{A+B}(x) = \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x)$, *and*
$\nu_{A+B}(x) = \nu_A(x) \cdot \nu_B(x)$
$A \cdot B = \{(x, \mu_{A\cdot B}(x), \nu_{A\cdot B}(x)) \,|\, x \in X\}$, *where*
$\mu_{A\cdot B}(x) = \mu_A(x) \cdot \mu_B(x)$, *and*
$\nu_{A\cdot B}(x) = \nu_A(x) + \nu_B(x) - \nu_A(x) \cdot \nu_B(x)$

In [2] it is proved that the operations $\cap$ and $\cup$ are commutative, associative, distributive among themselves, idempotent and satisfy De Morgan's law; the operations $+$ and $\cdot$ are commutative, associative, satisfy De Morgan's law, and are distributive with respect to $\cap$ and $\cup$.

To measure hesitancy of membership of an element to a intuitionistic fuzzy set, Atanassov [2] used a third function.

Definition 3. *Given an IFS $A = \{(x, \mu_A(x), \nu_A(x)) \,|\, x \in X\}$ over an non-empty universe of discourse X, the degree of indeterminacy of x to A is given by*

$$\pi_A(x) = 1 - \mu_A(x) - \nu_A(x) \tag{4}$$

The function $\pi(x)$ is also called the intuitionistic fuzzy index, the hesitancy, or the ignorance degree of x to A. Clearly, $0 \leq \pi_A(x) \leq 1, \forall x \in X$. If $\pi_A(x) = 0, \forall x \in X$, then $\nu(x) = 1 - \mu(x)$ and the intuitionistic fuzzy set A is reduced to an ordinary fuzzy set A:

$$A = \{(x, \mu_A(x), 1 - \mu_A(x)) \,|\, x \in X\} \tag{5}$$

Some authors (Yusoff et al. [3], Zeng and Li [4]) consider that the third parameter $\pi(x)$ cannot be omitted from the definition of an AIFS:

$$A = \{(x, \mu_A(x), \nu_A(x), \pi_A(x)) \,|\, x \in X\} \tag{6}$$

and so an AIFS can be depicted as in Figure 2. A line parallel to the $\mu_A(x) + \nu_A(x) = 1$ diagonal describes a crisp set of elements x with the same level of hesitancy to A.

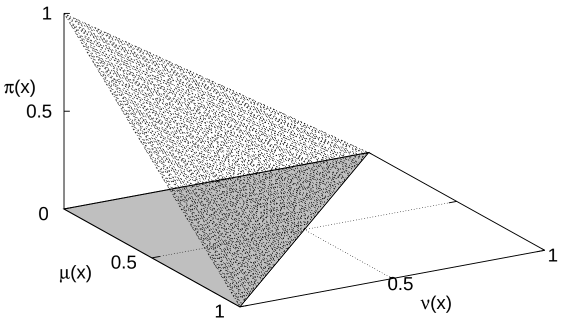


Fig. 2. AIFS with explicit fuzzy index: $\pi_A(x) = 1 - \mu_A(x) - \nu_A(x)$

IFSs are not a trivial generalization of ordinary Fuzzy Sets (FS) because they can be represented in the form $[A, B]$, where A and B are ordinary fuzzy sets or, even more, one can define modal operators *necessity* and *possibility* over IFS (see Atanassov [5]):

$$\Box A = \{(x, \mu_A(x), 1 - \mu_A(x)) \,|x \in X\} \tag{7}$$

$$\Diamond A = \{(x, 1 - \nu_A(x), \nu_A(x)) \,|x \in X\} \tag{8}$$

such that

$$\begin{aligned} \Box A \quad &\subset \quad A \quad \subset \quad \Diamond A && (9)\\ \Box \overline{A} \quad &= \quad \overline{\Diamond A} && (10)\\ \Diamond \overline{A} \quad &= \quad \overline{\Box A} && (11)\\ & && (12) \end{aligned}$$

while in ordinary fuzzy sets we have

$$\Box A = A = \Diamond A \tag{13}$$

Of course, all FS results can be easily generalized for IFS. Deschrijver and Kerre ([6], [7]) proved that AIFSs can also be seen as L-fuzzy sets in the sense of Goguen [8] by taking the lattice ${}_\star L = \{(x_1, x_2) \in [0, 1]^2 \,|\, x_1 + x_2 \leq 1\}$ with the partial order $\leq_{{}_\star L}$ defined as

$$(x_1, x_2) \leq_{{}_\star L} (y_1, y_2) \iff x_1 \leq y_1 \wedge x_2 \geq y_2$$

In [6] it is proved that $({}_\star L, \leq_{{}_\star L})$ is a complete and bounded lattice with the smallest element $0_{{}_\star L} = (0, 1)$ and the greatest element $1_{{}_\star L} = (1, 0)$. This lattice (and the similar ones we'll introduce later in this section) can then be used to define intuitionistic fuzzy negation [9]:

Definition 4. *A function $\mathcal{N} : L \to L$, where $\mathcal{N}$ is strictly decreasing, continuous, and $\mathcal{N}(0_L) = 1_L$, $\mathcal{N}(1_L) = 0_L$ is called an intuitionistic fuzzy negation.*
$\mathcal{N}$ is a strong fuzzy negation if it is involutive, that is $\mathcal{N}(\mathcal{N}(x)) = x$ holds for all $x \in L$.

Recall that a function $\varphi : [0, 1] \to [0, 1]$ that is continuous and strictly increasing, such that $\varphi(0) = 0$ and $\varphi(1) = 1$, is called automorphism.

Deschrijver and Kerre [6] also proved that any strong intuitionistic fuzzy negation $\mathcal{N}$ is characterised by a strong negation $N : [0, 1] \to [0, 1]$ such that, for all $(x_1, x_2) \in L$, $\mathcal{N}(x_1, x_2) = (N(1-x_2), 1-N(x_1))$. Trillas et al. [10] proved that $N : [0, 1] \to [0, 1]$ is a strong negation if and only if there exists an automorphism φ of the unit interval such that $N(x) = \varphi^{-1}(1 - \varphi(x))$.

II. Generalised Intuitionistic Fuzzy Sets

In the sequel, let X be a non-empty set and let us consider $A = \{(x, \mu_A(x), \nu_A(x)) \,|\, x \in X\}$, where $\mu_A : X \to [0, 1]$ and $\nu_A : X \to [0, 1]$, are used to define the degree of membership and the degree of non-membership, respectively, of x to A. Given an element $x \in X$, the condition $\mu(x) + \nu(x) \leq 1$ included in the definition of AIFSs suggests that if one of the two membership/non-membership functions has a big value (close to 1), the other function should have a very small value (close to 0) such that their sum is less than one. But it is possible that both functions have small values, that is membership degree and non-membership degree are quite insignificant. As one can see on Figure 1, both $\mu(x)$ and $\nu(x)$ have small values (less than 0.5) in the square with opposite corners $(0; 0)$ and $(0.5; 0.5)$ and only one of them has a big value (bigger than 0.5) in the two remaining triangles. The two cases are equal possible, in the sense that they cover surfaces of same size. The definition of an AIFS shows proneness to many generalisations. Atanassov's definition assumes that the membership and non-membership functions must have their sum smaller than or equal to one for every element of the universe of discourse. While it is a good hypothesis in many practical situations, there are cases when this constraint does not work and it must be replaced by other relations.

The definition of an AIFS shows proneness to many generalisations. A first extension was proposed by T. K. Mondal and S. K. Samanta [11], where the functions μ and ν satisfy the condition $\mu(x) \wedge \nu(x) \leq 0.5$, $\forall x \in X$. A second extension to both Atanassov and Mondal-Samanta models was proposed by H.C. Liu [12], by using a constant $L \in [0,1]$ such that the functions μ and ν satisfy the condition $\mu(x)+\nu(x) \leq 1+L, \forall x \in X$ and $L \in [0,1]$.

The main contribution of this paper is the replacement of the original Atanassov relation by some other conditions. We think that both functions can take any values in $[0,1]$ as long as the ignorance degree of x to A is non-negative and less than or equal to one (after we reshape it in an appropriate way). The condition $0 \leq \mu_A(x)+\nu_A(x) \leq 1, \forall x \in X$ is just a choice and it can be replaced by others. If Atanassov's original definition dealt with the left bottom triangle of the unit square, we will consider all four right angle triangles in the unit square (with the right angle a corner of the square), plus some other combinations of them, obtained by combining triangles between the square's diagonals, as well as the inscribed circle in the square. Therefore, our first generalisation (GAIFS1) is given by:

Definition 5. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS1) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \,|\, x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ satisfy the condition*

$$\mu(x)+\nu(x) \geq 1, \forall x \in X \tag{14}$$

The degree of indeterminacy of x to A is defined as

$$\pi_A(x) = \mu_A(x)+\nu_A(x)-1 \tag{15}$$

and, once again, clearly $0 \leq \pi_A(x) \leq 1, \forall x \in X$.
If $\pi_A(x) = 0, \forall x \in X$ then $\nu(x) = 1-\mu(x)$ and the intuitionistic fuzzy set A is reduced to an ordinary fuzzy set $A = \{(x, \mu_A(x), 1-\mu_A(x)) \,|\, x \in X\}$.

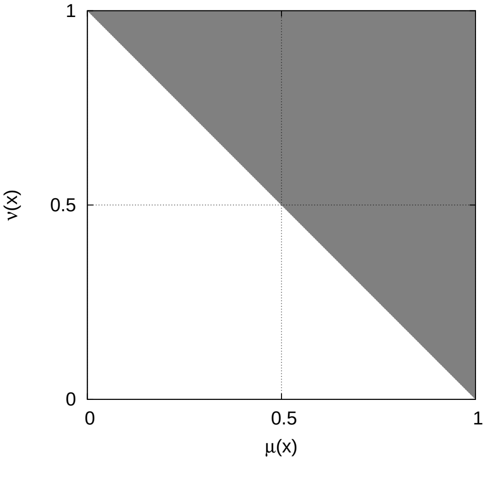


Fig. 3. First (GAIFS1) new definition of an AIFS

If we take the set

$$L^\star = \{(x_1,x_2) \in [0,1]^2 \,|\, x_1+x_2 \geq 1\}$$

with the partial order $\leq_{L^\star}$ defined as

$$(x_1,x_2) \leq_{L^\star} (y_1,y_2) \iff x_1 \geq y_1 \wedge x_2 \leq y_2$$

and , for each $A \subset L^\star$, we define:

$$\begin{aligned}\sup A &= (\inf\{x_1 \in [0,1] \,|\, \exists x_2 \in [0,1], (x_1,x_2) \in A\},\\ &\quad \sup\{x_2 \in [0,1] \,|\, \exists x_1 \in [0,1], (x_1,x_2) \in A\})\\ \inf A &= (\sup\{x_1 \in [0,1] \,|\, \exists x_2 \in [0,1], (x_1,x_2) \in A\},\\ &\quad \inf\{x_2 \in [0,1] \,|\, \exists x_1 \in [0,1], (x_1,x_2) \in A\})\end{aligned}$$

then $(L, \leq_{L^\star})$ is a complete lattice. The lattice can be defined as an algebraic structure $(L^\star, \wedge, \vee)$ where the meet and join operators are defined respectively

$$\begin{aligned}(x_1,x_2) \wedge (y_1,y_2) &= (\max(x_1,y_1), \min(x_2,y_2))\\ (x_1,x_2) \vee (y_1,y_2) &= (\min(x_1,y_1), \max(x_2,y_2))\end{aligned}$$

The smallest element is $0_{L^\star} = (1,0)$ and the greatest element is $1_{L^\star} = (0,1)$. Therefore, an GAIFS1 A is a L-fuzzy set whose L-membership function $\chi^A \in (L^\star)^X = \{\chi : X \to L^\star\}$ is defined such that for each $x \in X$, $\chi^A(x) = (\mu_A(x), \nu_A(x))$. The shaded area in Figure 4 is the set of elements $x = (x_1, x_2)$ belonging to $L^\star$.

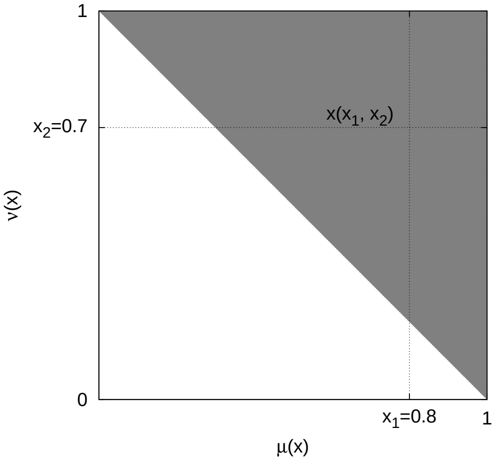


Fig. 4. New Intuitionistic Fuzzy Set as a L-fuzzy Set

The order $\leq_{L^\star}$ of $L^\star$ induces a natural partial order on $(L^\star)^X$: given $\chi^A, \chi^B \in (L^\star)^X$, we say that $\chi^A \leq_{L^\star} \chi^B$ if and only if $\chi^A(x) \leq_{L^\star} \chi^B(x)$ for all $x \in X$.
Thus, $((L^\star)^X, \leq_{L^\star})$ is a bounded and complete lattice in which the least and greatest elements are $\chi^{0_{L^\star}}$ and $\chi^{1_{L^\star}}$, respectively. Of course, $\chi^{0_{L^\star}}(x) = 0_{L^\star}$ and $\chi^{1_{L^\star}}(x) = 1_{L^\star}$, for all $x \in X$. The same considerations apply to all other L lattices we will define in the sequel. By using

$$A \subset B \text{ iff } \forall x \in X\, \mu_A(x) \geq \mu_B(x) \text{ and } \nu_A(x) \leq \nu_B(x)$$

all AIFS original operations can be re-written for GAIFS1.

The second generalisation (GAIFS2) is given by:

Definition 6. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS2) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \,|\, x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ satisfy the condition*

$$\mu(x) \leq \nu(x), \forall x \in X \tag{16}$$

ISBN: 978-1-61208-254-7

The degree of indeterminacy of x to A is defined as

$$\pi_A(x) = \nu_A(x) - \mu_A(x) \tag{17}$$

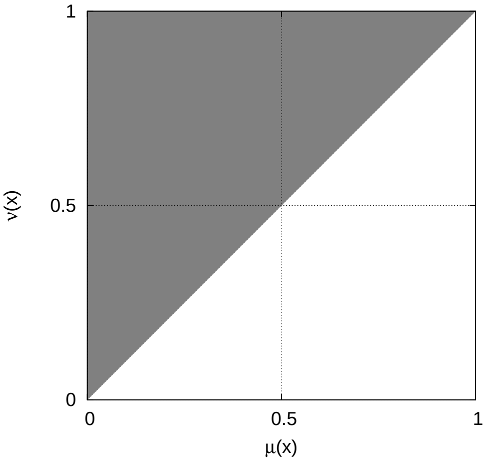


Fig. 5. Second (GAIFS2) new definition of an AIFS

The corresponding complete lattice in this case is

$$^{\star}L = \{(x_1, x_2) \in [0,1]^2 \mid x_1 \leq x_2\}$$

with the partial order $\leq_{^{\star}L}$ defined as

$$(x_1, x_2) \leq_{^{\star}L} (y_1, y_2) \iff x_1 \leq y_1 \wedge x_2 \leq y_2$$

As described by Deschrijver in [13], if $x = (x_1, x_2) \in {}^{\star}L$, then the length $x_2 - x_1$ is called the uncertainty and is denoted by x_π. The interval $[x_1, x_2]$ gives the “range between a pessimistic and an optimistic truth evaluation of a proposition” [13]. The smallest and the largest elements in $^{\star}L$ are $0_{^{\star}L} = (0, 0)$ and $1_{^{\star}L} = (1, 1)$, respectively. By using

$$A \subset B \text{ iff } \forall x \in X\ \mu_A(x) \leq \mu_B(x) \text{ and } \nu_A(x) \leq \nu_B(x)$$

all AIFS original operations can be re-written for GAIFS2. The third generalisation (GAIFS3) is given by:

Definition 7. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS3) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \mid x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0, 1]$ and $\nu_A : X \to [0, 1]$ satisfy the condition*

$$\mu(x) \geq \nu(x),\ \forall x \in X \tag{18}$$

The degree of indeterminacy of x to A is defined as

$$\pi_A(x) = \mu_A(x) - \nu_A(x) \tag{19}$$

The corresponding complete lattice in this case is

$$L_\star = \{(x_1, x_2) \in [0,1]^2 \mid x_1 \geq x_2\}$$

with the partial order $\leq_{L_\star}$ defined as

$$(x_1, x_2) \leq_{L_\star} (y_1, y_2) \iff x_1 \geq y_1 \wedge x_2 \geq y_2$$

The interval $[x_2, x_1]$ gives, once again, the “range between a pessimistic and an optimistic truth evaluation of a proposition”, as stated in [13]. The smallest and the largest elements in $L_\star$ are $0_{L_\star} = (1, 1)$ and $1_{L_\star} = (0, 0)$, respectively.

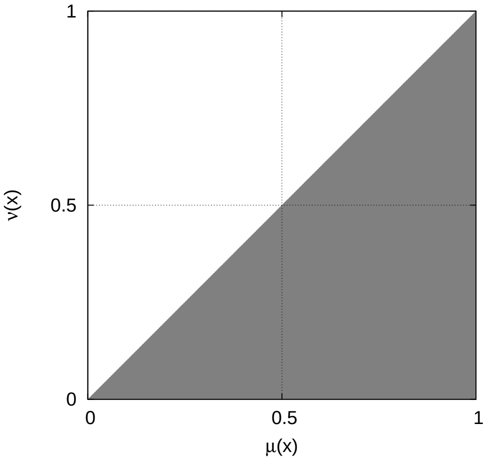


Fig. 6. Third (GAIFS3) new definition of an AIFS

By using

$$A \subset B \text{ iff } \forall x \in X\ \mu_A(x) \geq \mu_B(x) \text{ and } \nu_A(x) \geq \nu_B(x)$$

all AIFS original operations can be re-written for GAIFS3.

The fourth generalisation (GAIFS4) is given by:

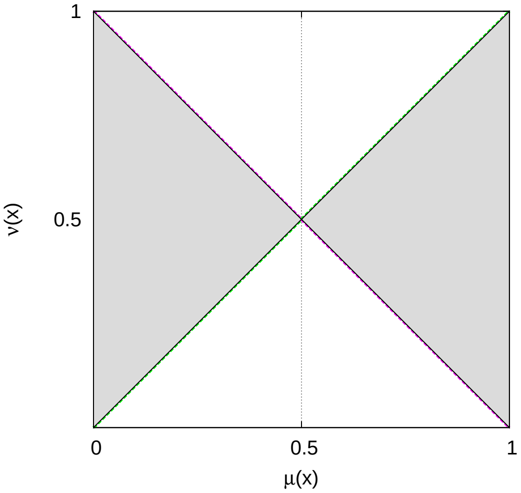


Fig. 7. Fourth (GAIFS4) new definition of an AIFS

Definition 8. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS4) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \mid x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0, 1]$ and $\nu_A : X \to [0, 1]$ satisfy the condition*

$$\begin{gathered}\mu(x) \geq \nu(x),\ \textit{and}\ \mu(x) + \nu(x) \geq 1,\ \textit{or} \\ \mu(x) \leq \nu(x),\ \textit{and}\ \mu(x) + \nu(x) \leq 1, \forall x \in X\end{gathered} \tag{20}$$

The fifth generalisation (GAIFS5) is given by:

Definition 9. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS5) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \mid x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ satisfy the condition*

$$\begin{aligned} \mu(x) \leq \nu(x), \ \ \textit{and} \ \ \mu(x)+\nu(x) \geq 1, \ \ \textit{or} \\ \mu(x) \geq \nu(x), \ \ \textit{and} \ \ \mu(x)+\nu(x) \leq 1, \forall x \in X \end{aligned} \tag{21}$$

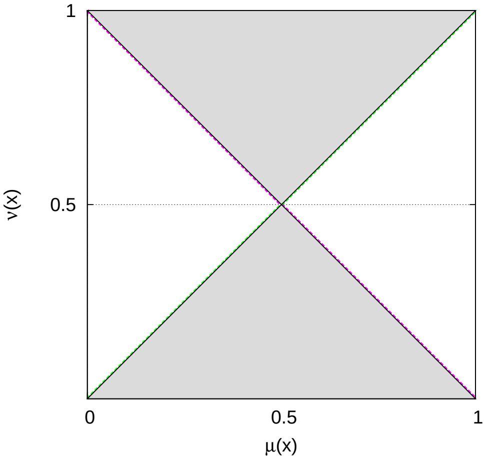


Fig. 8. Fifth (GAIFS5) new definition of an AIFS

It is also possible to consider the case when functions μ and ν cannot take values in the neighbourhoods of the four corners of the square $[0,1]^2$, that is giving our sixth generalization (GAIFS6):

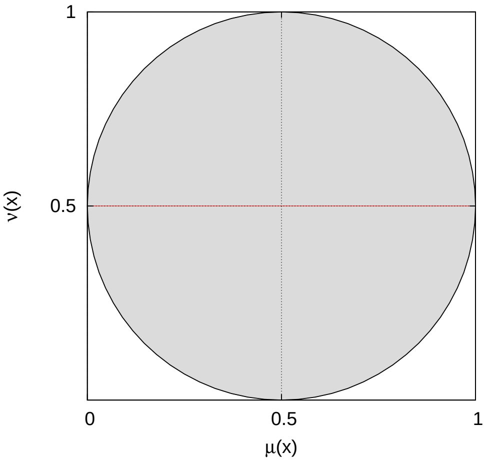


Fig. 9. Sixth (GAIFS6) new definition of an AIFS

Definition 10. *Let X be a non-empty universe of discourse. Then a generalised Atanassov intuitionistic fuzzy set (GAIFS6) is described by $A = \{(x, \mu_A(x), \nu_A(x)) \mid x \in X\}$, where the membership/non-membership functions $\mu_A : X \to [0,1]$ and $\nu_A : X \to [0,1]$ satisfy the condition*

$$\mu^2(x) + \nu^2(x) \leq 1, \forall x \in X \tag{22}$$

III. Generalised Interval-valued Fuzzy Sets

Interval-valued fuzzy sets were introduced by Zadeh [14], Grattan-Guiness [15], Jahn [16], and Sambuc [17]. Because it is hard in real life problems to assign a precise membership degree to elements in fuzzy sets, this was replaced by an interval $[\mu_1, \mu_2]$, with $0 \leq \mu_1 \leq \mu_2 \leq 1$ to which the membership degree belongs. The length of the interval is a measure of uncertainty of the membership of an element $x \in X$ to an interval-valued fuzzy set (IVFS) A. It is similar to the degree of indeterminacy of x to A in AIFS. The FS (AIFS) standard operations (union, intersection, complementation) can be extended to IVFS in the canonical way. If $M = [\mu_1, \mu_2]$ and $N = [\nu_1, \nu_2]$ are two IVFS, then for all $x \in X$

$$(M \cap N)(x) = [\min(\mu_1(x), \nu_1(x)), \min(\mu_2(x), \nu_2(x))] \tag{23}$$
$$(M \cup N)(x) = [\max(\mu_1(x), \nu_1(x)), \max(\mu_2(x), \nu_2(x))] \tag{24}$$
$$\bar{M}(x) = [1-\mu_2(x), 1-\mu_1(x)] \tag{25}$$

The equivalence between AIFS and IVFS has been studied in [18] and [6]. If $A = [\mu_1, \mu_2]$, $0 \leq \mu_1 \leq \mu_2 \leq 1$ then $\mu_1 - \mu_2 \leq 0$ so $\mu_1 + 1 - \mu_2 \leq 1$. By defining $\mu_1 = \mu$ and $\nu = 1 - \mu_2$, we obtain an AIFS. Conversely, starting with an AIFS $A = (\mu, \nu), \mu + \nu \leq 1$, we can create the interval $[\mu, 1-\nu]$ to correspond to an IVFS.

In the case of our first generalisation (GAIFS1), where $\mu(x) + \nu(x) \geq 1$ for all $x \in X$, the above equivalence still holds. If $A = [\mu_1, \mu_2]$, $0 \leq \mu_1 \leq \mu_2 \leq 1$ then $0 \leq \mu_2 - \mu_1$ so $1 \leq \mu_2 + 1 - \mu_1$ and, by defining $\mu = 1 - \mu_1$ and $\nu = \mu_2$, we obtain an AIFS.

The above equivalence also holds trivially for GAIFS2, GAIFS3, GAIFS4, and GAIFS5 generalizations. For instance, in the case of GAIFS2, $\mu(x) \leq \nu(x)$, so the corresponding IVFS should be characterised by $[\mu, \nu]$. In the case of GAIFS3, $\nu(x) \leq \mu(x)$ holds, so the corresponding IVFS should be characterised by $[\nu, \mu]$, etc. For GAIFS6 case, we take the IVFS to be given by the interval $[\mu, 1-\nu]$. Then $\mu^2 \leq (1-\nu)^2$ and $\mu^2+\nu^2 \leq (1-\nu)^2+\nu^2 \leq 1-2\nu+2\nu^2 \leq 1+2\nu(\nu-1) \leq 1$.

IV. Automorphisms

We deal with our first generalization GAIFS1 only, the other cases are treated in a similar way. GAIFS1 is equivalent to the lattice $(L^\star, \wedge, \vee, 0_{L^\star}, 1_{L^\star})$, where

$$L^\star = \{(x_1, x_2) \in [0,1]^2 \mid x_1 + x_2 \geq 1\}$$

with the partial order $\leq_{L^\star}$ defined as

$$(x_1, x_2) \leq_{L^\star} (y_1, y_2) \iff x_1 \geq y_1 \wedge x_2 \leq y_2$$

and $0_{L^\star} = (1,0)$ and $1_{L^\star} = (0,1)$. The operations are defined as

$$(x_1, x_2) \wedge (y_1, y_2) = (x_1 \vee y_1, x_2 \wedge y_2) \tag{26}$$
$$(x_1, x_2) \vee (y_1, y_2) = (x_1 \wedge y_1, x_2 \vee y_2) \tag{27}$$

and one can easily verify that

$$\begin{aligned}(x_1,x_2)\wedge(1,0)&=(x_1\vee 1,x_2\wedge 0)=(1,0)=0_{L^\star}\\(x_1,x_2)\vee(1,0)&=(x_1\wedge 1,x_2\vee 0)=(x_1,x_2)\\(x_1,x_2)\wedge(0,1)&=(x_1\vee 0,x_2\wedge 1)=(x_1,x_2)\\(x_1,x_2)\vee(0,1)&=(x_1\wedge 0,x_2\vee 1)=(0,1)=1_{L^\star}\end{aligned}$$

As in any lattice, the meet operator $\wedge$ and the join operator $\vee$ are related to the ordering $\leq_{L^\star}$ by the following equivalences: for every $x=(x_1,x_2), y=(y_1,y_2)\in L^\star$

$$x\leq_{L^\star} y \iff x\vee y=y \iff x\wedge y=x$$

Indeed, if $(x_1,x_2)\leq_{L^\star}(y_1,y_2)$, then

$$\begin{aligned}(x_1,x_2)\vee(y_1,y_2)&=(x_1\wedge y_1,x_2\vee y_2)=(y_1,y_2)\\(x_1,x_2)\wedge(y_1,y_2)&=(x_1\vee y_1,x_2\wedge y_2)=(x_1,x_2)\end{aligned}$$

Definition 11. *An* `automorphism` *of $L^\star$ is a bijection $f: L^\star\to L^\star$ such that $f(x)\leq_{L^\star} f(y)$ if and only if $x\leq_{L^\star} y$, for all $x,\, y\in L^\star$.*
An `anti-automorphism` *of $L^\star$ is a bijection $f: L^\star\to L^\star$ such that $f(x)\leq_{L^\star} f(y)$ if and only if $y\leq_{L^\star} x$, for all $x,\, y\in L^\star$.*

It is obvious that an automorphism takes $0_{L^\star}$ and $1_{L^\star}$ to themselves, while an anti-automorphism interchanges these elements.

We denote by $\mathbf{A}ut(L^\star)$ the set of all automorphisms of $L^\star$ and by $\mathbf{M}ap(L^\star)$ the set of all automorphisms and anti-automorphisms of $L^\star$. They are groups under the composition of morphisms and $\mathbf{A}ut(L^\star)$ is a normal subgroup of order 2 of $\mathbf{M}ap(L^\star)$. [19]

Let $f\in\mathbf{A}ut(L^\star)$. Since $f(x)\leq_{L^\star} f(y)$ if and only if $x\leq_{L^\star} y$, for all $x,\, y\in L^\star$, then

$$\begin{aligned}f(x\vee y)&=f(x)\vee f(y)\\f(x\wedge y)&=f(x)\wedge f(y)\\f((1,0))&=(1,0)\\f((0,1))&=(0,1)\end{aligned}$$

If f is an automorphism of $[0,1]$, then, for $(x_1,x_2)\in L^\star$, $(x_1,x_2)\mapsto(f(x_1),f(x_2))$ is an automorphism of $L^\star$.

An anti-automorphism $\mathcal{N}$ such that $\mathcal{N}(\mathcal{N}(x))=x$, for all $x\in L^\star$ is an `involution` or `negation`. Obviously, $\mathcal{N}(0_{L^\star})=1_{L^\star}$ and $\mathcal{N}(1_{L^\star})=0_{L^\star}$.

All elements but identity of $\mathbf{A}ut(L^\star)$ are of infinite order; all anti-automorphisms are of infinite order or of order two. The order two anti-automorphisms are `involutions` and their set is denoted by $\mathbf{I}nv(L^\star)$. One (classical) involution is $\alpha: L^\star\to L^\star$ given by $\alpha(x_1,x_2)=(1-x_2,1-x_1)$, all other involutions are of the form $f^{-1}\alpha f$, for any $f\in\mathbf{M}ap(L^\star)$.

In [20] it is proved that if $\mathcal{N}$ is an involutive negator on $L^\star$ (negation) and $N:[0,1]\to[0,1]$, $N(a)=pr_1\mathcal{N}(a,1-a)$, for all $a\in[0,1]$, then $\mathcal{N}(x_1,x_2)=(N(1-x_2),1-N(x_1))$, for all $(x_1,x_2)\in L^\star$.

A triangular norm on $L^\star$ (`t-norm`) is any increasing, commutative, associative mapping $T: L^\star\times L^\star\to L^\star$ satisfying $T(1_{L^\star},x)=x$ for all $x\in L^\star$. A triangular co-norm on $L^\star$ (`t-conorm`) is any increasing, commutative, associative mapping $S: L^\star\times L^\star\to L^\star$ satisfying $S(0_{L^\star},x)=x$ for all $x\in L^\star$.

V. Conclusion

We introduced six possible new definitions for intuitionistic fuzzy sets by challenging the base condition in Atanassov's definition. While keeping the two membership functions, we extended the range of possible combinations between them and showed some interesting properties. We intend to further develop the approach for measuring similarity and compatibility between different sorts of intuitionistic fuzzy sets.

References

[1] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.

[2] K. Atanassov, "Intuitionistic fuzzy sets," *Fuzzy Sets Syst.*, vol. 20, pp. 87–96, August 1986. [Online]. Available: http://dx.doi.org/10.1016/S0165-0114(86)80034-3

[3] B. Yusoff, I. Taib, L. Abdullah, and A. F. Wahab, "A new similarity measure on intuitionistic fuzzy sets," *World Academy of Science,Engineering and Technology*, vol. 78, pp. 36–40, 2011.

[4] W. Zeng and H. Li, "Correlation coefficient of intuitionistic fuzzy sets," *Journal of Industrial Engineering International*, vol. 3, pp. 33–40, July 2007.

[5] K. T. Atanassov, "Intuitionistic fuzzy sets: past, present and future," in *EUSFLAT Conf.*, M. Wagenknecht and R. Hampel, Eds. University of Applied Sciences at Zittau/Görlitz, Germany, 2003, pp. 12–19.

[6] G. Deschrijver and E. E. Kerre, "On the relationship between some extensions of fuzzy set theory," *Fuzzy Sets Syst.*, vol. 133, no. 2, pp. 227–235, 2003.

[7] G. Deschrijver, C. Cornelis, and E. Kerre, "On the representation of intuitionistic fuzzy t-norms and t-conorms," *Fuzzy Systems, IEEE Transactions on*, vol. 12, no. 1, pp. 45 – 61, feb. 2004.

[8] J. Goguen, "L-fuzzy sets," *J. Math. Anal. Appl.*, vol. 18, pp. 145–174, 1967.

[9] B. R. C. Bedregal, "On interval fuzzy negations," *Fuzzy Sets and Systems*, vol. 161, no. 17, pp. 2290–2313, 2010.

[10] E. Trillas, C. Alsina, and J. Terricabras, *Introducción a la Lógica Borrosa*, ser. Ariel Matemática. Ariel, 1995. [Online]. Available: http://books.google.com.au/books?id=W1wOPQAACAAJ

[11] T. K. Mondal and S. K. Samanta, "Generalized intuitionistic fuzzy sets," *Journal of Fuzzy Mathematics*, vol. 10, pp. 839–861, 2002.

[12] H. C. Liu, "Liu's generalized intuitionistic fuzzy sets," *Journal of Educational Measurements and Statistics*, pp. 69–81, 2010. [Online]. Available: http://gsems.ntcu.edu.tw/center/public-year2-pdf/year18/18_1_4_Liu's%20%Generalized%20Intuitionistic%20Fuzzy%20Sets69-81.pdf

[13] G. Deschrijver, "Generalized arithmetic operators and their relationship to t-norms in interval-valued fuzzy set theory," *Fuzzy Sets and Systems*, vol. 160, no. 21, pp. 3080–3102, 2009.

[14] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning - i," *Inf. Sci.*, vol. 8, no. 3, pp. 199–249, 1975.

[15] I. Grattan-Guiness, "Fuzzy membership mapped onto interval and many-valued quantities," *Z. Math. Logik. Grundladen Math*, no. 22, pp. 149–160, 1975.

[16] K. Jahn, "Intervall-wertige mengen." *Math. Nach.*, vol. 68, pp. 115–132, 1975.

[17] R. Sambuc, "Fonctions -floues. application laide au diagnostic en pathologie thyroidienne," Univ. Marseille, Tech. Rep., 1975.

[18] K. Atanassov and G. Gargov, "Interval valued intuitionistic fuzzy sets," *Fuzzy Sets Syst.*, vol. 31, no. 3, pp. 343–349, Jul. 1989. [Online]. Available: http://dx.doi.org/10.1016/0165-0114(89)90205-4

[19] M. Gehrke, C. Walker, and E. Walker, "Some comments on interval valued fuzzy sets," *Int. Jour. Intelligent Systems*, no. 11, pp. 751–759, 1996.

[20] C. Cornelis, G. Deschrijver, and E. Kerre, "Implication in intuitionistic fuzzy and interval-valued fuzzy set theory: construction, classification, application," *International journal of approximate reasoning*, vol. 35, no. 1, pp. 55–95, 2004.

Distributed Cognition in Software Engineering

A Mapping Study

Mathieu Lavallée, Pierre N. Robillard, Samuel Paul
Département de génie informatique et logiciel
Polytechnique Montréal
Montreal, Canada
{mathieu.lavallee, pierre.robillard, samuel.s-paul}@polymtl.ca

Abstract—**This paper presents a mapping of the current research in distributed cognition in software engineering, using the systematic literature review approach. The result of the review shows that the literature focuses on the situational awareness of the software development team, mostly through the identification of team experts and the dissemination of task details. Research on cognitive support tools are mostly speculative, with little validation of the recommendations provided. Research on the impact of spatial disposition on team cognition is emerging, along with research on the impacts of certain emotional states. Very few papers are however concerned on the impacts of project, process and organizational constraints on team problem solving.**

Keywords; Distributed cognition; software development team; literature review; team meta-cognition; team situation awareness; team problem solving

I. Introduction

The concept of distributed cognition was first introduced in 1995 by Edwin Hutchins et al. [1] to explain how an individual can resolve problems through means beyond his internal cognitive processes. Distributed cognition observes how problems are resolved through the cognitive system around one or more minds.

The observation of distributed cognition can be applied to one individual in his/her environment. In that case, the researcher observes how the person interacts with tools around him (work documents, written notes, software, etc.).

Distributed cognition becomes especially interesting when applied to the study of teamwork. The observation of distributed cognition in team settings shows how information is transferred within the team and how solutions are created, judged and transformed by teammates.

The objective of this mapping study is to categorize the main answers given by the literature, along with potentially interesting future research avenues.

The selection process used for the literature review is presented in Section II. Section III presents an overview of the selected papers. Section IV presents a discussion of the conclusions of the selected papers as they relate to the concept of distributed cognition. Finally, Section V presents the overall conclusions of the review and introduces future research avenues.

II. Methodology

As a mapping study [4], this review is based on a lightweight version of the systematic literature review process described in the works of Barbara Kitchenham et al. [2, 3]. This section describes how the databases were searched in order to find the relevant papers and the criteria used for the paper selection and finally how the mapping and the conclusions were obtained.

A. Databases and Search String

The objective of the search is to find the published papers relevant to the subject of distributed cognition research in software engineering. The search was limited to the "Compendex" and "Inspec" databases of the "Engineering Village". The resulting search string, shown in Figure 1, returned 171 papers.

("software development" OR "development process" OR "software design" OR "software process" OR "software implementation")
AND
("distributed knowledge" OR "collaborative decision" OR "distributed decision" OR "distributed cognition" OR "collaborative problem solving" OR "collaborative knowledge" OR "team knowledge" OR "distributed problem solving" OR "team cognition" OR "team decision" OR "team understanding" OR "team problem solving" OR "collaborative understanding")

Figure 1. Final search string.

B. Selection Process

The selection process adds three more steps to the initial search, which are based on the title, the abstract, and the full text. The selection from the titles is limited to the removal of duplicate papers and conference proceedings introductions. The selection from the abstracts kept papers containing both software and cognition concepts in their abstracts. The selection from the full text removed low quality papers. The quality was evaluated through the identification of context descriptions and data collection methodologies. Papers without these elements were removed. Some theoretical papers were kept, based on the apparent validity of the model presented.

The selection process retained 24 papers. The documents produced by the process are available on request.

C. *Data Synthesis Process*

To perform an accurate synthesis based on our extracted data, we need to manage various types of qualitative data. From the thirteen synthesis approaches described by Cruzes and Dyba [5], we chose the Grounded Theory approach, because it is designed to work with a wide spectrum of qualitative data. We limited ourselves to the three following steps from the Anselm L. Strauss [6] works:

- Associate one keyword to each extracted conclusion,
- Regroup the keywords into concepts,
- Describe how the conclusions complete or contradict each other.

For a more thorough description of the application of Grounded Theory to the software engineering domain, the reader is invited to read the works of O'Connor et al [7, 8] and Lavallée et al [9].

III. RESULTS

This section present the results of the mapping study, where the selected papers are identified by the letter 'S', as described in Appendix A.

A. *Study Methodology*

Table I shows that most of the selected papers describe empirical and academic research, with a single paper whose context is labeled "Open".

Industrial context studies describe real software development projects performed in professional organizations. Academic context observes the work of students performing a formative task. The open context refers to a study performed on an open source development project. This open source community can include both professionals and academics. Note that some papers are purely theoretical and therefore do not present any study context.

Table II presents the many approaches used for data collection in the various selected papers. Some papers used multiple data collection approaches, therefore the total does not add up to 24.

TABLE I. RESEARCH CONTEXT OF THE SELECTED PAPERS

Context	#	Papers
Industrial	17	[S1], [S2], [S3], [S4], [S5], [S9], [S10], [S13], [S16], [S18], [S20], [S21], [S23], [S24].
Academic	10	[S8], [S11], [S12], [S15], [S17], [S19].
Open	1	[S22]

The survey questionnaire is mainly used to confirm or refute the conclusions obtained with other types of data, although some papers base their conclusions on the survey questionnaire alone. The artifact evaluation consists in the analysis of documentation issues of the software development process. This evaluation is often used to evaluate the quality of the work done, and thus the performance of the team. The semi-structured interview describes face-to-face meetings, which is often used to obtain feedback from the software developers. The non-participatory observation occurs when researchers observe the work done by software developers without interfering directly with them, a technique called "shadowing". The audio-video approach consists in the recording of work sessions performed by the software development team. These recordings can include conversations between team partners, computer screen capture videos, keystroke logging records, etc. Usage data consists in statistical measurements obtained from the use of specific software tools. These measurements show the usage frequency of the different functionalities available. These data help researchers in understanding how the software developers adapt software tools to their tasks. Participatory observation occurs when the researcher actively participates in the observed task. This approach enables a more accurate recording of the internal cognitive processes required to perform the task, at the cost of a significant bias.

TABLE II. DATA COLLECTION METHODOLOGY OF THE SELECTED PAPERS

Approach	#	Papers
Survey questionnaire	11	[S8], [S11], [S12], [S17], [S18], [S22], [S24].
Artefact evaluation	10	[S1], [S2], [S3], [S5], [S8], [S11], [S12], [S13], [S15].
Semi-structured interview	7	[S3], [S5], [S16], [S19], [S20], [S24].
Non-participatory observation	6	[S3], [S5], [S9], [S13], [S24].
Audio-video	3	[S1], [S2], [S9].
Usage data	3	[S21], [S23].
Participatory observation	2	[S4].

IV. DISCUSSION

This section presents the conclusions of the selected papers, based on the concepts found through the Grounded Theory approach.

The results of the synthesis can be related to Vygotsky's triangular model of mediated interaction [10], which stated that the activities performed by the software developers within their teams are always mediated by their environment. As Engestrom [11] elaborates, this mediation can take the first four forms presented in Table III.

The new "Emotion" form was motivated by the multiple studies evaluating the emotional state of the team. The importance of emotions during problem-solving has been deemed critical by recent research. Damasio insist on the fact that *"the presumed opposition between emotion and reason is no longer accepted without question"* [12]. Emotion must be considered alongside cognition.

A. *Community: The Software Development Team*

Distributed cognition in software engineering is closely related to how the team assists the individual developer. To be an effective mediator, the team must have coherent situational awareness. Situational awareness is defined as the knowledge a person has of himself/herself and his/her surroundings [13]. In software engineering research, this awareness is oriented along two axes: team meta-cognition and task awareness.

TABLE III. FORMS OF MEDIATIONS

Form	Description	In software engineering
Community	Team interaction	Development team
Instruments	Tools and artefacts used	Individual tools, groupware.
Division of Labor	Tasks performed	Team topology and team structure.
Rules	Impact of disciplined	Process, project, organization
Emotion	Emergent state	Motivation

Team meta-cognition is defined as what each team member knows on the knowledge of their teammates. Good team awareness implies that team members know who the experts are and who are reliable sources of information. It is related to the "who knows what". Task awareness is related to the team's shared mental model of the work to do. The more this mental model is coherent between team members, the team's environment and the relevant stakeholders, the better is task awareness.

1) *Team Meta-Cognition*

The importance of team meta-cognition has been outlined by the works of Kraut and Streeter [14], cited by [S7]:

"Experimentation has shown that developers valued other people as their most used source of help when developing software."

This observation has also been reported by Glor and Hutchins [S1]. When one team member is stuck on a problem, he can present it to one of his partner in order to start a discussion on the most appropriate solution.

The best source of information for software developers is their teammates. It is therefore important for the developer to know who hold this information within the team. This becomes problematic when the team meta-cognition is weak: Sub-optimal choices can be made because the decision-makers are not aware that better solutions exist. Similarly, team performance can be affected when the identified sources of information are not appropriate. Walz et al. [S2] report a case where the two developers with the most influence on decision-making where the ones with the less experience. The team had a poor perception of its own knowledge because the appropriate experts were not identified, resulting in a poor choice of solutions.

To resolve meta-cognition problems, many studies present specific methods [S8, S10, S11, S14, S16, S17]. For example, Kettunen [S10] recommends the identification of "knowledge dependencies" within the team. He presents the importance of information change propagation within the team: A developer must be aware of the people around him capable of providing information changes relevant to his work.

Ye's paper [S14] recommend the identification of expert related to the number and size of modifications made in a code module. Such a tool could enable a developer to contact directly the person most susceptible to know how this code works.

Hause et al. [S8] demonstrate the importance of efficient communications. Their research shows that high performance teams communicate *less* than lower performing ones, because their exchanges are better targeted and better structured. A better knowledge of who are the experts within the team could, for example, limit the communication exchanges to the person most susceptible to provide a relevant answer. Their conclusion [S8] is confirmed by Espinosa et al. [S16]. The later shows that a software development team distributed on distant sites possesses a better knowledge of its own experts. The difficulty of exchanging information over distant sites forces developers to have a better knowledge of the reliable information sources. Sarker et al. [S11] paper shows that sources providing large amounts of accurate information have the greatest impact on knowledge transfer.

Finally, He et al. [S17] show that team meta-cognition is essentially a matter of time. Their paper presents a significant correlation between the self-evaluation of the performance of the teams and the quality of the software product as the project progresses. The impact of familiarity between team partners, initially very strong, diminishes as the team members learn to know themselves better. The team has therefore a better vision of the strengths and weaknesses of their partners, and thus obtains a better self-evaluation of their performance.

2) *Task Awareness*

Better task awareness is mostly useful for the planning and coordination of the work. As Espinosa et al. [S16] explain, a shared knowledge of the task helps team coordination. For example, the use of a public media like the wall board of Sharp et al. [S13] improves team coordination by publicizing immediately any change in the state of the cognitive system. This immediate propagation of changes enables better team situational awareness. This immediate propagation also ensures that the mental model of the task remains synchronized throughout the team, as shows Kettunen [S10]. The presence of a synchronized mental model also diminishes the need to communicate, and thus improve the performance of the team Hause et al. [S8].

De-Franco Tomarello [S12] also shows that if an initial model of the task is imposed upon the team, it improves the problem comprehension. An initial model enables the team to start with a shared mental model better structured and a better organized.

The works of Flor and Hutchins [S1] and Spinuzzi [S5] outline the adaptation of the information received to the context of the task. They show that developers reuse and adapt the information obtained according to their immediate needs. Spinuzzi adds that the artifacts given to the developers are not used in the manner planned, but they are rather adapted to the nature of the task. Information must therefore be designed to be compatible to the needs of the task. Spinuzzi notes that important information resources are ignored because their usability in the context of the task is weak. Developers have therefore diminished task awareness because they do not have all the relevant information in hand. Spinuzzi's concerns are confirmed by Conradi and Dingsoyr [S4], who warn that inadequate data repositories become data cemeteries.

B. *Instruments: Cognitive Support Tools*

The mediating instruments are the various artifacts and tools used by the developers. Among the many tools available, some have an explicit objective to support individual and group cognitive tasks. Cognitive support tool research in software engineering is oriented along two axes: Tool supporting individual cognition, and tool supporting team cognition ("groupware").

1) *Individual Cognition Support Tools*

The papers on individual cognition present two tools common in software development environments: code completion [S14] and compilers' error list [S9].

Code completion, presented by Ye as a cognitive support tool [S14], is a feature of most modern integrated development environments (IDE). This tool recall to the developers all the words understood by the compiler. This enables them to speed up their works by giving them context-aware information. Given the large size and complexity of software components this tool is an essential asset of the software developer.

Walenstein [S9] shows that compilers' error list assists developers in their debug planning by providing a list of the problems found with links to the relevant code snippet. This tool facilitates the developer's work, who only needs to identify the reason for the problem, and not where the problem is located.

2) *Team Cognition Support Tools*

Groupware tools contains the management of public communication channels like wall boards, wiki software, shared calendars, web forums and audio-videoconference tools [S13]. The main characteristic of these tools is that they are transparent as to the origin and destination of the information transmitted. The drawback of this is that users of the system have access to data which do not concern them. In one specific case, a discussion forum had to be moved from a public to a private space, because it created exaggerated expectations from some of its users [S21]. However, private communication channels are also essential for efficient information exchange. Software developers can exchange intermediate steps of a work-in-progress to a team partner without concern for public judgment [S24].

Many papers on team cognition support present the required functionalities for collaborative tools ("groupware"). For example, De Franco-Tomarello et al [S7] list the following key functionalities to ensure that groupware offer a support for collective decision-making, team situation awareness, and sharing mental models :

- Ability to support the team communication channel,
- Ability to support the team collective tools, like planning tools, design tools and knowledge bases,
- Ability to support a collaborative approach to modeling.

Walz et al. [S2] add the necessity to document the rationale behind the choices made by the team. They blame current groupware solutions which report information without reporting how the information was obtained. They argue that the decisions made must be documented with more details.

Whittaker and Schwarz [S3] compare the advantages of a planning tool like Microsoft Project to a *kanban*-style wall board of tasks. They show that the wall board is beneficial because of its public aspect and its flexibility. They argue that current groupware are too restrictive and do not enable different planning approaches. They show however that the wall board is difficult to transmit to stakeholders on distant sites, and that it is difficult to make major changes. It is also not possible to follow the version changes on the wall board, contrarily to a software tool. Finally, it is not possible to present different views of the data when using the wall board.

Research on groupware took a different turn with the emergence of the "Web 2.0". A software team can now cook up a collaborative framework of tools from a plethora of tools available on the Cloud. For example, a team can use a knowledge base managed with Drupal (www.drupal.org), track its development issues with Bugzilla (www.bugzilla.org), plan their tasks with Trac (www.trac.edgewall.org), and keep contact with each others with Pidgin (www.pidgin.im).

C. *Division of Labor: The Structure of the Team*

The division of labor mediator describes the actual tasks performed by the different members of the team. It also considers how the team is spatially disposed, as the physical workspace can have an important impact on the interactions taking place.

For example, it is important to plan the disposition of team members and their communication channels when the team is distributed. The theoretical model of Kubasa and Heiss [S6] proposes and optimization of information flows based on geographical distances, hierarchies, cultural difference and personal familiarity (friendship, rivalry). This model also enables the calculation of a communication cost and of the probability of a delivered message without error.

Meneely and Williams [S23] focused instead on the modeling of a real case; a software development forum. Through a statistical analysis of its usage data, they identified the people performing the roles of "solution providers" and "solution approvers". They noted that approvers, those who choose a solution and implement it in code, are central in the communication network of the forum. Their statistical approach enables an evaluation of the state of an open-source community.

Bass et al. [S20] recommends that the physical disposition of team members across distant sites must consider team meta-cognition. The identification of domain experts during team construction ensures that every developer knows the reliable sources of information (see section IV.A.1). They also say that it is important that each distant site has one person acting as developer in the team.

This ensures a proper dissemination of information across the multiple sites despite the distance.

D. Rules: Project, Process and Organizational Constraints

Lavallee et al. [15] work on the impacts of processes on individual developers concluded that the impacts are not often considered despite having serious detrimental effects. The conclusions are similar at the team level: The impacts of mediating rules on the software development team are rarely observed. Stubblefield and Carson [S18] outline this concern by urging managers not to impose strict rules on the use of a groupware tool: Usage must be adapted to the cognitive needs of the task, and not the other way around.

Hause et al. [S8] note that decision-making mechanisms can be different from one team to another. As Falessi et al. [S15] note, having different process at the individual level is not problematic, but it can become critical at the team level. They show that to impose a decision-making approach with explicit alternative research improve the decision quality from 11% to 67%.

However, decision-making cannot be delayed indefinitely. As Walz et al. [S2] show, software development projects are split into two phases: A decision-making phase and an execution phase. The acquisition of new knowledge must occur at the beginning of the project; if this information arrives only after the midpoint of the project, it is typically ignored. Walz et al. report that adding experts after this knowledge acquisition phase has no impact on the decisions made beforehand. The development team has already made its decisions and does not want to roll back.

However, this capacity to roll back decisions is one characteristics of good working teams, as observed by Hause et al. [S8]. Good teams, having a better shared mental model of the work to do and a better knowledge of the experts in their midst, make less decisions than other teams, but are more ready to roll back and change previous decisions. Good teams changed 20% of their decisions, against only 9% for the bad teams.

E. Emotion: Team Emotional States and Motivation

One of the aspects uncovered by research in cognitive psychology is the fact that individual performance changes when the person's emotional state changes. We can observe the same fact at the team level, and thus the emotional state of the team can also affect its performance. Marks and Mathieu describes these emotional states as "emergent states":

> *"Emergent states describe cognitive, motivational, and affective states of teams, as opposed to the nature of their member interaction."* [16]

For example, team topology can stay stable for the duration of the project, but the emotional state can change within a single day, or even a single meeting. Whittaker and Schwarz [S3] studies the impact of a material wall board of tasks on the team sense of belonging. This wall board requires developers to cut pieces of paper detailing their estimations and to stick it to the wall. The manual aspect and the public nature of the board improve the perceived responsibility of the developers toward their task estimations. There is a certain shame in having to correct the content of the wall board; therefore the estimations are more carefully made. By contrast, the software tool is very often "write-only": Developers enter data in the tool, but they never read it. The quality of the estimations in the software tool is much weaker.

Parsarnphanich and Wagner [S22] study the motivation of important contributors to the Wikipedia knowledge base. They show that the greatest motivator to contribution is the quick feedback they receive from their contributions, even when the change is minor. This quick feedback outlines the public aspect of their contribution and incites them to continue. These two papers show that pride can affect performance [S3] and productivity [S22]. They also show that pride can be controlled by the public aspect of the task.

Trust is also an important emotion for certain team cognition elements, like meta-cognition. An initial face-to-face meeting seems to have an important impact on what the developers perceived of their supervisors. Bass et al. [S20] show that a visit from the manager to all the distant sites can improve communication for the project duration. Richardson et al. [S19] note that if the supervisors did not meet the other team members face-to-face, the developers did not ask them questions.

Motivation is also an essential emotion required for the success of any project. Wikis and other collective memory knowledge bases are dependent on the volunteer work of motivated individuals. A pragmatic altruism or idealism has been identified has a major factor of Wikipedia's success [S22]. Managers must encourage a cooperation culture within their teams. Tools must be able to support the implication of the team partners by promoting a good compatibility with the work to perform [S5].

V. Conclusions

The main conclusion from the synthesis is that there is no consensus on how to manage distributed cognition in software development: Many papers describe what should be done, but very few describe how to do it. The use of varied practices in varied contexts means that comparison between studies is very difficult, since contradictory conclusions abound.

There are also many suggestions for new functionalities for cognitive-support software tools. There are therefore few empirical studies of the software functionalities considered as important. One study notes that the main weakness of the existing groupware tools is that they lack flexibility [S3]: Software developers prefer using many tools more adapted to their task rather than one generic all-purpose tool.

Additionally, there are very few studies on the ergonomics of software tools. The papers describe what collaborative tools ("groupware") must support, but they do not describe how this support can be ensured. There is therefore not enough research on the affordance of collaborative tools available for software development teams.

There are also no papers on the mental workload of software developers. We do not know the impact of collaborative tools on the mental workload of software developers. There is no information on the global mental workload of the software team, nor whether cognitive effort is appropriately spread across team partners.

ACKNOWLEDGMENT

We thank Professor Jean-Marc Robert for the initial idea of a literature review on the subject of distributed cognition.

REFERENCES

[1] E. Hutchins, Cognition in the wild. Cambridge, MA: MIT Press, 1995.

[2] B. Kitchenham, “Procedures for Performing Systematic Reviews,” 2004.

[3] J. Biolchini, P. Gomes Mian, A. Candida Cruz Natali, and G. Horta Travassos, “Systematic Review in Software Engineering,” Rio de Janeiro, 2005.

[4] D. Budgen, M. Turner, P. Brereton, and B. Kitchenham, “Using Mapping Studies in Software Engineering,” in Proceedings of Psychology of Programming Interest Group, 2008, vol. 2, pp. 195–204.

[5] D. S. Cruzes and T. Dybå, “Research synthesis in software engineering: A tertiary study,” Information and Software Technology, vol. 53, no. 5, pp. 440–455, May 2011.

[6] A. L. Strauss, Qualitative Analysis for Social Scientists. Cambridge, UK: Cambridge University Press, 1987, p. 319.

[7] G. Coleman and R. O’Connor, “Software Process in Practice: A Grounded Theory of the Irish Software Industry,” in Software Process Improvement, vol. 4257, I. Springer Berlin / Heidelberg, 2006, pp. 28–39.

[8] S. Basri and R. V. O’Connor, “Understanding the Perception of Very Small Software Companies towards the Adoption of Process Standards,” in Systems, Software and Services Process Improvement, vol. 99, Springer Berlin Heidelberg, 2010, pp. 153–164.

[9] M. Lavallée and P. N. Robillard, “The Impacts of Software Process Improvement on Developers: A Systematic Review,” in 34th Intl Conf on Software Eng (ICSE 2012), 2012.

[10] L. S. Vygotsky, Mind in Society: The Development of Higher Psychological Processes. Cambridge University Press, 1978.

[11] Y. Engeström, “Activity theory as a framework for analyzing and redesigning work.,” Ergonomics, vol. 43, no. 7, pp. 960–74, Jul. 2000.

[12] A. Damasio, The Feeling of What Happens: Body and Emotion in the Making of Consciousness. Houghton Mifflin Harcourt, 2000, p. 400.

[13] M. R. Endsley, “Toward a theory of situation awareness in dynamic systems,” Human Factors, 37 (1), pp. 32–64, 1995.

[14] R. Kraut and L. Streeter, “Coordination in Software Development,” Communications of the ACM, 38, no. 3, 1995.

[15] M. Lavallée and P. N. Robillard, “The Impacts of Software Process Improvement on Developers: A Systematic Review,” in International Conference on Software Engineering, 2012.

[16] M. A. Marks, J. E. Mathieu, and S. J. Zaccaro, “A temporally based framework and taxonomy of team processes,” Academy of Management Review, vol. 26, no. 3, pp. 356–376, 2001.

APPENDIX A. THE SELECTED PAPERS

#	Reference
S1	N.V. Flor and E.L. Hutchins, (1991). *Analyzing distributed cognition in software teams: a case study of team programming during perfective software maintenance*. Empir. Stud. of Program.: 4th Workshop.
S2	D. Walz et al., (1993). *Inside a software design team: Knowledge acquisition, sharing, and integration*. Comm. of the ACM.
S3	S. Whittaker and H. Schwarz, (1999). *Meetings of the board: the impact of scheduling medium on long term group coordination in software development*. Computer Supported Cooperative Work.
S4	R. Conradi and T. Dingsoyr, (2000). *Software experience bases: a consolidated evaluation and status report*. PROFES 2000.
S5	C. Spinuzzi, (2001). *Software development as mediated activity: Applying three analytical frameworks for studying compound mediation*. ACM SIGDOC Intl Conf. on Comp. D.
S6	G. Kubasa and M. Heiss, (2002). *Distributed face-to-face communication in bottom-up driven technology management - A model for optimizing communication topologies*. IEEE Intl Eng. Mngmnt Conf.
S7	J. DeFranco-Tommarello and F.P. Deek, (2002). *Collaborative software development: a discussion of problem solving models and groupware technologies*. 35th Hawaii Intl Conf. Syst & Science.
S8	M. Hause et al., (2003). *Performance in international computer science collaboration between distributed student teams*. 33rd Annual Frontiers in Education.
S9	A. Walenstein, (2003). *Observing and measuring cognitive support: steps toward systematic tool evaluation and engineering*. 11th IEEE Intl Workshop on Program Compreh.
S10	P. Kettunen, (2003). *Managing embedded software project team knowledge*. IEE Proceedings: Software.
S11	Sa. Sarker et al., (2003). *Knowledge transfer in virtual information systems development teams: an empirical examination of key enablers*. 36th Hawaii Intl Conf. on Systems Sciences.
S12	J. DeFranco-Tommarello, (2003). *A study of collaborative software development using groupware tools*. Proceedings. ITRE.
S13	H. Sharp et al., (2006). *The role of story cards and the wall in XP teams: a distributed cognition perspective*. AGILE .
S14	Y. Ye, (2006). *Supporting software development as knowledge-intensive and collaborative activity*. ICSE 2006.
S15	D. Falessi et al., (2006). *Documenting design decision rationale to improve individual and team design decision making: An experimental evaluation*. ISESE’06.
S16	J.A. Espinosa et al., (2007). *Team knowledge and coordination in geographically distributed software development*. J. Mngmnt Inf. Syst.
S17	J. He et al., (2007). *Team cognition: Development and evolution in software project teams*. J. Mngmnt Inf. Syst.
S18	W.A. Stubblefield and T.L. Carson, (2007). *Software design and engineering as a social process*. Conf. on Human Factors in Computing Systems.
S19	I. Richardson, S. Moore, D. Paulish, V. Casey and D. Zage, (2007). *Globalizing software development in the local classroom*. Software Engineering Education Conf.
S20	M. Bass et al., (2007). *Collaboration in global software projects at siemens: An experience report*. ICGSE.
S21	P.-H. Cheng et al., (2008). *collaborative knowledge management process for implementing healthcare enterprise information systems*. IEICE Trans. on Inf. and Syst.
S22	P. Prasarnphanich and C. Wagner, (2009). *The role of wiki technology and altruism in collaborative knowledge creation*. J. Comput. Inf. Syst.
S23	A. Meneely and L. Williams, (2011). *On the Use of Issue Tracking Annotations for Improving Developer Activity Metrics*. Adv. Softw. Eng.
S24	S. Patil et al., (2011). *Methodological reflections on a field study of a globally distributed software project*. Inf. Soft.Tech.

Approaching Regular Polysemy in WordNet

Abed Alhakim Freihat, Fausto Giunchiglia
Dept. of Information Engineering and Computer Science
University of Trento,
Trento, Italy
e-mail: {fraihat,fausto}@disi.unitn.it

Biswanath Dutta
Documentation Research and Training Centre
Indian Statistical Institute (ISI)
Bangalore, India
e-mail: bisu@drtc.isibang.ac.in

Abstract— **WordNet has been used widely in natural language processing and semantic applications. Despite the reputation of WordNet, the polysemy problem that leads to insufficient quality of applications results is still unsolved. Many approaches have been suggested. However, none of them give a comprehensive solution to the problem. In this paper, we introduce a pattern based approach that solves the polysemy problem in the case of nouns. To achieve this result we introduce a set of novel relations that represent polysemy types and a set of new operations that allow us to organize the specialization polysemy cases.**

Keywords— Lexical databases; WordNet; Homonomy; Polysemy; regular Polysemy; Polysemy Reduction; Lexical semantics; Semantic Search; Knowledge Engineering

I. INTRODUCTION

Polysemy in WordNet [1][9] is considered to be the main reason that makes it hard to use for natural language processing (NLP) and semantic applications [12][17]. Differentiating between the types of polysemy should be possible through explicit semantic relations between the senses of polysemous terms. Unfortunately, relations between polysemous terms are not provided in WordNet [2]. For instance, WordNet does not provide the distinction between homographs, and complementary terms [6].

In the last, decades many approaches have been introduced to solve the polysemy problem through merging the similar meanings of polysemous terms [4]. These approaches are sometimes helpful in cases, where terms have meanings that are similar enough to be merged. However, polysemous terms with similar meanings are a sub-case of the solution of specialization polysemy [14]. They represent only a small portion of the polysemy problem. In fact, a significant portion of the polysemous senses should not be merged, as they are just similar in meaning [5] and not redundant. In another approach, CORELEX [6] has been introduced as an ontology of systematic polysemous nouns extracted from WordNet. However, CORELEX deals only with the upper level ontology of WordNet that corresponds mainly to the metonymy cases and does not provide a solution for other polysemy types [16].

In this paper, we introduce a pattern based approach that combines several ideas to solve the polysemy problem. Our approach follows the idea that the polysemy problem is a problem of semantic organization [3]. Thus, the goal of our approach is to reorganize the semantic structure of the polysemous terms in wordNet, where we transform the implicit relations between the polysemous terms at lexical level to explicit relations at the semantic level. This includes extending WordNet by adding new hierarchical and associative relations between the synsets to explicitly denote the polysemy type occurring between the meanings of each polysemous term, as suggested in [2]. To achieve this goal, our approach deals with all polysemy types at all ontological levels of WordNet. It deals with the lower level ontology of WordNet and it extends the merge operation suggested by the polysemy reduction approaches [4][15] by providing new operations that organize the relations between the meanings of polysemous terms. Our approach also deals with polysemy in the middle level, as it is the case in regular polysemy approaches [14] and also in the upper level ontology as in systematic polysemy approaches [6].

This paper is organized as follows: In Section II, we describe the polysemy problem in WordNet. In Section III, we describe the current approaches for solving the polysemy problem in WordNet. In Section IV, we present the semantic relations that denote polysemy types and the operations that reorganize the structure of polysemous terms in WordNet. In Section V, we introduce a pattern based approach for solving the polysemy problem in the case of polysemous nouns. In Section VI, we discuss the results and evaluation of our approach. In Section VII, we conclude the paper and describe our future research work.

II. POLYSEMY IN WORDNET

WordNet is a lexical database that organizes synonyms of English words into sets called synsets, where each synset is described through a gloss. For example, the words *happiness* and *felicity* are considered to be synonyms and grouped into one synset {*happiness, felicity*} that is described through the gloss*: state of well-being characterized by emotions ranging from contentment to intense joy*.

WordNet organizes the relations between synsets through semantic relations, where each word category has a number of relations that are used to organize the relations between the synsets of that grammatical category. For example, the hyponymy relation (X is a type of Y) is used to organize the ontological structure of nouns. WordNet 2.1 contains 147,257 words, 117,597 synsets and 207,019 word-sense pairs. Among these words there are 27,006 polysemous words, where 15776 of them are nouns.

From linguistics, a term is polysemous if it has more than one meaning [17]. Linguists differentiate between contrastive polysemy, i.e. terms with completely different and unrelated meanings - also called homonyms or homographs; and complementary polysemy, i.e. terms with different but related meanings. Complementary polysemy is classified in three sub types: Metonymy, specialization polysemy and metaphors. Following the above, we can classify the various forms of polysemy as follows:

1) Complementary polysemy: terms that have the same spelling and related meanings. Complementary polysemy can be:
   a. Metonymy: substituting the name of an attribute or feature for the name of the thing itself, such as in the following example:
      Peter caught *a chicken* in his garden.
      Peter prepared *chicken* for the dinner.
   b. Specialization polysemy: a term is used to refer to a more general meaning and another more specific meaning, such as in the following example the term *methodology*:
      1**. methodology,** methodological analysis: the branch of philosophy.
      2. **methodology**: the system of methods followed in a particular discipline.
   c. Metaphors: terms that have the same spelling and have literal and figurative meanings. Consider, for instance, the term *parasite*:
      1. **parasite**: an animal or plant that lives in or on a host (another animal or plant).
      2. leech**, parasite**, sponge, sponger: a follower who hangs around a host (without benefit to the host) in hope of gain or advantage.
2) Homographs: terms that have the same spelling and different unrelated meanings, such as the term *bank*:
   Peter sat on the *bank* of the river.
   Peter deposited money in the *bank*.

In WordNet, the number of senses a polysemous term may range from 2 senses to more than 30 senses. In some rare cases may more; for instance, the noun *head* has 33 senses.

Nevertheless, 90% of the polysemous nouns have less than 5 senses. Table I shows the distribution of these polysemous nouns according to the number of senses they have. Notice that, in this paper, we are concerned with polysemous nouns only and not the verbs, adverbs and adjectives.

The fact that a term has more than two senses implies that the meanings of the term belong to more than one type of polysemy. For example, the term *food* has 3 senses as mentioned below, where the polysemy type between the first and the second meanings is specialization polysemy, while the third meaning is metaphoric.

TABLE I. POLYSEMOUS NOUNS IN WORDNET

# of synsets	# of nouns (in percentage)
2	10186 (≈ 64%)
3	2968 (≈ 19%)
4	1186 (≈ 7%)

1. **food**, nutrient: any substance that can be metabolized by an organism to give energy and build tissue.
2. **food**, solid food: any solid substance that is used as a source of nourishment.
3. **food**, food for thought: anything that provides mental stimulus for thinking.

III. APPROACHES FOR SOLVING POLYSEMY IN WORDNET

The approaches of polysemy can be classified in two main approaches. The first is polysemy reduction, where the focus is on complementary polysemy to produce more coarse-grained lexical resources of existing fine-grained ones such as WordNet. The second type of polysemy approaches focuses on classifying polysemy into systematic or regular polysemy and homographs. Based on this classification, CORELEX was introduced as ontology of systematic polysemous nouns extracted from WordNet. Other approaches, such as in [13][14], were introduced to extract semantic relations between regular polysemous terms in WordNet.

In the following, we summarize Polysemy reduction approaches and CORELEX, the most famous systematic polysemy approaches. Notice that neither polysemy reduction approaches nor CORELEX could solve the polysemy problem in WordNet. In general, Polysemy reduction approaches could not solve the problem of the upper level ontology where CORELEX did not provide a solution for polysemy in the middle and lower level ontology of WordNet.

A. Polysemy Reduction Approaches

In polysemy reduction, the senses are clustered such that each group contains related polysemous words [18][15]. These groups are called homograph clusters. Once the clusters have been identified, the senses in each cluster are merged. To achieve this task, several strategies have been introduced. These strategies can be mainly categorized in semantic-based and statistics-based strategies [17]. Some approaches combine both strategies [15]. Although results of applications of these approaches are reported, these results are taken usually from applying them on sample data sets and there is no way to verify these results independently. Polysemy reduction approaches typically rely on the application of some detection rules such as: *If S1* *and S2 are two synsets containing at least two words, and if S1 and S2 contain the same words, then S1 and S2 can be collapsed together into one single synset* [10]. However, applying this rule may wrongly result in merging two different senses as in the following example:

1. **smoke, smoking**: a hot vapor containing fine particles of carbon being produced by combustion.
2. **smoke, smoking**: the act of smoking tobacco or other substances.

In general, polysemy reduction can neither predict the polysemy type occurring between the senses of polysemous words nor can deal with metonymy or metaphors. Polysemy reduction does not solve the polysemy problem in linguistic resource. Nevertheless, it can be potentially used to solve part of the problem, namely the identification and merging of genuine redundant synsets.

B. Regular Polysemy Approaches

J. Apresjan defined regular polysemy as follows: "*A polysemous Term T is considered to be regular if there exists at least another polysemous T' that is semantically distinguished in the same way as T*" [8]. Systematic polysemy approaches rely on this definition. CORELEX, the first systematic polysemy lexical database, follows the generative lexicon theory [3] that distinguishes between systematic (also known as regular or logic) polysemy and homographs. Systematic polysemous words are systematic and predictable while homonyms are not regular and not predictable. The type of polysemy of the word *fish* for example is systematic since the meaning *food* can be predicted from the *animal* meaning and so the word *fish* belongs to the systematic class *animal food*. The two meanings of fish describe two related aspects of *fish*: fish is an animal and fish is a food. That a word is systematic polysemous means that the meanings of this word are not homonyms and they describe different aspects of the same term. Following this distinction, CORELEX organizes the polysemous nouns of WordNet 1.5 into 126 systematic polysemy classes. The systematic polysemy classes in CORELEX have been determined in a top down fashion considering the patterns in the upper level ontology of wordNet only. It does not consider the metaphoric cases. Also, there is no cleaning process carried out on WordNet by CORELEX construction. Another important point is related to the fine grained nature of WordNet where the meanings of some CORELEX classes are very difficult to disambiguate and indistinguishable even for humans [11].

IV. Denoting polysemy types and organizing polysemy in WordNet

Making WordNet a more coarse grained lexical resource does not solve the polysemy problem, although there are some fine grained polysemous cases in WordNet. We believe that the polysemy problem in WordNet is primarily a problem of organizing the senses of polysemous terms. In the cases of homographs, metonymy, and metaphors, we need semantic relations that denote the polysemy type of corresponding cases. The cases of specialization polysemy on the other hand require reorganizing the semantic structure to reflect the (implicit) hierarchical relation between such senses. In the following, we introduce the relations to denote homographs, metonymy, and metaphors and then we present the operations for solving specialization polysemy cases.

A. Polysemy Type Relations

In the following, we explain the suggested relations to denote the polysemy types:

Homographs: There is no relation between the senses of a homograph term. Nevertheless, differentiating homographs from other polysemy types is very important improvement in wordNet. We use the relation *is_homograph* to denote that two synsets of a polysemous term are homographs. For example, this relation holds between the synsets *{saki as alcoholic drink}* and *{saki as a monkey}*.

Metonymy: In metonymy cases, there is always a *base meaning* of the term and other *derived meanings* that express different aspects of the base meaning [19]. For example, the term chicken has the base meaning {*a domestic fowl bred for flesh or eggs*} and a derived meaning {*the flesh of a chicken used for food*}. To denote the relation between the senses of a metonymy term, we use the relation *has_aspect*, where this relation holds between the base meaning of a term and the derived meanings of that term. To set up the relation we need to determine the base meaning and then relate the other derived meanings to it.

Metaphors: In metaphoric cases, we use the relation *Is_metaphor* to denote the metaphoric relation between the metaphoric meaning and literal meaning of a metaphoric term. For example this relation is used to denote that {*cool as great coolness and composure under strain*} is metaphoric meaning of the literal meaning {*cool as the quality of being at a refreshingly low temperature*}. In the cases, where this relation is applicable, we need to specify the literal meaning and the metaphoric meaning.

B. Operations for Specialization polysemy

Analysis of specialization polysemy cases shows that such cases can be classified based on the synset synonyms into the following three groups. To explain our idea, we have chosen cases, where the synsets of each term share the same common parent.

Let T be a polysemous term that occurs in two synsets S1 and S2. We consider T in the following three cases:

Case 1: T has synonyms in S1 and has synonyms in S2 as in the case of *kestrel*:

1. *kestrel*, falco sparverius: small American falcon.
2. *kestrel*, Falco tinnunculus: small Old World falcon.

Case 2: T has synonyms in S1 or in S2 but not in both as in the case of *dorsum*:

1. back, *dorsum*: the posterior part of a human (or animal) body from the neck to the end of the spine.

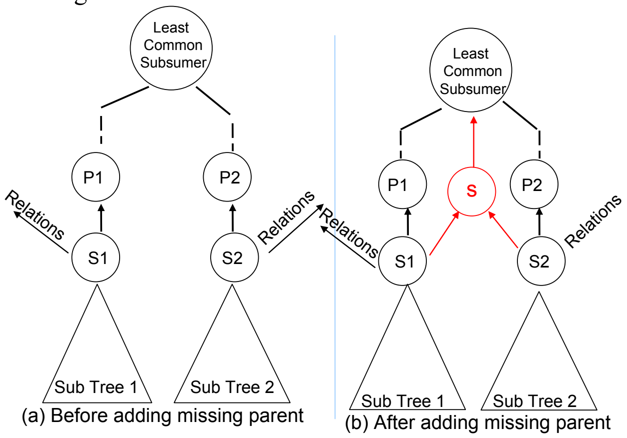
2. *dorsum:* the back of the body of a vertebrate or any analogous surface.

Case 3: T has no synonyms in S1 or S2 as the in the case of *compatible software*:

1. *compatible software*: application software programs that share common conventions.
2. *compatible software:* software that can run on different computers without modification.

In case 1, T has synonyms in S1 which means that T is exchangeable with the other synonyms of S1 and at the same time is also exchangeable with the synonyms of S2. Let T1, T2 be non polysemous synonyms of T in S1 and S2 respectively. T1 is synonymous with T but not with T2. Otherwise T1 and T2 should appear in the same synset. The fact that T1 and T2 appear in two different sibling synsets indicates that they are not the same. We think that the semantic relatedness between S1 and S2 is encoded at lexical level rather than semantic level. We have the same observation in case 2. The fact that one synset contains T only and the other synset contains additional terms indicates that the synset that contains T only is a more general meaning of the synset that have additional terms. We consider the terms in case 3 as candidates to be merged. Accordingly, we suggest the following operations to organize the relations between the senses in specialization polysemy cases:

Solution for Case 1: We add a new (missing) parent in cases, where the polysemous meanings of a term T can be seen more specific meanings of an absent more general meaning. In such cases, we create the new missing more general meaning and connect the more specific meanings to the new created new parent. This operation is schematized in the figure 1.



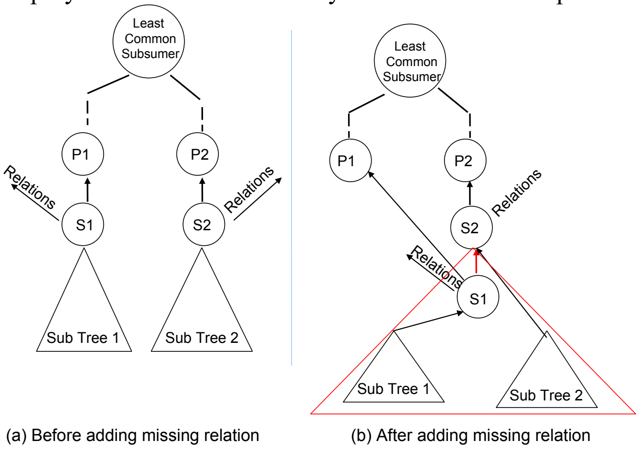
(a) Before adding missing parent (b) After adding missing parent

Figure 1. Adding missing parent

Solution for Case 2: In such cases, we establish a new (missing) ***is_a*** relation to denote that a sense of a polysemous term T is more specific than another more general meaning of T. We schematize this operation as illustrated in figure 2.

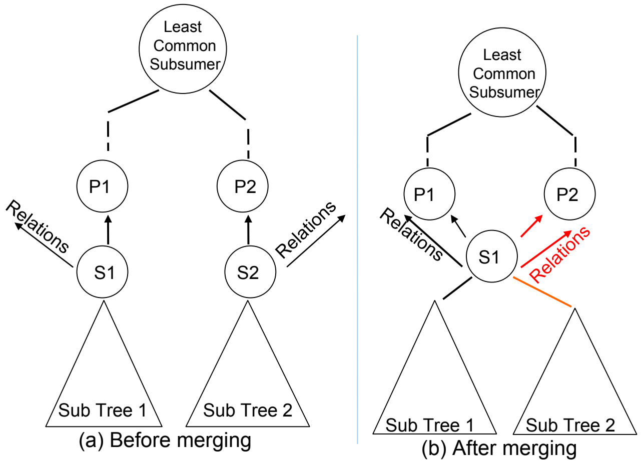
Solution for case 3: In such cases, we merge the meanings. The merge operation is schematized as in figure 3.

At the term level, we disambiguate the polysemous terms as follows: in case (1) We remove the polysemous terms from both child synsets and keep the polysemous words in the new added parent synset only. In case (2) We remove the polysemous term from the synset with the more specific meaning and keep it in the synset with the more generic meaning. The Merge operation in case 3 unifies the terms of both synsets in one synset. Thus, applying the three operations results in reducing the number of polysemous words in WordNet.



(a) Before adding missing relation (b) After adding missing relation

Figure 2. Adding missing relation



(a) Before merging (b) After merging

Figure 3. Merge operation

V. PATTERN BASED APPROACH FOR SOLVING POLYSEMY

In this section, we describe our approach for solving polysemy in WordNet. The approach has the following four phases. The first and the third phases are automatic, while the second and fourth are manual:

A. Patterns Identification

B. Patterns Classification

C. Polysemy type Assignment

D. Validation

A. Patterns Identification

We apply a pattern extraction algorithm that computes the regular patterns for the polysemous terms. The algorithm returns the following lists:

1. a list of regular patterns: contains the regular patterns, where at least two terms belong to each pattern.
2. a list of sub patterns: contains the sub patterns of the patterns identified in the regular patterns list.
3. a list of common parent terms: contains the terms, where the synsets or part of the synsets of these terms share the same hypernym.
4. a list of singleton patterns: This list contains the patterns that have less than two terms and are not sub patterns of any regular pattern.

Notice that it is possible for terms that have more than 2 senses to have more than one pattern. In the following, we illustrate the definitions, we used in our algorithm.

Definition 1: Regular Structural Pattern Let T be a polysemous term that has n meanings, $n > 1$. Let S be the set of the synsets of T. Let R be a subset of S. Let Q an ordered sequence of R, where $|R| = m, 2 \le m \le n$, and $Q = < s_1,..,s_m >, s_i \in R, s_i \neq s_j$, for $i \neq j$. A pattern *ptrn* of T is defined as $p\#< p_1,..,p_m >$, such that each p_i is a direct hyponym of p and subsumes $s_i, 1 \le i \le m$. A pattern is regular if there are at least two terms that belong to it. For example, the pattern *passerine#<oscine, tyrannid>* is regular since there are 3 terms that belong to it.

Definition 2: Sub pattern For a regular pattern *ptrn* = $p\#< p_1,..,p_m >$. A pattern *ptrn'* is a sub pattern of *ptrn* if $ptrn' = p\#< p_1^{'},..,p^{'}_k >$ and $\exists p_i, p_j^{'} (p_i = p_j^{'})$.

Sub patterns are important, since it is possible that the elements of a pattern and its sub patterns have the same polysemy type. For example, the pattern *passerine#<oscine, tyrannid> and its sub pattern passerine#<oscine,wren>* belong both to the specialization polysemy patterns.

Definition 3: Common parent class A term belongs to the common parent class if it has at least two synsets that share the same hypernym. For example, the synsets of the term *kestrel* in the previous section share the same hypernym. In polysemy reduction approaches, senses that have the common parent property are candidates to be merged. In our approach, such terms are candidates for specialization polysemy. Note that there are many terms that have this property, but they are not considered to be regular according to definition 1 since they have different hierarchical structures.

B. Patterns Classification

In this phase, we manually classify the patterns gained in the previous phase, where we assign each pattern the polysemy type, the terms of the pattern belong to. We classify the patterns into the following groups:

1. Specialization polysemy patterns
2. Metaphoric patterns
3. Metonymy patterns
4. Homonymy patterns
5. Singleton and mixed patterns

The singleton and mixed patterns group contains the singleton patterns and the patterns that contain patterns whose terms may belong to more than one polysemy type. For example, there are terms under the pattern *attribute#<quality, trait>* that belong metaphoric polysemy and others that belong to specialization polysemy. In the following, we describe our analysis in this phase according to the pattern position in the ontology of WordNet:

Top level patterns: the patterns at the top level ontology correspond to metonymy and metaphoric terms. It is unlikely to find specialization polysemy terms at the top level patterns. Although homonyms is not regular and, it is also possible to determine some homonymy patterns at the top level ontology. For example, the pattern *organism#<animal, plant>* is considered as homograph pattern, since we exclude the possibility of specialization polysemy, metonymy and metaphoric in the terms that belong to that pattern.

Middle level patterns: the patterns here correspond mainly to specialization polysemy and metaphoric cases. It is possible also to find homograph patterns at this level. To differentiate between specialization polysemy and other polysemy types, we use the following criteria:

- **Specialization polysemy/ metaphors**: specialization polysemy patterns indicate consistency between the pattern parts, while metaphoric patterns indicate meaning transfer from their literal meaning to a (metaphoric) meaning. For example, oscine and tyrannid are consistent since they belong to the type passerine in the pattern *passerine#<oscine, tyrannid>*, while we find meaning transfer from property {*a basic or essential attribute shared by all members of a class*} to trait {*a distinguishing feature of your personal nature*} in the pattern *attribute#<property, trait>* .
- **Specialization polysemy/ homographs**: In contrary to specialization polysemy, homograph patterns indicate inconsistency. For example, person is inconsistent with plant and the metaphoric link is excluded in the pattern *organism#<person ,plant>*.

Lower level patterns: the patterns at the lower level ontology are those patterns that belong to common parent class and they correspond mainly to specialization polysemy. It is possible to find metaphors and/or homographs at the lower level ontology. Such cases are determined and excluded in the validation phase.

C. Polysemy type Assignment

In this phase, the terms are assigned to the polysemy type of the pattern they belong to. The terms that belong singleton and mixed patterns are not assigned and they are subject to manual treatment in the validation phase.

D. Validation

In this phase, we manually validate the assigned polysemy type. This phase includes three tasks:

1. **Validation of the assigned polysemy types**: we check whether each of the nouns belong to its assigned polysemy type.
2. **Assigning the polysemy type**: for the terms that belong to the singleton and mixed patterns.
3. **Excluding of false positives**: we exclude the false positives from the terms of the automatic assigned groups. Our judgments during the validation are based on knowledge organization. Word etymology and linguistic relatedness have secondary role.

In table II, we show the results of our validation for sample patterns. An Example for false positives that we found in the common parent group: the meanings of term *apprehender* are homographs:

knower, *apprehender*: a person who knows or apprehends.
apprehender: a person who seizes or arrests.

TABLE II. SAMPLE PATTERNS VALIDATION

# of instances	Pattern	Assigned polysemy Type	# of False positives
1002	Common Parent	Spec. polysemy	93
75	attribute#property,quality	Metaphoric	7
52	attribute#quality,trait	Metaphoric	22
30	vascular plant#herb,woody plant	Spec. polysemy	1
29	*abstraction#communication,group	Metonymy	11
28	*abstraction#attribute,measure	Metaphoric	10
21	artifact#commodity,covering	Spec. polysemy	10
19	attribute#property,trait	Metaphoric	0
18	animal#invertebrate,larva	Spec. polysemy	0
16	woody plant#shrub,tree	Spec. polysemy	0

VI. RESULTS AND EVALUATION

In Table III, we present the results of our approach after the manual validation.

TABLE III. VALIDATED RESULTS OF THE ALGORITHM

Polysemy type	**# of words**	**# of words in percentage (%)**
Metaphor	559	13.6
Homograph	1011	24.8
Spec. Polysemy	2139	52.5
Systematic and Others	361	7.9

The cases in the column systematic and others are the cases that we think that they should be processed in a subsequent phase of our approach in the framework of approaching CORELEX systematic polysemy or cases, were the presence of the polysemous term in one of the synsets is inappropriate and should be removed from one of the synsets. An example for such cases is the term *senate* that appears in the synset and its direct hypernym:

United States Senate, U.S. Senate, US Senate, *Senate: the upper house of the United States Congress*.
=> *senate: assembly possessing high legislative powers*

In Table IV, we present the classification of specialization polysemy. The total number of reduced polysemous words is 2139 words. The total number of merged synsets represents about 10% of the total processed cases. At the same time we have added 1045 new synsets and 2775 new *is a* relations, while have deleted 409 synsets and 409 *is a* relations. This means that in our approach we have increased knowledge rather than decreasing knowledge to solve the polysemy problem.

TABLE IV. SPECIALIZATION POLYSEMY RESULTS

	# of words	# of words in percentage (%)
Missing parent	1045	49
Missing relation	685	32
Merge	409	19.1

To evaluate our approach, 1020 cases have been evaluated by two evaluators. In the following Table V, we report the statistics of the evaluation, where the column polysemy type refers to homonymy, metaphoric, metonymy, or specialization polysemy and polysemy operation refers to creating missing parent, adding missing relation, or merging operation. Note that, polysemy operation is applicable in case of specialization polysemy. The table presents the agreement between the evaluators and our approach. The third row represents the number of cases, where at least one evaluator agrees with our approach.

TABLE V. EVALUATION RESULTS

	Polysemy type agreement	Polysemy operation agreement
Evaluator 1	979 ≈ 96%	924 ≈ 90.5%
Evaluator 2	945 ≈ 92.5%	855 ≈ 84%
Partial agreement	1006 ≈ 98.5%	978 ≈ 96%

As we can see from the results above, although the agreement with the approach is high, in many cases, the evaluators agree on the specialization polysemy type but disagree on the operation type. The explanation for this is that the operation is decided according to the nature of lemmas in both synsets as explained in section IV.

VII. CONCLUSION AND FUTURE WORK

In the present paper, we introduced a pattern based approach for solving the polysemy problem in WordNet. Our approach deals and covers all polysemy cases at all ontological levels of wordNet. Furthermore, it improves the ontological structure of WordNet by transforming the implicit relations between the polysemous senses at lexical level into explicit semantic relations. The manual treatment in two phases of the approach guarantees the quality of the approach result. We have tested our approach on polysemous nouns that have two senses and the results were promising.

Our next step is to apply the approach on all polysemous nouns in WordNet. In a subsequent phase, we are going to extend our algorithm to handle verbs, adjectives and adverbs.

The main contributions of this work are at two levels:
At the conceptual level, we are providing a new foundation towards the problem of polysemy. At the implementation level, we aim to improve the quality of NLP and knowledge-based applications, especially in the field of the semantic search.

REFERENCES

[1] G. A. Miller, “WordNet: a lexical database for English,” Communications of the ACM 38 (11), November 1995, pp. 39 – 41.

[2] J. Gonzalo, “Sense Proximity versus Sense Relations,” Proc. of the Second Global WordNet Conference, Brno, Czech Republic, January 20-23, 2004, pp. 5-6.

[3] J. Pustejovsky, The Generative Lexicon, Cambridge: MIT Press, 1995.

[4] R. Mihalcea and D. I. Moldovan, “EZ.WordNet: Principles for Automatic Generation of a Coarse Grained WordNet,” FLAIRS Conference, 2001, pp. 454-458.

[5] B. Nerlich and D. D. Clarke, “Polysemy and flexibility: introduction and overview,” B. Nerlich, Z. Todd, V. Herman and D. D. Clarke (Hg.), Polysemy: Flexible Patterns of meaning in Mind and Language, Berlin, New York: Mouton de Gruyter, 2003, pp. 3-29.

[6] P. P. Buitelaar, “CORELEX: Systematic Polysemy and Underspecification,” PhD thesis, Brandeis University, Department of Computer Science, (1998)

[7] R. Snow, S. Prakash, D. Jurafsky, and A. Ng, “Learning to merge word senses,” Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, 2007.

[8] J. Apresjan, “Regular Polysemy,” Linguistics, vol. 142, 1974, pp. 5-32.

[9] G. A. Miller, R. Beckwith, Ch. Fellbaum, D. Gross and K. Miller, “Introduction to wordnet: an on-line lexical database,” International Journal of Lexicography, 1990.

[10] N. Verdezoto, and L. Vieu, “Towards semi-automatic methods for improving WordNet”, Proc. of the 9th International Conference on Computational Semantics, Oxford, UK, 2011.

[11] N. Tomuro, “Systematic Polysemy and Inter-Annotator Disagreement: emirecal Examinations,” Proc. of the First International Workshop on Generative Approaches to Lexicon, 2001.

[12] F. Giunchiglia, U. Kharkevich and I. Zaihrayeu, “Concept Search,” ESWC, 2009, pp. 429-444.

[13] W. Peters, “Detection and Characterization of Figurative Language Use in WordNet,” PhD thesis, Natural Language Processing Group, Department of Computer Science, University of Sheffield, 2004.

[14] L. Barque and F. R. Chaumartin, "Regular Polysemy in WordNet", JLCL, vol. 24, no. 2, 2009, pp. 5-18.

[15] R. Mihalcea, “Turning WordNet into an Information Retrieval Resource: Systematic Polysemy and Conversion to Hierarchical Codes,” IJPRAI, vol. 17, no. 5, 2003, pp. 689-704.

[16] J. Gonzalo, I. Chugur and F. Verdejo, “Sense clusters for Information Retrieval: Evidence from Semcor and the EuroWordNet InterLingual Index,” ACL-2000 Workshop on Word Senses and Multi-linguality, Association for Computational Linguistics, pp. 10-18.

[17] R. Navigli, “Word sense disambiguation: a survey,” ACM Comput. Surv., vol. 41, no. 2, 2009.

[18] M. Palmer, H. T. Dang amd C. Fellbaum, “Making fine-grained and coarse-grained sense distinctions, both manually and automatically,” Natural Language Engineering (NLE), vol. 13, no. 2, 2007, pp. 137-163.

[19] W. Peters and I. Peters, “Lexicalized systematic polysemy in WordNet,” Language Resources and Evaluation, 2000.

A Proposition for Fixing the Dimensionality of a Laplacian Low-rank Approximation of any Binary Data-matrix

Alain Lelu
Université de Franche-Comté/ELLIADD
LORIA
Campus Scientifique BP 239 - 54506
Vandoeuvre-lès-Nancy Cedex, France
alain.lelu@univ-fcomte.fr

Martine Cadot
Université de Lorraine
LORIA
Campus Scientifique BP 239 - 54506
Vandoeuvre-lès-Nancy Cedex, France
martine.cadot@loria.fr

Abstract— **Laplacian low-rank approximations are much appreciated in the context of graph spectral methods and Correspondence Analysis. We address here the problem of determining the dimensionality K* of the relevant eigenspace of a general binary datatable by a statistically well-founded method. We propose 1) a general framework for graph adjacency matrices and any rectangular binary matrix, 2) a randomization test for fixing K*. We illustrate with both artificial and real data.**

Keywords-dimensionality reduction; intrinsic dimension; randomization test; low-rank approximation; graph Laplacian; bipartite graph; Correspondence Analysis; Cattell's scree; binary matrix.

I. Introduction and State-of-the-Art

Spectral methods are used for optimally condensing and representing a set of objects in a space of lower dimensionality than the number of their descriptors. In this way, new relevant, informative and composite features are set apart from noisy, non-informative ones. This is especially useful when the descriptor space is sparse, which is typically the case for data of "pick-any" type, such as words in text segments, or links between Web pages, or social networks. An ever-growing number of applications rely on this type of condensed representation: Latent Semantic Analysis (LSA), supervised or semi-supervised learning, manifold learning, linear or non-linear PCA, vector symbolic architectures, spectral graph clustering and many others. A recurrent problem in any low-rank approximation process consists of determining the "right" rank of this approximation. Methods have been proposed to deal with this problem in the case of pre-defined data distributions, such as in [1]. But in the general case, nothing but empirical rules have been proposed, to the best of our knowledge: relying on the empirical evidence of a "gap" in the scree-plot of the eigenvalue sequence, whether visual or based on numerical indices such as first or second differences [2], or more basically on the value sqrt(N), etc. In LSA, empirical recommendations are provided [3], such as keeping the 200 to 400 first components. We address here the problem of determining the relevant dimensionality of the simplest, very common type of tabular data, i.e. the sparse binary tables. As such tables include adjacency matrices of unweighted graphs, and as graphs are known to be a powerful and extensively studied representation of many classes of data, we aim at incorporating in our framework a state-of-the-art representation space of graphs, i.e. one in the Laplacian family of eigenspaces. In the prospect of a maximum generality, a pleasant observation is that any binary datatable may be considered as a part of the adjacency matrix of a bipartite graph: we will focus, without lack of generality, on determining the best representation space, and its optimal number of dimensions, for unweighted and unoriented graphs, and thus for any binary matrix.

In Section II we will recall basic results about eigenanalysis of graphs and Correspondence Analysis. Section III will bridge the gap between graphs and general binary tables, and Section IV will present a randomization test for fixing the dimensionality of the relevant eigenspace. Applications to graph and data matrices are the topic of Section V, while we will close by presenting some related approaches, conclusions and future work.

II. Eigen-Spaces for Graph Mining

To the best of our knowledge, the first application of eigen-analysis to graphs dates back to Benzécri [4], when Correspondence Analysis (C.A.) was applied to adjacency matrices. Let us recall that C.A. [5][6] relies on the eigen-analysis of a matrix **Q** issued from any two-way correspondence matrix **X** (in the case of an undirected and unweighted graph, **X** is binary and symmetric; **Q** is symmetric, too) with

$\mathbf{Q} = \mathbf{Dr}^{-1/2}\, \mathbf{X}\, \mathbf{Dc}^{-1/2}$,

where **Dr** and **Dc** are the diagonal matrices of the row and column totals. The eigen-decomposition of **Q** writes $\mathbf{Q} = \mathbf{U}\, \mathbf{\Lambda}\, \mathbf{V'}$ where $\mathbf{\Lambda}$ is the diagonal matrix of the eigenvalues ($\lambda_1,..., \lambda_L = 1$, L being the number of connected components; $1 > \lambda_{L+1} >... > \lambda_R > 0$, R being the rank of **X**). **U** and **V** are the eigenvector matrices for the rows and columns respectively, giving rise to several possible variants of C.A. factors, depending on the authors. Benzécri [4] has shown analytical solutions for simple graphs such as rings or meshes. Lebart [7] has generalized to contiguity analysis, and illustrated by showing that the (**F2, F3**) factor plane representation of the contiguity graph between French counties reconstitutes the allure of the France map.

An independent research track starting with [8] has defined two “normalized graph Laplacians”, namely the symmetric Laplacian (**I** – **Q**), where **I** is the identity matrix, and $\lambda_1,..., \lambda_L = 0$, L being the number of connected components; $0< \lambda_{L+1}<...< \lambda_R$, R being the rank of **X**. The “random walk” variant is $\mathbf{I} - \mathbf{D_r}^{-1}\mathbf{X}$.

Spectral graph clustering consists of grouping the similar nodes in a K-dimensional major eigen-subspace – for a review see [9] – and is an increasingly active research line. To our knowledge and up to now, the problem of determining the number K, when the distribution of degrees is non-standard, has not received more satisfactory answers than the scree-plot visual or second-difference heuristics [2], visually prominent in the case of small graphs, but difficult to put into practice in the case of large ones.

III. From Graphs to Binary Matrices Through Bipartite Graphs

A well-established result in data analysis states that the relevant, noise-filtered information lies in the dominant eigen-elements of a data matrix [8]. In the case of the **Q** matrix, Benzécri [4], Chung [8] and many others have shown that the value of its first eigenvalue, of multiplicity L (L being the number of connected components), is one (the same is true of the $\mathbf{D_r}^{-1}\mathbf{X}$ matrix).

In the case of a bipartite graph, whose adjacency matrix and symmetric Laplacian write respectively $\left|\begin{matrix}\mathbf{0} & \mathbf{M}\\ \mathbf{M'} & \mathbf{0}\end{matrix}\right|$ and $\left|\begin{matrix}\mathbf{0} & \mathbf{Q}\\ \mathbf{Q'} & \mathbf{0}\end{matrix}\right|$, a simplification follows from the property of this type of matrices to have their eigenvalue set composed of the singular values of their rectangular non-empty submatrix, stacked with their opposites – in our Q case, in the range [-1; 1]. It follows that the basic correspondence analysis of any binary matrix, i.e. the SVD of its “symmetric” Laplacian matrix **Q**, giving rise to the signed contributions of its rows and columns to the inertia accounted by each factor, constitutes a basic reference for comparing this matrix to random counterparts.

IV. A Randomization Test for Fixing the Dimensionality of the Relevant Eigenspace

We have set up a randomization method [10] for generating random versions of a binary datatable with the same margins as those of this table, and set up the ensuing test for validating any statistics conducted on it. It is to be noted that the principles of generation of random matrices with same margins as a reference matrix seem to have been discovered independently several times, in various application domains: ecology, psychometrics, combinatorics, sociology. Cadot [11] legitimates a rigorous permutation algorithm based on rectangular “flip-flops”, and shows that any Boolean matrix can be converted into any other one with the same margins in a finite number of cascading flip-flops, i.e. compositions of elementary rectangular flip-flops: at the crossings of rows i_1 and i_2, and columns j_1 and j_2, a rectangular flip-flop keeping the margins unchanged is possible if the (i_1, j_1) and (i_2, j_2) values are 1 whereas the (i_1, j_2) et (i_2, j_1) values are 0. To our knowledge it was the first time this principle was introduced in data mining.

As is the case for all other randomization tests [12], the general idea comes from the exact Fisher test [13], but it applies to the variables taken as a whole, and not pairwise. The flip-flops preserve the irreducible background structure of the datatable, but break up the meaningful links specific to a real-life data table. For example, most of texts×words datatables have a power-law distribution of the words, and a binomial-like one for the number of unique words in the texts. This background structure induces our “statistical expectation” of no links conditionally to the type of corpus. Getting rid of the background structure enables this method to process any type of binary data, both (1) taking into account the marginal distributions, (2) doing this without any need to specify any statistical model for these distributions.

When using this algorithm, one must fix the values of three parameters: the number of rectangular flip-flops for generating non-biased random matrices, the number of randomized matrices, the alpha risk. This test is akin to be applied to adjacency matrices of bipartite, unoriented, unweighted graphs, as the non-empty parts of such matrices are made up of two symmetric rectangular binary matrices, and this structure is akin to be reproduced when generating random versions as described above. For generating randomized versions of the adjacency matrix of an unoriented, unweighted graph, further constraints have to be imposed at the step of enabling or not a rectangular flip-flop: the square matrix must be kept symmetric and its diagonal empty. Note that the problem at stake is different from the one addressed by [19], i.e. generating a random matrix with prescribed margins, which does not necessarily supports an exact solution.

V. Applications

We have detected the relevant dimension K* and used the corresponding reduced dataspace in the context of both graph and non-graph problems. In a proof-of-concept perspective, we will present here one artificial and one real dataset for each category – for a more detailed but less general presentation, see [14]. Throughout these examples, we will put forward successive visual representations we found useful.

A. Graphs

First, we have built the adjacency matrix of an unoriented and unweighted graph of 66 vertices, with four noisy cliques (missing intra-clique links: 17%; inter-clique links: 12%). Figure1 shows that the computed sequence of the 66 eigenvalues of its Laplacian, whether positive or negative, ranked by decreasing value of their module, fits into the “confidence funnel” of its 200 randomized counterparts, except the first one (value: one, by construction) and the next three, which clearly delineate the relevant support space for representing the four clusters as vertices of a tetrahedron (Figure2).

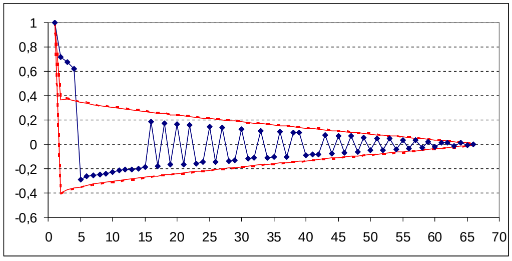


Figure 1. Graph with four noisy clusters: confidence funnel (red) of the eigenvalues (blue).

We have also processed the "Football league" data [15] which embed the "theoretical" social structure made of 12 regional "conferences", as well as the unsupervised structure emanating from the 115-node graph.

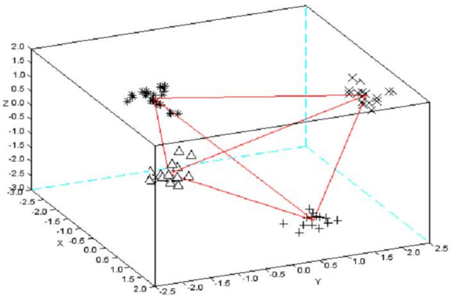


Figure 2. Graph with four noisy clusters: the 3-eigenvector relevant representation space of the 4 clusters.

According to our test with 200 randomized adjacency matrices, at the 99% confidence threshold, the ten "first" eigenvalues (N°2 to N°11, as there is a single connected component in the graph) of the original Laplacian matrix clearly dominate the confidence "funnel" of its 200 randomized counterparts.

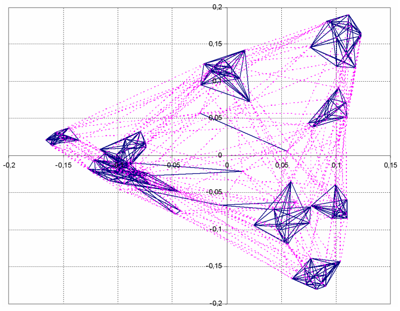


Figure 3. Football league: the U2 × U3 plane

An extra cluster analysis in this reduced space resulted in a quasi-perfect F-score measure for nine conferences, and a good or meager one for three of them, less geographically interrelated, summing up in a .956 global F-score. The U2 × U3 plane provides an overview of the structure (Figure 3), whereas the U4 to U11 dimensions offer more local points of view, and the subsequent ones show no recognizable structure.

B. *Binary Rectangular Tables*

We have designed a (1500; 836) binary matrix with a power-law distribution of the row sums and two fuzzy and overlapping column clusters, built by pasting twice the same (750; 836) Zipfian-distributed datatable, the second time with a random reordering of the columns. Our test results in two relevant eigenvalues, giving rise to a planar representation of the rows (Figure 4) showing off the orthogonality of two "logics", or "scales", and not a crisp opposition between two clusters.

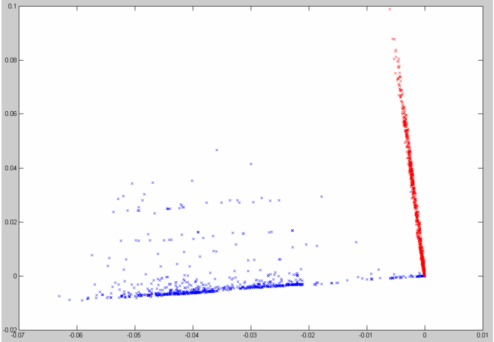


Figure 4. Data cloud with two overlapping "logics" projected onto the representation space of the two non-trivial and relevant eigenvectors U2 (vertical) and U1 (horizontal). In blue, the first 750 rows, in red the others.

Our experience is that this specific and rarely identified data structure is frequent in textual data; it takes here a concrete form when sorting the rows and columns of the datatable according to the dominant non-trivial eigenvector U2 (Figure 5).

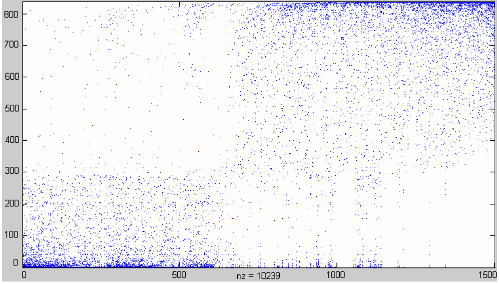


Figure 5. A binary datatable with two overlapping "logics" reordered by sorting the 1500 rows according to U2 and the 836 columns to V2.

We also processed a 753 per 11,567 Texts × Words matrix issued from a query to the Lexis-Nexis press database concerning three months of environmental controversies in the French press: it ensued that the relevant eigenspace was a 195-dimensional one, in which a cluster analysis of the words had put to the fore one clearly syntactical cluster from a hundred or so other ones devoted each to a particular press "story" noticeable in this period. In this case, the assessment criterion could be nothing but qualitative. Table 1 displays an example of such a news story.

TABLE I. EXAMPLE OF A CLUSTER IN THE PRESS DATABASE: THE NEWS STORY "THE EUROPEAN COMMISSIONER DACIAN CIOLOS NEGOTIATES AGRICULTURAL ISSUES IN WASHINGTON".

Rank	*Word*	*POS-tag*
1	antimicrobiens	U
2	spongiforme	U
3	Peterson	U
4	Ron	U
5	Mike	U
6	Dacian	U
7	CIOLOS	U
8	impression	SBC
9	09-févr	SBC
10	Lucas	U
11	US	SBC
12	durant	PREP
13	494	CAR
14	rappeler	PAR
15	préparation	SBC
16	répondre	PAR
17	conserver	VNCFF
18	subvention	SBC

VI. RELATED APPROACHES

While we have listed in Section I heuristic approaches for determining the relevant dimensionality of a data matrix, in [16] we presented a test in the same line as the one we develop here: we compared the singular values of a raw binary matrix to their counterparts in a set of randomized versions of this matrix. However this approach is subject to a major statistical concern: the singular-value scree does cross the upper bound of the singular-values of the randomized matrices, defining the desired relevant eigen-subspace, but it also crosses the lower bound, thus resulting in a difficult interpretation problem for the "significantly small" singular values. Moreover this approach offers no connection to Laplacian eigenspaces, nor Correspondence Analysis, as does the present one. Gionis et al. [17] deals, as we do, with the problem of finding out the number of relevant eigen-dimensions in a rectangular binary matrix, but presents a heuristic approach based on a unique randomized matrix, and no connection either to Laplacian eigenmaps nor Correspondence Analysis.

VII. CONCLUSION AND FUTURE WORK

We have presented a general framework for the dimensionality reduction of undirected and unweighted graphs, as well as of any rectangular binary table, a perspective covering both Laplacian eigenmaps and Correspondence Analysis of the said matrices. We have then shown that the number of dimensions of such an embedding space could be determined by a rigorous randomization test, contrasting with preceding heuristic approaches.

A major extension relates to scaling the procedure: whereas no efficiency issues arise for the data-class "n*1000 to m*10,000 vectors of n*1000 to m*10,000 dimensions", parallelization has to be set up beyond, both at the randomization level and the linear algebra computation one, which is well within the scope of the state-of-the-art. Another major extension, addressing both theoretical and practical difficult issues, is to generalize to any signed or unsigned integer matrix, if not any real-valued one.

ACKNOWLEDGMENTS

We thank the *CNRS/ISCC* for allowing us to use a part of its press data, and Azim Roussanaly for providing us with his efficient POS-Tagger [18].

REFERENCES

[1] Bouveyron C., Celeux G., and Girard S., "Intrinsic Dimension Estimation by Maximum Likelihood in Probabilistic PCA", Statistics and Computing, vol. 17(4), 2007.

[2] Cattell R. B., "The scree test for the number of factors", Multivariate Behavioral Research, vol. 1(2), 1966, pp. 245–276

[3] Deerwester S., Dumais S. T., Furnas G. W., Landauer T. K., and Harshman R. "Indexing by Latent Semantic Analysis". JASIS, vol. 41 (6), 1990, pp. 391–407 .

[4] Benzécri J.-P. L'analyse des données (3 tomes) Dunod, Paris, 1973

[5] Lebart L., Morineau A. and Warwick K., Multivariate Descriptive Statistical Analysis, John Wiley & sons, NY, 1984.

[6] Greenacre M., Correspondence Analysis In Practice, Chapman & Hall/crc Interdisciplinary Statistics Series, 2007.

[7] Lebart L., "Correspondence Analysis of Graph Structure", Comm. Meeting of the Psychometric Society, Bulletin Technique du CESIA, vol 2, 1984, pp. 5–19.

[8] Chung F.R.K., Spectral Graph Theory, (CBMS Regional Conference Series in Mathematics, No. 92), American Mathematical Society, 1997.

[9] Von Luxburg L., "A Tutorial on Spectral Clustering", Statistics and Computing, 2007, vol: 17(4).

[10] Cadot M., "A simulation technique for extracting robust association rules", CSDA'05, Chania, Greece, 2005.

[11] Cadot M., Extraire et valider les relations complexes en sciences humaines: statistiques, motifs et règles d'association. PhD thesis, Franche-Comté University, 2006.

[12] Manly B., Randomization, Bootstrap and Monte Carlo methods., Chapman and Hall/CRC, 1997.

[13] Fisher R., "The use of multiple measurements in taxonomic problems", Annals of Eugenics, 1936, pp. 179–188.

[14] Lelu A. and Cadot M., "Espace intrinsèque d'un graphe et recherche de communautés", Revue I3, CEPADUES, Toulouse, 2011, vol. 11, pp. 1–25.

[15] Girvan M. and Newman M. E. J., "Community structure in social and biological networks", Proc. Natl. Acad. Sci. USA vol. 99, 2002, pp. 7821–7826.

[16] Lelu A., "Slimming down a high-dimensional binary datatable: relevant eigen-subspace and substantial content", COMPSTAT'10, Paris, 2010.

[17] Gionis, A., Mannila, H., Mielikäinen, T., and Tsaparas, P., "Assessing data mining results via swap randomization", ACM Trans. Knowl. Discov. Data, 2007.

[18] Roussanaly, A., Morph POS-Tagger: www.loria.fr/~azim; accessed on 12/24/2012.

[19] Lancichinetti A. and Fortunato S., "Benchmark for testing community detection algorithms on directed and weighted graphs with overlapping communities", Physical Review. Vol. E 80, 2009.

Merging Capabilities of the Social Web and the Semantic Web to Support Knowledge Management in Small and Medium-Sized Enterprises

Raitis Sevelis, Alla Anohina-Naumeca
Department of Systems Theory and Design
Riga Technical University
Riga, Latvia
e-mail: raitis.sevelis@gmail.com, alla.anohina-naumeca@rtu.lv

***Abstract*—The research focuses on finding an interlinking solution for integration of the Social Web and the Semantic Web for knowledge management needs of small and medium-sized enterprises in respect to preserving their knowledge assets. The solution proposed includes a conceptual level framework for describing metadata of knowledge assets and a process for its usage. The framework can be used for introduction of a semi-structured knowledge repository which role is to act as a middleware between Social Web and Semantic Web tools used by small and medium-sized enterprises. Moreover, analysis of data acquired through online questionnaire and reveling trends in usage of Social Web environments and knowledge management activities within small and medium-sized enterprises is presented in the paper.**

Keywords-Semantic Web; Social Web; knowledge management; DBpedia; small and medium-sized enterprise

I. Introduction

Small and medium-sized enterprises (SMEs) are a significant subject in the European agenda, because they play a decisive role in the competitiveness and dynamic of the European economy [1]. Their central role is recognized by the Small Business Act for Europe [2] adopted in June of 2008. It puts into place a comprehensive SME policy framework for the EU and the Member States [3]. At the same time, since adoption of the Lisbon Strategy [4] the Europe is moving towards development of a knowledge-based economy that even more emphasizes significance of SMEs as providers of employment opportunities and key players for the well-being of local and regional communities [2]. It is logical that with the transition to a knowledge-based economy knowledge has became an important competitive factor and the most substantial value in all aspects of social and professional life. Therefore, today, staying competitive means relying upon knowledge of human resources and performing effective knowledge management with aim do not lose business-vital knowledge assets because usually know-how of enterprises is closely associated to tacit knowledge of their employees. In the paper, a knowledge asset is understood as any kind of knowledge like enterprise's know-how, employees' experience and competence, etc. used or held by a SME. In [5], two approaches to knowledge management are distinguished: 1) a product-oriented approach focusing on creating, storing, and re-using documents and 2) a process-oriented approach in which knowledge is tied to the person and is shared with other employees through communication. The paper focuses on the last approach.

Rapidly growing information and communication technology (ICT) provides a lot of solutions for effective knowledge management, but in the process-oriented approach the purpose of ICT is to help employees to communicate knowledge. However, typically SMEs use different environments, platforms, and collaborative tools (very often with different formats and structures of information stored), usually freely available in the Web, for support of knowledge management. This restricts collaborative use and sharing of knowledge assets in order to provide employees with knowledge needed in a particular time span. There is no mechanism in place which allows retrieving and transforming of knowledge assets from and between different environments without losing contextual meaning of these assets. One of the solutions for solving this issue could be introduction of technological capabilities offered by Web 3.0 (the Semantic Web). Web 3.0 allows describing and structuring of different knowledge assets using ontology which adds meaningful metadata to knowledge assets. Regardless that Web 3.0 technology suffers from drawbacks related to high-cost development and maintenance, as well as technological capabilities, it can be successfully used together with the already widely popular Social Web. Both technologies – the Semantic Web and the Social Web – are able to supplement each other in order to achieve knowledge integrity and ubiquity.

The current research was initially motivated by personal experience of one of the authors who is a CEO of a SME. Therefore, it seeks to answer the main research question: how both mentioned technologies can be integrated in an applicable way to support knowledge management activities within SMEs? The paper offers a conceptual level solution that can be adopted by SMEs on the top of their existing ICT infrastructure. It includes a framework for describing metadata of knowledge assets and a process for its usage. The framework can be used to introduce a semi-structured knowledge repository the role of which is to act as a middleware between Social Web and Semantic Web tools used by SMEs.

The paper is structured as follows. Section II considers related work in the field of usage the Semantic Web for needs of SMEs. General information on the Social Web and the Semantic Web is provided in Section III. Section IV presents the research methodology. Results of the analysis of data acquired through the online survey and revealing trends in usage of Social Web environments and knowledge management activities within SMEs are given in Section V. Then the solution for merging capabilities of the Social Web and the Semantic Web is described together with its testing results. Conclusions and directions of future work are given at the end of the paper.

II. Related Work

Breslin et al. [6] point out that the Semantic Web has became very popular last years and many large companies have started to experiment with it in order to understand the value of this technology for their business. The authors consider theoretical fundamentals of the Semantic Web and its usage in different application areas, inter alia the paper includes a quite comprehensive section on the Semantic Web and knowledge management. Rezgui et al. [7] concentrate on construction industry where SMEs are dominating and develop ontology for integration of disparate web-enabled applications and management of interactions between individuals and teams. The system which is based on the Semantic Web and supports knowledge workers in learning at workplace is presented in [8]. The authors report a case of deployment of the mentioned system in a network of SMEs and discuss issues which a company has to face, when it wants to deploy a modern learning environment relying on the Semantic Web technology. Goy and Magro [9] consider the problem of integration of the capabilities of knowledge based systems using ontologies (the Semantic Web) and possibilities of the Social Web and provide a design of a social web-based repository of software solutions offered by ICT companies for SMEs. It is necessary to note that regardless that there is a huge amount of publications and books on the Semantic Web itself, there are a few researches on potential of this technology for SMEs. The main reason could be related to drawbacks of this technology, especially high development and maintenance costs which could not be acceptable for SMEs usually operating within tight financial limits. At the same time, integration of the Semantic Web with the already widely distributed Social Web could provide significant advantages for SMEs without necessity to spend additional financial resources.

III. The Social Web and the Semantic Web

The Social Web, also referred as Web 2.0, assumes that content is created and managed by users populating a single environment (usually a social network). Users in this environment are related by the set of social relations which allow communication and content sharing. According to Rohani and Hock [10], social networking services offer users a space where they can maintain their relationships, chat with each other, share information, and build new relationships through existing ones. Kim et al. [11] define social websites as those websites that give people possibility to form online communities and share user-created contents. Therefore, the more users share their knowledge in a social network, the bigger knowledge base is created within this environment.

The Semantic Web, also referred as Web 3.0, introduces intelligence in the Web through representation of information in machine readable and understandable way leading to more adaptable and personalized environments. According to Bonilla-Morales, Medianero-Pasco, and Vargas-Lombardo [12], the Semantic Web is composed of a set of Web and knowledge representation technologies constituting what is known as a web of data where some human intelligence is integrated into the Web, making the search easier and more productive.

The architecture of the Semantic Web consists of the following layers [12][13]:

- Uniform Resource Identifier (URI) [14] used for location and identification of resources in the Web through giving them unique names;
- XML together with XML Schema [15] allowing transfer of different data between different environments;
- Resource Description Framework (RDF) together with RDF Schema [16] used to describe semantics of information; SPARQL query language allows expressing of queries for data that are stored in RDF format;
- Web Ontology Language (OWL) [17] allowing definition of a common vocabulary or, in other words, a common library of meanings which can be used between different environments.

The Semantic Web provides a number of benefits like possibility to describe all information in a semantic way, more consistent search queries, ability to transfer information between environments without losing its meaning, shifting tasks from humans to artificial intelligence, and improved decision making processes. However, regardless that the concept of the Semantic Web is very promising, it faces many issues such as high development costs due to necessity to rebuild ICT structure to get value from introduction of semantics, time consuming ontology building for a particular domain of interest, high maintenance costs due to changes, ageing, and appearance of knowledge in a dynamic way, etc.

At the current point of technological development and of users' overall ICT skill level, it is hard to maintain a computable Semantic Web-based environment alone, without introduction of any other already existing (or new) technologies. The research presented by Jovanović, Gašević, Torniai, Batemand, and Hatala [18] suggests combining the already popular and well adopted Social Web with the technological capabilities offered by the Semantic Web. In this case, it is possible to avoid drawbacks of the Semantic Web and also to improve the current Web infrastructure, which lacks interoperability from the perspective of knowledge transfer. Combination of these technologies does not mean "blind" merging of the capabilities as such, but taking the best independent components and/or concepts to bind them together. The same research also mentions that the Semantic Web cannot work alone in an available Web environment as it requires collaborative applications based on the Social Web that allow operating with shared knowledge. On the other side, the Social Web can benefit from structured and easily transferable knowledge because it can be used by multiple applications without special adoptions.

IV. Research Methodology

The research presented in the paper followed the following main steps:

1) Identification of trends in usage of the Social Web within SMEs and issues related to preserving knowledge assets within an enterprise;

2) Examination of capabilities of the Social Web and the Semantic Web in relation to the results of Step 1;

3) Finding an integration solution.

Step 1 was based on a questionnaire which development proceeded in two stages:

1) First of all, an initial questionnaire was developed and then used in the interview process with several management representatives of SMEs;

2) After the interview process, the initial questionnaire was modified to acquire a meaningful set of questions and then an online questionnaire was developed using services of the website [19].

The final questionnaire included 16 questions distributed between the following categories: 2 questions – information about the enterprise; 3 questions – usage of Social Web environments and tools within the enterprise; 3 questions – knowledge management tools used by the enterprise; 3 questions – enterprise's opinion about usage of the Social Web for knowledge management; 5 questions – trends in having employees with a unique knowledge set and existence of a scenario for preserving knowledge assets.

The final questionnaire was distributed to 90 managers of different SMEs in Latvia using personal contacts of the authors.

V. Usage of the Social Web and Knowledge Management Activities within Small and Medium-Sized Enterprises

The survey was organized with the following hypothesis in mind: despite active use of Social Web tools in SMEs, enterprises do not have a scenario in place for managing and preserving their knowledge assets.

During March of 2012, 50 respondents (from 90 to whom the questionnaire was sent) from different business areas took part in the survey. They included managers and CEOs of SMEs in the Republic of Latvia. The majority of the data were collected from enterprises in ICT field - 22,22% (11), followed by Production - 18,18% (9) and companies which provide B2C services 12,12% (6). The whole set of enterprises participated in the survey was distributed as follows: ICT – 22%; Other – 22%; Production – 18%; Services – 12%; Finance – 8%; Construction/Real estate – 6%; Wholesale – 4%; Retail – 4%; Media – 2%; Transport/Logistics – 2%.

The usage of Social Web environments and tools within an enterprise is very closely related to behaviour of employees and their willing to use such tools not only in their primary duties at work, but also in everyday life. Figure 1 presents data about usage of Social Web environments and tools among employees from the perspective of management. It is possible to note that 50% (25) of all enterprises assume that at least 60% of their employees use Social Web environments in their everyday life to perform different activities, such as communication, knowledge sharing, and learning. However, it is important to take into account that this data can be biased, because of the fact that they are provided by the management, not by employees themselves and can differ from the real situation, as some employees may hide their behaviour of using the Social Web.

Further analysis of the data shows that 74% (37) of all respondents use the Social Web in their enterprises, nevertheless only 58% (29) answered that they know what is knowledge management and its associated activities. Moreover, 50% (25) of respondents were able to choose specific Social Web tools they find beneficial for knowledge management within their enterprises. The managers chose the following tools: Messenger – 20%; File transfer – 18%; Multimedia – 12%; Chat – 11%; Forums – 8%; Wiki – 8%; Blogs – 7%; Bookmarks – 7%; Feedback forms – 6%; Other – 3%.

The data show that the most popular are Messenger with included electronic mail, File transfer, Multimedia, and Chat. The least popular are tools that require much more efforts in relation to content creation and its management. They are Forums, Wiki, Feedback forms, Bookmarks, and Blogs. Among other tools beneficial for knowledge management, the managers mentioned Twitter which has got enormous growth in recent years due to easy and simple interaction and user interface.

The whole spectrum of the tools chosen by the management allows making a conclusion that two main aspects of knowledge management are equally important for enterprises: a) communication and b) knowledge sharing. The managers who participated in the survey explained that they consider these tools as a mechanism which supports knowledge sharing in a more efficient manner and allows faster interexchanging (communication) of different knowledge assets between employees.

Analysis of the data concerning knowledge holders within enterprises reveals that 80% (39) of enterprises have key employees who own a unique knowledge set from the perspective of a particular enterprise. This is a common situation for small companies and startups in their early development stage, when there is a group of key employees who perform several crucial tasks and monitor main business processes. Regardless that having such employees is a big risk for any enterprise, the answers showed that the managers pay no effort in solving such a situation, as 74% (37) answered that they do not have a scenario developed for preserving knowledge in case if an employee leaves the enterprise. The reason for such a behaviour can be explained in two ways. The first one is that usually SMEs delegate

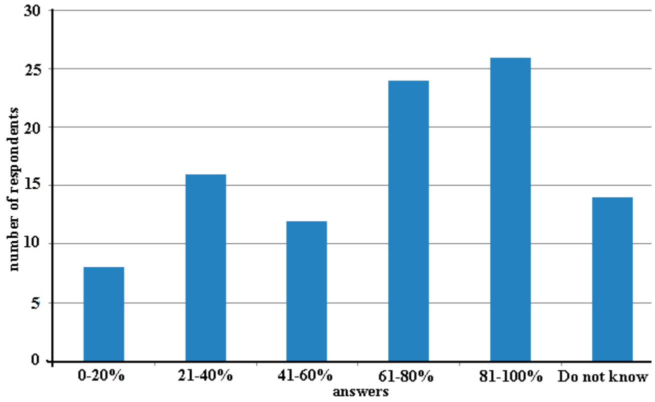


Figure 1. Usage of Social Web environments among employees from the perspective of enterprises' management.

several fields of competence to the same employee to save financial resources. This is a pure management problem which does not have relationship with technological issues. However, still effective business process optimization with the help of ICT can improve the situation. The second reason is that the most of the management are not aware or do not have a mechanism available to capture key knowledge and preserve them. Regardless that tools built on the concept of the Social Web allow collecting of knowledge, there is still a problem of developing efficient exchange of this knowledge across different applications in necessary context and provision of them to employees in the same manner as a unite knowledge asset. Variety of tools and data formats preferred and used by enterprises hinders efficient knowledge sharing across the enterprise, resulting in a huge risk for organizations to suffer from loss of key knowledge assets.

VI. Integration of the Semantic Web and the Social Web for Purposes of Knowledge Management

The analysis of the questionnaire presented in Section V revealed that the most part of SMEs use Social Web tools to ensure communication and sharing of knowledge assets. Therefore, to support both of the mentioned aspects in efficient way, it is necessary to integrate metadata of the Semantic Web into a Social Web environment to provide semantic meaning for knowledge assets of the enterprise.

A. The Proposed Solution

Taking into account that different types of tools and environments use different methods for representation of knowledge, it is necessary to find common patterns in order to introduce a standardized knowledge representation structure. For this purpose examination of Social Web tools for communication, knowledge sharing and management was made paying attention to their data sets. The results allowed assuming that it is possible to use a common metadata description framework which relies on the following attributes:

- title of a knowledge asset;
- abstract of the knowledge asset;
- description of the knowledge asset;
- data about the author;
- link to a knowledge source.

The title of a knowledge asset is a common and mandatory description for any type of knowledge provided and used within a SME. The abstract of the knowledge asset provides a short introduction into knowledge contained in the asset. The description of the knowledge asset includes content of the knowledge asset from a communication tool and it should not be mandatory because file transfer and multimedia solutions in most cases do not provide this kind of information. To ensure reliability of knowledge and to increase responsibility of employees within SMEs, it is necessary to store information about the author of the provided knowledge asset. Such information could also be useful in case of necessity to add additional knowledge to the asset. In case of using a knowledge sharing and management tool for storing or transforming non-textual knowledge assets, links to the original knowledge source show the exact location of the knowledge asset. This attribute should not be mandatory, as textual information does not require a link to a knowledge source because of its integration into the description area of the knowledge asset. Taking into account different formats of knowledge assets used within SMEs, it is recommended to use links to original knowledge sources. A knowledge asset itself is still kept within the storage repository of the Social Web environment. Following such a guideline allows bypassing implementation process of a complex and costless mechanism for managing different knowledge assets within ICT environment of SMEs.

The next step is to discover common patterns within the Semantic Web to create a linkage between Web 2.0 and Web 3.0 technologies. In order to define the research scope, it was decided to focus on technological capabilities and structure of the Semantic Web portal DBpedia [20], which is the biggest publicly available free semantic knowledge repository, containing more than 3.6 million descriptions of knowledge assets and over 1 billion of RDF triples. DBpedia structures and describes knowledge assets in a semantic way from Wikipedia which is a typical Social Web environment. DBpedia's centrality and cross-domain nature makes it one of the most important and most referred knowledge bases on the Web of Data, generally used as a reference for data interlinking [21]. Its knowledge assets contain information about persons, places, music, films, games, organisations, species, diseases, and many other subjects of interest that can be used to support knowledge management in SMEs by enriching their knowledge repositories with additional knowledge interlinked using an ontological description.

In order to access the knowledge repository of DBpedia, it is possible to use SPARQL endpoint offered by DBpedia for direct implementation of complex queries to retrieve description of a specific entity displayed within infobox. Depending on the type of an entity, DBpedia uses different infoboxes to display its content to the user; nevertheless there are still common properties for all knowledge assets. Examination of different types of content revealed the following set of properties:

- title of a knowledge asset;
- rdf:type – a type defined in DBpedia and assigned to each knowledge asset with aim to group knowledge assets and make their hierarchy;
- dbpedia-owl:abstract – abstract of each knowledge asset of DBpedia;
- dbpedia-owl:wikiPageExternalLink – a link to a description of a knowledge asset in Wikipedia;
- is dbpedia-owl:wikiPageRedirects – a link to synonyms of a knowledge asset.

Structure of the properties described above can be considered as similar to some degree with the metadata description proposed by the authors of the paper. Such similarity allows introduction of interlinking between unstructured knowledge assets of Social Web tools and the knowledge repository of DBpedia in order to add additional meaning to unstructured knowledge assets. Knowledge assets retrieved from the Social Web can benefit from such interlinking by acquiring additional knowledge and imposing structure through different assets of DBpedia. By introduction of the structuring mechanism for knowledge assets of the Social Web, it is possible to get semi-structured knowledge, which can be used to support knowledge management within SMEs, and to increase reliability and accessibility of particular knowledge. Moreover, it allows an enterprise to use additional knowledge available in DBpedia.

In order to introduce interlinking between knowledge assets of the Social Web and DBpedia, it is necessary to develop a middleware that will serve as a metadata description repository of knowledge assets within an environment. The purpose of the middleware is to tie knowledge assets from different Social Web tools with knowledge instances from the knowledge repository of DBpedia. Figure 2 illustrates interlinking of Social Web environments and DBpedia.

Taking into consideration the structure of properties in DBpedia and the one within the Social Web, it is possible to define a list of properties that should be introduced within the middleware for description of knowledge assets:

- rdf:title – title of a knowledge asset;
- rdf:category – a category of the knowledge asset allowing to group knowledge assets; categories can be freely defined by an enterprise;

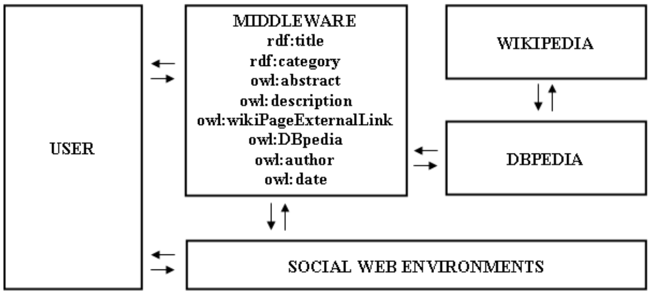


Figure 2. User interaction with the middleware and Social Web environments.

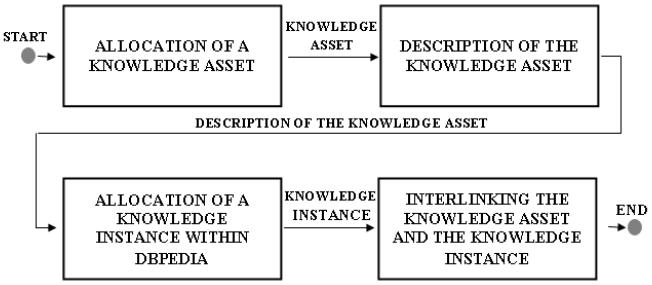

START
ALLOCATION OF A KNOWLEDGE ASSET
KNOWLEDGE ASSET
DESCRIPTION OF THE KNOWLEDGE ASSET
DESCRIPTION OF THE KNOWLEDGE ASSET
ALLOCATION OF A KNOWLEDGE INSTANCE WITHIN DBPEDIA
KNOWLEDGE INSTANCE
INTERLINKING THE KNOWLEDGE ASSET AND THE KNOWLEDGE INSTANCE
END

Figure 3. The process for interlinking a knowledge asset and its instance.

- owl:abstract – abstract of the knowledge asset;
- owl:description – description of the knowledge asset;
- owl:wikiPageExternalLink – a link to a Social Web tool, where the knowledge asset is located;
- owl:DBpedia – a link to a description of the knowledge asset in DBpedia;
- owl:author – data about the author (employee);
- owl:date – date when the knowledge asset was added; introduced with aim to fix when the last modifications were made and how actual is the knowledge asset.

In order to achieve better linking between the middleware and DBpedia, it is necessary to describe all properties and values using the RDF descriptive language. Such an approach is able to provide additional benefits in future by moving an enterprise's knowledge repository even closer to semantic structure. In order to perform interlinking of a Social Web environment and DBpedia, the middleware should be able to support a process proposed by the authors of the paper and illustrated in Figure 3. First of all, it is necessary to allocate a knowledge asset within a particular Social Web environment. This task is not mandatory for the middleware as it can be performed manually by an employee of the enterprise. This allows to lower development costs of the middleware. After the allocation process, it is necessary to describe the knowledge asset using the RDF descriptive language according to the property structure described previously. In order to interlink the knowledge asset with DBpedia, the employee must perform search procedure to allocate a corresponding knowledge instance within DBpedia. After retrieving information about the knowledge instance, the employee must add value to owl:DBpedia property to establish a link between the asset and the instance. After implementation of the described process, employees of the enterprise will be able to access the interlinked knowledge asset for performing learning or decision-making processes.

Following the process described above, it is possible to define any type of knowledge assets from the Social Web and interlink them with knowledge instances of DBpedia in order to provide additional meaning to a particular knowledge set and to introduce a semi-structured knowledge repository within SMEs without rebuilding their ICT infrastructure. The interlinking solution (see Figure 3) allows still relying on capabilities offered by the Social Web; nevertheless it allows using of concepts of Web 3.0

within an enterprise to improve knowledge management activities.

B. Testing the Solution

In order to prove feasibility of the solution developed and presented in this paper, it was decided to perform an experiment within a SME by linking its existing knowledge assets from a Social Web environment with external knowledge assets from the Semantic Web Portal DBpedia. The SME is using Wiki platform for knowledge management purposes by describing key knowledge used in its business processes. Regardless that Wiki allows grouping of knowledge assets using criteria developed by an organization, it lacks semantic structure which makes it difficult to structure knowledge in a meaningful way. The Wiki of the SME contains knowledge assets related to user interface and web design. Taking into account two mentioned groups of knowledge assets, it was decided to group knowledge assets presented in Wiki by linking them to two different instances of DBpedia about web design and user interface issues. Table I illustrates a description schema for interlinking knowledge assets of the SME with knowledge instances from DBpedia taking into account category of a particular knowledge asset.

After that, the following experiment was performed. A knowledge asset was chosen from the Social Web portal Behance Network (http://www.behance.net/) which serves as a creative community for communication between designers and storing of different data formats for building creative portfolio. Portfolio assets within Behance Network in most cases consist of graphic and textual materials presented using a common template for all sets of data. The chosen knowledge asset contains portfolio information about design of user interface for Latvian fashion designer Davids' Internet resource and can be located through the following [22]. Table 2 lists properties and values of the description of the knowledge asset.

In order to test the description of the knowledge asset and integration results of DBpedia, it was decided to share the knowledge asset with two employees of the enterprise. One of the employees received the knowledge asset with the attached description. The second partner received just the knowledge asset without the description. After that to test results, the meeting with all three parties (the manager and 2 employees) was arranged to discuss the project from the perspective of web design. Results of the meeting revealed that both employees were informed about the project and had no problems with involvement into discussion about the particular subject. Nevertheless when the topic of the meeting was changed in a favour of web design area itself, the employee who had received the knowledge asset with the additional description showed better understanding about the problem area and was taking more active part in the discussion process. Moreover, he had also acquired knowledge about graphic design as it was linked with the web design instance of DBpedia.

TABLE I. PROPERTIES AND VALUES FOR THE DESCRIPTION OF THE KNOWLEDGE ASSETS

Properties	Interlinking with 'Web Design'	Interlinking with 'User Interface'
rdf:title	[Knowledge asset title]	[Knowledge asset title]
rdf:category	Web design	User interface
owl:abstract	[Abstract]	[Abstract]
owl:description	[Description]	[Description]
owl:wikiPageExternalLink	[link to original knowledge source within Wiki]	[link to original knowledge source within Wiki]
owl:DBpedia	http://dbpedia.org/page/Web_design	http://dbpedia.org/page/User_interface
owl:author	[Data about author]	[Data about author]
owl:date	[DD-MM-YYYY HH:MM]	[DD-MM-YYYY HH:MM]

TABLE II. DESCRIPTION OF THE KNOWLEDGE ASSET USED IN THE EXPERIMENT

Properties	Values
rdf:title	David's user interface
rdf:category	User interface
owl:abstract	User interface design for Latvian fashion designer David's Internet resource
owl:description	User interface layouts for Latvian fashion designer David's Internet resource. User interface set includes layouts for introduction, homepage, and inner pages of web design project
owl:wikiPageExternalLink	http://www.behance.net/gallery/D-Fashion/149987
owl:DBpedia	http://dbpedia.org/page/Web_design
owl:author	Raitis Sevelis
owl:date	16-05-2012 13:38

The experiment showed that usage of the interlinking solution within SME allows, first of all, introduction of a semi-structured knowledge repository by adding additional knowledge grouping criteria based on the Semantic Web Portal DBpedia. Moreover, the proposed solution enlarges the knowledge repository of the SME by adding free available knowledge from DBpedia by linking it with the already existing and used knowledge assets and can be considered as a first step towards the Social Semantic Web.

VII. CONCLUSION AND FUTURE WORK

The research presented in the paper included analysis of data acquired through the survey aimed to uncover trends regarding usage of Web 2.0 environments and knowledge management tools within SMEs. As a result, the main conclusion made was related to the fact that SMEs actively use Social Web environments for knowledge communication and sharing. At the same time, the most of enterprises do not have a scenario for preserving their knowledge assets. Taking into account the findings mentioned above, the solution for interlinking the knowledge repository of the Semantic Web portal DBpedia with Social Web environments was proposed and tested. Integration of DBpedia, as a resource of Web 3.0, can help SMEs to structure their unstructured knowledge assets from different Social Web environments by introducing the semi-structured knowledge repository through adding contextual meaning to existing knowledge assets The proposed solution allows interlinking of different Social Web based knowledge communication and sharing tools and does not restrict enterprises in adding metadata descriptions to existing knowledge assets. Nevertheless, it requires introduction of the middleware between DBpedia and a Social Web environment, which serves as a bridge between the Social Web and the Semantic Web. Future work is related to practical implementation of the proposed solution. It is planned to introduce the mentioned repository in several SMEs in Latvia and to perform pilot testing with further identification of its usability, advantages, and drawbacks through a questionnaire offered to staff and managers of the SMEs.

REFERENCES

[1] European Commission, EU SME policy, European Small Business Portal, http://ec.europa.eu/small-business/policy-statistics/policy/index_en.htm 19.12.2012.

[2] “Think small first: a “Small Business Act” for Europe”, Communication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, COM(2008) 394, Brussels, 25.06.2008.

[3] European Commission, “Small and medium-sized enterprises (SMEs): Small Business Act for Europe, Enterprise and Industry”, http://ec.europa.eu/enterprise/policies/sme/small-business-act/ 19.12.2012.

[4] Presidency Conclusions, Lisbon European Council, March 23-24
2000, http://www.consilium.europa.eu/uedocs/cms_data/docs/pressdata/en/ec/00100-r1.en0.htm 19.12.2012.

[5] D. Apostolou, G. Mentzas, R. Young, A. Abecker, “Consolidating the product versus process approaches in knowledge management: the know-net approach,” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.21.5816 19.12.2012.

[6] J. G. Breslin, D. O’Sullivan, A. Passant, and L. Vasiliu, “Semantic Web computing in industry,” Computers in Industry, no. 61, 2010, pp. 729–741.

[7] Y. Rezgui, S. Boddyb, M. Wetherill, and G. Cooperc, “Past, present and future of information and knowledge sharing in the construction industry: towards semantic service-based e-construction?,” Computer-Aided Design, no. 43, 2011, pp. 502–515.

[8] C. Christl, C. Ghidini, J. Guss, S. Lindstaedt, V. Pammer, P. Scheir, and L. Serafini, “Deploying semantic web technologies for work integrated learning in industry. A comparison: SME vs. large sized company,” Proc. 7th International Semantic Web Conference (ISWC 2008), Springer, Oct. 2008, pp. 709-722.

[9] A. Goy and D. Magro, “Exploiting folksonomies and ontologies in an e-business application,” http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.154.9487 19.12.2012.

[10] V. Rohani and O. S. Hock, On Social network Web sites: definition, features, architectures and analysis tools,” Journal of Advances in Computer Research, no. 2, 2010, pp. 41–53.

[11] W. Kim, O.-R. Jeong, and S.-W. Lee, “On social Web sites,” Information Systems, no. 35, 2010, pp. 215–236.

[12] B. Bonilla-Morales, X. Medianero-Pasco, M. Vargas-Lombardo, “Survey: grid computing and Semantic Web,” International Journal of Computer Science Issues, vol. 7, 2010, pp. 1-6.

[13] M. M. Taye, “Understanding Semantic Web and Ontologies: Theory and Applications,” Journal of Computing, vol. 2, iss. 6, 2010, pp. 182-193.

[14] Semanticweb.org, "Uniform Resource Identifier", http://semanticweb.org/wiki/Uniform_Resource_Identifier 19.12.2012.

[15] W3C, "XML Schema", http://www.w3.org/XML/Schema 19.12.2012.

[16] W3C, "RDF Vocabulary Description Language 1.0: RDF Schema", http://www.w3.org/TR/rdf-schema/ 19.12.2012.

[17] W3C, "OWL 2 Web Ontology Language", http://www.w3.org/TR/owl2-overview/ 19.12.2012.

[18] J. Jovanović, D. Gašević, C. Torniai, S. Batemand, and M. Hatala, “The Social Semantic Web in intelligent learning environments: state of the art and future challenges,” Interactive Learning Environments, vol. 17, iss. 4, 2009, pp. 273-309.

[19] Webanketa, "Free creation of questionnaires,surveys, tests and polls!", http://webanketa.com/ 03.02.2012.

[20] DBpedia.org, "About DBpedia", http://wiki.dbpedia.org/About 13.04.2012.

[21] F. Orlandi and A. Passant, “Modelling provenance of DBpedia resources using Wikipedia contributions,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, iss. 2, 2011, pp. 149-164.

[22] Behance, Inc., "D-Fashion", http://www.behance.net/gallery/D-Fashion/149987 16.04.2012.

Negative Discourse and Emergence of Game Genres

A social networks exploration

Emanuela Marchetti
Centre for Design Learning and Innovation
Department of Learning and Philosophy
Aalborg University
Esbjerg, Denmark
ema@create.aau.dk

Andrea Valente
Centre for Design Learning and Innovation
Department of Architecture, Design and Media Technology
Aalborg University
Esbjerg, Denmark
av@create.aau.dk

***Abstract*— The fluidity of game genres can be considered an issue to be solved; however, we propose to look at it as a social value, from a positive and creative perspective. In this study, we analyze emergence of game genres, experimenting with a constructive use of negative discourse from online communities, about the controversial genre of mono-dimensional games (i.e., 1D). Analyzing online conversations, 1D games can be defined as a *non-genre*, since posts about them are almost exclusively negative, referring to it as synonymous of an impossible or meaningless effort. We decided to challenge this attitude towards 1D games, iteratively designing and coding a few 1D games, and then post about them on selected online communities. This exploration of the 1D games-design space was conducted through user centred design and netnography. Our results suggest that 1D games can be recognized as fun and challenging, from the perspective of both players and game designers.**

Keywords- netnography; game genre exploration; game design.

I. INTRODUCTION

Users of online forums sometimes start a conversation, for fun, about a particular concept or topic, as quintessentially meaningless. In some game development forums and online articles the idea of a 1 dimensional game has often been discussed as evidently silly, possibly useless or just clearly impossible. However, a few 1D games exist and have received positive reviews. Therefore, it was decided to conduct an investigation on 1D games, as example of emergent games genre, through the method of netnography, a form of ethnography applied to online communities, combined with user centred design (UCD) [1] [2] [16].

The initial survey that we conducted, involving online game development communities and online magazines, convinced us that 1D games might be more than just an evident impossibility, and that we might instead be witnessing the early stages of gestation of a game genre: the moment before a *non-genre* becomes an actual socially recognized class of games.

In order to discuss 1D games we have to investigate more closely what could a mono-dimensional game be like. Moreover, we need to find out whether a 1D game would actually be playable, engaging and if the 1D game space (i.e., the *design space* defined by all the possible 1D games) allows enough room for game developers to be creative.

In the next sub-sections, the motivation and related work are presented (A and B). In Section II the notion of negative discourse and non-genre are introduced, and then in Section III the method followed to conduct the study is presented. Section IV presents the 1D games we designed and in V data from the netnographic fieldwork. Finally in Section VI the outcome from the study is discussed and in VII discussion and future work.

A. Motivation

In order to explore how game genres emerge, intended as fluid social values, this study focuses on 1D games, approximately defined as a game that uses a single line of pixels (vertical or horizontal) as visualization. Working with such a limitation is a challenge, but there is a purity and simplicity in it that we find fascinating.

The emergence of one button games [4] as genre, might represent a similar case to 1D games. It was defined less than a decade ago [4], and perhaps the idea of a game with a single button, as its user interface, must have sounded useless and silly at first. The emergence of one button games seems to have started out as a technical challenge, but nowadays these games are acknowledged as a proper genre by many game developers (especially occasional and indie ones). Particular interest has been expressed by the disabled online community, that quickly adopted one button games and even encouraged developers to create more of this kind of games, to empower disabled players [17]. The recent diffusion of touch-sensitive devices gives even more relevance to the single-button concept, as exemplified by the many tapping-games in the android and iPhone market. There might have been single button games before the [4] article formalized them, but making the idea explicit, and discussing technical issues and aesthetic potentials, generated discussion and game designs, ultimately establishing one button games as a proper game genre.

This need of technical artefacts as well as social (online) discussion is central also in our attempt to assess the potential of 1D games: we decided in fact to poke at the online communities, and challenge their negative attitude towards the 1D non-genre. For that, we developed simple online games, using free and easy-to-use technologies such as javascript and HTML5 canvas, while keeping a discourse open with players, other developers and game designers. To explore the 1D games design space, we started by re-designing classic 2D games so they could be playable and recognizably related to the original game, even with only a 1D visualization. We call this redesign process *flattening*. Keeping a game recognizable after flattening it, requires to find out what is the spirit or the identity of that game; this identity should have to do with how the players feel, think and interact with the game itself. Flattening a game like Tetris [18] for instance, should result in a 1D game that, when players play it, brings to mind the classic 2D Tetris. Our exploration is therefore centered on the player's perception of games, and on the artistic expressivity that 1D games might offer as a medium for game design.

B. Related work-The social dimension of game genre

In this paper, we investigate the emergence of game genre, in relation to online communities, composed by players and designers. The social essence of ontology and the role of social networks have been discussed in relation to different fields, such as recommendation for academic conferences [5] and a web-based social bookmarking tool called del.icio.us [6].

According to Mika [6], the notion of ontology as an engineering artefact, is too simplistic, as it ignores the social aspect of ontology. Hence the traditional bipartite model for ontologies, based on concepts and instances, has been expanded with the third dimension of actors. The goal of this model is to discuss the emergence of ontologies, intended as folksonomies, which demand for social presence to be created and maintained [6].

Looking into current literature in the field of game genres, it seems that a debate has emerged in relation to the value of clustering games into genres and also to the emergence of game genres [7]. Games can be analyzed as both design objects and emergent culture [7][8], tokens in public conversations of broader societal issues within contemporary offline society [7]. In this sense, studying games through the lens of online communities can contribute to achieve a deeper understanding of contemporary culture and related social dynamics. However, when it comes to game genres, there seems to be a tendency to categorize games in relation to the dichotomy of *ludology*, the study of games as a system of rules [7] and *narratology*, the study of games as texts [7][9]. On the other hand, the tendency to classify games into market-based categories is seen as hiding the essence and meaning of games, as a new medium. In fact it seems that the many existing game categories are not based on the essence of games per se, but are based on previous media, as a consequence, a cohesive discussion of games as a new media, seems to be missing [9]. This situation seems to be originated, by the tendency to classify games in an uncritical way, based on their different representational strategies and not on other common features. Apperly [9] focuses his analysis on four genres:

- *Simulation*, games that simulate activities such as sports, flying and driving, or social dynamics in relation to towns or other small communities.
- *Strategy*, divided into real time or turn based, they have a similar aesthetics, with a god's eye view over the scene and photorealistic visualization.
- *Action*, divided into 1st person shooter, in which the screen represents the player's own view, and 3rd person, in which the player interacts through an avatar.
- *Role-Play or Adventure*, games inspired by the literary genre of fantasy.

All these genres are based upon visualizations, preexisting media and other leisure activities, leading to a fragmented analysis of games. On the other hand, interactivity, intended as the way a game should be played, is a common non-representative characteristic, which determines players' experience, is seen as a more promising framework [9].

A similar view is shared by Schell [8], who claims that games are *designed experience*, characterized by different forms of interaction. However, in [8] game genres are seen as a fluid, continually changing phenomenon, while the basic principles of game design are based upon human psychology, which is regarded as stable and more reliable set of knowledge, enabling designers to master all the different genres and even invent new ones.

This ever-changing character of game genre is also acknowledged by Arseneault [10], in his study about the emergence and evolution of game genres. Game genres are discussed here as a complex and incoherent notion, as they can be constructed based on different characteristics of games, and in some cases what is considered a genre by someone, is seen as a sub-genre, or "flavor", or even another medium [10]. This difficulty in creating coherent genres is an issue not only in relation to games, but with the very concept of *genre* itself. Interestingly, Arsenault quotes Apperly [9] in saying that he agrees on the genre issue, but he would conclude that there is no suitable solution. He then proposes the concept of the *Great Genre Illusion,* according to which genre is an umbrella word, grouping together disparate things with little relations to each other [10].

Connecting emerging perspectives on the relation between ontology and social networks, this study aims at analyzing the emergence of new game genres, intended as social values determined by a community. Moreover, the fluidity of genres is seen as an opportunity for innovation, providing new design inspirations and adapting to emerging needs.

II. Negative Discourse and Non-Genre

According to our study, the genre of 1D games is emerging as a meaningless and ludicrous concept, as a consequence there are very few 1D games and most of them are a mockery of the concept itself. Ironically, the negative discourse about 1D games is based on reviews of existing 1D games. Examples could be Tetris 1D [19], where the player always wins, since the only Tetris piece always falls at the right place and gives points. The author seems to be making fun of the fact that in 1D no 2D rotations are possible, leaving no room for the kind of gameplay typical of the classic 2D Tetris.

Another 1D game is Wolfenstein 1D [20], meant mostly as a tribute to the Wolfenstein 3D game (developed by id Software and published by Apogee Software in 1992). Players who know the original game should be able to move around in the 1D version, and enjoy trying to understand what happens in the over-simplified rendering of the original 3D game. Interestingly, having 3 dimensions is a very central element for Wolfenstein 3D, which is usually considered the game that started 3D first-person gaming on PCs, back in the early 1990s. In this sense, the re-design of this game is a meaningful experiment, especially for players of the original game, and provides a sense of nostalgia as well as fun. To reconnect the 1D game to its 3D original, sounds and color theme from the original game are kept.

There are, however, a few actual 1D games, which are often not even explicitly presented as mono-dimensional. Among the most interesting: Z-rox [20], Line [21] and Gauge [22] for iPhone. Z-rox is a game in only 1 dimension, in the sense described by Flatland [11], where a bi-dimensional shape (a character in this game) crosses a line, and the player has to guess which character it was, just by observing its 1D projection. The authors of the game Line talk about exploring the possibility of a one-dimensional shooter game; they decided to develop their game as a collection of mini-games, with minimalistic graphics and using mostly grays. The comments on the forum that followed the post about line are generally positive. Finally, Gauge is a commercial game for iPhone where the player tap on a single button (i.e., the entire screen) to control a horizontal gauge that changes size. The closer the gauge gets to the edge of the screen, the more points the player gets, but if the gauge exceeds the screen the player loses. While this game is clearly mono-dimensional, nothing about its 1D essence is explicitly written in its description. In this sense, the existence of such games confirm that a 1D genre is in the process of its definition and still in need to be acknowledged by the developer community.

III. Method

Based on our first encounter with the concept of 1D games, it was decided to conduct our investigation, through the method of netnography (form of ethnography applied to online communities [1] [2]) in combination with user centred design. Netnography is broadly applied within the marketing field, as it allows to conduct exploratory studies in an unobtrusive way [1], getting in touch with a large number of users. The procedure of applying netnography to a particular study has been formalized according to traditional ethnography. Hence according to Kozinets [1], netnography requires four essential steps, which can be reconducted to common ethnographic practice: 1. cultural entrée, 2. gathering and analysing data, 3. ensuring trustworthy interpretation, 4. conuducting ethical research [1] [2].

The first step requires to identify particular online forums, based upon the product or service to be investigated and the "specific research questions" [1]. Moreover, in order to gain rich data, the communities of interest should have exchanges on a focused segment or topic, high rate of posting, with detailed and descriptive messages [1]. The researchers can also choose if engaging in pure observations or in participant observations, joining the selected communities as active members [2]. Our study started in the opposite way, since we identified 1D games as an intriguing domain of investigation, after we read comments reported in online forums. Hence, we decided to follow six communities, including Facebook, through which it was possible to expand the basis of our participants, also addressing to people in our network. Hence the author of the games joined two communities of game design, which seem particularly focused on the topic, for participant observations, directly asking feedback on the games.

The second step required by netnography is to copy the conversation exchanges from the Internet and write annotations, regarding observations on a community and its members' interaction and meanings [1]. This step responds to transcriptions of verbal exchanges, in structured methods such as conversation analysis and grounded theory.

Regarding achieving trustworthy interpretations (step 3), netnography is based upon textual discourse, which on the Internet expresses the identity of their authors. In this sense, texts posted online probably represent a "controlled self-image" [1] of the community members, so that it may be more difficult to reason upon their motives. However, this is a risk also in ethnography, in which the focus is to study the behavior expressed by a group of people, and not to analyze the individuals expressing it [1]. Moreover, to provide a solid ground to netnographic data, Kozinets [1] and Seraj [2] suggest to combine other methods, such as in person interviews. In the present study, netnography is combined with user centred design, an approach to design that has become popular in different research communities. UCD is an iterative design method, which prescribes to involve users as informants, since the beginning of the design process, so to formulate design requirements based on the users' needs [16]. Hence a few prototypes are created and iteratively tested, so to fine tune design requirements and create better prototypes. This approach may include different methods, such as ethnography and interviews, supported by video and conversation analysis. In the present study, we engaged in designing 1D games, based upon the conversations we spotted on the net, the resulting games were then posted on the more focused communities, on Facebook and showed to a group of players from our University. We approached data collection with a "purposive sampling of material" [9], an approach to netnography used in marketing research, according to which noteworthy messages and conversation are selected and interpreted "in terms of a particular sample," [9], hence it is not necessary that the sample is representative of other populations. Our focus was on people interested in games and in exploring their essence, in terms of visualization and experience. Analysis of row data was inspired by grounded theory, as suggested by Seraj [2], so that we went through the conversations, copied on a separated file, and coded them. The aim of this analysis was to identify emerging themes, in relation to how people perceive 1D games and possible design inspirations. Recurring and interesting utterances were transcribed on post its, which were pasted on large sheets of papers, so to represent emergent themes into tangible and visual clusters, as it is common practice in design research [16]. Since the conversations we analyzed were concise and straight to the point, and we did not know exactly what to expect, this method revealed to be effective and well suited for the study. In this way, we were able to gather comments about the playability of our games, on their similarity with the original games, suggestions to improve them and create new ones.

Finally, as there is the possibility of psychological arm in inadequately reporting messages and utterances from online conversations, a researcher is supposed to act correctly with respect to privacy, and informed consent [1]. In respect of the rights of online communities members, it is recommended that researchers reveal their identities and purposes to the communities they follow. They should also be careful when reporting literal quotes, which online represent the individual identity of the members. Taking these recommendations into account, the communities that were followed through participant observations, were informed of this study and affiliation of the researchers. Moreover, in respect of the privacy of the communities, we decided not to report literal quotes from individual members. The data are reported in a descriptive way, in relation to emerging themes, which were identified from the observed conversations and used for the design of our games.

Application of netnography is supposed to require a long time commitment [1] [2], at this point our study is being running for five months, starting from May 2012, when we occasionally encountered conversations regarding 1D games. Most of our games were developed during the month of July, after the games were shared within the three communities, a systematic netnographic analysis was undertaken. Currently the study is still ongoing, focusing on a deeper reflection about non-genre of 1D games and the creation of new original games.

IV. GAMES

The games were created following a user centred and iterative development approach, inspired by agile development. The main goal with these games was to transpose classical 2D games in 1D, possibly in various different ways, and keep in contact with players, to validate the design as quickly as possible.

All the games are written in javascript, they all use HTML5 canvas for their graphics, and the code is willingly simple and portable, so that the games can run on most browsers and on most WIFI enabled mobile devices. Dropbox was used to distribute our 1D games. In line with the principles of agile development, we had daily code releases, in most of the cases few hours apart, to show and discuss online our games with the players and other developers.

The famous games we decided to flatten in 1D are: Tetris, pinball, Bloxorz [23], Sokoban [24] and Rogue [25]. For each of these games we designed different flattenings, so that different 1D games have been derived from the same original 2D game, and for each 1D game prototype many version have been developed, usually between 3 and 5. All our games are freely available [26].

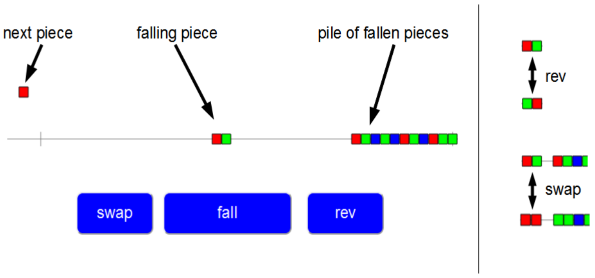


Figure 1 - 1Dminos, our transposition of Tetris onto 1D. As in the classic Tetris, a piece is falling (from left to right) and will join the pile of fallen pieces. The game also shows the next piece that will fall, to help the player better plan a strategy.

Figure 1 shows the game *1Dminos*, our 1D remake of Tetris. To decide how to reduce Tetris pieces and operations from 2D onto 1D, we followed a structuralist approach [3]. First we considered all Tetris pieces (as in figure 2) and the operations available to the user: horizontal translation, and clockwise and counter-clockwise 90 degrees rotation.

Moreover, according to our analysis, the key factors of 2D Tetris are:

- shape, i.e., distribution of squares in 2D to form connected figures,
- horizontal, vertical position and orientation of shapes.

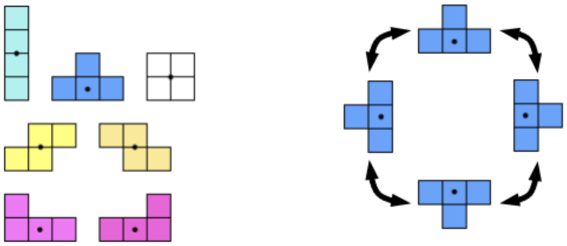


Figure 2 - On the left the basic 2D Tetris pieces. On the right the rotation group defined by the 'T' shape.

The goal of a game can then be expressed in relation to pieces and operations: in Tetris, connecting shapes in particular ways, a kind of *fitting*. Rotation and horizontal translation affect shapes, their fitting might change. The player has to find strategies to control the rotation and position of the one falling shape with respect to the shapes that already fell.

This structuralist analysis helps when defining a flattening of Tetris. We proceeded considering:

- how shapes would be in 1D
- which operations can be defined on the 1D pieces, that "remind of" the original operations; for example 1D operations that define symmetries similar to the rotation groups of the 2D game

Shape is a key factor in Tetris, we decided to represent them as *color patterns*, in 1D. A 1D shape is therefore a sequence of a few squares, each with a specific color; another way to consider a 1D shape is to see it as a string: the falling piece in figure 1, for example, is a red square followed by a green one, so it can be represented by "RedGreen" or simply "RG". As operations we choose **reverse** (reverse of "RGB" is "BGR") and **swap** of the right-most square of the falling piece and the top square on the top of the pile (see figure 1). Now we can rephrase Tetris' main goal in 1D, as string concatenation in particular ways, i.e., a kind of *color fitting*.

Figure 1 also shows the user interface of the 1Dminos game, composed of 3 buttons. The player can also press keys instead of use buttons, but the presence of buttons allows the game to be played on mobile devices that usually don't have a physical keyboard (and a software keyboard is not always available while running javascript applications).

V. CONVERSATION WITH DEVELOPERS AND PLAYERS

Four main themes emerged, from our analysis of online conversation about 1D games:

- non-genre
- essence of what is a 1D game
- experience
- design

These themes were identified analyzing general conversations, which focused around existing games like Super Mario Bros, but also newly developed 1D games. Afterwards the same themes were used as a framework to analyze the feedback we received for our games.

The first theme, non-genre, emerged from comments stating that a 1D game is impossible, this kind of comments inspired us to take the challenge to explore our own 1D games. These comments are of three kinds: either direct critics of the concept of 1D games, or ironic celebrations of a 1D game, which was found dull by the community, and its author, and finally through positive comments in relation to a new game that was found engaging to play. These last kind of comments, express surprise from the part of an individual, who started with a negative utterance about 1D games being nonsensical, but then change her mind, after playing a particular mono-dimensional game.

Other exchanges deal with the essence of what is a 1D game. In this regard two main points of view emerged, and we have coded them into two sub-themes: *geometry* and *view*. The comments clustered in the geometry sub-theme tend to claim that 1D games should respect the basic geometrical principle that a 1D space should be mapped on a computer screen, on a line of pixels. Hence the game should be represented only on one axis, vertical or horizontal, in general there should not be any change in the other axis, otherwise the game is 2D. On the other end, some comments reflect a different understanding of the essence of 1D games, more related to the visual rendering and movement of the character. Hence Super Mario Bros could be seen as a 1D game, as the character moves towards the same line, although he can jump and go underground. Moreover, some individuals commented that a 1D game could be the 1D view of a 2D world, some even recalled the book Flatland [11], a satirical novel that criticizes Victorian society mapping its hierarchical structure into a 2D world. Others accept that a designer can take the freedom to scale a line of pixels to a larger size, for usability purposes, such as facilitating view and understanding of the game. Furthermore, some comments are placed in between the geometry and view themes, a game was said to be 1D, in relation to the feeling provided by the game play, however, they felt the need to specify that the visual rendering of the game is, technically speaking, 2D. This metaphorical understanding of 1D, related to game play and the user experience, was applied to the design of our games.

Regarding the theme of experience, two main sub-themes emerged, which are called appreciation and critics, as the members of the communities we analyzed, expressed positive and negative judgments. On the positive side, these comments overlap with the absurd theme, as they express surprised, praising the game for being interesting and challenging, despite being 1D. Moreover, other comments praise the authors for being a creative and original thinker, in some cases the authors of such comments said that they would have liked to design a 1D game or that they have tried by themselves before. Similar comments were posted also regarding our games. These comments are particularly interesting as they show that 1D games can be engaging not only from a players' perspective, but also from a developer's one. However, there were several negative comments, stating that these games are “not so fun” or that the game play is vague and not very engaging. A sub-theme, called understanding, seemed to emerge within the category of experience, as several comments stated that these games are too difficult to understand, so that it takes a while before being able to enjoy the game or that it is too easy to die. The same comments appeared also in a positive light, as some individuals enjoyed the fact that they had to find out how to play by themselves, as it was part of the game play.

Finally some members of the forums were able to push further their critics or appreciation, providing interesting design guidelines to the designers. Most of these guidelines are aimed at supporting the player in understanding the game, so that tutorials or menus, explaining the function of the game features, were suggested. In other cases, given the difficulty of the game play, it was suggested to implement a button that allowed the player to start from the beginning of the level they were playing and not from the first level. Since these games occupy a little portion of the screen, some members suggested that they could be interesting apps for mobile platforms, such as smart phones and Nintendo DS. Other purists instead advised the authors, to make their metaphorical 1D game into a real 1D game, redesigning it with 1 line of pixels.

All these comments were carefully analyzed, as they provided inspiration in exploring what a 1D game can be, hence the same themes were applied as a framework to analyze the feedback we received for our games.

VI. A PROMISING NON-GENRE

We posted about our games in two online communities and we used Facebook to discuss them with friends, our students, colleagues and acquaintances. The same 4 themes (mentioned in V) also emerged in relation to our own 1D games.

Regarding the non-genre theme, our games elicited surprise. Some comments pointed at the fact that the 1D genre was unknown. Some people in our network, in the University and on Facebook commented that “only we” could have thought of them, in a teasing fashion. Other comments claimed that 1D games are interesting and that designing them does not seem an easy task.

Moreover, some community members expressed perplexity in relation to our definition of 1D; these comments are in between the non-genre and essence themes. From a purely geometrical point of view (and according to some posts) our games do not use only 1 dimension, i.e., a line of colored points. As visible for instance in figure 1, we have text and buttons in a 2D arrangement, and our points are scaled up to look like colored squares. According to this interpretation it is perhaps impossible to create an actual 1D game since screens are bi-dimensional. Interestingly, other community members answered for us that a 1D game could also be the 1D visualization of a 2D space or that designers are entitled to violate the geometrical definition. Regarding our games, people seem to be divides between a purist *geometrical* stand and a *visualization* one. Among the comments, we also got signaled games that could be classified as 1D, such as Gauge. Some comments even pushed the discussion further, questioning if text-based games could be considered 0D, since no pixels are used to visualize the state of the game. Finally, the domain of 1D games was said to be interesting and that it is surprising how many games could be flattened into a 1D visualization and still be playable. In a community in particular, we got in touch with a member, who after getting acquainted with the idea of 1D games, proposed to consider 1D a game he created years ago, during his studies. Other members claimed that after having realized of the existence of this genre, they will try to design new 1D games.

Under the theme of experience, we received many critical and positive comments. For instance, a few members of the same online community claimed that they tried all the games seem “nice”. However, they cannot see many possibilities for 1D games from a commercial point of view, since nowadays we are used to high quality graphics. Another comment was instead positive regarding the experience of playing a 1D game, but concluded that it seems to require too much effort to design a similar game, or that 1D might be too constrictive in terms of ideas and inspirations. Finally a few negative comments reported confusion and difficulty in the game play. Positive comments instead pointed out how inspiring it can be to design a 1D game, as the space limitation may push to think out-of-the-box. On the players’ side we received some enthusiastic reviews of our 1Dminos; the comments stated that our game looked more like a puzzle than the original Tetris, therefore it was more engaging. Another interesting post said that our games emphasize how the quality of the graphics is secondary to the game experience, supporting the idea that engaging games do not necessarily need glamorous graphics if the concept behind them is sound. Furthermore, some players enjoyed the difficulty of playing our 1D prototypes, saying that they liked to have to figure out what to do.

More directly usable and interesting comments were design suggestions, which we often took as requirements for the next versions of the prototypes. For instance, we were suggested to try to flatten a Pinball game and see what the outcome could be. Most comments focused instead in suggesting improvements; regarding our 1Dminos for example, different individuals provided interesting ideas, such as to make two blocks fall at the same time, to have a button to rotate colors in the falling piece, to allow for easier matches, or to alter the color palette because some colors were confusingly similar. Following a player suggestion, in one version of 1Dminos a *randomizing function* was added, with the effect of scrambling the colors of the falling piece. Later this randomization was criticized for making the game too easy, and we removed it from one line of prototypes. Finally it was mentioned that 1D games could work well for devices with small screens, such as tablets and smart phones. Hence, a few of our games have been rewritten with Android and IPhone in mind (e.g., changing the resolution required to see the whole game canvas, or rotating the graphics from horizontal to vertical), and tested by players in our network. In these cases, comments focused mainly on the usability, so that our 1Dminos for instance was said to work fine, but to be difficult, as there was some issues with the size of the buttons. It is clear that the user interface of our 1D games should work with gestures as well as buttons, to better integrate in tablets and mobiles.

VII. Discussion and conclusion

From the online discussions we had with players and developers, 2 views emerged circa the definition of 1D games: a purist *geometrical* interpretation and a *visualization* one. This lack of a clear and accepted definition might be an integral part of a non-genre, and perhaps the main reason why a new genre will never become established. But it might be also considered an opportunity to innovate by challenging apparently contradictory concepts, as in single-button games. For this reason we decided to develop an inclusive and socially acceptable definition for 1D games, and then use it to create games of this genre. We believe the reactions to our approach show that the game genres are not simply a way to classify existing games, but a tool to stimulate discussion and creative thinking.

Our revised definition of a 1D game is: a game that can be rendered and played, at least in principle, with a single line of pixels, either vertical or horizontal.

In our exploration of 1D games, we worked with 4 rendering styles for 1D games, from simple to complex, and also from literally 1 dimensional towards more artistically free renderings: *black and white*, *colors*, *rich* and *artistic* rendering. With black and white the state of the game can only be presented to the player by a line (vertical or horizontal) of black or white pixels, i.e., the result of rendering the game can be seen as a binary string. The color rendering is the same, but the pixels can be colored. The rich rendering is done using a finite set of icons, all of the same size; when rendering, a number of icons (eventually repeated) are arranged in a vertical (or horizontal) strip. The last rendering style is the artistic one, where freedom is allowed: icons, possibly animated, and of different size and shape can be used, still in a predominantly vertical or horizontal alignment. Special effects as rotation or wiggling in 2D are allowed, to draw the player's attention or highlight particular situations. The rendering should still result in a linear arrangement, but the line could be twisted in 2D, e.g., icons could be drawn along a sinusoidal line, instead of a straight line as in rich rendering.

According to our definition, even when a 1D game is developed with an artistic rendering, it must be possible to play the game with only a black and white (or color) rendering; it might be necessary to use 2 or more pixels in the color rendering, to properly represent an artistic one, but if it not possible at all or if it makes the game so hard to play to be meaningless, then the game should not be considered 1D.

So far we have flattened adventure and puzzle games, and 2 reusable techniques emerged to map 2D game in 1D, i.e., spatial and temporal mapping. A 2D space can be sliced in horizontal strips, with a single strip perceived while playing (i.e., time acts as extra spatial dimension) or all strips can be visible, in a long line, on the screen (i.e., more space to cope with the loss in dimensionality). The player can stay within a strip by moving left/right, and moving up/down she can jump from strip to strip.

Another general technique, that worked for us, is to perform a structural analysis before designing a flattening of a 2D game. The results of the analysis suggest, but do not dictate, what the flattening should focus on, still leaving plenty of artistic freedom to the 1D game designer.

In conclusion, our artistic and social exploration of the non-genre of 1D games provided us with a working definition of what a 1D game should be. We have also shown that different genres can be flattened (re-conceptualized so to be visualized and played in 1D), and that some games leave room for creative level design (e.g., our Sokoban in 1D).

A typical reaction in developers, who play our 1D games is starting to suggest improvements or new directions to explore. Hence it seems that 1D games, even when fully functional and with refined graphics, are perceived as being at a prototypical stage, usually associated with low-fidelity prototypes [16]. Therefore, these games might be useful as designerly tools to think out-of-the-box when exploring new types of games.

Results from this study suggested also new research directions, such as to study how the presence of constraints forces designers and players to analyze the identity of a game, in relation to "central" elements, operations and dynamics as game identity. Moreover, design of minimalistic games could be used as a challenging task, to support students in learning of game design and development, focusing on the aspects characterizing the essence of a game.

Following an approach similar to [12], we are currently planning to involve a group of students in the co-design a 1D Super Mario game, as part of a participatory game design event; the flattening of a platform-type game could offer the opportunity to reflect in depth upon typical genre mechanics, game play and visualizations. In this sense, we believe that challenging a non-genre (or an emergent game genre) in participatory design workshops can become a viable tool when teaching game design and game programming.

REFERENCES

[1] Kozinets, Robert V. "*The field behind the Screen: Using Netnography for Marketing Research Online Communities*." Journal of Marketing Research, vol. 39, no. 1, pp. 61-72, February 2002.

[2] Seraj, Mina, "*We Create, We Connect, We Respect, Therefore We Are: Intellectual, Social, and Cultural Value in Online Communities*." Journal of Interactive Marketing, vol. 26, pp. 209-222, 2012.

[3] Piaget, Jean, "*Structuralism*." Harper & Row, 1971

[4] Gamasutra's article about one button games. www.gamasutra.com/view/feature/2316/one_button_games.php last visited 20/10/2012.

[5] Hamasaki Masahiro, Matsuo, Yutaka, Nishimura, Takuichi, Takeda, Hideaki. "*Ontology Extraction using Social Network.*" SWeCKa 2007 Workshop on Semantic Web for Collaborative Knowledge Acquisition, The Twentieth International Conference on Artificial Intelligence, pp. 32-38, Hyderabad, India, 6-8 January 2007.

[6] Mika, Peter. "*Ontologies are us: A unified model of social networks and semantics*." Proceedings of ISWC2005, pp. 522-536, 2005.

[7] Steinkuehler, Constance A. "*Why Game (Culture) Studies Now?*" Games and Culture, vol. 1, no. 1, pp. 97-102, Sage Publications, January 2006

[8] Schell, Jesse. "*Art of Game Design. A book of lenses.*" Morgan Kaufmann Publishers, 2008.

[9] Apperly, Thomas H. "*Genre and Game studies: Toward a critical approach to video game genres.*" Simulation and Gaming, vol. 37, no. 1, pp. 6-23, Sage Publications, March 2006.

[10] Arsenault, Dominic. "*Video Game Genre, Evolution and Innovation.*" Eludamos. Journal for Computer Game Culture, vol. 3, no. 2, pp. 149-176, 2009.

[11] Abbott, Edwin. *"Flatland, a Romance of Many Dimensions."* 1884, freely available at: http://en.wikisource.org/wiki/Flatland_(first_edition) last visited 20/10/2012.

[12] Valente Andrea, Marchetti Emanuela, "*Programming Turing Machines as a Game for Technology Sense-Making*." Proceedings of the 11th IEEE International Conference on Advanced Learning Technologies (ICALT 2011), pp. 428-430, 2011.

[13] GameDev.Net, topic "exploration-of-1d-games" http://www.gamedev.net last visited 20/10/2012 last visited 20/10/2012.

[14] Gameprog.it, topic "Esplorazione di giochi ad 1 dimensione" http://gameprog.it last visited 20/10/2012.

[15] Brown, Stephen, Kozinets, Robert V. and Sherry, John F. Jr., "*Teaching Old Brands New Tricks: Retro branding and the Revival of Brand Meaning*." Journal of Marketing, vol. 67, no. 3, pp. 19-23, July 2003.

[16] Sharp Helen, Rogers Yvonne, and Preece Jenny. "*Interaction Design: Beyond Human-Computer Interaction*." Wiley, 2 edition, March 2007.

[17] Oneswitch: http://www.oneswitch.org.uk/ last visited 20/10/2012.

[18] Tetris: http://www.tetris.com/ last visited 20/10/2012

[19] Tetris 1D: http://www.tetris1d.org/ last visited 20/10/2012.

[20] Wolfstein 1D: http://www.kongregate.com/games/EvilDog/z-rox last visited 20/10/2012.

[21] Line: http://forums.tigsource.com/index.php?topic=17734.0 last visited 20/10/2012.

[22] Gauge: http://www.148apps.com/reviews/gauge-review/ last visited 20/10/2012

[23] Bloxorz: http://www.bloxorzgame.com/ last visited 20/10/2012.

[24] Sokoban: http://www.sokoban.jp/ last visited 20/10/2012

[25] Gamasutra's article. "The History of Rogue: Have @ You, You Deadly Zs" at http://www.gamasutra.com last visited 20/10/2012.

[26] 1D games https://dl.dropbox.com/u/1518199/1D%20games/index.html last visited 20/10/2012.

Choosing a BPMN 2.0 Compatible Upper Ontology

Ludmila Penicina
Department of Systems Theory and Design
Riga Technical University
Riga, Latvia
ludmila.penicina@rtu.lv

***Abstract* — Nowadays, linkage of BPMN 2.0 business process models with ontologies to achieve consistency and semantic compatibility is still a challenge. This paper addresses a question of finding BPMN 2.0 meta-model compatible upper ontology for the analysis of the completeness of BPMN 2.0 model. Upper ontologies are meta-structures for domain ontologies and based on the correspondence between BPMN 2.0 meta-model and upper ontology a link between BPMN 2.0 models and domain ontology can be provided. A comparison of 5 existing upper ontologies showed that the Bunge-Wand-Weber ontology is the most compatible with the BPMN 2.0 meta-model.**

Keywords-BPMN 2.0; upper ontologies; BWW ontology.

I. INTRODUCTION

Business processes are one of the most valuable assets of any organization. Business processes require applying existing business process knowledge. According to Grant [1], knowledge is the most strategically important resource of the firm and primary role of any organization is application of knowledge in its everyday activities. However, the application of existing knowledge has always been a sophisticated task. A holistic view of end-to-end business process knowledge is required because knowledge of cross-functional processes is distributed across departments, documents, regulations and applications. Different stand-alone applications and documents contain explicit process knowledge and tacit knowledge is "stored" in heads of employees. Business process knowledge must be reusable and applicable across many business processes.

According to Xiao et al. [2], ontologies use a formal way to represent knowledge as a set of concepts and relationships among the concepts. As described by Xiao et al. [2] ontologies are widely used for knowledge representation and sharing. There exist many definitions about what ontology is; however, in the scope of this paper, ontology is a formal specification of a shared conceptualization, as described by Gomez-Perez et al. [3]. According to Gomez-Perez et al. [3], there exist different types of ontologies identified in the literature based on their conceptualization.

Ontologies exist at several levels of abstraction. According to Semy et al. [4], upper ontology is defined as a high-level, domain-independent ontology from which more domain-specific ontologies may be derived. Domain ontologies are reusable in a given specific domain (e.g., medical, law, enterprise, engineering, etc.) providing vocabularies about the activities taking place in that domain and their relationships, as described by Gomez-Perez et al. [3]. As described by Mascardi et al. [5] upper ontology contains general concepts that are the same across all domains. Thus, upper ontology can be used as a meta-structure for defining domain ontologies.

Motivation for this research is described as follows. BPMN 2.0 (or Business Process Model and Notation 2.0 [6]) is the de-facto standard for representing in a very expressive graphical way the processes occurring in virtually every kind of organization, as described by Chinosi et al. [7]. However, the goal of any modelling activity is a complete and accurate understanding of the real-world domain. Hence business process modelling requires a background knowledge e.g., domain ontology that complements behavioural aspect of an information system. Providing linkage between BPMN 2.0 and domain ontology will facilitate consistency between information system models and domain requirements. Nowadays, linkage between BPMN 2.0 and domain ontology is still a challenge. However the new BPMN 2.0 specification [6] allows integration with third party components using XML-based representation languages (e.g., OWL, RDF) [8]. This new BPMN 2.0 "plug-and-play" feature opens the potential for linking domain ontologies represented as XML structures with BPMN 2.0 models. But, firstly, it is necessary to provide consistency and semantic compatibility between BPMN 2.0 and ontology at the meta-level, namely, linking BPMN 2.0 meta-model with upper ontology that is used as a basis for deriving domain ontology.

Figure 1 depicts the idea of linking BPMN 2.0 process models with domain ontology that is based on compatibility between BPMN 2.0 meta-model and upper ontology that is used as a meta-structure for defining domain ontology. In order to implement the proposed approach for linking BPMN 2.0 process models with domain ontology it is necessary to choose an upper ontology that is compatible with BPMN 2.0 meta-model. The chosen BPMN 2.0 compatible upper ontology will be used as a meta-structure for deriving domain ontology that will be linked with BPMN 2.0 process models in order to sustain the consistency and semantic compatibility between business process models and ontology. Hence, the goal of this research is to choose a BPMN 2.0 meta-model compatible upper ontology by evaluating the most popular upper ontologies described in the literature.

Linking BPMN 2.0 process models with domain ontology will contribute to:

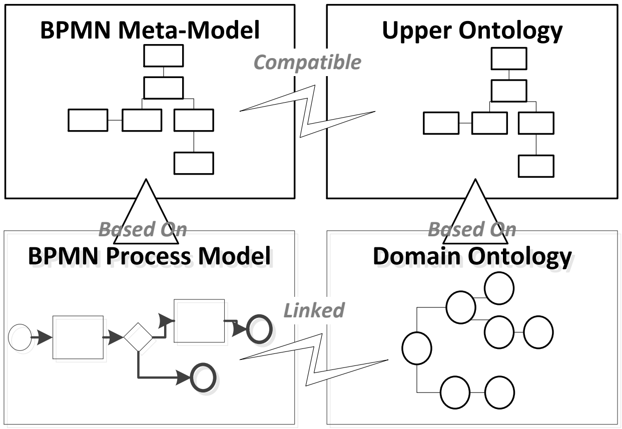


Figure 1. Proposed approach for BPMN and Domain Ontology compatibility

- Consistency between process models and domain ontology - as a result domain ontology and business process models can be validated against each other.
- Analysis of the completeness of BPMN 2.0 models.
- Monitoring of changes introduced to process models or domain ontology and the effects of these changes.
- Establish a semantic consistency and interoperability between process models and domain ontology.
- Gaining better understanding of processes and reasoning capabilities (as ontologies play one of the most important roles in semantic web).

The paper is structured as follows. Section II presents related works. Section III describes the procedure of comparing existing upper ontologies. Section IV describes BPMN 2.0. Section V describes candidate upper ontologies. Section VI presents comparison of upper ontologies. Section VII presents conclusion and future works.

II. Related works

Semy et al. [4] examine standard upper ontologies and assess their applicability for a U.S. Government or U.S. Military domain. In this research authors evaluate the state of the art and applicability of upper ontologies using consideration of the ontology purpose, ontological content decisions, licensing restrictions, structural differences, and maturity [4]. Mascardi et al. [5] are finding correspondences between entities belonging to different ontologies describing a set of algorithms that exploit upper ontologies. The analysis presented by Mascardi et al. [5] shows under which circumstances the exploitation of upper ontologies gives significant advantages with respect to traditional approaches that do not use them.

Mascardi et al. [9] are analysing 7 upper ontologies namely BFO, Cyc, DOLCE, GFO, PROTON, Sowa's ontology and SUMO, according to a set of standard software engineering criteria. Rosemann et al. [10] address the issue of modelling information systems by presenting a meta model of the BWW ontology using a meta language that is familiar to information systems professionals facilitating the application of the BWW theory to other modelling techniques that have similar meta models defined.

Francescomarino et al. [11] propose an automated technique to support the business designer both in domain ontology creation/extension and in the semantic annotation of process models expressed in BPMN 2.0. Natschläger et al. [12] present BPMN 2.0 ontology. The defined BPMN 2.0 ontology can be used as a knowledge base for learning BPMN, as a syntax checker to validate separate BPMN 2.0 models and to identify contradictions in specification.

sBPM (or Semantic Business Process Management) was introduced to solve the problem of inconsistency between various process models in a domain using semantic annotating of process models with concepts from ontology. That facilitates reusing of process model parts and unambiguity of the domain concepts. Francescomarino et al. [13] show how semantic web techniques can be applied to formalize, verify and integrate the domain knowledge in BPMN 1.1 diagrams. Wang et al. [14] propose the approach of ontological descriptions of semantics of supply chain processes. Nicola et al. [15] propose the approach of representing a BPMN diagram by using ontology based formalism.

The SUPER EU project (or Semantics Utilised for Process Management within and between Enterprises) created the technological framework constituting BPM enriched with machine readable semantics by employing Semantic Web technology [16].

This research is based on the results of related works and related works have encouraged this research and showed that linkage between BPMN 2.0 and ontologies is an important issue to facilitate information system modelling consistent with the real-world domain. However, to the best of author's knowledge there is no research that compares existing upper ontologies for the compatibility with BPMN 2.0 meta-model.

III. Procedure for Comparison of Upper Ontologies

To compare existing upper ontologies and choose upper ontology compatible with BPMN 2.0, the following steps were carried out:

- During the mapping of BPMN 2.0 elements to elements of upper ontologies the correspondence link *MAPS* introduced by Etien et al. [17] has been applied. Etien et al. [17] define two correspondence links - *MAPS* and *REPRESENTS*, *MAPS* link was selected because it is defined as following: "one class X maps another class Y if there exist an isomorphism between the set of properties of X. In other terms, each property of X corresponds to one of Y even (domains being eventually different)." *REPRESENTS* link is defined as an association when two constructs of different nature can be linked. In this research *MAPS* link is applied to obtain an upper ontology that is compatible with BPMN 2.0.
- Meta-model for BPMN 2.0 analytical level or level 2, as defined by Silver [8], is created using UML

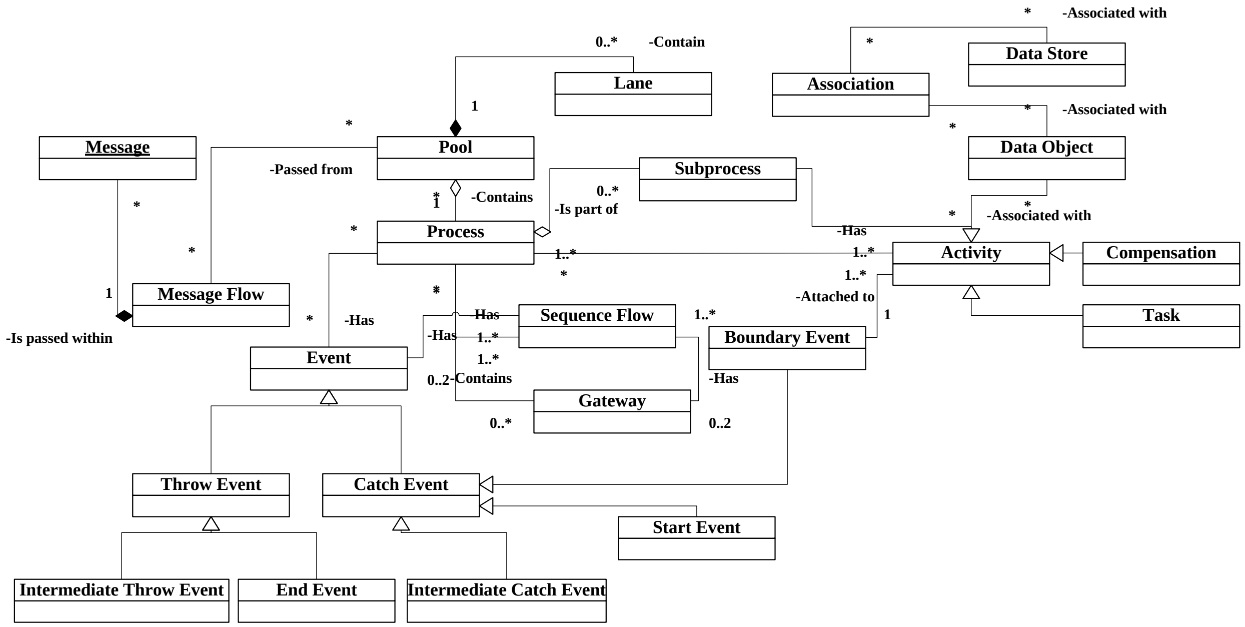


Figure 2. Simplified BPMN 2.0 meta-model

class diagram. BPMN 2.0 meta-model is built to use it as a base for comparison of upper ontologies.

- Candidate upper ontologies are chosen based on whether upper ontology is free to use and whether upper ontology is still being maintained.
- Meta-models for chosen upper ontologies are created using UML class diagrams in order to explicitly compare them with created BPMN 2.0 meta-model.
- A table showing compatibility between chosen upper ontologies and BPMN 2.0 meta-model is presented.
- The upper ontology the meta-model of which supports all main BPMN 2.0 elements is chosen.

IV. BPMN 2.0

Business Process Model and Notation (BPMN 2.0) [6] is the de-facto standard for representing in a very expressive graphical way the business processes occurring in virtually every kind of organization, as described by Chinosi et al. [7].

BPMN 2.0 core elements can be grouped in the following groups of elements [6]:

1. *Swimlanes* – pools and lanes allow grouping BPMN 2.0 model elements according to participants of the process, information systems, organization structure, etc.

2. *Flows* – message and sequence flows between BPMN 2.0 elements.

3. *Data* – data in BPMN 2.0 is represented through data objects and data stores.

4. *Flow objects* – events, activities, and gateways are main BPMN 2.0 flow objects.

According to Silver [8] BPMN 2.0 allows integrating business process model with third party components (e.g., database, web services etc.). BPMN 2.0 defines formal mechanisms to link business process data with a process model using XML Schema Definition language (XSD) or Web Service Definition language (WSDL), as described by Silver [8]. BPMN 2.0 allows linking, sharing and re-using existing business process data across BPMN 2.0 models. This BPMN 2.0 feature can be extended to provide not only process data linkage with a BPMN 2.0 model, but also linking a domain ontology with a BPMN 2.0 model to enable semantic compatibility and consistency between process models and domain ontology. To achieve this linkage it is necessary to represent domain ontology as a BPMN 2.0 compatible structure in order to be able to associate it with related BPMN 2.0 model elements.

Based on a method described in Section II, in this section, a simplified BPMN 2.0 level 2 (as defined by Silver [8]) meta-model is presented in Figure 2.

V. UPPER ONTOLOGIES

The concepts expressed in upper ontologies are basic and universal concepts and are used to ensure generality and, for a wide range of domains, represent common sense.

Two main parameters were established for choosing candidate upper ontologies - openness of upper ontology (meaning whether ontology is free available) and continuing development of upper ontology (whether upper ontology is still maintained). Based on these criteria the following 5 upper ontologies were chosen for the assessment:

- Basic Formal Ontology (BFO) [18]
- Sowa’s Top level ontology, as described by Sowa [19]
- Bunge-Wand-Weber ontology (BWW), as described by Allen et al. [20]
- Suggested Upper Merged Ontology (SUMO) [21]
- Cyc’s Upper Ontology [22].

A. Basic Formal Ontology (BFO)

The BFO project was initiated in 2002 and is maintained to this day [18]. BFO is focused on the task of providing a

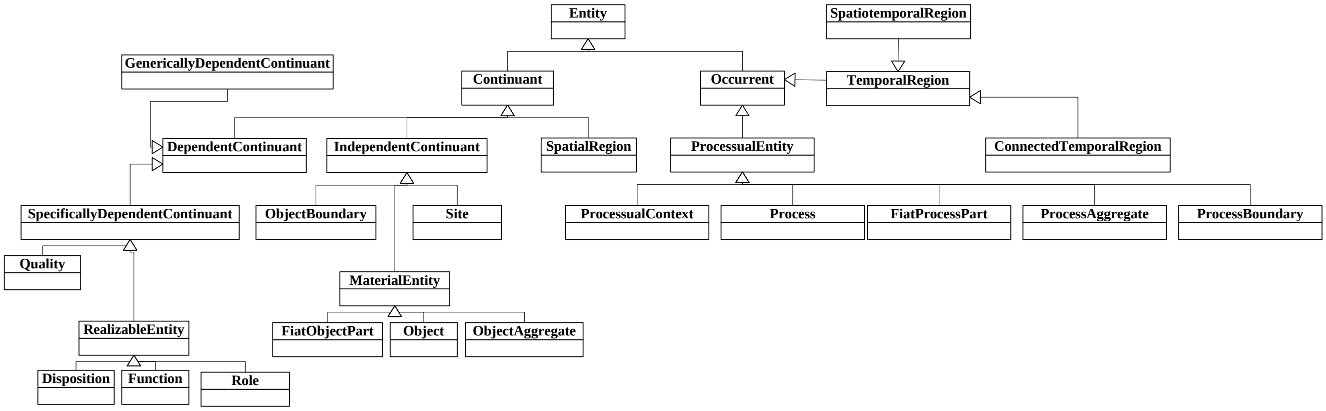


Figure 3. BFO Meta-model.

genuine upper ontology which can be used in support of domain ontologies [18]. BFO consists in a series of sub-ontologies but in this research the upper ontology of BFO is addressed. Figure 3 represents BFO upper ontology meta-model created using UML class diagram.

At the core of BFO consists of is *Entity*. *Entities* are either *continuants* or *occurrents*. A *continuant* is something existing at an instant in time, an *occurrent* is something that has temporal parts. A *spatial region* is three-dimensional. A *processual entity* is something that occurs or happens.

B. *Sowa's Top Level Ontology*

Sowa's top level ontology includes the basic categories and distinctions that have been derived from a variety of sources in logic, linguistics, philosophy and artificial intelligence, as described by Gomez-Perez et al. [3]. Sowa's top-level ontology includes 12 central categories which are generated from primitive categories. Figure 4 represents Sowa's top level ontology meta-model created using UML class diagram.

C. *BWW Ontology*

Rosemann et al. [10] describe BWW ontology as useful for description of information systems. As described by Davies et al. [23] an ontology presented by Bunge has been extended and applied to the modelling of information systems. Figure 5 represents BWW Ontology meta-model created using UML class diagram.

D. *SUMO Ontology*

The Suggested Upper Merged Ontology (SUMO) [21] is an upper level ontology that has been proposed as a starter document for The Standard Upper Ontology Working Group, an IEEE-sanctioned working group [24] of collaborators from the fields of engineering, philosophy, and information science, as described by Niles et al. [25]. Figure 6 represents SUMO meta-model created using UML class diagram.

E. *Cyc's Upper Ontology*

Cyc's Upper Ontology is contained in the Cyc Knowledge Base, which holds huge amount of common sense knowledge, as described by Gomez-Perez et al. [3]. According to Mascardi et al. [9], the Cyc Knowledge Base is a formalized representation of facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life. Figure 7 represents Cyc's Upper Ontology meta-model created using UML class diagram.

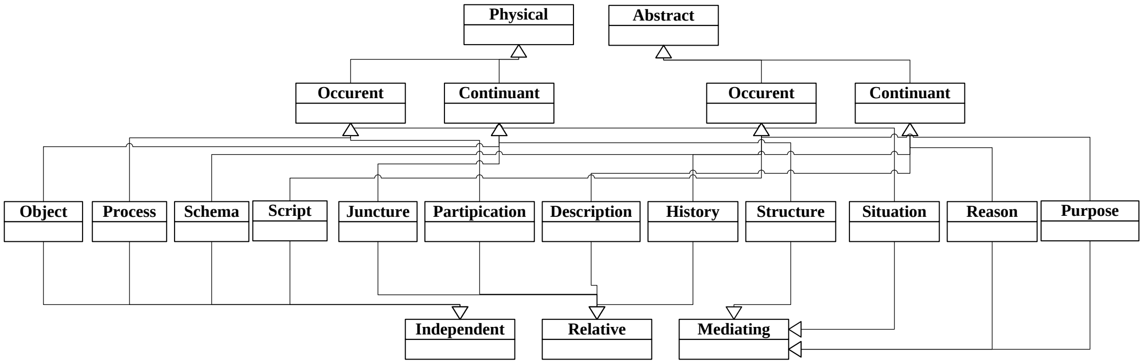


Figure 4. Sowa's Ontology meta-model.

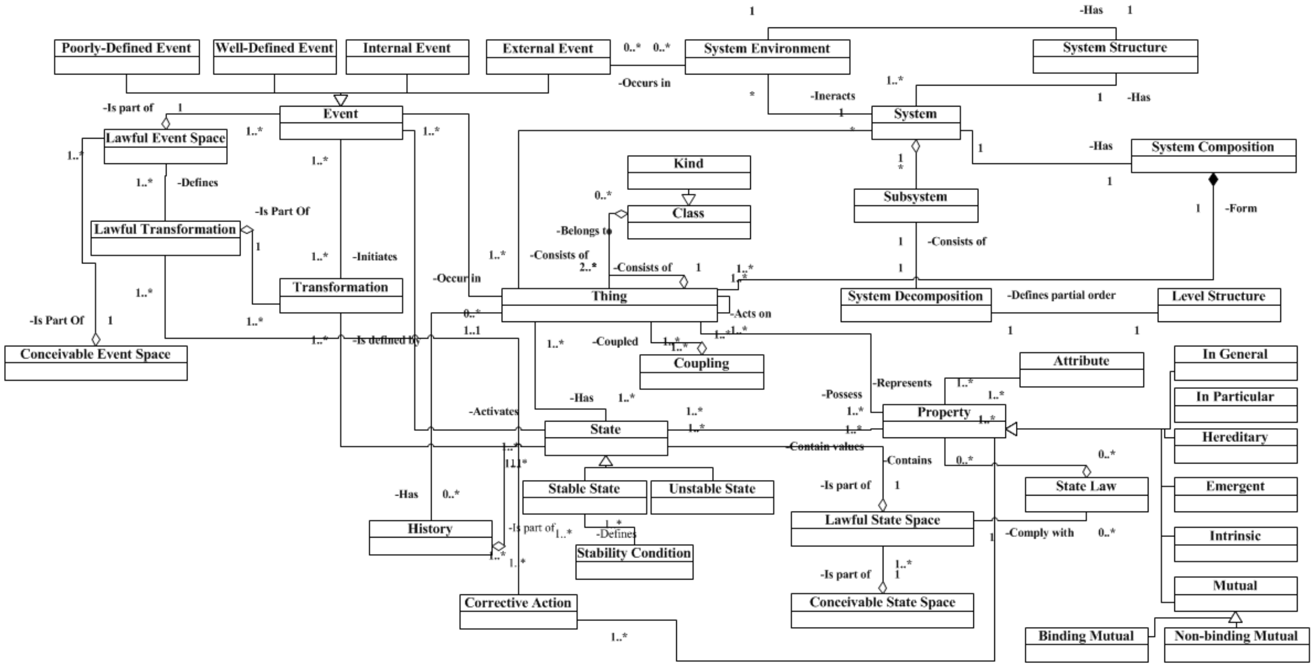


Figure 5. BWW Ontology.

F. Requirements for Upper Ontologies

The requirements that should be fulfilled by an upper ontology that is compatible with BPMN 2.0 can be summarized as follows:

- Ability to represent the notion of a process.
- Ability to represent the notion of an atomic activity.
- Ability to represent the performer of activities and processes.
- Ability to represent artifacts processed.
- Ability to represent internal and external events occurring in the process.
- Ability to represent the sequence flow and logic of activities.
- Ability to represent message flows between various processes.

VI. COMPARISON OF UPPER ONTOLOGIES FOR COMPATIBILITY WITH BPMN 2.0

This section presents a comparison of upper ontologies for their compatibility with BPMN 2.0 meta-model. The analysis is presented in a Table I showing which elements of upper ontologies described in Section V correspond to BPMN 2.0 meta-model elements.

From Table I, the following can be concluded:

- BWW upper ontology supports most of the presented BPMN 2.0 elements.
- SUMO upper ontology is not compatible with BPMN 2.0 meta-model because one of the most important notions of process modelling - the Event notion - is not supported by SUMO upper ontology.
- BPMN 2.0 meta-model element Gateway is supported only by BWW upper ontology with its State Law element which restricts the values of the properties of a thing to a subset that is deemed

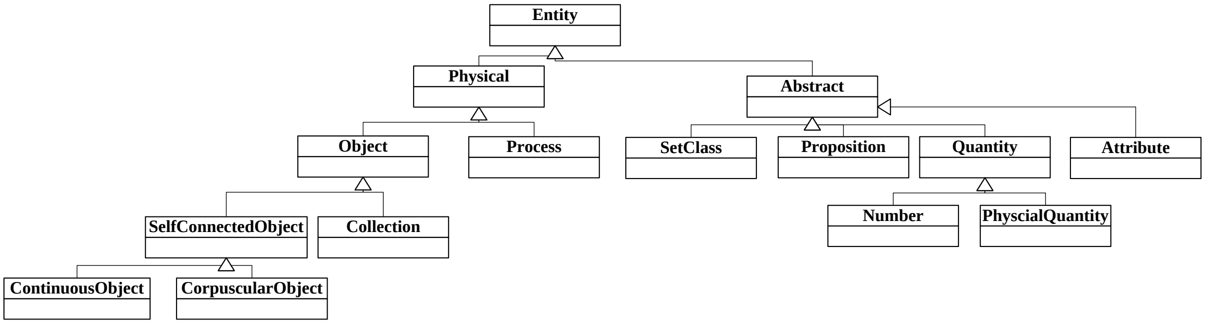


Figure 6. SUMO Ontology.

lawful, as described by Rosemann et al. [26]. According to Silver [8], a Gateway element in BPMN 2.0 has conditions attribute defined controlling the flow of the process.

- BPMN 2.0 elements *Process, Subprocess, Activity, Call Activity*, *Loop Activity, Compensation, and Task* are supported by all reviewed upper ontologies.
- For particular BPMN 2.0 elements relationship to upper ontology elements is 1 to many - one BPMN 2.0 element can be associated with several upper ontology elements, e.g., BPMN 2.0 *Event* element can be mapped to BFO ontology elements *ProcessBoundary*, *TemporalRegion* and *ConnectedTemporalRegion*.
- BPMN 2.0 element *Loop Activity* is supported only by BFO ontology, which defines *ProcessualContext* element as "(..) consisting of a characteristic spatial shape inhering in some arrangement of other occurrent entities" [18].
- Some elements of upper ontologies are not represented in BPMN 2.0 meta-model.

Based on the comparison presented in Table I, BWW upper ontology is concluded to be the upper ontology supporting BPMN 2.0 meta-model at most.

VII. CONCLUSIONS AND FUTURE WORK

In order to link BPMN 2.0 models with domain ontology to provide consistency it is necessary to ensure compatibility between BPMN 2.0 meta-model and upper ontology that domain ontology is derived from. The paper presented a comparison of existing upper ontologies in order to choose BPMN 2.0 meta-model compatible upper ontology. As a result BWW upper ontology was concluded to be BPMN 2.0 compatible upper ontology supporting most of the basic BPMN 2.0 elements.

By linking BPMN 2.0 process models with domain ontology, the enterprise may achieve the consistency and semantic compatibility between process models and existing ontology. This will help business process modellers across organization to identify, share and reuse existing knowledge explicitly and conduct qualitative process analysis to make decisions concerning new process development. With ontologies supplying the context of process, this contextual information can be exploited to perform semantic analyses of the process.

Practical implications of the presented research can be summarized as follows. Connecting BPMN 2.0 models with ontology will contribute to more precise requirements definition and possibly reducing the time of development and implementation of changes. The BWW representation might be used to analyse the completeness of BPMN 2.0 for software requirements.

However, the BWW ontology does not fully comply with the BPMN 2.0 meta-model. In the future research extensions of BWW and BPMN 2.0 will be addressed to tackle this issue. The author does not propose a new custom upper ontology, because the reviewed upper ontologies are largely recognized, especially BWW ontology in the IS modelling domain, as described by Rosemann et al. [10].

The conducted research has mostly been of a purely theoretical nature. Technical linkage and consistency checking between BPMN 2.0 and upper ontology is a concern of further research. The future work will also address building of algorithms for evaluating the completeness of business process models based on the metrics developed by Etien and Rolland [17]. The future research includes development of the prototype of the proposed solution using existing Open Source solutions, as well as validation of the implemented prototype in the real information systems projects.

In the future work implementation of the proposed approach will be addressed by using BPMN 2.0 existing capability to connect to third party components (e.g., to connect to domain ontology represented as a XML based structure - OWL). The paper has some limitations, namely, no verification or test for validity of this mapping is considered, which will be addressed in the future research.

ACKNOWLEDGMENT

The author acknowledges Professor Dr.sc.ing. Marite Kirikova for valuable comments on the draft of the paper.

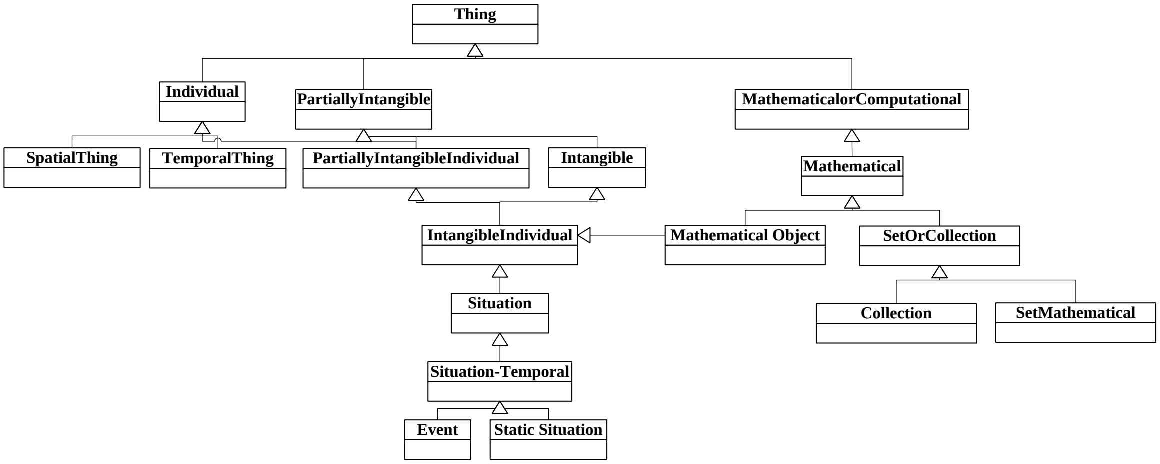


Figure 7. Cyc's Upper Ontology meta-model [5].

REFERENCES

[1] R. M. Grant, “Toward a Knowledge-Based Theory of the Firm,” *Management*, vol. 17, pp. 109–122, 1996.

[2] H. Xiao, B. Upadhyaya, F. Khomh, Y. Zou, J. Ng, and A. Lau, “An Automatic Approach for Extracting Process Knowledge from the Web,” in *2011 IEEE International Conference on Web Services*, 2011, pp. 315–322.

[3] A. Gomez-Perez, M. Fernandez-Lopez, and O. Corcho, *Ontological Engineering*. Springer-Verlag London Limited, 2004.

[4] S. K. Semy, M. K. Pulvermacher, and L. J. Obrst, “Toward the Use of an Upper Ontology for U . S . Government and U . S . Military Domains : An Evaluation,” 2004.

[5] V. Mascardi, A. Locoro, and P. Rosso, “Automatic Ontology Matching via Upper Ontologies: A Systematic Evaluation,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 5, pp. 609–623, May 2010.

[6] OMG, “BPMN,” *Business Process Model and Notation*, 2011. [Online]. Available: www.bpmn.org. [Accessed: 30-Nov-2012].

[7] M. Chinosi and A. Trombetta, “BPMN: An introduction to the standard,” *Computer Standards & Interfaces*, vol. 34, no. 1, pp. 124–134, Jan. 2012.

[8] B. Silver, *BPMN Method and Style with Implementer’s Guide*, 2nd Editio. Cody-Cassidy Press, 2011.

[9] V. Mascardi, V. Cordì, and P. Rosso, “A Comparison of Upper Ontologies (Technical Report DISI-TR-06-21),” 2008.

[10] M. Rosemann and P. Green, “Developing a meta model for the Bunge–Wand–Weber ontological constructs,” *Information Systems*, vol. 27, no. 2, pp. 75–91, Apr. 2002.

[11] C. Di Francescomarino and P. Tonella, “Supporting Ontology-Based Semantic Annotation of Business Processes with Automated Suggestions,” pp. 211–223, 2009.

[12] Ch. Natschläger, “Towards a BPMN 2.0 Ontology,” *Lecture Notes in Business Information Processing*, vol. 95, no. Business Process Model and Notation, pp. 1–15, 2011.

[13] C. Di Francescomarino, C. Ghidini, M. Rospocher, L. Serafini, and P. Tonella, “Semantically-Aided Business Process Modeling,” *8th International Semantic Web Conference (ISWC2009)*, pp. 114–129, 2009.

[14] X. Wang, N. Li, H. Cai, and B. Xu, “An Ontological Approach for Semantic Annotation of Supply Chain Process Models,” *On the Move to Meaningful Internet Systems: OTM 2010 Lecture Notes in Computer Science*, vol. 6426, pp. 540–554, 2010.

[15] A. De Nicola, T. Di Mascio, M. Lezoche, F. Taglino, and I. Iasi, “Semantic Lifting of Business Process Models,” *12th Enterprise Distributed Object Computing Conference Workshops*, 2008.

[16] “SUPER project.” [Online]. Available: http://www.ip-super.org/. [Accessed: 18-Dec-2012].

[17] A. Etien and C. Rolland, “Measuring the fitness relationship,” *Requirements Engineering*, vol. 10, no. 3, pp. 184–197, Aug. 2005.

[18] IFOMIS, “Basic Formal Ontology,” 2012. [Online]. Available: http://www.ifomis.org/bfo. [Accessed: 20-Oct-2012].

[19] J. F. Sowa, “Sowa’s Top Level Ontology,” *Top-Level Categories*. [Online]. Available: http://www.jfsowa.com/ontology/toplevel.htm.

[20] G. Allen and S. March, “A Critical Assessment of the Bunge-Wand-Weber Ontology for Conceptual Modeling,” *16th Annual Workshop on Information Technolgies & Systems (WITS) Paper*, 2007.

[21] “Suggested Upper Merged Ontology (SUMO).” [Online]. Available: http://www.ontologyportal.org/. [Accessed: 20-Oct-2012].

[22] “OpenCyc Upper Ontology.” [Online]. Available: http://www.cyc.com/cycdoc/upperont-diagram.html. [Accessed: 20-Oct-2012].

[23] I. Davies, P. Green, S. Milton, and M. Rosemann, “Using Meta Models for the Comparison of Ontologies,” 2003.

[24] “Standard Upper Ontology Working Group (SUO WG) - Home Page.” [Online]. Available: http://suo.ieee.org/. [Accessed: 20-Oct-2012].

[25] I. Niles and A. Pease, “Towards a standard upper ontology,” *Proceedings of the international conference on Formal Ontology in Information Systems - FOIS ’01*, vol. 2001, pp. 2–9, 2001.

[26] M. Rosemann and J. Recker, “A study of the evolution of the representational capabilities of process modeling grammars,” *Advanced Information Systems Engineering*, pp. 447–461, 2006.

TABLE I. COMPARISON OF UPPER ONTOLOGIES

BPMN 2.0 Element	BFO Element	Sowa's Element	BWW Element	SUMO Element	Cyc's Element
Process	ProcessualEntity	Process	Transformation	Process	TemporalThing
Subprocess	Process	Process	Transformation	Process	TemporalThing
Activity	Process Function	Process	Transformation	Process	TemporalThing
Compensation	Process	Process	Transformation Well-Defined event State	Process	TemporalThing
Task	Function Disposition	-	Transformation	Process	TemporalThing
Event	ProcessBoundary TemporalRegion ConnectedTemporalRegion	Situation Reason	Event	-	Event
Throw Event	ProcessBoundary	Situation Reason	Internal Event Poorly-Defined Event	-	Event
Catch Event	ProcessBoundary	Situation Reason	External Event Poorly-Defined Event	-	Event
Intermediate Throw Event	ProcessBoundary	-	Internal Event Poorly-Defined Event	-	Event
End Event	ProcessBoundary	Reason	Event Well-Defined Event State	-	Event
Start Event	ProcessBoundary	Purpose Reason	Event Poorly-Defined Event	-	Event
Intermediate Catch Event	ProcessBoundary	-	External Event Poorly-Defined Event	-	Event
Boundary Event	ProcessBoundary	-	Event Poorly-Defined Event	-	Event
Message	-	Reason	External Event	-	-
Message Flow	-	Juncture	Coupling Acts on	-	
Pool	Role Site	Object	Thing Kind Class System	Object Collection	Individual PartiallyIntangible
Lane	Role Site	Object	Subsystem Kind Class	Object Collection	Individual PartiallyIntangible
Association	-	Juncture	-	-	-
Data Store	ObjectAggregate	Object	Thing	Object	Thing
Data Object	MaterialEntity Object FiatObjectPart	Object	Thing	Object	Thing
Sequence Flow	-	Juncture	Lawful transformation	-	-
Gateway	-	-	State Law	-	-

Publishing Multidimensional Statistical Linked Data

Airton Zancanaro, Leandro Dal Pizzol, Rafael de Moura Speroni, José Leomar Todesco, Fernando O. Gauthier
Department of knowledge engineering – Universidade Federal de Santa Catarina (UFSC)
Florianópolis - Brazil
{airtonz; leandro; speroni; tite; gauthier}@egc.ufsc.br

Abstract—**The access information law approved by the Brazilian government regulates the provision of open government data in the Web. However, they are heterogeneous, unstructured and derived from independent sources, making it difficult to interconnect. This paper presents a process of identifying sources, ontology generation, mapping and publishing statistical linked data in the form of multidimensional cubes, represented by the RDF Data Cube Vocabulary. In this process, data are transformed and assigned semantic meaning through its connection with domain ontologies. Through a web application, the publication of these data is automated, allowing for future analysis operations with the use of Online Analytical Processing (OLAP). As a result, the approach is expected to increase the scale in the publication of statistical linked data, and therefore, increasing the potential for analysis.**

Keywords–Linked Data; OLAP; Open Data; RDF Data Cube; Statistical Data.

I. INTRODUCTION

In November, 18th, 2011, Brazil approved the Information Access Law (“Lei de Acesso à Informação”, law number 12.527), regulating and granting the right to access public information of the Brazilian government, which is assured by its Federal Constitution. Taking effect in May, 16th, 2012, such law represents a big leap towards transparency and citizenship, forcing public agencies to consider openness a rule and confidentiality an exception, broadening citizen participation in governmental actions.

Such initiative gave rise to a higher availability of data on the *Web*, being originated from several sectors of public administration. Even so, such data is typically heterogeneous and has no integrated statistic treatment [1]. Moreover, there are no available means to expose, share or link the data so that they can add more information [2].

Statistical data are important sources of information for: a) the Government, as a way of verifying the Strong and weak points of their administration, therefore contributing for better policymaking decisions; b) science, as an important tool for accepting or rejecting a theory; and c) for business, as a way of supporting strategic decisions of the administration. For that matter, it is paramount that such data are semantically linked to ontologies or knowledge databases [3].

Two of the main challenges mentioned by Kämpgen, O'Rain and Harth [4] towards the use of OLAP in statistical linked data are: a) OLAP requires a data cube model, with dimensions and measurements; b) OLAP queries are complex and require specialized data models, such as the star model in relational databases, to run efficiently.

On that basis, the Federal University of Santa Catarina’s (UFSC) Knowledge Engineering and Management Graduation Program (EGC) researchers, through the Knowledge Engineering Laboratory (LEC), developed a supporting tool for the publishing of statistical data series in an open multidimensional model pattern, using OLAP data cubes. The main contribution of this paper is to automate the process of cube construction that is extremely complex when running on a non-automated way. Furthermore, enables semantic search while using SPARQL [5] language query over these datasets. Furthermore, it enables semantic search while using SPARQL language query over these datasets.

Therefore the following technologies were combined: OLAP Data Cube, ontologies and linked open data, resulting in a functional tool for generating statistical data. This method uses an ontology named cube and makes the Extraction, Transformation and Loading (ETL) in an OLAP structure that is mounted according to this ontology.

The research related to this work is further presented, in Section II. In Section III, RDF language is described. In Section IV, we explain the pattern for exchanging and sharing SDMX data. In Section V, the RDF Data Cube vocabulary is described. Section VI addresses the obtainment, processing and publication of data. The final considerations are in Section VII.

II. RELATED WORKS

Researches, such as Hull [6], point out the integration of different databases for the purpose of adding more content to what is being searched on the Web. Thus, the primary intention of linked open data [7] is to publish, share and connect different databases openly available on the Web. In order to make this possible, Berners-Lee [8] identified four principles that standardize the publication of the data that form the so-called Web of data: use Uniform Resource Identifiers (URIs) to identify things, use HTTP to find these names on the web, provide useful information in the form of Resource Description Framework (RDF) [9], and include links to other URIs so that you can discover more things. These principles, associated with the use of ontologies [10] and the SPARQL query language [11] form a set of technologies that have been established for the publication and the integration of data [12], [13] and [14].

These principles can be observed in the work of Kämpgen and Harth [15], which aimed at using linked data from various databases available on the web in an OLAP system. For that, the transformation and integration of data were made into an appropriate format using OLAP operations and SPARQL queries.

The OLAP, also known as "OLAP cube" [10], was officially described in an article submitted to Arbor Software Corp. in 1993 entitled “Providing OLAP (Online Analytical Processing) to User-Analysts: An IT Mandate”, written by W.H. Inmon, R. Kimball, and E.F. Codd [16], although its concept is known more earlier [17]. Date [17] defines OLAP as an interactive process of creating, managing, analyzing and reporting data. Typically its operations are roll-up (increases the level of aggregation) and drill-down (decreasing aggregation and increases the breakdown providing a smaller granularity) [18] along one or more dimensions. Slice and dice (selection and projection) are responsible for working with the information, changing positions whenever necessary and pivoting (reorientation and multidimensional view of data), with the ability to summarize and group data in various formats [19].

Works, such as Kämpgen, O'Rain and Harth [4], suggest a new way of interacting with linked statistical data in an RDF-modeled cube, a format of structured data that enables querying to multiple data sources through the use of SPARQL language. To this end Zapilko and Mathiak [1] developed an approach for the purpose of assisting researchers to statistically analyze the linked data with the aid of SPARQL.

The main advantage for the publication and consumption statistical data, according to Kämpgen [20], is the ease of integration and enrichment of data using other sources. In addition Cyganiak et al. [21] shows a number of benefits to publish statistics using RDF standards: a) the ability to access the annotations generated by third parties, b) statistical data can integrate a wider range of linked data, c) the possibility to perform operations of slice and dice in the datasets and the granularity of the information, and d) the flexibility offered by RDF in publishing statistical data.

For Salas et al. [3], statistical data are the main source of information for governments, researchers and administrators. In this sense, the author proposes two tools that use the RDF Data Cube vocabulary in order to provide the representation of statistical data in the multidimensional format. The first is the *OLAP2DataCube*, which allows the analysis of a large amount of data and its efficient transformation into RDF. The second tool, which is the *CSV2DataCube*, offers conditions to transform data available in CSV format to RDF.

On the other hand, the RDF data cube vocabulary was, according to Follenfant, Trastour and Corby [22], introduced by Cyganiak, Reynolds and Tennison [23] with the purpose of allowing the publication of statistical data on the Web, providing a metamodel for a set of multidimensional data [2].

Finally, Casanova et al. [24] highlights the lessons learned in the conversion of government data in RDF and graphically presents a comparison between the American and Brazilian data.

III. RDF

The RDF is a language originally created to represent and identify semantic content and information on Web pages [9]. However, in a generalized concept of Web, RDF can be used to identify any information or resource existing in the Web. Normally, it is used when the information that is wanted to be retrieved will be processed by machine rather than only displayed to the user.

According to Manola and Miller [9], RDF provides a common way of expressing information enabling the exchange of information among the applications without losing the meaning.

Fundamentally, the RDF vocabulary is fully extensible and consists of identifying objects via URIs [25]. URIs are used to name things, identify resources and properties in RDF.

The RDF properties may be thought of as attributes of resources and thus corresponds to traditional attribute-value pairs [26]. They also represent relationships between resources, resembling to an entity-relationship diagram where resources correspond to objects and properties correspond to instance variables.

The RDF data model consists of three basic types of objects [27] they are:

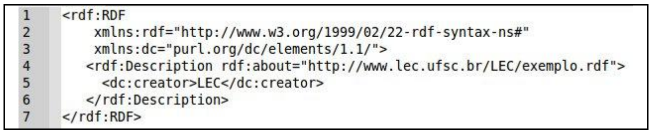
a) Resources: all that can be described by RDF expressions and identified by a URI (whole or part of web pages, a figure, or even an object that is not directly accessible via the Web, for example: a printed book);

b) Properties: specific aspects, characteristics, attributes or relations used to describe a resource. Each property has a specific meaning, defines its permitted values, the types of resources that can describe, and its relationship with other properties [26] and;

c) Sentences: structured information composed of subject (resource), predicate (property) and object (property value). The object of a statement can be another resource or it can be a literal, that is, a resource (specified by a URI), a simple string or other primitive data type defined by XML.

Figure 1 illustrates the sentence "the archive ‘exemplo.rdf’ was created by the Knowledge Engineering Laboratory LEC" by using the syntax and RDF.

ISBN: 978-1-61208-254-7



```
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dc="purl.org/dc/elements/1.1/">
   <rdf:Description rdf:about="http://www.lec.ufsc.br/LEC/exemplo.rdf">
     <dc:creator>LEC</dc:creator>
   </rdf:Description>
</rdf:RDF>
```

Figure 1. Sentence "the archive exemplo.rdf was created by LEC", expressed in RDF syntax.

The same sentence presented by Figure 1 can be expressed graphically by using arcs (properties) and nodes (resources or objects) as illustrated in Figure 2.

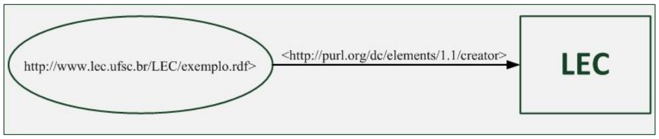

http://www.lec.ufsc.br/LEC/exemplo.rdf>
<http://purl.org/dc/elements/1.1/creator>
LEC

Figure 2. Sentence of Figure 1 graphically expressed.

Another way to express the sentences RDF is the Notation 3 (N3) [28], in which the three elements are listed in order: subject, predicate and object. Its objectives are to optimize the expression of logic and data in the same language and to allow the RDF can be expressed and rules to be integrated seamlessly to the RDF to be the most readable, natural and symmetrical as possible.

In order to promote interoperability and comparability between datasets using RDF, Milošević et al. [2] describes a syntax for the exchange and sharing of statistical data and metadata known as SDMX-RDF which will be presented in sequence.

IV. SDMX

Proposed in 2001 by the International Organization for Standardization (ISO), Statistical Data and Metadata eXchange (SDMX) is a standard for exchanging and sharing of statistical data and metadata between organizations. The SDMX standard for exchange of such data is a joint proposition of seven organizations worldwide, among which the U.S. Federal Reserve Federal Reserve, European Central Bank, the World Health Organization (WHO), the International Monetary Fund (IMF) and the United Nations (UN).

The SDMX has the SDMX-RDF syntax which, according to Cyganiak, Reynolds and Tennison [23], consists of the same model information specified in SDMX, but with information expressed in RDF, allowing the simple discovery and publication of linked data to the Web. SDMX-RDF defines classes and predicates to represent RDF statistical data compatible with the SDMX information model.

The key component of the SDMX standards, according to Cyganiak, Reynolds and Tennison [23], is the Content-Oriented Guidelines (COGs), a set of domain concepts, code lists and categories that support interoperability and comparability between data sets, providing a common language SDMX between applications. The RDF versions of these components are available as part of SDMX-RDF, and should be reused where possible.

V. RDF DATA CUBE VOCABULARY

The concept of multidimensional modeling, as it is known today, was proposed by Kimball [29] and subsequently deepened and enhanced in Kimball [30]. According to the Kimball [30], the great advantage of the multidimensional model is its simplicity, which is essential to enable users to understand databases, and allow recovery in an efficient way.

Multidimensional models are designed to store statistical data sets that, according to Cyganiak, Reynolds and Tennison [23], comprise a collection of observations made at some points across a logical space. This collection is characterized by a set of dimensions which define the scope of each observation along with metadata describing what was measured, as was measured and how the observations are expressed.

Statistical data can be set in a multidimensional way in space, that is, as a hypercube. A cube is arranged according to a set of dimensions, attributes and measures.

The dimensions are used to identify the observations, that is, the set of values for each dimension represents a single observation. Examples of dimensions include the time that the observation applies, or the geographic region that the observation covers.

The attributes, for its part, qualify and interpret the observed values. They allow the specification of measurement units and scale factors.

Lastly, the measures represent the phenomenon to be observed, for example, the population growth of a municipality.

The formalization of the understanding of data cubes based on Data Cube vocabulary (QB) is presented in the following and illustrated in Figure 3.

With QB vocabulary as points Kämpgen, O'Rain and Harth [4], there is greater ease in handling and ability to capture the statistical semantics of the linked data. As examples of projects that use the same vocabulary we have the UK government program *Combined Online Information System* (COINS) and the North American program of the U.S. Security and Exchange Commission.

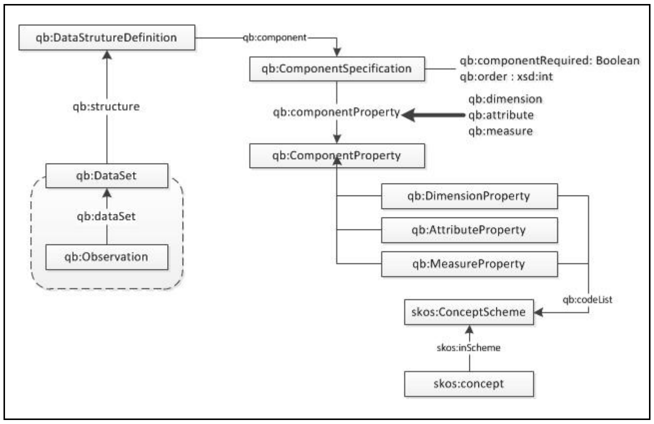


Figure 3. The data cube core [23].

Figure 3 illustrates, in a general way, the division of the *QB* structure in classes and properties which include the datasets used - its dimensions, attributes and measures - besides the observations and operations that will be applied on the data.

The main classes and properties of the structure of the data cube are:

a) DataSets: Class *qb:DataSet* - represents a collection of observations, possibly organized into multiple slices, as the common dimensional structure is determined;

b) *Observations*: Class *qb:Observation* - represents a single observation of the cube, which can have one or more measurement values associated. Related to this class we found properties *qb:dataset*, which indicates what set of data that observation belongs, and *qb:observation*, that indicates an observation contained within a slice of the data set;

c) *Data Structure Definitions* (DSD): Class: *qb:DataStructureDefinition* - defines the structure of data set (*Dataset*) or a slice. Associated to this class are the properties: *qb:structure*, which indicates the structure to which this data set belongs, and *qb:component*, responsible for the specification of the component that is included in the structure of the dataset;

d) *Dimensions, Attributes and Measures*: Class: *qb:ComponentProperty* - subclass of *rdf:Property*, a super abstract property of all properties that represent dimensions, attributes and measures. Class *qb:DimensionProperty* - represents components that form the cube dimensions. Class *qb:AttributeProperty* - formed by components which represent the attributes of the observations in the cube, for example, units of measure. Class *qb:MeasureProperty* - represents the measured values for the observed phenomenon. Associated to this class have the property *qb:measureType*, a generic measure of size. The value of this dimension indicates how far (within the set of measures of DSD) is provided by the value of observation, or other primary measure;

e) *Slices***:** Class *qb:Slice* –denotes a subset of a dataset that is set by setting a subset of dimensional values. Its property *qb:slice*, indicates the subset defined by setting a subset of the values of the dimension.

A. Multicubes

Cubes which share dimensions constitute a a dimensional model of multiple cubes. In *QB*, a cube corresponds to multiple cubes that use instances of *qb:ComponentProperty* with equivalence, and thus can be connected using the property *owl:sameAs*.

Similarly, members of a cube can be equivalent as in the case of Figure 4 which shows the relationship afforded by binding property *owl:sameAs*.

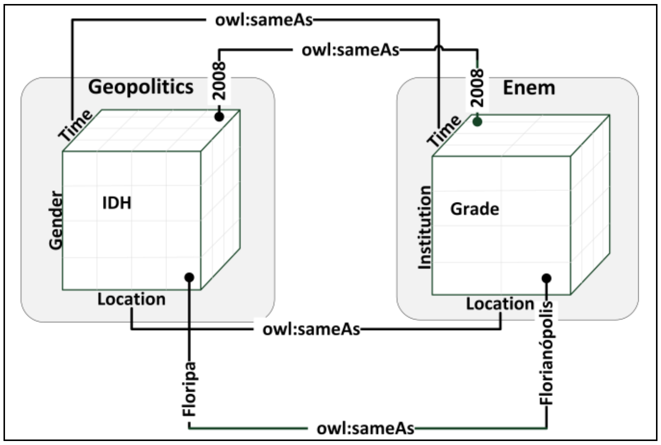


Figure 4. Example of a multicube.

In Figure 4, there are two datasets, one representing Brazilian Geopolitics and other representing the scholar data in Brazil (Enem [31]). Both have a dimension which denotes a geographical entity and has a member which denotes a city, such as Florianópolis. Also, both may use the same time dimension with literal values. If there is a statement *owl:sameAs* between the geographical dimensions, the two cubes can be represented as a multicube.

B. Relating OLAP operations to SPARQL in QB

The set of terms in an RDF Triple Store consists of URIs, blank nodes and literals. Triple store management system is a database for RDF, in which a triple (s, p o) is called RDF triple, where "s" is the subject, "p" represents the predicate and "o" is the object. These systems provide management and access to data through APIs and query languages for RDF data. Many Triple Quad Stores Stores are indeed due to the need to maintain the provenance of RDF data within the system. Any Triple Store that supports graphs will probably be a Quad Store [32] and [33].

Given a Triple Store with statistical linked data, we use basic queries in SPARQL about this dataset to return a specific set of multidimensional elements comprising their respective URIs or a blank node. In this section we present common OLAP queries on a multidimensional model using SPARQL in QB. The similar operations (projection, slice, dice and roll-up) can be found in [4], [34] and [35] and are illustrated in Figure 5.

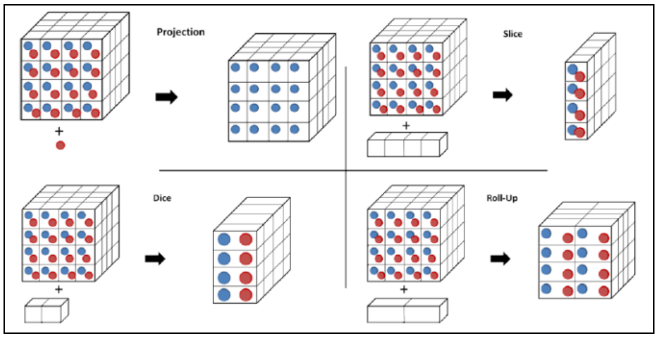


Figure 5. Representation of OLAP operations with inputs e outputs. Source: [4].

a) Projection: DataCube x Measure → DataCube - removes a measure of the entry cube and allows you to see only a specific measure. In the example above, all triple referring to a measurement are removed, resulting in a query subcube;

b) Slice: DataCube x Dimension →DataCube - removes a dimension from the entry cube and all its contents added over the members of a dimension;

c) Dice: DataCube x Dimension x Value → DataCube - allows to filter and aggregate on certain dimension members. Note that Dice is not an selection operation, but a filter combined with the Slice operation;

d) Roll-up: DataCube x Dimension x Level → DataCube - allows to create a cube that contains instance data at a high level of aggregation. Remark: the Drill-Down operation has not been set yet, since it can be viewed as an inverse operation to roll up.

VI. COLLECTION, PROCESSING AND PUBLICATION OF DATA

The proposed for publication of data consists of four main steps: identification of sources identifying, ontology generation, mapping and publication of data; shown in Figure 6. Using the data from outside sources, it is intended to add semantic meaning by connecting them to other sources of data, thus publishing them in the form of a multidimensional cube.

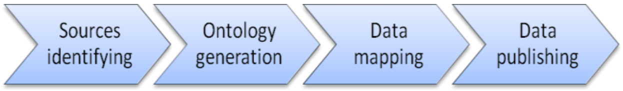


Figure 6. Steps in the process of obtaining, processing and publication of data.

Sources Identifying: the composition of logical cubes composed of dimensions, measures, attributes, hierarchies and levels, allows to represent in a simple way real-world entities. Furthermore, it can facilitate the analysis of measures, to define which dimensions and attributes can represent significant data, and organize the dimensions of a given scope in levels and hierarchies.

In the choice of sources, some criteria must be observed such as the accuracy that indicates whether the values are stored in accordance with the actual values; the temporality, which indicates whether the recorded values are updated; the completeness, which indicates whether the needed values are stored and that these have an appropriate depth and width; and the consistency of the data [36]. Last but not least, the quality of the data must meet the requirements of its use.

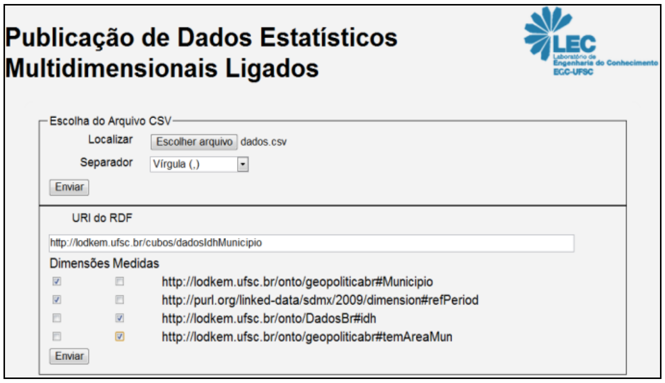
To illustrate this, we used Brazilian governmental open data sources [37]. This choice is justified for these data sets have features such as: historical time series and the division by municipalities, considering also their spatial location, essential in the construction of cube.

Ontology Generation: the governmental data sources available have no semantic meaning and for this it is necessary that these sources are represented by a domain ontology. An ontology provides an explicit specification of a conceptualization [38] and its objective is to define which primitives are necessary for the representation of knowledge in a given context. In this process, the ontologies are used to provide a semantic representation of the dimensions and observations that describe the data to be published. In this stage, it is necessary to generate a proper ontology or to use an existing one to represent the data set.

Data Mapping: After the identification of the different data sources the ETL process is performed, through which data are integrated to form a single assembly. The data from this set are analyzed and linked to concepts represented in the ontologies. Thus, we want to clarify this relationship through a mapping that indicates, for each column of data, the URI of the corresponding concept. Also in this step the values contained in the dataset to the corresponding URIs are mapped, based on the representation of domain ontologies. Thus, each data value will be represented by the URI that identifies it.

Data Publishing: The publication step is the transformation of the data set already properly mapped to a multidimensional OLAP cube model. The adopted vocabulary for the publication is the RDF Data Cube, which represents, in the form of RDF, structures and standardized data appropriately for subsequent processing through OLAP operations. Furthermore, the fact of using RDF and URIs ensures integration of the dataset with the ontology and other datasets.

Figure 7 shows the interface through which the user makes the choice of the file containing the dataset as well as the definition of what are the dimensions that will form the multidimensional cube, and the measures that will make the facts stored in it.



Publicação de Dados Estatísticos Multidimensionais Ligados

Escolha do Arquivo CSV

Localizar: Escolher arquivo dados.csv

Separador: Vírgula (,)

Enviar

URI do RDF

http://lodkem.ufsc.br/cubos/dadosIdhMunicipio

Dimensões	Medidas	
☑	☐	http://lodkem.ufsc.br/onto/geopoliticabr#Municipio
☑	☐	http://purl.org/linked-data/sdmx/2009/dimension#refPeriod
☐	☑	http://lodkem.ufsc.br/onto/DadosBr#idh
☐	☑	http://lodkem.ufsc.br/onto/geopoliticabr#temAreaMun

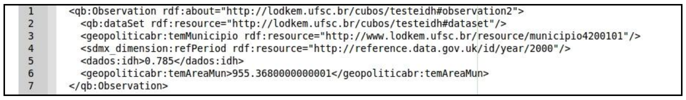
Enviar

Figure 7. Web application interface for publishing data.

Through the Web application developed, the dataset obtained is mapped, linked to other data sources and then converted into the format cube RDF.

In the processing each data set corresponds to an instance of the class *qb:Dataset*, which has its structure defined by an instance of *qb:DataStrucureDefinition*. This, in turn, is described in dimensions (represented by the property *qb:dimension*) and measures (represented by the property *qb:measure*).

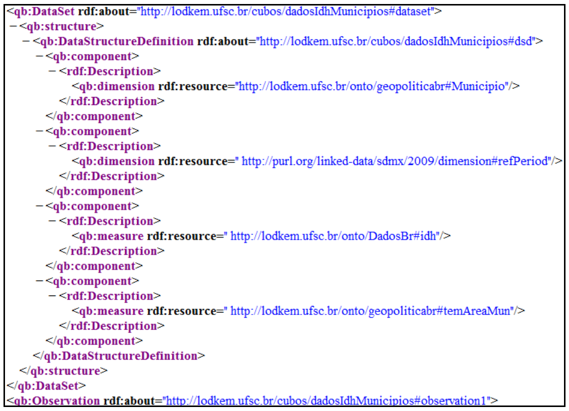
Each of the observations that make up the cube are represented by an instance of the class *qb:Observation* (Figure 8), and its properties are associated with references to the *qb:Dataset* and for each of the dimensions and measures.



```
1   <qb:Observation rdf:about="http://lodkem.ufsc.br/cubos/testeidh#observation2">
2     <qb:dataSet rdf:resource="http://lodkem.ufsc.br/cubos/testeidh#dataset"/>
3     <geopoliticabr:temMunicipio rdf:resource="http://www.lodkem.ufsc.br/resource/municipio4200101"/>
4     <sdmx_dimension:refPeriod rdf:resource="http://reference.data.gov.uk/id/year/2000"/>
5     <dados:idh>0.785</dados:idh>
6     <geopoliticabr:temAreaMun>955.3680000000001</geopoliticabr:temAreaMun>
7   </qb:Observation>
```

Figure 8. Example of an observation.

Figure 8 presents an example of representation of a observation of the cube, in RDF form. Lines 1 and 7 delimit the observation indicating that it is a by stating that it is an instance of the class *qb:Observation*. The line 2 indicates the observation this cube belongs to, represented by the property *qb:dataSet*. Lines 3 and 4 indicate the references for the dimension values, while lines 5 and 6 present the measurements relating to the observation. The result of the transformation is a file containing an RDF graph as shown in Figure 9.



```
<qb:DataSet rdf:about="http://lodkem.ufsc.br/cubos/dadosIdhMunicipios#dataset">
 -<qb:structure>
   -<qb:DataStructureDefinition rdf:about="http://lodkem.ufsc.br/cubos/dadosIdhMunicipios#dsd">
     -<qb:component>
       -<rdf:Description>
           <qb:dimension rdf:resource="http://lodkem.ufsc.br/onto/geopoliticabr#Municipio"/>
        </rdf:Description>
      </qb:component>
     -<qb:component>
       -<rdf:Description>
           <qb:dimension rdf:resource=" http://purl.org/linked-data/sdmx/2009/dimension#refPeriod"/>
        </rdf:Description>
      </qb:component>
     -<qb:component>
       -<rdf:Description>
           <qb:measure rdf:resource=" http://lodkem.ufsc.br/onto/DadosBr#idh"/>
        </rdf:Description>
      </qb:component>
     -<qb:component>
       -<rdf:Description>
           <qb:measure rdf:resource=" http://lodkem.ufsc.br/onto/geopoliticabr#temAreaMun"/>
        </rdf:Description>
      </qb:component>
    </qb:DataStructureDefinition>
  </qb:structure>
</qb:DataSet>
<qb:Observation rdf:about="http://lodkem.ufsc.br/cubos/dadosIdhMunicipios#observation1">
```

Figure 9. Excerpt of the generated RDF file.

The RDF graph generated can then be loaded into a Triple Store, so that data are made available on a linked data infrastructure and to allow SPARQL queries.

VII. CONCLUSION AND FUTURE WORK

The provision of open government data has been gaining momentum in many countries. The standardization and structure, however, are not yet a concern of the agencies responsible for disseminating them, complicating the analyzes on them, especially by machines.

In this work it was presented a process for publication of statistical data related to a multidimensional cube format. The use of a standard vocabulary representing the cube structure allows publication of a large amount of statistical data from different sources.

The choice of RDF Data Cube vocabulary as standard allows not only the publication scale as well as increases the potential for analysis, since the operations are also standardized. It also permits that the published data may be used by others applications.

The concern with the mapping of data ensures semantics and connect them with external sources and with other cubes that are already stored on Triple Store. In the case of cubes that share semantic concepts and have dimensions in common, multiple cubes are materialized.

Upcoming efforts that follow this work are the development of tools for visualizing multidimensional statistical data, using standard OLAP operations. These tools will provide a more powerful analysis of the data, besides allowing identifying links between different datasets.

Among the expected benefits of this proposal is the publishing of large-scale statistical series of linked data, which would serve as a base for open data portals, whether governmental or not.

The publication of the data in the form of OLAP cube allows the use of standard operations (e.g., slice, dice, roll-up, drill-down), which facilitates the analysis.

The published data, along with analysis tools enable a more agile application for data presentation. New mashups can be created according to the needs, only requiring the specification of which cubes and which operations to be applied.

In further research, integration tests will be performed with tools that enable graphical visualization of data from the data cubes developed from the process presented in this work, culminating in the publication of results in a portal public domain.

REFERENCES

[1] B. Zapilko and B. Mathiak, “Performing Statistical Methods on Linked Data,” in *Proc. Int’l Conf. on Dublin Core and Metadata Applications 2011*, 2011, pp. 116–125.

[2] U. Milošević, V. Janev, M. Spasić, J. Milojković, and S. Vranes, “Publishing Statistical Data As Linked Open Data,” in *Proceedings of the 2nd International Conference on Information Society TechnologyInformation Society of the Republic of Serbia*, 2012, no. September, pp. 182–187.

[3] P. E. R. Salas, M. Martin, F. M. Da Mota, S. Auer, K. Breitman, and M. A. Casanova, “Publishing Statistical Data on the Web,” in *International Conference on Semantic Computing*, 2012, 6th ed., pp. 285–292.

[4] B. Kämpgen, S. O’Rain, and A. Harth, “Interacting with Statistical Linked Data via OLAP Operations,” in *Proceedings of the International Workshop on Interacting with Linked Data*, 2012, pp. 36–49.

[5] E. Prud’Hommeaux and A. Seaborne, “SPARQL Query Language for RDF,” *SPARQL Query Language for RDF, Technical Report*, 2004.

[6] R. Hull, “Managing Semantic Heterogeneity in Databases : A Theoretical Perspective,” in *Proc. ACMSymposium on Principles of Databases*, 1997, pp. 51–61.

[7] J. Umbrich, K. Hose, M. Karnstedt, A. Harth, and A. Polleres, “Linked data-the story so far,” *World Wide Web*, vol. 14, no. 5–6, pp. 495–544, Jan. 2011.

[8] T. Berners-Lee, “Linked Data: Design Issues,” 2006. [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html. [Accessed: Jul. 2012].

[9] F. Manola and E. Miller, “RDF Primer W3C Recommendation,” 2004. [Online]. Available: http://www.w3.org/TR/rdf-primer/ . [Accessed: Aug. 2012].

[10] M. Niinimaki and T. Niemi, “An ETL Process for OLAP Using RDF / OWL Ontologies,” in *Journal on Data Semantics XIII*, 2009, pp. 97–119.

[11] E. Prud’hommeaux and A. Seaborne, “SPARQL Query Language for RDF,” 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/. [Accessed: Jan. 2013].

[12] T. Bray, “RDF and Metadata,” 1998. [Online]. Available: http://www.xml.com/pub/a/98/06/rdf.html. [Accessed: Jun. 2012].

[13] A. Kerzazi, O. Chniber, I. Navas-Delgado, and J. F. Aldana-Montes, “A Semantic Mediation Architecture for RDF Data Integration,” in *SWAP 2008*, 2008.

[14] I. F. Cruz and H. Xiao, “The Role of Ontologies in Data Integration,” *Journal of Engineering Intelligent Systems*, vol. 13, pp. 245–252, 2005.

[15] B. Kämpgen and A. Harth, “Transforming statistical linked data for use in OLAP systems,” in *Proceedings of the 7th International Conference on Semantic Systems - I-Semantics ’11*, 2011, pp. 33–40.

[16] J. M. Pérez, R. Berlanga, M. J. Aramburu, and T. B. Pedersen, “Integrating Data Warehouses with Web Data : A Survey,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 7, pp. 940–955, 2008.

[17] C. J. Date, *Introdução a sistemas de bancos de dados*. Rio de Janeiro: Campus, 2004, p. 865.

[18] R. Elmasri and S. B. Navathe, *Sistemas de banco de dados*, 4 ed. São Paulo: Addison Wesley, 2005, p. 724.

[19] S. Chaudhuri and U. Dayal, “An overview of data warehousing and OLAP technology,” *ACM SIGMOD Record*, vol. 26, no. 1, pp. 65–74, Mar. 1997.

[20] B. Kämpgen, “DC proposal: online analytical processing of statistical linked data,” *The Semantic Web–ISWC 2011*, vol. 7032, pp. 301–308, 2011.

[21] R. Cyganiak, C. Dollin, and D. Reynolds, “Expressing Statistical Data in RDF with SDMX-RDF,” 2010. [Online]. Available: http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/index.html. [Accessed: Oct. 2012].

[22] C. Follenfant, D. Trastour, and O. Corby, “A Model for Assisting Business Users along Analytical Processes,” *files.ifi.uzh.ch*, 2004.

[23] R. Cyganiak, D. Reynolds, and J. Tennison, “The RDF Data Cube vocabulary,” 2010. [Online]. Available: http://publishing-statistical-data.googlecode.com/svn/trunk/specs/src/main/html/cube.html. [Accessed: Jul. 2012].

[24] M. A. Casanova, K. Breitman, P. Salas, D. Saraiva, V. Gama, J. Viterbo, R. Pires, E. Franzosi, and M. Chaves, “Open Government Data in Brazil,” *IEEE Intelligent Systems*, no. May-June, 2012, pp. 45–49, 2012.

[25] G. Klyne and J. J. Carroll, “Resource Description Framework (RDF): Concepts and Abstract Syntax,” 2004. [Online]. Available: http://www.w3.org/TR/rdf-concepts/. [Accessed: Aug. 2012].

[26] O. Lassila and R. R. Swick, “Resource Description Framework (RDF) Model and Syntax Specification,” 1998. [Online]. Available: http://www.w3.org/1998/10/WD-rdf-syntax-19981008. [Accessed: Aug. 2012].

[27] D. Brickley and R. V. Guha, “RDF Vocabulary Description Language 1.0: RDF Schema,” 2004. [Online]. Available: http://www.w3.org/TR/rdf-schema. [Accessed: Aug. 2012].

[28] T. Berners-Lee and D. Connolly, “Notation3 (N3): A readable RDF syntax,” 2011. [Online]. Available: http://www.w3.org/TeamSubmission/n3/. [Accessed: Jan. 2013].

[29] R. Kimball, *The data warehouse toolkit: practical techniques for building dimensional data warehouses*, 1^a ed. 1996, p. 388.

[30] R. Kimball, L. Reeves, M. Ross, and W. Thornwaite, *The Data Warehouse Lifecycle Toolkit: Expert Methods for Designing*, 1^a ed. 1998.

[31] INEP, “ENEM 2012 - Passo a passo,” 2013. [Online]. Available: http://www.enem.inep.gov.br. [Accessed: Jan. 2013].

[32] Virtuoso, “RDF Triple Store FAQ,” 2012. [Online]. Available: http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSRDFFAQ. [Accessed: Aug. 2012].

[33] J. Rusher and R. Networks, “Triple Store,” 2001. [Online]. Available: http://www.w3.org/2001/sw/Europe/events/20031113-storage/positions/rusher.html. [Accessed: Aug. 2012].

[34] J. Pardillo, J.-N. Mazón, and J. Trujillo, “Bridging the semantic gap in OLAP models,” in *Proceeding of the ACM 11th international workshop on Data warehousing and OLAP - DOLAP ’08*, 2008, pp. 89–96.

[35] O. Romero and A. Abelló, “Automating multidimensional design from ontologies,” in *Proceedings of the ACM tenth international workshop on Data warehousing and OLAP - DOLAP ’07*, 2007, pp. 1–8.

[36] D. P. Ballou and H. L. Pazer, “Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems,” *Management Science*, vol. 31, no. 2, pp. 150–162, Feb. 1985.

[37] Brasil, “Portal brasileiro de acesso a informação,” 2013. [Online]. Available: http://data.gov.br. [Accessed: Jan. 2013].

[38] T. Gruber, “Towards principles for the design of ontologies used for knowledge sharing,” *International Journal of Human-Computer Studies*, vol. 45, no. 5–6, pp. 907–928, 1995.

Towards Recovering Provenance with Experiment Explorer

Delmar B. Davis, Hazeline U. Asuncion
Computing and Software Systems
University of Washington, Bothell
Bothell, WA, USA
{davisdb1, hazeline}@u.washington.edu

Ghaleb M. Abdulla, Christopher W. Carr
Lawrence Livermore National Laboratory
Livermore, CA, USA
{abdulla1, carr19}@llnl.gov

***Abstract*—In this work, we present Experiment Explorer (EE), a framework for recovering provenance that uses provenance-compatible research processes and a lightweight, user-friendly metadata search tool. EE also captures recovered provenance along with file relationships and incorporates new files to support provenance recovery over time. Our case study at a research laboratory suggests that EE is effective in connecting distributed provenance information and in increasing the accessibility of related experiment files. Our scalability analysis also indicates that EE's tool support can scale to hundreds of thousands of heterogeneous files.**

Keywords-data provenance; metadata; provenance recovery; information management.

I. Introduction

More scientific research now involves the generation and analysis of heterogeneous data sets. Multiple instruments, analysis tools, and scripts are used to collect and analyze data. The origin of a data set or the processing applied to a data set is referred to as data provenance. Data provenance is necessary in assessing a data set's integrity and in supporting repeatability of analyses or experiments. However, obtaining provenance is a difficult task, especially for data sets that have already been reduced or aggregated from their raw form—some of the context of the data may have been lost. Data provenance may be obtained from experimental conditions and data processing or reduction. Deficiencies in data provenance related to experimental conditions may arise from mundane technical reasons (e.g., a loose cable) or fundamental reasons (e.g., not monitoring a critical parameter because its importance is not known at the time of the experiment). For example, the growth of sites exposed to laser was studied for many decades before it was discovered that the temporal pulse shape was important. In experiments conducted prior to this discovery, only the duration of the pulse was recorded.

Current provenance techniques often capture provenance *during* an analysis or experiment run by logging the steps that were followed or the analysis modules that were invoked [1, 2, 3]. Many of these techniques use scientific workflows where researchers pre-specify the experiment design as a workflow and a log of the workflow execution provides the provenance of the data produced at the end of the execution [1, 4]. Other techniques record commands or events while the dataset is being processed [2, 3, 5] .

To support the use case of recovering provenance where processing logs may not be available, we present Experiment Explorer (EE). EE allows users to recover provenance by incorporating provenance-compatible research processes and by enabling researchers to search for experiment-related information, possibly scattered across heterogeneously-represented files, that provide clues to the provenance of a data set. EE leverages a lightweight and user-friendly metadata search tool to aid researchers in uncovering provenance, which requires minimal training time and usage overhead from them. EE also allows researchers to capture the recovered provenance and to incrementally support provenance recovery over time. Moreover, EE can be used in conjunction with recording provenance techniques in cases where incomplete provenance has been captured (e.g., recording took place only in some parts of the entire data processing). We previously introduced EE as a metadata provenance search [6]. In this paper, we elaborate on how EE can be used to support the recovery of data provenance.

The contributions of this paper are as follows: 1) a technique for recovering provenance, 2) a means of increasing the accessibility of related experiment files, and 3) a means of capturing recovered provenance and file relationships for future reference. In addition, our case study at a research lab suggests the effectiveness of our technique in locating distributed provenance information and in raising the visibility of relevant experiment files.

The rest of the paper is organized as follows. In the next section, we compare our work with existing techniques. Section 3 covers challenges in providing provenance support to research in general, with an example of support for the optics inspection and analysis group at Lawrence Livermore National Laboratory (LLNL). We then introduce our approach in Section 4. We provide details regarding our tool support in Section 5 and discuss evaluation results in Section 6. We close the paper with future work.

II. Related Work

Provenance techniques generally record provenance as the data is being processed. These techniques are often associated with workflows [1, 4, 7]. Other recording techniques include capturing using interactions [2, 5] or listening to system-level events [3]. Different levels of provenance may also be recorded [8]. Recovering provenance complements both prospective provenance, i.e., specification of the process as in a workflow, and retrospective provenance, e.g., log of execution [9].

One key aspect to EE is the crafting of research processes that are provenance compatible. Others have also suggested

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-254-7

process-centered approaches that revolve around the use of services and tools [10] or around collaboration steps [4]. Another technique collects provenance based on coordination points in distributed enterprise workflows [11].

Metadata has also been used to represent or link different types of data. These links between different types of data may be represented as a Resource Description Framework (RDF) graph [12]. Provenance metadata may also be represented as an RDF graph that can be queried [13]. Discovery techniques exist for discovering metadata via recommender systems [14] or discovering data using metadata [15]. Data provenance which has been registered to a service may also be later discovered [8]. Tools that capture metadata include XMC-Cat [16] and Taverna [17]. XMC-Cat relies on workflow cyber infrastructure to capture metadata as associated data is generated [16]. Taverna captures metadata by observing processing units and it imports metadata from existing entities [17]. In EE, we use metadata to search for experiments and their corresponding artifacts. EE can also generate metadata for experiment files.

There are also various techniques for searching for documents. These involve using string matching techniques [18] or incorporating user activity into the search for documents [19]. These techniques fall short of searching for files that do not have a text representation (e.g., image files). Other techniques use a database [20] or map-reduce [21] to increase the efficiency. As we demonstrate in our scalability analysis, EE's search tool is highly scalable. Document management systems like Placeless Documents allows users to search for documents according to metadata, such as file properties (e.g., file size = 100MB) or user-defined properties (e.g., priority = high) [22]. Instead of users specifying properties on a file-by-file basis, EE uses the information embedded in the directory structure and the provenance template to automatically assign provenance metadata to files. Configuration management systems also allow users to search for files based on commit records or change entries [23]. EE, on the other hand, allows users to search for files based on their relationship to an experiment or based on an experiment attribute.

Recovering provenance has been discussed outside the field of eScience. One technique discusses how provenance can be recovered from executable software [24].

III. Motivation

The utility of scientific data is ultimately limited by its provenance. Aggregating data into sets requires that the provenance of the individual data points is compatible. When data are collected at the same time, compatibility is typically ensured by good experimental hygiene, even if critical aspects of provenance are not documented, or even not known to exist as it is the same for each datum. However if data from different experiments or different experimenters are to be combined, then it becomes critical to ensure the provenance is sufficient and compatible. We now provide an overview of challenges encountered in scientific research which effective data provenance can assist.

Fast-paced and adaptive research environment: Research efforts in highly competitive fields or in support of a larger project must be highly adaptive. Though most programs have well defined deliverables with due dates many months or years in the future, situations emerge from time to time which must be addressed immediately. A researcher often responds to new project needs or publication of related discoveries by repurposing existing data and analysis. Such adaptability is greatly facilitated by provenance techniques that are lightweight, with minimal setup and training time. In addition to short term priority shifts, research labs are also highly adaptive to advances in current technology and changes to tools. Using new architectures and tools requires documenting the old and new environments as part of the metadata.

Numerous heterogeneous experiment files: Each experiment run produces hundreds of files and multiple experiments are conducted for each hypothesis. Individual researchers working on related topics conduct their own experiments with techniques and analysis algorithms optimized to isolate particular experimental parameters. Thus, the number and type of experiment files quickly grows, making it difficult to manually search through files. Using off-the-shelf search tool for some of these formats, such as binary or image files, is inadequate because these tools are designed to search across distributed and unrelated files and there is no utilization of the relationships between the different files and experimental objects.

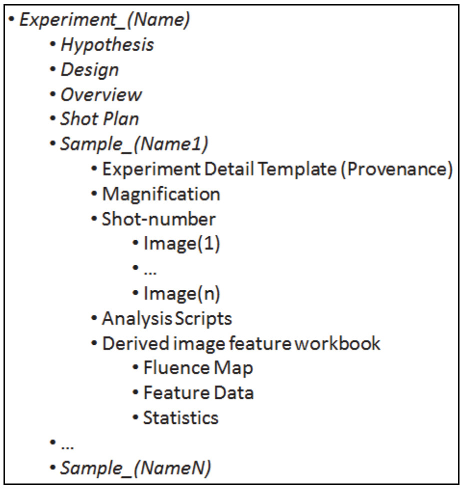
Accessing experiment files: Often researchers would like to amalgamate data from several data sets to create and test hypotheses. Compiling data sets without the aid of a provenance recovery tool requires researchers to search manually for the appropriate files. If the researchers are fortunate enough to be looking for an attribute that was used to categorize the experiment, they are able to gather a large data set and only incur the time it took to locate the data file. Otherwise, the experimental details must be manually retrieved from lab books, scripts, or correlated experiment files, which can be a time-consuming process.

IV. Approach & Application

To address the challenges discussed in the previous section, we present EE and apply it to the specific case of correlating laser induced damage data generated in a number of labs for optics inspection and analysis group at LLNL. Provenance recovery is a technique that approximates the provenance of a data set based on available information (e.g., researcher notes, scripts, hypotheses). EE facilitates the recovery of provenance by incorporating provenance-compatible research processes and enabling researchers to piece together distributed provenance information. While the provenance is an approximation, it may be "close-enough" for researchers to fill-in the missing gaps. Once provenance has been recovered, it can be captured for future reference. Additionally, EE supports incremental provenance recovery over time.

A. Incorporating Provenance-Compatible Research Practices and Conventions

The analysis process, which is part of the overall experiment process, is a collaborative effort, requiring the



- *Experiment_(Name)*
    - *Hypothesis*
    - *Design*
    - *Overview*
    - *Shot Plan*
    - *Sample_(Name1)*
        - Experiment Detail Template (Provenance)
        - Magnification
        - Shot-number
            - Image(1)
            - ...
            - Image(n)
        - Analysis Scripts
        - Derived image feature workbook
            - Fluence Map
            - Feature Data
            - Statistics
    - ...
    - *Sample_(NameN)*

Figure 1. Hierarchical relationship between experiment files

contributions of material science specialists, data analysts, and physicists. The workflows as well as the set of artifacts produced remain consistent across experiments.

Figure 1 shows the artifacts produced over the course of an experiment. (Note: an artifact is any file produced during a research process.) The process begins with a hypothesis that is used to derive an experiment design. Along with the design, a detail template, or provenance template, is created providing a central overview of the experiment attributes as data is analyzed. As scientists move through the workflow, raw experiment data are produced, labeled as image in the figure. When the images are analyzed, a derived image feature workbook is created, which contains information such as fluence map.

As part of the process, this provenance template is filled-in by researchers. Some of this information is obtained from the workflow and hypotheses, while others are derived from analysis. Attributes, which may be experiment design parameters (e.g., average fluence or average site separation), are also recorded in a provenance template.

All the artifacts produced for an experiment are stored in a folder labeled with the experiment sample name, to support provenance recovery. For example, hundreds or thousands of image files can result from one experiment. Each file is named in a way that captures metadata such as location of the damage image. In addition, a metadata file is generated for each experiment file, which includes author, date, experiment sample name, keywords, and category. This metadata provides a connection between each file and the associated experiment sample or among related files based on provenance-specific fields. Because the provenance template is stored at the top level sample name folder, the keywords, category, author, and date can be automatically extracted from the provenance template. This template can be manually used to locate experiments performed at certain dates, by specific scientists, or used fluence ranges between X and Y, etc.

B. *Using a Lightweight Metadata Search*

Once the metadata is created for every experiment file, researchers may now proceed with recovering provenance using a metadata search tool. First, we index the provenance template and the metadata associated with each file. The provenance template is an Excel workbook containing an experiment summary sheet, a high level data overview sheet, and the server location for the experiment sample. In order to search for experiment artifacts that span the entire data server, we first provide a means of relating the files. As we mentioned, all the experiment artifacts are stored within a folder with the experiment sample name. Thus, one way files can be related is by examining the path of an experiment file to obtain the experiment sample file name. Once the sample file name is obtained, the metadata for each file can include the experiment sample name. This metadata can then be re-indexed.

To address the challenge of searching through heterogeneous artifacts, we index different types of data using artifact-specific indexing components. We use artifact-specific indexing components to extract the metadata. The metadata follows a uniform format, enabling the metadata from various artifact types to be indexed in a similar fashion. Consider the following example:

Shot plans are documents requesting a particular set of laser exposures including the sample to be exposed, the location on the sample, and the number and type of laser exposures. A new file is detected by the indexing component as a result of comparing the past directory structure with the current directory structure. The artifact

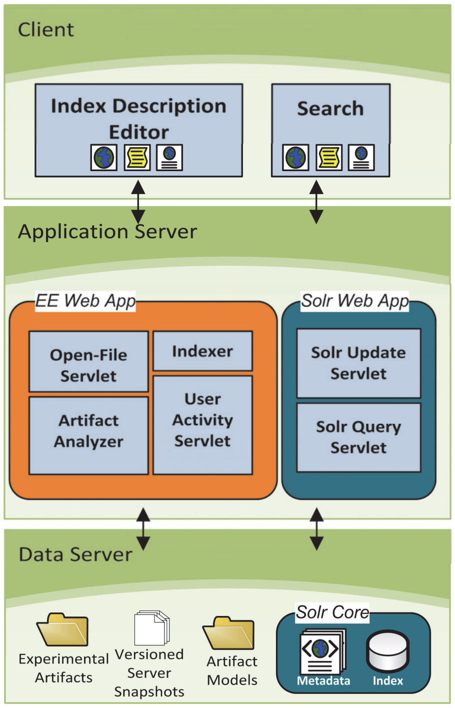


Figure 2. Architecture of EE's search tool

analyzer consults a set of models that reflect the conventions of the researchers. The file location, file naming pattern, extension, and structure match those of a shot plan.

The analyzer extracts the experiment name and the indexer searches for it. In this case, one result is expected because a shot plan file belongs to only one experiment. Upon success a link to this file is added to the experiment index document and the experiment is re-indexed.

We can also obtain the provenance of derived features. As the derived feature image workbook is modified on the server, it is re-indexed as described in our example above. By utilizing the file hierarchy shown in Figure 1, features from the workbook can be related to the experiment, which can then be related to the provenance template.

C. *Capturing Recovered Relationships and Provenance*

After the metadata for various files has been generated and indexed, researchers may search for experiment files for a given experiment sample name or a given attribute or even files created on the same date as the experiment. When search results are returned, researchers may directly access the experiment file by following a link from EE's search result, making the relevant files accessible to researchers.

In addition, since the search and access to files are integrated, it becomes straightforward to capture the search terms used and the files opened by users. The search terms can be captured to log how data sets are correlated. Capturing the correlations and the provenance of experiment files allows researchers to reflect upon their past search activities. This record can help them redo their past search activities or identify unexamined correlations.

D. *Supporting Provenance Recovery Over Time*

As time goes on, new hardware and software tools may be used to process the data. EE can accommodate these changes over time since it depends on the research processes and the metadata generated for each file, not the technology used for processing or analyzing data. New software tools may also produce new file formats that EE must accommodate. This can be handled by EE by creating a new component to index the new file format.

Over time, new experiment files will also be added into the file system. A version control system [25] can be used to track new files within the server folder. This way, new or modified artifacts can be indexed.

V. Tool Support

We now describe the design and current implementation.

A. *Recovering Provenance with EE*

In order to match the distributed system and collaborative work within the optics inspection and analysis group, Experiment Explorer is comprised of components following a 3-tier client server architecture, as shown in Figure 2. (Note: an earlier version of this design was described in [6]). Two web applications are integrated with the data server to form a lightweight metadata search tool. The data server contains experimental artifacts and metadata describing them, as well as an index of the metadata and a set of version controlled server snapshots. Experimental artifacts are found on the lab server, but accessible by the components shown in the application server. The metadata that describes these artifacts are stored as XML documents formatted as Apache Solr update instructions [26]. Fields within the metadata documents are recorded to match the Solr schema, enabling directed indexing. The Solr core folder contains the index as well the metadata. The version controlled snapshots are used to detect changes on the server and trigger indexing of new or modified experiment associated files.

At the client, two pages allow users to manually index and search for experiments. In order to maintain extensibility and adapt to the addition of new types of experiment related files, both pages are constructed dynamically based on indexed fields. When new fields are added to the index they become available for user input. Currently, those fields corresponding to an experiment overview are the only ones indexed, as shown in Figure 3. In order to keep newly added fields from overrunning the page, a selection drop down will provide access to the additional fields.

When a user clicks the search button, scripts gather the field-specific input and format a query that is posted to the Solr query servlet of the application server shown in Figure 2. Results are indicated by the fields that best differentiate experiments from one another. Some of the fields point to files located on the server. These server locations are formatted into links. When clicked, they call a servlet responsible for delivering the specified file to the client.

The index description editor allows researchers to manually index artifacts. Many of the functions found on the search page are also used in the editor. Additionally the editor is able to load fields from an indexed experiment and those found in an experiment overview located on the server.

Components in the application server represent components called by the client and an indexing component that responds to changes on the data server. As mentioned above, EE integrates two web applications. The first, Apache Solr, is used to manipulate and access the index. The second is composed of the remaining java servlets, which are responsible for manipulating and accessing the data server. One exception to this is the servlet responsible for logging user actions, such as queries issued and files requested. A record of files opened from links in the client is used to direct priority in the indexer.

The indexer is used to find and index new or modified artifacts on the server and is called at regular intervals. To handle frequent changes to the server, files accessed from the client are given the highest priority when responding to differences between versions of the server snapshots. In addition, a queue of outstanding un-indexed files is maintained in order to throttle server access by the indexer. This queue is produced by making comparisons between the different version controlled server snapshots found in the data server area of Figure 2. The server snapshots do not contain actual files but file system information starting at the highest level folder designated for experiment files.

As we discussed, the process of indexing new artifacts relies on conventions followed by the researchers in formatting and naming data on the server. These conventions are specified by the models in the data server area of Figure 2. The models are used by the indexer to link the artifact to the experiment overview by sample name.

B. State of Implementation

Currently, EE's search tool supports the following: searching for experiment metadata (e.g., provenance), linking experiment files included in the search results, and capturing search terms and files opened by researchers. We plan to implement the following functionality: relating files to the experiment sample name with an indexer, displaying captured relationships between experiments' files based on artifact models, and integrating a version control system to automate the capture of future artifact metadata.

VI. Evaluation

We now discuss evaluation of EE's search tool through a case study with a group at LLNL and a scalability analysis.

A. Case Study at the LLNL

An evaluation was performed within the optics inspection and analysis group at LLNL. Five subjects participated in the study, including scientists and data analysts. The files indexed by EE were spreadsheets which contain the overview of the various experiments conducted in the lab and pointers to the locations of the various files. For the study, only a subset of the overview files was indexed in EE. Prior to the study, subjects were provided with training, including documentation and video tutorials on using the tool. During the study, subjects were asked to perform a search task that they might perform while conducting their research. Feedback was obtained via interviews.

We sought answers to the following research questions:

Q1: Is EE's search tool easy to use? Can it be incorporated into the research process at LLNL?

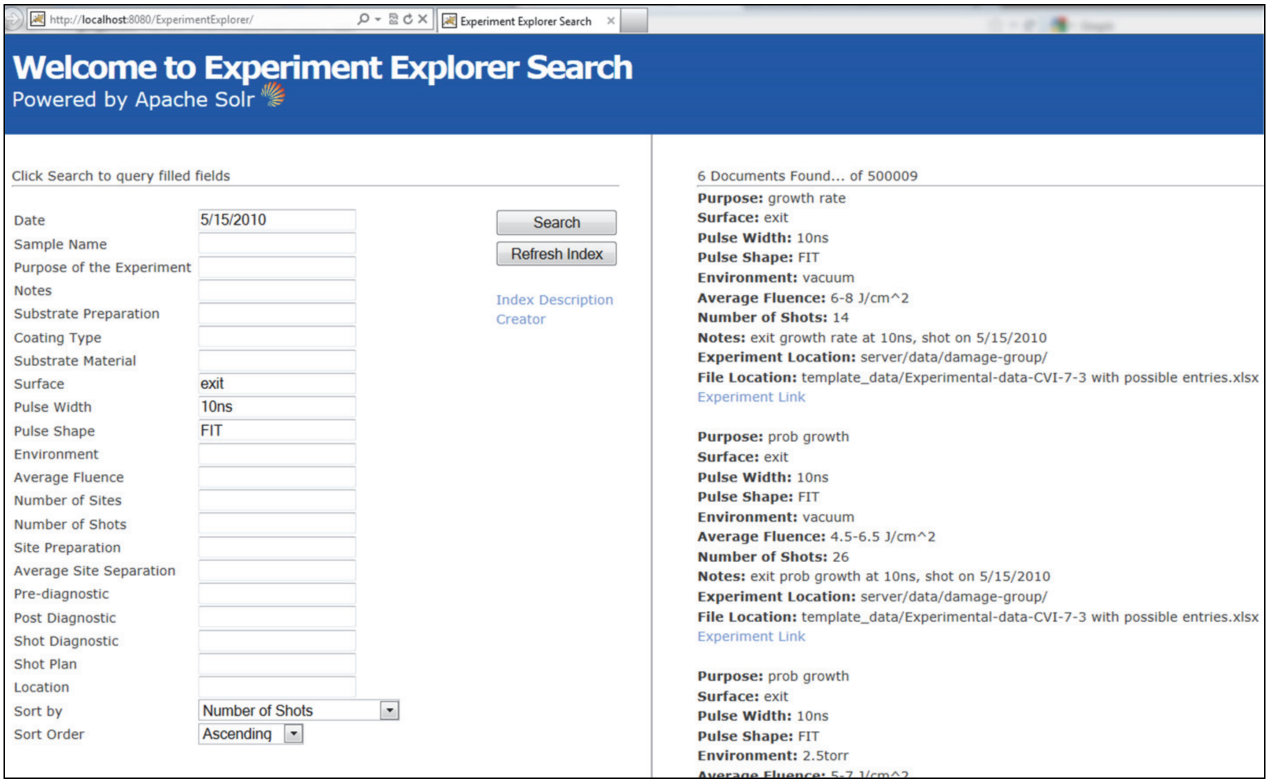
Q2: Does EE's search tool provide relevant experiment files?

Q3: Does EE enable researchers to determine which experiments were related to which files?

Q1: Subjects were asked to compare EE's search tool with previous search techniques. Subjects were asked to rate EE on a scale of 1 to 5, with 1 as exceptional and 5 as unacceptable. On average researchers rated the acceptability of time spent learning the software at 1.6, the time spent using the software at 1, and the time spent finding relevant files at 1.2. Three subjects even pointed out that the feature they like about the tool is ease of use.

The subjects were generally pleased with EE's search capability. Three of the five subjects said that they would use the tool to perform their research, while one user said that he would use the tool if his suggested changes were incorporated. Previous search techniques involved manually traversing folders on a server or asking a colleague regarding the location of files. Since there are numerous files, these previous techniques were time-consuming.

Q2: On the average, subjects found the experiment they were searching 84% of the time. When asked how often links to the actual overview files were missing, they answered 0%.


http://localhost:8080/ExperimentExplorer/
Experiment Explorer Search
Search
Refresh Index
Index Description
Creator

Welcome to Experiment Explorer Search
Powered by Apache Solr

Click Search to query filled fields

Date	5/15/2010
Sample Name	
Purpose of the Experiment	
Notes	
Substrate Preparation	
Coating Type	
Substrate Material	
Surface	exit
Pulse Width	10ns
Pulse Shape	FIT
Environment	
Average Fluence	
Number of Sites	
Number of Shots	
Site Preparation	
Average Site Separation	
Pre-diagnostic	
Post Diagnostic	
Shot Diagnostic	
Shot Plan	
Location	
Sort by	Number of Shots
Sort Order	Ascending

6 Documents Found... of 500009

Purpose: growth rate
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: vacuum
Average Fluence: 6-8 J/cm^2
Number of Shots: 14
Notes: exit growth rate at 10ns, shot on 5/15/2010
Experiment Location: server/data/damage-group/
File Location: template_data/Experimental-data-CVI-7-3 with possible entries.xlsx
Experiment Link

Purpose: prob growth
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: vacuum
Average Fluence: 4.5-6.5 J/cm^2
Number of Shots: 26
Notes: exit prob growth at 10ns, shot on 5/15/2010
Experiment Location: server/data/damage-group/
File Location: template_data/Experimental-data-CVI-7-3 with possible entries.xlsx
Experiment Link

Purpose: prob growth
Surface: exit
Pulse Width: 10ns
Pulse Shape: FIT
Environment: 2.5torr
Average Fluence: 5-7 J/cm^2

Figure 3. Experiment Explorer Search Tool

Q3: While the entire experiment artifacts were not available for the study, the overview files which were indexed, contains pointers to the experiment artifacts. According to one subject, once she obtains the experiment overview from EE's search tool, she can find a given experiment artifact at least twice as fast without the tool.

All of the interviewees agreed that the experiment overview was important information regardless of what type of exploration they performed. In terms of direct links from the results page, three of the five researchers felt that the addition of the data overview for an experiment would support a more efficient exploration. The fifth wanted all of the artifact links embedded directly into the overview file.

Discussion: In general the researchers were pleased with EE's tool support. One of the researchers said, "Like this idea of organizing and indexing my experiments. Productivity and sharing is much easier this way." Another researcher said, "It was very fast and the fields were relevant to the searches that we normally make." In addition, the results show that the tool is easy to use (with average ratings of less than 2.0). The results also indicate that all (except for one subject) would use the tool. This suggests that the tool would benefit the fast-paced adaptive research environment of the optics inspection and analysis group at LLNL.

Areas for improvement within the search software included support for range queries, richer logical support for interpreting user input, and limiting possible input for fields that have a small set of possible field data. All of the researchers expected to have the system show a preview of the most likely input as they typed, indicating that support for fast incremental exploration be supported directly from the search page. The users would also like to be able to directly link to the experiment files once an experiment sample name has been obtained from EE.

B. *Scalability Analysis*

A scalability analysis was also performed. Over five hundred thousand metadata documents were created and indexed, twenty five thousand documents at a time, on a machine with an i7 processor and 6 GB of RAM. Indexing time at each step took around 5 minutes. However, even at five hundred thousand documents, the search still performed in near real-time, taking 40-70ms to deliver results. Restarting the server increased the search time to 120ms and leveled back down to 40-70ms after 3 searches. Searching with two integer fields and a date field resulted in an average of 12 documents being returned from over 500,000 possibilities, which are the correct documents. This was done over 20 runs with different known numbers.

VII. Conclusion and Future Work

In this paper, we provided a technique for recovering provenance for data sets that have already been analyzed or processed. This technique, referred to as EE, incorporates a provenance-compatible process with a lightweight metadata search tool. EE complements existing techniques which record provenance while a data set is being analyzed. We conducted two types of evaluation to assess EE's tool support: a case study at LLNL and a scalability analysis. The case study suggests that EE is effective in connecting information which provides clues regarding the provenance of a given experiment file. The scalability analysis reveals that the tool can easily handle large sets of experiment files.

In the future, we will investigate incorporating ontologies into the metadata generation and search to enable sharing research files with researchers outside the optics inspection and analysis group. We plan to provide a visualization to help researchers more efficiently find related data sets. We will also examine other automated techniques for determining the related experiment sample name for a given experiment file. Finally, we will investigate how EE can be applied to other settings, such as recovering the provenance of software development files or the provenance of business reports.

Acknowledgment

We thank David Cross, Zhi Liao, Raluca Negres, Mary Norton, and Rajesh Raman at LLNL for valuable insights. This work is partly funded by University of Washington's Royalty Research Fund No. A65951.

References

[1] I. Altintas, O. Barney, and E. Jaeger-Frank, "Provenance collection support in the Kepler Scientific Workflow System," in *Proc of Int'l Provenance and Annotation Workshop (IPAW)*, 2006.

[2] D. Bourilkov, "The CAVES project - Collaborative Analysis Versioning Environment System, the CODESH project — COllaborative DEvelopment SHell," *International Journal of Modern Physics*, vol. A20, no. 16, pp. 3889–3892, 2005.

[3] D. A. Holland, M. I. Seltzer, U. Braun, and K.-K. Muniswamy-Reddy, "PASSing the provenance challenge," *Concurrency and Computation: Practice and Experience*, vol. 20, pp. 531–540, 2008.

[4] P. Missier, B. Ludascher, S. Bowers, S. Dey, A. Sarkar, B. Shrestha, I. Altintas, M. Anand, and C. Goble, "Linking multiple workflow provenance traces for interoperable collaborative science," in *Workshop on Workflows in Support of Large-Scale Science*, 2010.

[5] H. U. Asuncion, "*In Situ Data* provenance capture in spreadsheets," in *Proc of the Int'l Conf on e-Science*, 2011.

[6] D. B. Davis, H. U. Asuncion, and G. Abdulla, "Experiment explorer: Lightweight provenance search over metadata," in *Proc of the USENIX Workshop on the Theory and Practice of Provenance*, 2012.

[7] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo, "Managing rapidly-evolving scientific workflows," in *Proc of the IPAW*, 2006.

[8] P. Yue, J. Gong, L. Di, L. He, and Y. Wei, "Semantic provenance registration and discovery using geospatial catalogue service," in *Proc of the Int'l Workshop on the Role of Semantic Web in Provenance Mgmt*, 2010.

[9] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," in *Proc of Int'l Conf on Mgmt of Data*, 2008.

[10] L. Chen, X. Yang, and F. Tao, "A semantic web service based approach for augmented provenance," in *Proc of Int'l Conf on Web Intelligence*, 2006.

[11] M. Allen, A. Chapman, B. Blaustein, and L. Seligman, "Capturing provenance in the wild," in *Provenance and Annotation of Data and Processes*, vol. 6378, pp. 98–101, Springer Berlin / Heidelberg, 2010.

[12] U. Marjit, K. Sharma, and U. Biswas, “Provenance representation and storage techniques in linked data: A state-of-the-art survey,” *Int’l Journal of Computer Applications*, vol. 38, no. 9, pp. 23–28, 2012.

[13] A. Chebotko, X. Fei, C. Lin, S. Lu, and F. Fotouhi, “Storing and querying scientific workflow provenance metadata using an RDBMS,” in *Proc of the Int’l Conf on e-Science and Computing Grid*, 2007.

[14] M. S. Aktas, M. Pierce, G. C. Fox, and D. Leake, “A web based conversational case-based recommender system for ontology aided metadata discovery,” in *Proc of the Int’l Workshop on Grid Computing*, 2004.

[15] S. M. S. Da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso, “A provenance-based approach to resource discovery in distributed molecular dynamics workflows,” in *Proc of Int’l Conf on Resource Discovery*, 2010.

[16] S. Jensen and B. Plale, “Trading consistency for scalability in scientific metadata,” in *Proc of Int’l Conf on e-Science*, 2010.

[17] K. Belhajjame, K. Wolstencroft, O. Corcho, T. Oinn, F. Tanoh, A. William, and C. Goble, “Metadata management in the Taverna workflow system,” in *Int’l Symposium on Cluster Computing and the Grid*, 2008.

[18] C. Duda, D. Kossmann, and C. Zhou, “Predicate-based indexing for desktop search,” *VLDB Journal*, vol. 19, no. 5, pp. 735–758, 2010.

[19] J. Chen, H. Guo, W. Wu, and W. Wang, “iMecho: an associative memory based desktop search system,” in *Proc of Conf on Information and Knowledge Mgmt*, 2009.

[20] S. M. S. Da Cruz, P. M. Barros, P. M. Bisch, M. L. M. Campos, and M. Mattoso, “A provenance approach to trace scientific experiments on a grid infrastructure,” in *Proc of Int’l Conf on e-Science*, 2011.

[21] E. Dede, Z. Fadika, C. Gupta, and M. Govindaraju, “Scalable and distributed processing of scientific XML data,” in *Proc of Int’l Conf on Grid Computing*, 2011.

[22] P. Dourish, W. K. Edwards, A. LaMarca, J. Lamping, K. Petersen, M. Salisbury, D. B. Terry, and J. Thornton, “Extending document management systems with user-specific active properties,” *ACM Transactions on Information Systems*, vol. 18, no. 2, pp. 140–170, 2000.

[23] “Git.” http://git-scm.com/, Accessed Dec 13, 2012.

[24] N. Rosenblum, B. P. Miller, and X. Zhu, “Recovering the toolchain provenance of binary code,” in *Proc of Int’l Symposium on Software Testing and Analysis*, 2011.

[25] J. Estublier, D. B. Leblang, G. Clemm, R. Conradi, A. van der Hoek, W. Tichy, and D. Wiborg-Weber, “Impact of the research community on the field of software configuration management,” *Trans on Software Engineering Methodology (TOSEM)*, vol. 14, no. 4, pp. 383–430, 2005.

[26] Apache Software Foundation, “Apache Solr.” http://lucene.apache.org/solr/, Accessed Dec 13, 2012.

Security of Information System in View of Business Continuity Management

Kiyoshi Nagata
Faculty of Business Administration
Daito Bunka University
Tokyo, Japan
Email: nagata@ic.daito.ac.jp

Dieter Hertweck
Electronic Business Institute
Heilbronn University
Heilbronn, Germany
Email: dieter.hertweck@hs-heilbronn.de

***Abstract*—In any type of company, the information system is one of the core systems in order to accomplish their business objectives. Exposure of its malfunction or defect sometimes causes critical damages to the company in view of business continuity, and the company should form a plan consist of several measures to prevent, to reduce, to transfer, to avoid risks and also to recover the system. Since the company's information system is closely related to their business type and strategy, the plan should be laid considering them. In this paper, we propose a methodology to ameliorate the present state of the company's information system in business continuity perspective.**

***Keywords*-IT system; business continuity; security controls; business process management**

I. Introduction

As a core system of a company, the information system and its management system is critical, and their malfunction directly affects the business performance. There are several systems proposed for evaluation and management of the information related system, and some companies acquired a kind of certificates of the information security, such as ISO/IEC27001 or BS7799-2. These certificates may give companies a guarantee on their information management system. However, the approaches of these certification sometimes tend to formal and stereotype, and does not reflect the companies characteristics. Especially in Small or Medium-size company, the information system and its management system should be evaluated based on their own assets, activities, and strategy. Thus a self-directed, business oriented, and assets based evaluation is recommended, and some systems such as OCTAVE [2] [10], ENISA's Information Package for SMEs (Small or Medium-size Enterprise) [13], and MEHARI [12] can be applied to SMEs.

Although the business oriented, self-directed evaluation system matches the company's characteristics and strategic goal, it sometimes requires to compose a relatively small team, called an analysis team, whose members are from several important sections of the company. The analysis team leads the evaluation process by acquiring information on their system all over the company. So the top managements should be in sympathy with the importance of information security evaluation, then consensus of staff members are necessary. SMEs neither have sufficient human resources nor have diligent intention to assign them to such a job, even if they know the malfunction of their information system causes serious problem on their business performance.

In this paper, we propose a methodology for information security evaluation and management which reduces company's workload by adopting evaluation process in ENISA's Information Package for SMEs, and by referencing their security measures which are actually in OCTAVE. In order to reflect company's characteristic or business objectives, the consensus with company staffs on the evaluation values are made in several steps.

Since our methodology is business oriented, essentially asset based, we need to find out from 3 to 5 critical assets. After the specification, we propose to describe the internal process related to each of asset using business process management tool such as ADONIS. This type of system has performance indicators, and we can see the bottle neck in the total process related to the asset. Then comparing the measures should be implemented with those of the output from ENISA, we can suggest a risk mitigation plan.

The rest of this paper is organized as follows; we refer to the ENISA's Information Package for SMEs in the next section, some methods to choose critical assets are described, some references on ADONIS as a business process management tool, then the total process of our proposed methodology comes up.

II. ENISA's Information Package for SMEs

ENISA (European Network and Information Security Agency) developed and delivered the Information Package for SMEs. The method is highly structured and one can obtain a set of several controls considered to be effective to solve the organization's information security problem or to improve their current condition.

The system procedure includes four phases as follows:

Phase1. Select Risk Profile
Output: Identified risk area, risk profile table with risk level labels, organizational risk profile

Phase2. Identify Critical Assets
Output: Five most critical assets, security requirement selection table with rationales for selection

Phase3. Select Control Cards
Output: Organizational controls, asset based controls
Phase4. Implementation and Management
Output: Gaps between recommended controls and current status, risk management plans

The risk profile table in Phase1 includes only four risk area which roughly correspond to the impact classification of OCTAVE, whose details are coming up in Section V. The risk should be evaluated for each of these areas, however the evaluation is very simple and automatically performed. Here we notice that the evaluation of "legal and regulatory" is dependent on the handling level of customers' personal information defined in the EU Data Protection Law [4, pp.104-105].

The personal data means any information relating to an identified or identifiable narural person, and the details are slightly different in countries. For example, according to the German Federal Data Protection Act, the definition of the sensitive personal data are as follows;

- Racial or ethnic,
- Political opinions,
- Religious or philosophical beliefs,
- Trade union membership,
- Health or sex life.

The assets are classified into four categories, System, Network, People, and Applications. In this phase, the analysis team, a team of small number of personnel from various sectors of organization also introduced in OCTAVE, has to choose five critical assets from many of possible assets. Evaluation is done by considering the impact to the organization when "Disclosure" or "Modification" or "Loss and Destruction" or "Interrupted Access" occurs. These scenarios are just the set of outcomes appears in OCTAVE's threat profile worksheet.

Like as many other security evaluation systems recommend, assets are evaluated in the usual three perspectives, that is Confidentiality, Integrity, and Availability (CIA). In the process of choosing five critical assets, we need to evaluate each assets from each perspective, then aggregate the resulted values or establish a method for giving priorities to each of the assets according to their values. We always have this kind of problem when performing an asset-based evaluation system, also in OCTAVE, and several methods can be applied to solve this kind of problem. For instance, AHP and FSM are very popular, where pair-wise comparisons of alternatives are performed and the priority value is expressed as the weight of each alternative. By identifying five critical assets with references of security requirement in three perspectives, the security requirements selection table is completed.

The control cards choosing phase, the Phase3, has two processes. One is for the organizational control cards corresponding to the Strategic practice in OCTAVE, and the selection of controls depends only on the risk levels of each risk area described in Phase1. In TableI, SP1, SP2, SP3, SP4, SP5, and SP6 are sets of controls related to "Security Awareness and Training", "Security Strategy", "Security Management", "Security Policies and Regulations", "Collaborative Security Management", and "Contingency Planning/Disaster Recovery", respectively, [13].

The other is for the asset based control cards corresponding to the Operational practice in OCTAVE, and possible and effective controls are listed in the asset control card whose selection depends on the level of total risk profile, the asset category, and the asset's risk level. Once selecting a card, one can find out controls to be adopted according to three perspectives and security requirements of "Physical security", "System and network management", "System authentication", "Monitoring and auditing IT security", "Authentication and authorization", "Vulnerability management", "Encryption", "Security architecture and design", "Incident management", and "General staff practices".

Table I
ORGANIZATIONAL CONTROL CARDS

Risk Area	High	Medium	Low
Legal and Regulatory	(SP1) (SP4)	(SP1) (SP4)	SP1.1
Productivity	(SP3), (SP4) (SP6), (SP5)	(SP4) (SP6)	SP4.1
Financial Loss	(SP2), (SP1), (SP4)	(SP4)	SP4.1
Productivity	(SP1) (SP5)	(SP4) (SP1)	SP4.1

The last phase consists of Gap analysis and planning the risk management. From the previous phase, recommended controls are proposed and one can see the gap from currently performed controls. Then make a plan in order to fill in the gap to compromise the present risk.

III. METHOD FOR FINDING CRITICAL ASSETS

In any decision making process, choosing one or a few critical alternatives from a large set of them is an important and difficult task. There are many methods proposed from theoretical point of view, and some are applied to practical cases. Here we refer to pairwise comparison based methods, like as AHP(Analytic Hierarchy Process), FSM(Fuzzy Structural Modeling). In AHP [8] or FSM [9], weights are obtained by computing the principal eigenvector of a subordination matrix with pairwise comparison values entries. The Perron-Frobenuis theorem guarantees the principal eigenvector to be considered as the importance weight vector. In order to apply the theorem to the matrix obtained by pairwise comparison, the (j, i)-entry value should be set as the inverse value of the corresponding (i, j)-entry in AHP, and the reachability matrix should be computed in FSM. Instead of computing the principal eigenvector of reachability matrix in FSM, we have more simplified method by which the weight of evaluation factors can be found on the basis of the ratio calculation [1] [7]. The relationships between evaluation factors are transitive regarding the contextual relation “Importance degree”.

At first, put all the alternatives in sequential oder, then give values $f_{i,i+1}$ $(i = 1, 2, \ldots, n-1)$ as the importance degree of i-th alternative compared with $(i+1)$-st one, where n is the number of all the alternatives. The corresponding symmetrical value of $f_{i,i+1}$, can be calculated by $f_{i+1,i} = 1 - f_{i,i+1}$ $(i = 1, 2, \ldots, n-1)$. These values are carefully given on the basis of experience and knowledge of the decision makers and/or specialists.

From these relative comparison values, we compute the evaluation value E_i of i-th alternative so as to satisfy following ratio equations:

$$E_k : E_{k+1} = f_{k,k+1} : f_{k+1,k} \qquad (1 \leq k \leq n). \tag{1}$$

A set of answer values of the simultaneous ratio equations is given by following formulae:

$$E_k = \prod_{i=1}^{k-1}(1 - f_{i,i+1}) \prod_{i=k}^{n-1} f_{i,i+1} \quad (1 \leq k \leq n), \tag{2}$$

where the empty product is set to be 1 for $k = 1$ or $k = n$.

If the values for each $f_{i,i+1}$ are carefully chosen to satisfy the transitivity, we actually do not need other comparison value $f_{i,j}$ $(i < j)$. We have only to evaluate each alternative to the adjacent one, and the total number of essential values is just $n - 1$.

In other words, we need to be careful that the set of values $\{f_{i,i+1}\}_{i=1,\ldots,n-1}$ should be transitive, which means that the importance degree of i-th alternative to j-th one $(i < j)$ should be approximately equal to the value calculated from $\{E_i\}$ in a certain degree of error. However in some practical applications, it is not so easy to guarantee that condition, and we will propose some modified pragmatic methods.

A. Checking system to guarantee the transitivity

Supposing that $\{E_i\}$ is the set of importance weights, the relative important degree of i-the alternative to j-th alternative should satisfies a ratio equation $E_i : E_j = f_{i,j} : f_{j,i}$, which is solved to have the following formula for $i < j$;

$$\frac{E_i}{E_i + E_j} = \frac{\prod_{k=i}^{j-1} f_{k,k+1}}{\prod_{k=i}^{j-1} f_{k,k+1} + \prod_{k=i}^{j-1}(1 - f_{k,k+1})}. \tag{3}$$

When this value seems to be considerably different from the value evaluated directly, we need to reconsider initial comparison values. If we notice that the directly given value should be modified, it will be all right.

B. Averaging over All or Some of Sequences

Sometimes it may happen that values given by formulae (2) widely vary depending on the way of setting alternatives in order, and an adjustment seems to be difficult. We can take the average of the evaluation values corresponding to each sequence given by a permutation of $\{1, \ldots, n\}$. When several principal sequences can be distinguished, the average can be taken only over them. For a subset S of the permutation group S_n, the formula for the normalized averaged weight are given by the followings:

$$\begin{aligned} E_i &= \sum_{\sigma \in S} \frac{1}{t_{\sigma^{-1}}} e_{\sigma(i)\sigma^{-1}} \\ &= \sum_{j=1}^{n} \left(\sum_{\sigma^{-1} \in S, \sigma(j)=i} \frac{1}{t_\sigma} e_{j\sigma} \right), \end{aligned} \tag{4}$$

where $e_{j\sigma} = \prod_{l=1}^{j-1} f_{\sigma(l+1),\sigma(l)} \prod_{l=j}^{n-1} f_{\sigma(l),\sigma(l+1)}$, $t_\sigma = \sum_{j=1}^{n} e_{j\sigma}$, and $S^{-1} = \{\sigma : \sigma^{-1} \in S\}$.

C. Hierarchical Block-wise Computing

If it seems that there are small set of invisible attributes behind alternatives, classify them according to these attribute. Then compute weights of alternatives in each class separately, and the weight of attribute class should be calculated. The aggregation process is done by multiplying the weight of super set to that of each attribute after normalization.

The merit of this option is that we may have small number of comparison candidates, and our weight computing method will effectively work. If the number of first blocks is large, we will try to find out some attribute factors which are common to several blocks.

IV. Business Process Management Tool

In this paper, as we are concerned about risks related to each of chosen critical assets, the management process on which the asset related process are mapped should be assessed carefully. Fortunately, there are some good computerized application of business process management tool. Here, we introduce ADONIS which is developed and provided by BOC Group [11].

ADONIS is composed of several model type such as “Company map”, “Business process diagram”, “Choreography diagram”, “Conversation diagram”, “Business process model”, “Document model”, “IT system model”, “Product model”, “Working environment model”, “Risk model”, “Control model”, and “Use case diagram”, and each model includes some objects with own attributes. Among them, here we refer to following models:

Business process model

> This process model is essential model of this system containing several type of objects such as Activity, Subprocess, Decision, Performance Indicator, etc. which are combined with each other.

This can also contain Risk and its Control objects when an activity has a risk, and the risk and its control are described in each corresponding model. Figure 1 is a part of “Accept transfer model” in Example files down loadable from BOC group homepage, [11].

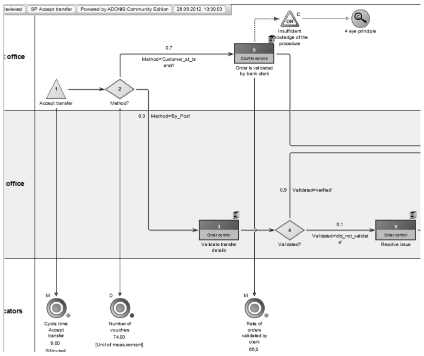


Figure 1. ADONIS Business Process Model

IT system model

This model contains Application, Service, Infrastructure element, and Operation. The model describes physical and logical relationship between objects by “has”, “uses”, “Is dependent on”, “has note”, and “has cross-reference” indicators.

Document model

This model is composed of several documents referred by personnel especially in case of emergency or trouble. The document’s attribute has a reference link to a Word, or Excel, or PowerPoint file.

Working environment model

This model contains Organizational unit, Performer, Role, Position, etc. The position and primary roles of personnel, and their command structure is embedded in this model.

Risk model

The main object of this model is Risk whose attribute has several risk types such as “operational riks”, “strategic risk”, “market price and liquidity risk”, “credit risk”, “quality risk”, and “other risk”.

Control model

The main object is Control, and the control process is described as a business process model which is referred in the attribute of this object.

V. Proposed System

Before going on the detail of our proposed methodology, we just refer to the business performance in general. In the research area of evaluation of business performance, there are several models or methodologies proposed by many researchers or consultants. But most methodologies refer to “financial”, “customer”, “productive or internal process”, and “learning and growth”, which are emphasized especially in the Balanced Score Card (BSC) [3]. Of course, BSC insists that these four perspectives should be balanced and equally treated.

Although BSC and some other methodologies for business performance are tailored not only for evaluating the performance status but also for improve the future state of business, it seems that they do not consider any collapse caused in short span. Here we are interested in business continuity when some disasters occur or serious attacks, e.g. DDoS, are exposed. In most cases, problem seemed to be serious to the business continuity occurs in productive or internal process which causes big financial impact. Thus we explicitly focus on financial and internal process perspective among these perspectives.

Now we explain our proposed methodology for enforcing company’s information system in view of business continuity. Figure 2 describes the total flow of our systematic methodology composed of four main phases.

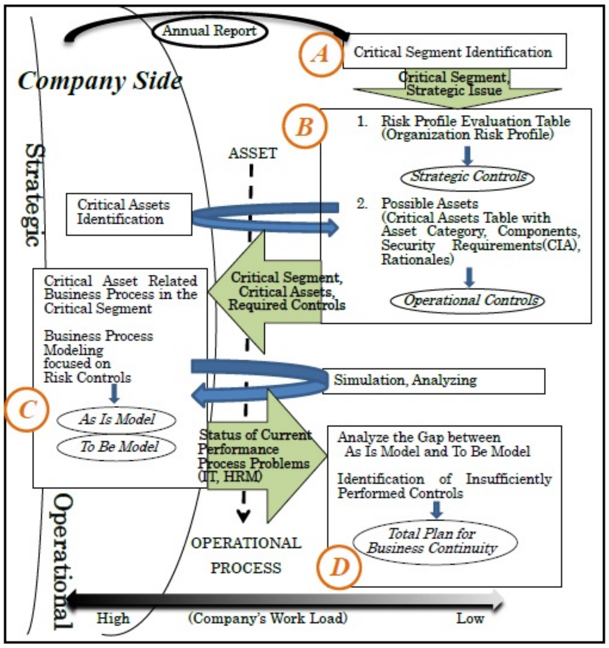

Annual Report
A
Critical Segment Identification
Company Side
Critical Segment, Strategic Issue
B
1. Risk Profile Evaluation Table (Organization Risk Profile)
Strategic Controls
ASSET
Strategic
Critical Assets Identification
2. Possible Assets (Critical Assets Table with Asset Category, Components, Security Requirements(CIA), Rationales)
Critical Asset Related Business Process in the Critical Segment
Critical Segment, Critical Assets, Required Controls
Operational Controls
Business Process Modeling focused on Risk Controls
Simulation, Analyzing
C
As Is Model
To Be Model
Status of Current Performance Process Problems (IT, HRM)
Analyze the Gap between As Is Model and To Be Model
Identification of Insufficiently Performed Controls
Operational
OPERATIONAL PROCESS
Total Plan for Business Continuity
D
High
(Company's Work Load)
Low

Figure 2. Flow of Total System

In the figure, the region on the left side is the company’s region where workload by the company’s personnel is high. The workload set in the right side region should be done by researcher’s group.

The total work flow proceeds in the following way. First of all, find out the most important and critical business segment from financial perspective and the company’s strategy referring the annual report. Next, apply ENISA system to evaluate company’s information related property which lead to strategic practices required for the company. We also need to identify a small set of critical assets which would lead a set of necessary operational practices. In the third phase, business process modeling related to each of critical asset is constructed using application software like as ADONIS. Then try to find out serious problems in the process by performing recursive simulation, considering suggestive operational and strategic practices. The final phase is a phase for making up a total risk control plan.

A. Critical Business Segment

If the company is very small and/or there is unique business segment, the critical segment might be clear. Even if the company has several segments, the critical business segment might be trivial just from the company's strategy. In these cases, this phase is seemed be unnecessary. However, we recommend to perform the total evaluation of segments in financial perspective in case that some disastrous incidents cause serious impact on the business continuity.

From the pyramid of financial ratios, ROCE(Return On Capital Employed; $= \frac{“NetProfit”}{“Capital\ employed”}$) is the starting point, then it is initially decomposed into the product of the net profit margin ($= \frac{“NetProfit”}{“Sales”}$) and the sales on capital employed ($= \frac{“Sales”}{“Capital\ employed”}$). The former index is on profitability, and the later one is again expressed as the product of the turnover of total assets ($= \frac{“Sales”}{“Total\ assets”}$) and the inverse value of the capital employed ratio($= \frac{“Total\ assets”}{“Capital\ employed”}$).

Focusing on the net profit margin derives other indices such as the break-even point ration, EBIT, EBITDA and their margins. The turnover of total assets is an index on investment effectiveness, and the CCC (Cash Conversion Cycle;$= 365 \times (\frac{“Receivable”}{“Sales”} + \frac{“Inventory”}{“Cost\ of\ sales”} - \frac{“Accounts\ payable”}{“Cost\ of\ sales”}$) is one of serious index for business continuity. The capital employed ration is an index on financial leverage, which derives DE ratio (Debt Equity ratio;$= \frac{“Debt(with\ interest)”}{“Equity”}$), and ICR (Instance Coverage Ratio; $= \frac{“Operating\ profit” + “Financial\ income”}{“Interest\ expense”}$) for example.

Fortunately almost all the information necessary to calculate them are on the company's annual report, consisting of PL, BS, Cash flow statements. The annual report also contains information on the "strategy" which helps us to find out not only the critical segment but also critical assets.

B. Applying ENISA

This phase includes main two process. First one is the evaluation of company's, or segment's, proper condition by the risk profile evaluation table, then the process output a set of strategic practices from the organizational(Strategic) control cards. Second process is choosing a small set of critical assets in the assigned segment. Then the asset based(Operational) control card table is referred to have a set of controls according to the evaluation value of each asset from CIA points of view.

1) Risk Profile Evaluation Table: ENISA's risk profile evaluation table is very simple composed of four risk areas, and the evaluation is fairy automatically done as follows.

Legal and Regulatory

"High" is marked if the company's business handles customer information of sensitive and personal nature including medical records and critical personal data as defined by the EU Data Protection Law. "Medium" if the handled customer information is not sensitive. "Low" if the business does not handle customer information.

Productivity

"High" is marked if the business employs more than 100 employees having a daily need to access business applications and services. "Medium" if the number of such employees is between 50 and 100. "Low" if the number is less than 50.

Financial Stability

"High" is marked if yearly revenues are of excess of 25 million Euros or/and financial transactions with 3rd parties or customers are taking place as part of the business as usual process. "Medium" if the yearly revenue are between 5 million and 25 million Euros. "Low" if the revenue are less than 5 million Euros.

Reputation and Loss of Customer Confidence

"High" is marked if unavailability or service quality directly impact business profile or/and more than 70% of customer base have online access to business products and services. "Medium" if the impact is indirect and/or less than 5% of customer base have online access. "Low" if there are no impact on business profile or on loss of revenue.

According to the set of evaluation values, strategic controls are determined from the Table I in Section II.

2) Critical Assets: Task of choosing a few, at most five, critical assets is very important and difficult. As we mention in Section II, it is performed generally considering the case of "Disclosure" or "Modification" or "Loss and Destruction" or "Interrupted Access" from view point of each of CIA. When focusing on the business continuity, we should mainly consider the "Loss and Destruction", and the "Interruption of Access" to assets.

First listing possible information related assets in the business segment, then choose any two of them and compare them as the important level with values in the open interval $(0, 1)$ from each of CIA point of view. If we apply the ratio based method, the number of pairs to be evaluated is just $n - 1$, but adjustment or modification process might be needed.

After calculating the set of three weight vectors, they are aggregated according to the importance degree of perspectives, CIA. Although the confidentiality has high degree in usual information evaluation activity because of nowadays increasing concerns on personal data protection, the availability should be the highest from the business continuity point of view.

Once a few critical assets are distinguished, the asset base control card is assigned according to the asset category, “Application”, or “System”, or “Network”, or “People”, and the company’s risk level (“High”, “Medium”, or “Low”). We also need to find out components related to each of critical asset and security requirement as confidentiality, and/or integrity, and/or availability.

C. Business Process Mapping

Before starting the business process mapping using an application soft ware, list up possible impacts on each of critical asset caused by any disaster or threats. Then map all the components to the application’s model. For example, the total process for management or recovery of the asset related system is described by the Business Process Model whose activity is dependent on IT system and HRM system.

If the company has own model for this process, map it on the application and give the precise information on processing time, possibilities of decision, working time of each employee, and performance indicators should be carefully constructed. Once the process model is completed, several cases are simulated and results are stored for the next phase.

D. Risk Control/Mitigation Plan

In the last phase, we make a plan for business continuity considering the result comes out of the previous phase. When the company has formalized plan and process against critical asset affecting exposure of risk, we will set up the total plan which reinforcing the process using the result from the simulation and controls as the output of ENISA system.

In case of no current effective process, we need to establish it considering the company’s IT and HR condition which are already investigated in the previous phase.

The risk is usually treated in four types of way; retention, reduction, transfer, and avoidance. It is very important to investigate which activity has the key indicator mainly affect the business continuity, then determine controls in one or some of these types.

VI. Conclusion and Future Work

In this paper, we proposed a systematic methodology for evaluation of current status of critical business factor and for suggesting effective risk controls and mitigation plan. Since our main concern is the business continuity, we have considered the total business performance evaluation system, and try to incorporate the financial factor as the most critical business segment. As we mainly concern about the reduction of company’s workload, a simple and systematical method, ENISA for SMEs, is adopted for the preliminary evaluation. Although usual security evaluation is essentially based on the asset based method, we also focus on the process against the exposure of risks, and propose to use a kind of business process management application whose simulation functions help us to review or compose an effective risk control/mitigation plan.

Instead of ENISA, we might use OCTAVE-S, a version for relatively small enterprises, if some method for extracting effective mitigation controls are established, see [5] [6]. We will apply our methodology to some of real company and see how it works.

References

[1] M. Amagasa, *Performance Measurement System for Value Improvement of Services*, Bulletin of The Australian Society for Operations Research Inc., Vol.29, No.1, pp.35-52, 2010.

[2] C. Alberts and A. Dorofee, *Management Information Security Risks*, Addison-Wesley, 2003.

[3] R. S. Kaplan, D. P. Norton,*The Balanced Scorecard-Measures that Drive Performance-*, Harvard Business Review. Vol. 70, No.1, pp.71-79. 1992.

[4] C. Kuner, *European Data Protection Law*, Oxford University Press, 2nd ed. 2007.

[5] K. Nagata, Y. Kigawa, D. Cui, and M. Amagasa, *Method to Select Effective Risk Mitigation Controls Using Fuzzy Outranking*, Proceedings of the 9th International Conference on Intelligent Systems Design and Applications, pp. 479-484, 2009.

[6] K. Nagata, *On Clustering of Risk Mitigation Controls*, Proceedings of 2011 International Conference on Network-Based Information Systems, pp. 148-155, 2011.

[7] K. Nagata, M. Amagasa, and H. Hirose, *Multi-attribute Decision Making Based on Fuzzy Outranking*,Proceedings of 13th IEEE International Symposium on Computational Intelligence and Informatics, pp.169-174, 2012.

[8] T. L. Saaty, *Decision Making for Leaders; the Analytical Hierarchy Process for Decisions in a Complex World*, Wadsworth, Belmont, Calif., 1982.

[9] E. Tazaki and M. Amagasa, *Structural Modelling in a Class of Systems Using Fuzzy Sets Theory*, International Journal of Fuzzy Sets and Systems, Vol.2, No.1, pp.87-103, 1979.

[10] C. Alberts, A. Dorofee, J. Stevens, and C. Woody,(2005). *OCTAVE-S Implementation Guide*, Version 1.0, CMU/SEI-2003-HB-003. Available from http://www.cert.org/octave/octaves.html, 18.12.2012.

[11] *ADONIS Community Edition: Taking BPMN 2.0 one step further*. Available from http://www.adonis-community.com/, 18.12.2012.

[12] *MEHARI 2010: Fundamental concepts and functional specifications*. Available from http://www.clusif.asso.fr/fr/zz production/ouvrages/type.asp?id=METHODES, 18.12.2012.

[13] *Risk Management: Information Package for SMEs*. Available from http://www.enisa.europa.eu/act/rm/cr/risk–management–inventory/downloads, 18.12.2012.

Exploring the Effect of Cognitive Map on Decision Makers' Perceived Equivocality and Usefulness in the Context of Task Analyzability and Representation

Soon Jae Kwon
Department of Business Administration
Daegu University
Kyong San 712-714, Republic of Korea
kwonsj72@gmail.com

Emy Elyanee Mustapha
Department of Business Administration
Daegu University
Kyong San 712-714, Republic of Korea
elyanee@gmail.com

***Abstract*—Cognitive map (CM) has been widely accepted as a robust decision support mechanism with which decision makers can analyze causal relationships existing among relevant variables, and represent tacit knowledge explicitly in a form of causal relationships. In literature, there are many successful cases with CM. Nevertheless, there is no study that clearly investigates the potentials of CM in reducing decision maker's perceived equivocality, and enhancing perceived usefulness. To pursue the research objective like this, we organized an experiment in which participants are given two types of tasks (analyzable vs less-analyzable) and two types of task representation (text-based vs CM-based). Results clearly showed that the CM can provide significantly improved performance in decision making support.**

Keywords-cognitive map; task analyzabilit; perceived equivocality; perceived usefulness

I. INTRODUCTION

Cognitive map (CM) is used to capture perception of decision makers (DMs) faced with complex and unstructured decision problems. Many relevant literatures showed that CM can be used for solving many kinds of decision problems [1], [2], [3], [4] most of which belong to unstructured decision problems. And also, CM can describe and facilitate elaboration of real world for individuals. Elaboration is the cognitive process whereby individuals consciously or subconsciously establish paths between nodes in a semantic network representing newly learned material and nodes representing already known material [5.

This study attempt to verify that CM is an effective methodology by conducting an experiment which examines that CM is more effective than Text under analyzable tasks by comparing CM with Text. And further analyzes the difference in problem solving between analyzable tasks and less-analyzable tasks within the framework of equivocality of information in CM and Text. The focus is that that while CM method knowledge is important in solving all such tasks, the role of application domain knowledge is contingent upon the type of understanding task under investigation. We use the theory of cognitive fit to establish theoretical differences in the role of application domain knowledge among the different types of schema understanding tasks.

II. THEORY AND HYPOTHESIS

A. *Semantic Network and Problem Solving*

CM proposes that individuals will be able to better understand domain knowledge that complies with its criteria. This is supported by two bodies of theories of cognition. Firstly, the semantic network theory proposes that CM lead analysts to construct efficient mental representations of a domain [6]. Semantic network theory states that individuals store concepts in memory as nodes connected by paths [7]. In order to perform cognitive tasks, individuals must recall concepts from memory; which follows a process of spreading activation: a node is primed in memory, which leads to paths connecting to it being activated [7]. Activation has to be strong enough for a search to reach a connected node. Empirical tests show that greater activation strength enables faster and more accurate recall [7]. CM leads to efficient mental representations by reducing activation strength and excluding relevant nodes.

Secondly, the problem-solving theories suggest that the quality of a person's mental representation of a domain is a key driver of his/her ability to reason about the domain [8]. Specifically, problem solving theories suggest that a person reasons about a domain is by drawing on his/her mental representation of the domain together with his/her mental representation of the problem s/he faces about the domain to construct a "problem space" in memory [8]. Tests show that problem solving performance is driven by a person's ability to search his/her problem space [8], [9]. Since semantic network theory suggests that CM leads to efficient mental representations, we can therefore propose that CM reduce analysts' ability to construct inefficient problem spaces in memory and thereby increase analysts' ability to search their problem space when reasoning about the domain.

B. *Conceptual Schema Understanding Task*

Schema understanding tasks can be viewed as either read-to-do (with access to the schema) [10] or read-to-recall tasks (without access to the schema) [11]. Recall tasks have been used to investigate problem solvers' knowledge structures, that is, chunks of knowledge that are stored in internal memory and reused when appropriate [12].

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-254-7

Two types of comprehension tasks that have been employed in prior Information Systems (IS) researches are supported in the education literature, which identifies two different types of knowledge, syntactic and semantic [13], [14]. Syntactic knowledge involves understanding the vocabulary specific to a modeling formalism and syntactic comprehension tasks are those that assess the understanding of just the syntax of the formalism associated with a schema. Semantic knowledge involves understanding the meaning, or the semantics, of the data embedded in the conceptual schema. Thus, semantic comprehension tasks are those that assess the understanding of the data semantics conveyed through constructs in the schema [15]. More recently, researchers have investigated tasks that require a deeper level of understanding than comprehension tasks, tasks that are referred to as problem-solving tasks [16].

C. *Cognitive Fit*

The notion of task-technology cognitive fit is viewed as an important factor determining whether the use of technology would result in performance improvement [17], [18], [19]. Briefly, the task-technology fit hypothesis argues that for an IS to have a positive impact on performance, it must be designed and utilized in such a way that it fits with the tasks it supports. When the information emphasized by the presentation matches the task, DMs can use the same mental representation and decision processes for both the presentation and the task, resulting in faster and more accurate solutions [19]. When a mismatch occurs, one of two processes will occur. Firstly, DMs may transform the presented data to better match the task, which might increase the time needed and might decrease accuracy because any transformation can introduce errors [19]. Secondly, DMs may adjust their decision processes to match the presentation [20], decreasing accuracy and increasing time because the information does not match the ultimate needs of the task.

To better understand this relationship, we first need to explain the key concept equivocality of information. High equivocality means confusion and lack of understanding [21]. Note that at times the literature uses the term equivocality to describe the characteristics of tasks. In this paper, the term exclusively uses to describe information characteristics. Furthermore, less-analyzable task is consists of syntactic and semantic knowledge. By contrast, problem solving task is presented as analyzable task. Therefore, we will examine whether quality in decision making can be changed by the task type (analyzable vs less-analyzable) of Text and CM and its equivocality.

D. *Hypotheses*

In this study, three hypotheses will be verified through one experiment. In this experiment, CM-based method is proposed to be more effective than text-based method under analyzable tasks by comparing CM-based method with Text-based method. This was done by analyzing the difference in problem solving between analyzable tasks and less-analyzable tasks within the framework of equivocality of information in CM and Text. For analyzable tasks, since the information needed to perform the task is known and clear guidelines about how to perform the task exist, the DM does not have to rely on subjective judgments or contextual information to interpret the situation or task. CM can capture perception of decision makers' knowledge in real world, and describe and facilitate elaboration. CM is to support a "what-if" and "goal seeking" analysis. In this regard, CM-based method can decrease the equivocality. It is because CM-based method can provide more accurate information than text-base method in solving analyzable tasks. Therefore, a CM-based representation is more effective than a text-based representation in supporting the information needs of analyzable tasks.

H1: For analyzable tasks, the CM-based representation, when compared to the text-based representation, will lead to a lower level of perceived equivocality.

In this study, less-analyzable task is consists of syntactic and semantic knowledge. In other words, less-analyzable task means that DMs need knowledge of surface level that asks simply the true and the false of the facts. In this case, it is assumed that there will be little difference between Text-based method and CM-based method.

H2: For less-analyzable tasks, there will be no difference between the multimedia and the text-based representation in terms of the perceived equivocality level.

Davis and his colleagues [22], [23] observed that if users perceive a system to be useful, they are more likely to use it. Other studies [24], [25] also found further support for the impact of perceived usefulness on system use. These studies established the theoretical and practical importance of perceived usefulness. Extending from the research in perceived usefulness, one can argue that users only perceive the system to be useful if the system helps them to perform the tasks it was designed for.

H3: The CM-based representation will be perceived as more useful than the text-based representation.

III. OVERVIEW OF THE EXPERIMENTS

The experiment involved less-analyzable tasks and analyzable tasks, and two representations, text-based and CM-based information, representing two levels of richness. The two were equivalent in terms of text and diagram information content. It is hypothesized that the task required only surface-level understanding of the domain. Thus, it is predicted that the text effect would dominate elaborative and inferential diagram effects. Participants who used CM would therefore outperform participants who used text only. In analyzable tasks, it is hypothesized that this task requires a deep level understanding of a domain if it is to be performed effectively. Thus, it is predicted that the elaborative and inferential effects would dominate the text effect. Participants who used causal map only would therefore outperform participants who used text. The results supported this prediction.

IV. EXPERIMENT

To test the first hypotheses, we conducted a laboratory experiment. The experiment employs a 2 x 2 x 2 design. The within-subject factors are representation type (CM-based vs. text-based representation) and task type (analyzable vs. less-analyzable task). The between subject factor is the order (text-CM vs. CM-text).

A. Task Setting

1) Representation Type

Two representations, text-based and CM-based information, representing two levels of richness were compared. The two were equivalent in terms of text and diagram information content. Appendix A shows the corresponding subjects the text-based method when the same selection was made. In both methods, subjects could reexamine the information which was presented with CM-based method and text-based method.

2) Analyzable and Less-analyzable Task Type

In this study, the less-analyzable task consisted of syntactic and semantic comprehension tasks. By contrast, problem solving task was presented as analyzable task. The task employed consisted of evaluating the ambiguity/equivocality level of the information relating to 17 statements. Subjects were asked to evaluate the degree to which they felt that the information needed to evaluate the 17 statements, as provided by the method, was equivocal. Ten of these statements were related to facts that were surface-level understanding stated in the method; these formed the less-analyzable task. The other seven required subjects to make judgments about the deep-level understanding of problem solving task stated in the method; they formed the analyzable task. All subjects performed both tasks.

3) Order

Subjects performed the experimental task twice: once with the text-based method and once with the CM-based method. Half of the subjects were randomly assigned to use the text-based method first (text-CM condition) while the other half started with the CM-based system (CM-text condition). Helson's Adaptation-Level Theory [26], [27] suggests that a subject's response to a judgmental task depends on three things: (1) sum of the subject's past experiences, (2) the context or background (for making comparison judgments), and (3) the stimulus given (the representation type in this study). To the extent that there is no context/background given, the subject will make a judgment using the sum total of all his/her previous experiences about what he/she perceived as ambiguous. Given that each of the subjects has different experiences, when no context/background is provided there is no common frame of reference to make a judgment. The closer a context is provided to the judgment, the more it will be made within that context rather than based on the sum of all past experience. Following Helson's argument, we used the first representation to allow our subjects to establish a frame of reference. The second representation was then presented and subjects were asked to evaluate the ambiguity level of the second as compared to that of the first. To take into consideration the potential learning effect, as is customarily done, we used a counter-balanced design by asking half of the subjects to first evaluate using the text-based and the other half to first evaluate using the CM-based representation.

B. Participants

The participants are college students who took one or two of the five undergraduate computer science courses offered by the School of Business Administration. The participants were organized into two groups. The 34 pre-test responses from the first group were used to pretest the manipulation check between analyzable and less-analyzable tasks. Pair t-test was conducted to check analyzability after reviewing 17 analyzable and less-analyzable tasks with students. The results show that there is a difference between analyzable and less-analyzable tasks ($t(33) = 4.52$, $p<0.001$). These results suggest that the experimental manipulation between analyzable and less-analyzable tasks was successful.

The 64 responses of the second group were used to prove research hypotheses. They participated on a voluntary basis following the instruction that bonus points will be given for those who completed surveys successfully within the time limit. Of the respondents, 38 were male and 26 were female. On average, they were 24.5 years old, and they used the Internet for 21 hours per week.

C. Dependent Variables

Two dependent variables, perceived equivocality and perceived usefulness of the system, were used in the analyses.

1) Perceived Equivocality

Two Likert-type scales, adapted from [28], were used to measure perceived equivocality of the information used for evaluating the 17 statements which comprised the experimental task. Since this study focused on the equivocality level of information rather than solution, one item was not applicable to this study and was dropped. The internal reliability of the original instrument, as reported by Daft and Macintosh, was 0.73. The modified scale used in this study has a higher reliability score (0.86). For each of the 17 statements, subjects were asked to indicate (1) if the information used to evaluate the statement could be interpreted in several ways, and (2) if the information used to evaluate the statement could mean different things to different members of the website design team. The response scale ranged from –3 (full disagreement) to +3 (full agreement). For items (1) and (2), the higher score meant higher perceived equivocality.

2) Perceived Usefulness

The 10 item scale proposed by [22] was adapted for this present study. The reliability of the instrument, as originally reported by Davis, was 0.92. Subjects were asked to rate the perceived usefulness of the second method relative to using the first method.

D. *Experimental Procedures*

Subjects were run through the experiment one at a time. Prior to the experiment, subjects were trained on how to use the CM method. The training session lasted about 30 minutes. Next, subjects were handed a task description (see Appendix A) and the questionnaire for this experiment, which contained 17 statements. They were told to read through the questionnaire before examining the information in the method. This procedure was used to help them focus on the information needed to respond to the 17 questions. Depending on their assignment, subjects first used either the text-based method or the CM-based method to examine the described with information. After they had completed the questionnaire, the subjects were given a five minute rest break. After the break, subjects proceeded to second experiment. Second experiment was a repeat of first experiment, except that subjects used the other method (either text-based or CM-based system). At the end of the experiment, subjects were given the perceived usefulness questionnaire. The entire experiment took about an half and hour to complete.

E. *Results*

Data associated with perceived equivocality was analyzed using a repeated-measures ANOVA test with the three independent variables, representation, task, and order. Table I reports the results. The mean values and standard deviations are shown in Table II.

TABLE I. RESULTS OF THE REPEATED-MEASURES ANOVA FOR PERCEIVED EQUIVOCALITY LEVEL

	DF	Mean Squared	F-value	P
Between-Subjects				
Order Type	1	1.071	0.443	0.508
Error (Order Type)	62	2.419		
Within-Subjects	192			
Task Type	1	80.492	80.073	0.000***
Task × Order	1	9.492	9.443	0.003**
Error (Task)	62	1.005		
Representation type	1	38.392	19.773	0.000***
Representation × Order	1	35.750	18.413	0.000***
Error (Representation)	62	1.942		
Task × Representation	1	28.226	38.323	0.000***
Task × Representation × Order	1	2.507	3.404	0.070
Error (Representation × Task)	62	0.737	-	-

There are several significant outcomes: task (F = 80.073, p < 0.000), representation (F = 19.773, p < 0.000), representation x order (F = 18.413, p < 0.000), and task x representation (F = 38.323, p < 0.000). The focus is on the task x representation interaction effect, which provides direct evidence for testing H1 and H2.

TABLE II. MEANS (STANDARD DEVIATIONS) FOR PERCEIVED EQUIVOCALITY LEVEL

Analyzable Task		**Less-Analyzable Task**	
First Rating	***Second Rating***	***First Rating***	***Second Rating***
CM: -2.08 (1.07)	Text: -0.13 (1.45)	CM: -1.96 (0.77)	Text: -2.1 (0.99)
Text: 0.31 (1.24)	CM: -0.62 (1.57)	Text: -1.23 (1.28)	CM: -1.66 (1.29)

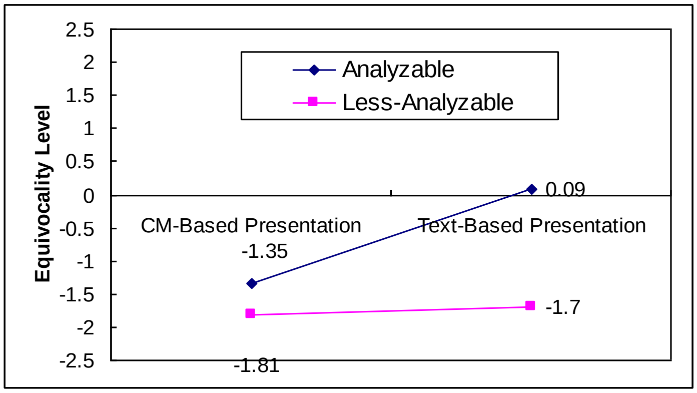


Figure 1. Task x Representation Interaction Effect

The task x representation interaction effect is depicted in Figure 1. For analyzable tasks, the perceived equivocality ratings (summarized over the first and second set of ratings) associated with CM is lower than that of text (–1.35 for CM and 0.09 for text; t = 5.93, p = 0.000). This supports H1, which states that for analyzable tasks, CM-based representation will lead to lower level of perceived equivocality than the text-based representation. For less-analyzable tasks, the perceived equivocality rating associated with the two representations are about the same (-1.81 for CM and -1.7 for text; t = 0.536, p < 0.594). This supports H2, which states that for less-analyzable tasks, the CM-based representation, when compared to the text-based representation, will be no difference in level of perceived equivocality.

The order x representation interaction effect shows that the perceived equivocality reduced only when the CM-based representation was used after the text-based representation (t = 4.01, p < 0.000) but not vice versa (t = 0.78, p = 0.542). Consistent with H1 and H2, this suggests that only the CM-based representation led to lower perceived equivocality.

Recall the earlier discussion of Helson’s Adaptation Theory [26], [27]. How then can we interpret these findings within the context of this theory? Given that the first ratings were made in the absence of an established frame of reference for comparison, based on Helson, we would expect that the theoretical differences between text-based and CM-based representations will be less evident for the first set of ratings, but more so for the second set, as they were made with a clear basis for comparison. In short, the differences as delineated in H1 and H2 should be more evident in the second ratings than the first.

Therefore, for H1 (analyzable tasks), we expect to observe significant differences in the second ratings between CM-based and text-based representations. For the first pair of ratings, there is a major difference (0.304 for text vs. -0.28 for CM; t = 8.23, p < 0.000). However, such differences are expected to be weaker in the second ratings, due to lack of a frame of reference. For the second pair of ratings (-0.127 for text vs. -0.621 for CM; t = 1.292, p = 0.200), there was a no significant difference between the text and CM conditions.

For H2 (less analyzable tasks), as stated previously, we will not expect to see any differences between CM and task in either the first or second ratings. But, for less analyzable tasks, for both the "text-CM" and the "CM-text" conditions, the first equivocality ratings are significant difference between the text and CM conditions (-1.295 for text versus –1.955 for CM; t = 2.451, p = 0.02). For the second time (-2.10 for text and –1.661 for CM; t = 1.53, p = 0.13), the equivocality level are no significant difference between the text and CM conditions. t-test on the aggregated score on perceived usefulness.

TABLE III. RESULTS OF THE INDEPENDENT SAMPLE T-TEST ON PERCEIVED USEFULNESS

Perceived Usefulness	**Representation Type**		**t-Value for difference (Text-Based vs CM-based)**	**p-Value**
	Text-Based representation	***CM-Based representation***		
n	32	32		
Mean	1.55	5.89	18.174	0.000***
S.D	0.56	1.23		

The results of the independent sample t-test on perceived usefulness, together with the means and the standard deviations for the two conditions, are summarized in Table III. Subjects perceived the CM-based representation as being more useful than the text-based representation in helping them to perform the task (t(62) = 18.174, p < 0.000; mean score 5.89 vs. 1.55, for the range of 1 to 7). The results on perceived usefulness support H3, which states that the CM-based representation will be perceived as more useful than the text-based representation.

V. DISCUSSION AND CONCLUDING REMARKS

In this study, we observed a task cognitive fit relationship with regard to perceptions of equivocality. For analyzable tasks, only CM-based representation led to lower perceived equivocality levels. When subjects were given a second representation to perform the task, only those subjects who used CM method as the second representation reported a level of perceived equivocality that was lower than that reported after the first representation was used. This result indicates that conventional text-based representation is inferior in reducing equivocality for analyzable tasks compared to CM-based representation. For less-analyzable tasks, whether subjects use a text-based representation or a CM-based representation for the second task, their perceived equivocality level is the minor difference. This is because both representations are effective in conveying the information needed to perform less-analyzable tasks. However, it should be emphasized that this conclusion is based on subjects' self-reported perceived equivocality levels rather than actual task performance, which is often difficult to measure when dealing with less-analyzable tasks [20]. Overall, we set forth to test our theory in the context of an individual decision maker interacting with a CM method to decision that was previously experienced and found support for the theory.

There are two limitations of this experiment that warrant further discussion and need to be kept in mind when interpreting the results observed. The first limitation relates to the two tasks used in the experiment. The two tasks were chosen based on the construct level definition of task analyzability and to maximize the treatment effect variance [30]. Since these are two specific operationalizations of the construct, more tasks need to be tested to further validate the theory. The second limitation relates to the choice of the two representations used. To maximum the treatment variance, we chose to use text and CM in our operationalization of the rich versus lean representation construct. As such, these operational definitions only represent two specific instances. One interesting future research direction is to test a system that has various combinations of text and CM-based representation.

ACKNOWLEDGMENT

This research was supported by the Daegu University Research Grant, 2012.

REFERENCES

[1] Lee, K.C. and Kwon, S.J., "The use of cognitive maps and case-based reasoning for B2B negotiation", Journal of Management Information Systems, 22 (4), 2006, pp. 337-376.

[2] Clarke, L. and Mackaness, W., "Management 'Intuition': An Interpretative Account of Structure and Content of Decision Schemas Using Cognitive Maps", Journal of Management Studies 38 (2), 2001, pp. 147-172.

[3] Liu, Z.Q. and Satur, R., "Contextual fuzzy cognitive map for decision support in geographic information systems", IEEE Transactions on Fuzzy Systems 7 (5), 1999, pp. 495–507.

[4] Satur, R. and Liu, Z.Q., "A Contextual Fuzzy Cognitive Map Framework for Geographic Information Systems", IEEE Transactions on Fuzzy Systems 7 (5), 1999, pp. 481-494.

[5] Bradshaw, G.L. and Anderson, J.R., "Elaborative encoding as an explanation of levels of processing", J. Verbal Learning and Verbal Behavior 21, 1982, pp. 165–174.

[6] Collins, A.M. and Quillan, M.R., "Retrieval time from semantic memory", J. Verbal Learn. Behavior 8, 1969, pp. 240–247.

[7] Ashcraft, M.H., Cognition. Prentice Hall, Upper Saddle River, NJ, 2002.

[8] Newell, A. and Simon, H.A., Human Problem Solving. Prentice Hall, Englewood Cliffs, NJ, 1972.

[9] Pretz, J.E., Naples, A.J. and Sternberg, R.J., Recognizing, defining, and representing problems. J. E.Davidson, R. J. Sternberg, eds. The Psychology of Problem Solving. Cambridge University Press, Cambridge, U.K., 2003, pp. 3–30.

[10] Khatri, V., Ramesh, V., Vessey, I. Clay, P. and Park, S.J., “Understanding conceptual schemas: Exploring the role of application and IS domain knowledge”, Inform. Systems Res 17 (1), 2006, pp. 81–99.

[11] Burkhardt, J.M., Détienne, F. and Wiedenbeck, S., „Object-oriented program comprehension: Effect of expertise, task and phase”, Empirical Software Engineering, 7(2), 2002, pp. 115–156.

[12] Bodart, F., Sim, M., Patel, A. and Weber, R., “Should optional properties be used in conceptual modelling? A theory and three empirical tests”, Information Systems Research 12 (4), 2001, pp. 385–405.

[13] Schneiderman, B. and R.E. Mayer., “Syntactic/semantic interactions in programmer behavior: A model and experimental results” Internat. J. Comput. Inform. Sci 8, 1979, pp. 219–238.

[14] Mayer, R. E., Thinking, Problem Solving, Cognition. W.H. Freeman and Company, New York, 1991, pp. 560–578.

[15] Elmasri, R. and S. B. Navathe., Fundamentals of Database Systems, 2nd ed. Benjamin/Cummings Publishing Co., Redwood City, CA, 1994.

[16] Gemino, A., “Empirical comparisons of animation and narration in requirements validation”, Requirements Engineering 9 (3), 2004, pp. 153–168.

[17] Goodhue, D.L. and Thompson, R.L., “Task Technology Fit and Individual Performance”, MIS Quarterly 19, 1995, pp. 213-236.

[18] Tan, J.K.H. and Benbasat, I., “The Effectiveness of Graphical Presentation for Information”, Decision Sciences 24, 1993, pp. 167-191.

[19] Vessey, I., “Cognitive Fit: A Theory-Based Analysis of the Graphs Versus”, Decision Sciences 22, 1991, pp. 219-240.

[20] Perrig, W. and Kintsch, W., „Propositional and situational representations of text”, Journal of Memory and Language, 24, 1985, pp. 503-518.

[21] Daft, R.L., Lengel R.H. and Trevino, L.K. “Message Equivocality, Media Selection and Manager Performance”, MIS Quarterly (11), 1987, pp. 355- 364.

[22] Davis, F.D., “Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology”, MIS Quarterly 13, 1989, pp. 319-340.

[23] Davis, F.D., Bagozzi, R.P. and Warshaw, R.P., “User Acceptance of Computer Technology: A Comparison of Two Theoretical Models”, Management Science (35), 1989, pp. 982-1003.

[24] Venkatesh, V., Morris, M.G., Davis, G.B. and Davis, F.D., “User acceptance of information technology: toward a unified view”, MIS Quarterly 27 (3), 2003, pp. 425-478.

[25] Taylor, S. and Todd, P., “Assessing IT usage: the role of prior experience”, MIS Quarterly 19 (4), 1995, pp. 561-570.

[26] Mayer, R.E. and Gallini, J.K., “When is an illustration worth a thousand words”, J. Ed. Psych 82, 1990, pp. 715-726.

[27] Helson, H., Adaption-Level Theory, Harper & Row, New York, 1964.

[28] Streitfeld, B. and Wilson, M., “The ABCs of Categorical Perception”, Cognitive Psychology 18, 1986, pp. 432-451.

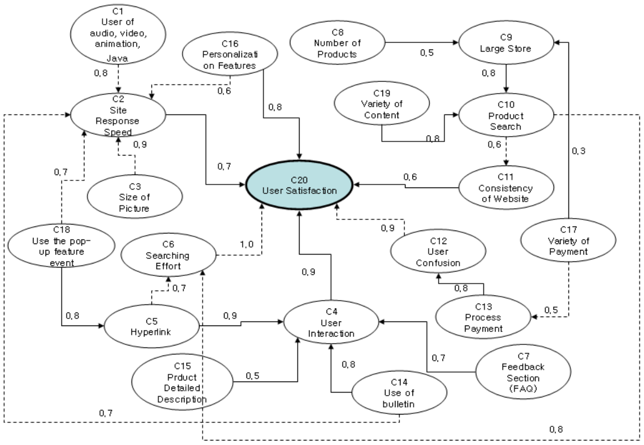
[29] Daft, R. L. and Macintosh, N. B., “A Tentative Exploration into the Amount and Equivocality of Information Processing in Organizational Work Units”, Administrative Science Quarterly 26, 1981, pp. 207-224.

[30] Cook, T.D. and Campbell, D.T., Quasi Experimentation: Design and Analysis Issues for Field Settings. Houghton Mifflin, Boston, MA, 1979.

APPENDIX A. CM-BASED METHOD

Graph below [Appendix_Figure 1] shows a variety of factors affecting user’s satisfaction in a web store that was derived through interviews of the experts. The graph below also shows a variety of possibility outcome.

[Appendix_Figure 1] Web site design and usability assessment of cognitive



The table below shows a variety of factors affecting user’s satisfaction in a web store that was derived through interviews of the experts. In addition, [Appendix_Table 4-1] displayed the input node for various values, and this value shows a variety of possibility outcome.

[Appendix_Table 1] Table of Website Development

Audio, video, animation, Java-enabled node(C1)		**Size of Picture (C3)**		**Feedback Section (C7)**	
1st use per page	0.6	Big	1.0	Within 24 hours	0.7
2nd use per page	0.7	Medium	0.8	Within 1 to 2 days	0.8
3rd use per page	0.8	Small	0.6	Within 2 to 5 days	0.9
4th use per page	0.9	-	-	After 5 days	1
Personalization Features (C16)		Using events popup		Number of Products	
Total 1st use	0.3	1 number	0.5	Less than 100	0.3
Total 2nd use	0.5	2 number	0.8	101-300	0.5
Total 3rd use	0.7	3 number	1.0	301-500	0.8
Total 4th use	0.9	-	-	More than 500	1.0

Grouped Queries Indexing For Relational Database

Radosław Boroński
Dept. of Electronics and Computer Science
Koszalin University of Technology
Koszalin, Poland
radoslaw.boronski@tu.koszalin.pl

Grzegorz Bocewicz
Dept. of Electronics and Computer Science
Koszalin University of Technology
Koszalin, Poland
bocewicz@ie.tu.koszalin.pl

Robert Wójcik
Institute of Computer Engineering, Control and Robotics
Wrocław University of Technology
Wrocław, Poland
robert.wojcik@pwr.wroc.pl

***Abstract*-This paper discusses the problem of minimizing the response time for a given database workload by a proper choice of indexes. We propose to look at the database queries as a group and search for good indexes for the group instead of an individual query. We present condition for applying the concept of grouped queries index selection. Such condition is illustrated by three practical examples.**

Keywords-database;index;ISP;grouped queries;related queries

I. INTRODUCTION

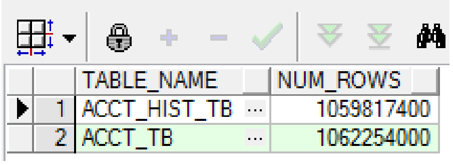
Getting database search result quickly is one of the crucial optimization problems in a relational database processing. The major strength of relational systems is their ease of use. Users interact with these systems in a natural way using nonprocedural languages that specify what data are required, but do not specify how to perform the operations to obtain those data [8]. Online Internet shops, analytics data processing or catalogue search are examples of structures where data search must be processed as quick as possible with minimal hardware resources involved. Common practice is to minimize the database search process at minimal cost. A database administrator (or a user) may redesign the physical hardware structure or reset the database engine parameters, or try to find suitable table indexes for a current query. Most vendors nowadays offer automated tools to adjust the physical design of a database as part of their products to reduce the DBMS's total cost of ownership [3]. As adding more CPUs or memory may not always be possible (i.e. limited budget) and maneuvering within hundreds of database parameter may lead to a temporary solution (wrong settings for other database queries), index optimization should be considered as being foremost.

Indexes are optional data structures built on tables. Indexes can improve data retrieval performance by providing a direct access method instead of the default full table scan retrieval method [7]. In the simple case, each query can be answered either without using any index, in a given answer time or with using one built index, reducing answer time by a gain specified for every index usable for a query [14]. Hundreds of consecutive database queries together with large amount of data involved lead to a very complex combinatorial optimization problem. Two sample tables in a data warehouse of an international automobile factory contain over 1 billion records each (Fig. 1). Time needed to obtain result of both index-less tables joined together may be up to 45 minutes. Such delays are not acceptable for production environment processes. Indexes in such cases may reduce the response time of 50% (depending on which columns are used for the indexing). The classic index selection method focuses on a tree data structure, which could limit the search area as much as possible. Literature acknowledges us with such B-tree types as:

- Sorted counted B-trees, with the ability to look items up either by key or by number, could be useful in database-like algorithms for query planning [5],
- Balanced B*-tree that balances more neighboring internal nodes to keep the internal nodes more densely packed [12],
- Counted B-trees with each pointer within the tree and the number of nodes in the subtree below that pointer [19].

The B-tree and its variants have been widely used in recent years as a data structure for storing large files of information, especially on secondary storage devices [11]. The guaranteed small (average) search, insertion, and deletion time for these structures makes them quite appealing for database applications.

The topic of current interest in database design is the construction of databases that can be manipulated concurrently and correctly by several processes. In this paper, we discuss a simple variant of the B-tree (balanced B*-tree, proposed by Wedekind [20] especially well-suited for use in a concurrent database system [15].



		TABLE_NAME	NUM_ROWS
▶	1	ACCT_HIST_TB ···	1059817400
	2	ACCT_TB ···	1062254000

Figure 1. Example of large number of rows for two data warehouse tables

ISBN: 978-1-61208-254-7

While the selection of indexes structure have a very important role in the design of database management tools so far avoided interference in the structure of indexes at the stage of the database operation. In such situations more important is to ask a question “how to choose a set of indexes for the selected query sets?”. It turns out that the proper selection of indexes can bring significant benefits for the database query execution time. Typical approaches found in the literature mainly focus on the search indexes only for single column or single query [16], [10], [9], [17], [4]. In this paper, an approach associated with the search query indexes for groups called blocks is presented.

In this case we will consider B-tree indexes. A B-tree index allows fast access to the records of a table whose attributes satisfy some equality or range conditions, and also enables sorted scans of the underlying table [18].

The rest of the paper is organized as follows: in Section II we describe a problem statement. In Section III, we briefly present classic index selection approach together with simple examples that will illustrate the subject. In Section IV, we demonstrate new method of grouped queries index selection and compare examples results with the classic approach. Section V and VI present our conclusions and future works.

II. Problem Statement

Motivation for this work is to suggest an approach of multi-queried SQL block where sub-optimal or optimal solution is to be found that gives decision makers some leeway in their decisions. The main goal is to choose a subset of given indexes to be created in a database, so that the response time for a given database workload together with indexes used to process queries are minimal.

The index selection problem has been discussed in the literature. Several standard approaches have been formulated for the optimal single-query and multi-query index selection. Some past studies have developed rudimentary on-line tools for index selection in relational databases, but the idea has received little attention until recently. In the past year, on-line tuning came into the spotlight and more refined solutions was proposed. Although these techniques provide interesting insights into the problem of selecting indexes on-line, they are not robust enough to be deployed in a real system [18]. The problem is known in a literature as Index Selection Problem (ISP) According to [8] it is NP-hard. Note that in practice the space limit in the ISP is soft, because databases usually grow, thus the space limit is specified in such way that a significant amount of storage space remains free [13].

In a real life scenario, for thousands database queries (Fig. 2) compromising hundreds of tables and thousands of columns, the search space is huge and grows exponentially with the size of the input workload.

Considered case of Index Selection Problem can be defined in following way.

Given is a set of tables:

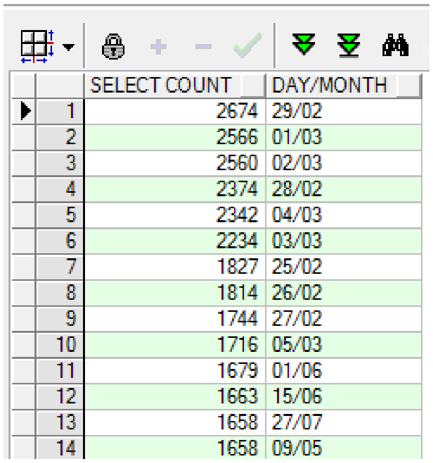
$$T = \{T_1, \dots, T_i, \dots, T_n\}, \qquad (1)$$

described by a set of columns included in the tables:

$$K = \{k_{1,1}, \dots, k_{1,l(1)}, \dots, k_{i,j}, \dots, k_{n,1}, \dots, k_{n,l(n)}\}\ , \qquad (2)$$

where: $k_{i,j}$ is a j-th column of table T_i.

Each column $k_{i,j}$ corresponds to set of values $V(k_{i,j})$ (tuples set) included in this column.



	SELECT COUNT	DAY/MONTH
▶ 1	2674	29/02
2	2566	01/03
3	2560	02/03
4	2374	28/02
5	2342	04/03
6	2234	03/03
7	1827	25/02
8	1814	26/02
9	1744	27/02
10	1716	05/03
11	1679	01/06
12	1663	15/06
13	1658	27/07
14	1658	09/05

Figure 2. Example of number of database queries in a given day for a production data warehouse

For set of tables T various queries Q_i can be formulated (in SQL these are SELECT queries). These queries are put against the specified set of columns $K^* \subseteq K$. The result of query Q_i is set as:

$$A_i \subseteq \prod_{k_{i,j} \in K^*} V(k_{i,j})\ , \qquad (3)$$

where: $\prod_{i=1}^{n} Y_i = Y_1 \times Y_2 \times \dots \times Y_n$ is a cartesian product of sets $Y_1, \dots, Y_n$.
For a given database DB it is taken into account that A_i is a result of following function:

$$A_i = Q_i(K^*, Op(DB))\ , \qquad (4)$$

where: K^* is a subset of available indexes, $Op(DB)$ is set of operators available in database DB of which relation describing query Q_i is built.

The time associated with the determination of the set A is depended on the DB database used (search algorithms, indexes structures) and adopted set of indexes $J \subseteq \mathcal{P}(K^*)$ (where $\mathcal{P}(K^*)$ - is a power set of K^*). It is therefore assumed that the query execution time Q_i in given database DB, is determined by the function: $t(Q_i, J, DB)$. In short the value of execution time for query Q_i, data base DB and set of indexes J will be define as: $t_i(J)$.
In the context of the so-defined parameters, a typical problem associated with the ISP responds to the question:

What set of indexes $J \subseteq \mathcal{P}(K^*)$ minimizes the query Q_i execution time: $t_i(J) \rightarrow min$?

When a multi-component set of queries $Q = \{Q_1, \dots, Q_m\}$ is considered, question takes the form:

What set of indexes $J \subseteq \mathcal{P}(K^*)$ minimizes the queries block Q execution time: $\sum_{Q_i \in Q} t_i(J) \to min$?

III. CLASSIC INDEX SELECTION APPROACH

Classic index selection approach focuses on individual query and tries to find good index or indexes set for tables in a single query in a given block. Such approach does not take into consideration queries in a block as a whole. By doing so, a database user may expose database to create excess number of indexes which could be redundant or not used for more than one query in an examined block. This could also result in utilizing too much disk space and time needed for the indexes creation. Finding good index group for a large database queries' block was never an easy task to do and usually users and database administrators rely on their experience and good practice. In the commercial use one may find tools that support the index selection process, such as SQL Access Advisor (Fig. 3) [6], Toad, SQL Server Database Tuning Advisor [1].

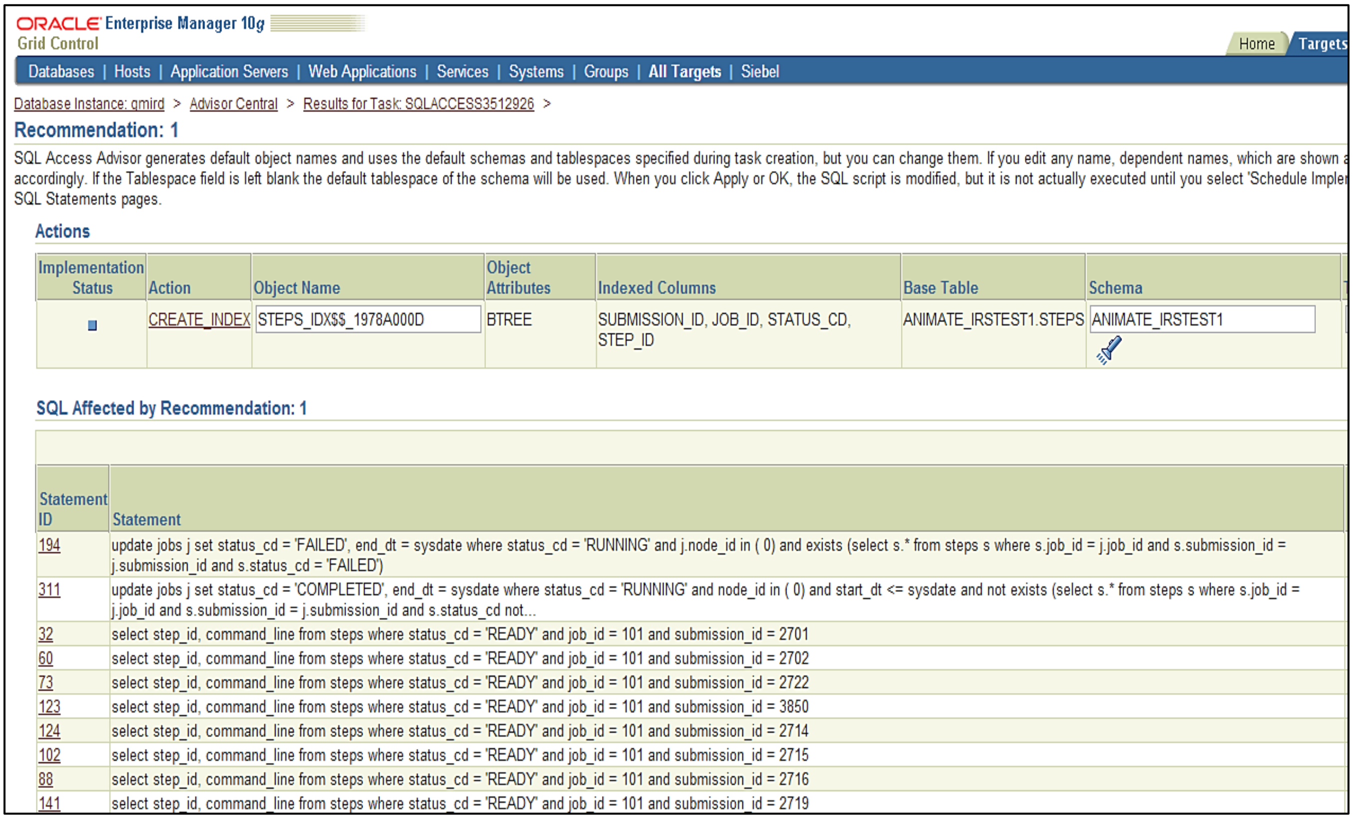
Let us consider three examples where given is a group of three database queries $Q = \{Q_1, Q_2, Q_3\}$:

Q_1: *SELECT * FROM T_1, T_2 WHERE $k_{1,1}$ < $k_{2,2}$ AND $k_{1,3}$=[const],*
Q_2: *SELECT * FROM T_2, T_3 WHERE $k_{2,2}$ = $k_{3,2}$,*
Q_3: *SELECT * FROM T_2 WHERE $k_{2,1}$ > [const].*

Interpretation of this type of queries (according to (4)) is as following:

Q_1: searching for a set of triples: $A_i = \{(a, b, c): a \in V(k_{1,1}), b \in V(k_{2,2}), c \in V(k_{1,3}); a < b, c = [const]\}$,
set $K^* = \{k_{1,1}, k_{2,2}, k_{1,3}\}$.
Q_2: searching for a set of pairs: $A_i = \{(a, b): a \in V(k_{2,2}), b \in V(k_{3,2}); a = b\}$,
set $K^* = \{k_{2,2}, k_{3,2}\}$.
Q_3: searching for a set: $A_i = \{a: a \in V(k_{2,1}); a = [const]\}$,
set $K^* = \{k_{2,1}\}$.

Tables T_1, T_2, T_3 contain 1*10^6 records each. No indexes are built on either table: $J = \emptyset$. With the first test run, database returned following response times: $t_1(J) = 2040$s, $t_2(J) = 3611$s, $t_2(J) = 345$s respectively, resulting in full table scans for each Q. Queries Q ran on database Oracle 11.2.0.1 installed on server with Redhat 6 operating system with 64GB memory and ASM used for disk storage.


ORACLE Enterprise Manager 10g
Grid Control
Home
Targets
Databases | Hosts | Application Servers | Web Applications | Services | Systems | Groups | All Targets | Siebel

Database Instance: gmird > Advisor Central > Results for Task: SQLACCESS3512926 >

Recommendation: 1

SQL Access Advisor generates default object names and uses the default schemas and tablespaces specified during task creation, but you can change them. If you edit any name, dependent names, which are shown a accordingly. If the Tablespace field is left blank the default tablespace of the schema will be used. When you click Apply or OK, the SQL script is modified, but it is not actually executed until you select 'Schedule Impler SQL Statements pages.

Actions

Implementation Status	Action	Object Name	Object Attributes	Indexed Columns	Base Table	Schema
■	CREATE_INDEX	STEPS_IDX$$_1978A000D	BTREE	SUBMISSION_ID, JOB_ID, STATUS_CD, STEP_ID	ANIMATE_IRSTEST1.STEPS	ANIMATE_IRSTEST1

SQL Affected by Recommendation: 1

Statement ID	Statement
194	update jobs j set status_cd = 'FAILED', end_dt = sysdate where status_cd = 'RUNNING' and j.node_id in (0) and exists (select s.* from steps s where s.job_id = j.job_id and s.submission_id = j.submission_id and s.status_cd = 'FAILED')
311	update jobs j set status_cd = 'COMPLETED', end_dt = sysdate where status_cd = 'RUNNING' and node_id in (0) and start_dt <= sysdate and not exists (select s.* from steps s where s.job_id = j.job_id and s.submission_id = j.submission_id and s.status_cd not...
32	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2701
60	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2702
73	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2722
123	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 3850
124	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2714
102	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2715
88	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2716
141	select step_id, command_line from steps where status_cd = 'READY' and job_id = 101 and submission_id = 2719

Figure 3. Oracle's 10g2 SQL Access Advisor

The classic approach requires treating every database query individually. Hence indexes are built: $k_{1,1}$ and $k_{1,3}$ on table T_1; $k_{2,1}$, $k_{2,2}$ on table T_2; $k_{3,2}$ on table T_3. This kind of indexes are represented by the set: $J = \left\{\{k_{1,1}, k_{1,3}\}, \{k_{2,2}\}, \{k_{3,2}\}, \{k_{2,1}\}\right\}$ containing four sets. Each element (set) of J contains the columns which are used to build the indexes. For example, the set $\{k_{1,1}, k_{1,3}\}$ means that we have to build one index for columns $k_{1,1}, k_{1,3}$.

The set of indexes J is built for three different tables, resulting in use of 2GB of additional disk space. With the second test run, database returned following response times: $t_1(J) = 2612$s, $t_2(J) = 2580$s, $t_3(J) = 5$s respectively. As the response time is better by approximately 10%, there is still unreasonable disk space used and time needed for creating 4 large indexes. Creating 4 indexes forced query optimizer to use them, and instead of decreasing Q_1 execution time, it got increased. This is because optimizer decided to read $k_{1,1}$ column index content first and because it couldn't find values for $k_{1,3}$ column, it performed full table scan for table T_1. Examples shows that selected indexes may increase the query execution performance where in other cases may have the opposite effect.

IV. Grouped queries approach

In this paper we focus on related queries group and because of this relation and the number of indexed columns. We take into account the search for a good index for the entire queries' block. We propose a new approach by using multi-query SQL block selection. Such block consists tabular relations between queries, meaning that the number of tables columns used in previous query is present in other queries. The proposed approach could be an alternative to the classic index selection method, where one common index set could be found. Grouped queries approach has to be studied for its effectiveness and authenticity via a series of numerical tests. Furthermore, to compare the performance of the method commercial tools will have to be used and results compared.

For previous examples, we suggested to create a pool of all columns taking part in all queries in a group and build sub-optimal indexes set for queried tables. Such task will involve creating the weighted list that will include all the index candidate query-related columns and their number of occurrence in the examined queries block:

$$KW = \left((k_{1,1}, 1\,), (k_{1,3}, 1), (k_{2,1}, 1), \boxed{(k_{2,2}, 2)}, (k_{3,2}, 1)\right). \quad (5)$$

Of course, only $k_{2,2}$ column (marked by the box in (5)) is a query-related candidate column that could be used for the index creation. Nevertheless, other columns from remaining tables could also be revised. In that context, we suggest to create composite index for the same table T_2 on columns $k_{2,1}$ and $k_{2,2}$: $J = \left\{\{k_{2,1}, k_{2,2}\}\right\}$. By doing so, user not only speeds up block execution but also saves significant volume of disk space. With the third test run, database returned following response times: $t_1(J) = 1235$ s, $t_2(J) = 2430$ s, $t_3(J) = 5$s, respectively, decreasing total execution time of 35% and saving disk space of 60%. This is due to the fact that only index is used or full table scan for non-indexed table resulting in smaller response times for Q_1 and Q_2. Database optimizer does not need to perform an additional read operation (separate for index and if values not found and separate for a table). This proves that indexes should be selected with care.

Determining the answers to a set of queries can be improved by creating some indexes.

Classic index selection focuses on each query individually and final indexes set is a sum of indexes sub-sets for each query.

We show that groups of queries, one can get better indexes set if such group is treated as a whole.

Grouped queries index search can only benefit and have an advantage over single query search, only if queries in the group satisfy the condition of mutual dependence. Queries Q_1, Q_2, Q_3, from previous examples are dependent so below statement applies. Such dependency must be clearly defined.

In the present case, the dependence set of queries Q is determined by connectivity of hypergraph $G(Q)$.

Example of a hypergraph for considered queries Q is presented on Fig. 4.

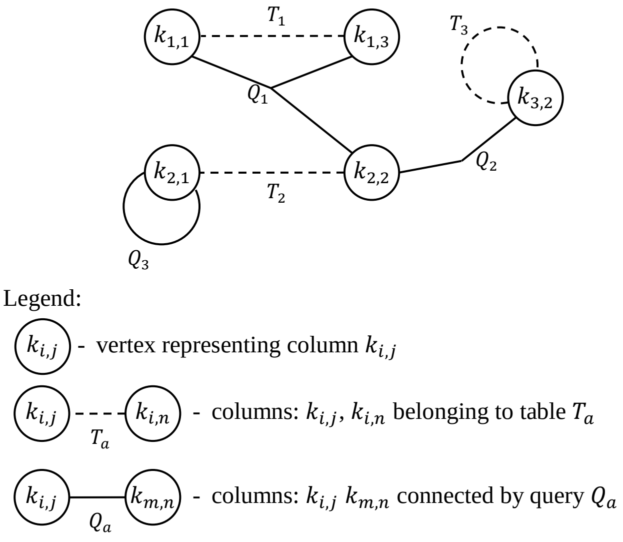

T_1
$k_{1,1}$
$k_{1,3}$
T_3
Q_1
$k_{3,2}$
$k_{2,1}$
$k_{2,2}$
Q_2
T_2
Q_3
Legend:
$k_{i,j}$ - vertex representing column $k_{i,j}$
$k_{i,j}$ T_a $k_{i,n}$ - columns: $k_{i,j}$, $k_{i,n}$ belonging to table T_a
$k_{i,j}$ Q_a $k_{m,n}$ - columns: $k_{i,j}$ $k_{m,n}$ connected by query Q_a

Figure 4. Hypergraph for considered set of queries Q

In this type of graph vertices represent the columns used in queries Q, edges connect those vertices which combined make table T_a(dashed line hyper edge) or related queries Q_i (solid line hyper edge). For example, hyper edge connecting vertices $k_{1,1}, k_{2,2}, k_{1,3}$ represents relation with query Q_1.

It is assumed that the query set Q is related if corresponding hypergraph $G(Q)$ is consistent.

In this context, the group queries indexes set creation can benefit compared to classic index selection only for related sets.

As a counterexample, given is a group of three database queries $Q^* = \{Q_1^*, Q_2^*, Q_3^*\}$:

Q_1^*: *SELECT * FROM* T_1, T_2 *WHERE* $k_{1,1} > k_{1,2}$,
Q_2^*: *SELECT * FROM* T_2, T_3 *WHERE* $k_{2,1} = k_{3,2}$,
Q_3^*: *SELECT * FROM* T_4 *WHERE* $k_{4,1} >$ *[const].*

Example of a hypergraph for considered queries Q^* is presented on Fig. 5. This kind of hypergraph presented is inconsistent. For this reason queries Q^* are treated as the unrelated queries.

Unrelated queries for index selection process means they cannot be treated as a group. In such cases best index set is a set determined for each query individually:

$$J^* = \{\{k_{1,1}, k_{1,2}\}, \{k_{2,1}\}, \{k_{3,2}\}, \{k_{4,1}\}\}. \qquad (6)$$

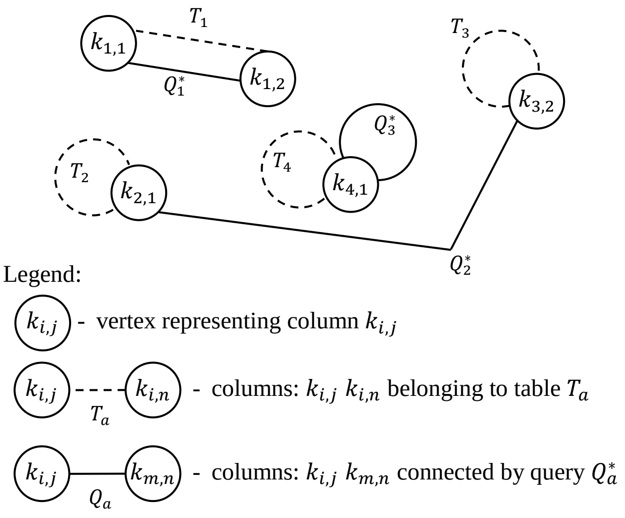

T_1
$k_{1,1}$
$k_{1,2}$
Q_1^*
T_3
$k_{3,2}$
Q_3^*
T_2
$k_{2,1}$
T_4
$k_{4,1}$
Q_2^*
Legend:
$k_{i,j}$ - vertex representing column $k_{i,j}$
$k_{i,j}$ T_a $k_{i,n}$ - columns: $k_{i,j}$ $k_{i,n}$ belonging to table T_a
$k_{i,j}$ Q_a $k_{m,n}$ - columns: $k_{i,j}$ $k_{m,n}$ connected by query Q_a^*

Figure 5. Hypergraph for considered set of queries Q^*

Weighted list for Q^* that that includes all the index candidate columns:

$$KW^* = ((k_{1,1}, 1), (k_{1,2}, 1), (k_{2,1}, 1), (k_{3,2}, 1), (k_{4,1}, 1)).(7)$$

One can notice there are no query-related candidate columns (single column occurrence) that could be used for the grouped queries index set creation. Each table T_i will have to be indexed separately for each individual query Q^*.

V. Conclusion

Finding a good index or indexes set for a table is very important for every relational database processing not only from the performance point but also cost aspect. Indexes can be crucial for a relational database to process queries with reasonable efficiency, but the selection of the best indexes is very difficult.

Presented examples shows that there is a need for finding an automatic index selection mechanism with grouped queries-oriented rather than a classic (single query) approach. Practice shows that index focus on grouped queries gives better results and enables user to save time needed for index creation. It also saves system hardware resources. In the examples we show that grouped queries indexes set are more effective than individual queries indexes because queries Q_1, Q_2, Q_3 satisfy the relation condition (Table 1).

For the automatic index selection, the system continuously monitors queries block and gathers information on columns used in queries. The administrator (or user) can summon the automatic system at any time to be presented with the current index recommendation, or tune it to the queries' block needs. The system also presents the user index set and allows user to choose best option. User decides whether to reject or accept proposed set. Due to index interactions, the user's decisions might affect other indexes in the configuration, so the recommendation would need to be regenerated, taking the user's constraints into account.

In the presented examples we considered three situations of database queries block execution, one without indexes, one with classic separate queries indexing and one with grouped queries indexing. Examples showed that one should create grouped indexes only for related queries. In that context presented relationship may be treated as sufficient condition for the evaluation of grouped queries indexing.

TABLE I. CLASSIC AND GROUPED QUERIES APPROACH FOR CORRELATED DATABASE QUERIES

Database queries:
Q_1: *SELECT * FROM* T_1, T_2 *WHERE* $k_{1,1} < k_{2,2}$ *AND* $k_{1,3}$*=[const];*
Q_2: *SELECT * FROM* T_2, T_3 *WHERE* $k_{2,2} = k_{3,2}$*;*
Q_3: *SELECT * FROM* T_2 *WHERE* $k_{2,1} >$ *[const];*
	Classic approach:
	CREATE INDEX k1_col1_idx ON $T_1(k_{1,1})$;
	CREATE INDEX k1_col3_idx ON $T_1(k_{1,3})$;
	CREATE INDEX k2_col1_idx ON $T_2(k_{2,1})$;
	CREATE INDEX k2_col2_idx ON $T_2(k_{2,2})$;
	CREATE INDEX k3_col2_idx ON $T_3(k_{3,2})$;

	Grouped queries approach: CREATE INDEX k2_col1_col2_idx ON $T_2(k_{2,1}, k_{2,2})$;

VI. Future work

Our current works are focused on grouped queries index selection method with the use of genetic algorithm [2] that analyzes database queries, suggests indexes' structure and tracks indexes influence on the queries' execution time. We work on the system that will be used in an attempt to find better indexes for a critical part of long-running database queries in testing and production database environment.

Recording queries with good indexes together with their total execution time is a starting point for broader searches in the future. Simple test presented in this article proves reasonableness of this method. The developed system is scalable: there is a potentiality of combining smaller queries' blocks into larger series and finding better solution based on execution history.

REFERENCES

[1] S. Agrawal, S. Chaudhuri, L. Kollar, A. Marathe, V. Narasayya, and M. Syamala, "Database Tuning Advisor for Microsoft SQL Server 2005". In Proceedings of the 30th International Conference on Very Large Databases, 2004.

[2] T. Back, "Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms", Oxford University Press Oxford, UK, 1996.

[3] N. Bruno and S. Chaudhuri, "Automatic physical database tuning: a relaxation-based approach", SIGMOD '05 Proceedings of the 2005 ACM SIGMOD international conference on Management of data, ACM New York, NY, USA, 2005, pp.227-238.

[4] S. Chaudhuri and V. Narasayya, "An efficient Cost-Driven Index Selection Tool for MS SQL Server", Very Large Data Bases Endowment Inc, 1997.

[5] D. Comers, "The Ubiquitous B-Tree", Computing Surveys 11 (2), doi:10.1145/356770.356776, pp. 123–137.

[6] B. Dageville, D. Das, K. Dias, K. Yagoub, M. Zait, and M. Ziauddin, "Automatic SQL Tuning in Oracle 10g". In Proceedings of the 30th International Conference on Very Large Databases, 2004.

[7] C. Dawes, B. Bryla, J. Johnson, and M Weishan, "OCA Oracle 10g Administration I", Sybex, 2005, pp.173.

[8] S. Finkelstein, M. Schkolnick, and P. Tiberio, "Physical database design for relational databases", ACM Trans. Database Syst. 13(1), (1988), pp.91–128.

[9] M. Frank and M. Omiecinski, "Adaptive and Automated Index Selection in RDBMS", Proceedings of EDBT, 1992.

[10] H. Gupta, V. Harinarayan, A. Rajaraman, and J. D. Ullman, "Index Selection for OLAP", In Proceedings of the Internatoinal Conference on Data Engineering, Birmingham, U.K., April 1997, p. 208-219.

[11] D. Knuth, "The Art of Computer Programming", vol. 3, Sorting and Searching. Addison- Wesley, Reading, Mass., 1973.

[12] D. Knuth, "Sorting and Searching, The Art of Computer Programming", Volume 3 (Second ed.), Addison-Wesley.

[13] P. Kołaczkowski and H. Rybiński, "Automatic Index Selection in RDBMS by Exploring Query Execution Plan Space", Studies in Computational Intelligence, vol. 223, Springer, 2009, pp.3-24

[14] J. Kratica, I. Ljubic, and D. Tosic, "A Genetic Algorithm for the Index Selection Problem", EvoWorkshops'03 Proceedings of the 2003 international conference on Applications of evolutionary computing, 2003.

[15] P.L. Lehman, "Efficient locking for concurrent operations on B-trees", ACM Transactions on Database Systems (TODS), Volume 6 Issue 4, Dec. 1981, pp.650-670.

[16] Y. Maggie, L. Ip, L. V. Saxton, and Vijay V. Raghavan, "On the Selection of an Optimal Set of Indexes", IEEE Transactions on Software Engineering, 9(2), March 1983, p.135-143.

[17] M. Schkolnick, "The Optimal Selection of Indices for Files", Information Systems, V.1, 1975.

[18] K. Schnaitter, "On-line Index Selection for Physical Database Tuning", ProQuest, UMI Dissertation Publishing, 2011.

[19] S. Tatham, "Counted B-Trees", http://www.chiark.greenend.org.uk/~sgtatham/algorithms/cbtree.html, 11.02.2013.

[20] H. Wedekind, "On the selection of access paths in a data base system. In Data Base Management", J.W. Klimbie and K.L. Koffeman, Eds. North-Holland, Amsterdam, 1974, pp. 385-397.

Integration of Healthcare Information Systems:
Improving Data Quality in a Diagnostic Imaging Department

Conceição Granja
Faculty of Engineering, University of Porto
Porto, Portugal
granja.conceicao @fe.up.pt

Zafeiris Kokkinogenis
LIACC and IDMEC – FEUP Campus
Faculty of Engineering, University of Porto
Porto, Portugal
pro08017@fe.up.pt

Joaquim Gabriel
IDMEC – FEUP Campus
Faculty of Engineering, University of Porto
Porto, Portugal
jgabriel@fe.up.pt

Terje Solvoll
Norwegian Centre for Integrated Care and Telemedicine
University Hospital of North Norway
Tromsø, Norway
terje.solvoll@telemed.no

***Abstract*—The existence of multiple information systems in the healthcare environment causes data integrity issues when the information flow architecture is not implemented considering the specificities of each department, the multiple data sources and the data input does not follow the systems requirements. This paper presents a case-study of a diagnostic imaging department information flow analyses and optimization, using the DICOM standard embedded on most equipment, to improve data integrity, availability and structure. The existing patient data structure was redefined and the existing information flow was re-engineered in accordance with the DICOM and IHE guidelines. This study focuses in a diagnostic imaging department and resulted in the identification of a new information flow that permits a significant and positive reduction of information inconsistencies, thus improving data quality.**

Keywords- Healthcare Information Systems; Information Flow Re-engineering; Medical Records; Data Quality

I. INTRODUCTION

The development of information technologies (IT) has made available, namely to healthcare providers (HcP), a better flow and processing of information that supports the clinical activity. The clinical activity has been brought closer to the inherent administrative and financial actions increasing the capacity to plan, monitor and evaluate the performed activities which has a positive impact in the financial management, production capacity and in the provided services quality [1].

Patient information is spread through several information systems (IS) that gather different kinds of data, such as demographic, medical, financial or managerial, each system having its own idiosyncrasies. In radiology departments, the IS comprised in the information flow are the Hospital Information System (HIS), Radiology Information System (RIS) and the Picture Archiving and Communication System (PACS). HIS is a management system that has three main functions: i) support clinical and medical care actions; ii) administer the hospital operation such as financial, resources scheduling and patient admissions; and iii) evaluate the hospital performance. A radiology department has specific operational requirements and, therefore, requires its own management system, acting under the umbrella of HIS. Such management systems need separate information which has to be integrated with the data from HIS, this being the role of RIS. PACS is the image management system, it acts as an archiving server which receives studies/images from the acquisition gateway, inserts/appends the study information to a database and stores the images. Together, these systems allow the HcP to manage the information flow and share it using different communication protocols in order to handle their heterogeneity.

The existence of multiple IS may cause data integrity issues when the information flow architecture is not implemented considering the specificities of each department, the multiple data sources and the input data does not follow the IS requirements.

A. Digital Imaging and Communications in Medicine

The Health Level 7 (HL7) [2]and the Digital Imaging and Communications in Medicine (DICOM) [3] are the most common communication Standards used to interface the healthcare IS in order to collect and integrate information from different sources and types facilitating its distribution and availability where it is needed.

The development of the DICOM Standard dates back to 1982 to a joint committee of the American College of Radiology (ACR) and the National Electrical Manufacturers (NEMA) [2]. Such cooperation intended to create the possibility of data transfer in healthcare regardless of the manufacturer's Standards. The first version of the DICOM Standard, at the time denominated ACR-NEMA Standards Publication No. 300-1985, was published in 1985 and revised by version 2.0 in 1988. The first version named DICOM was in 1993 (v.3.0), and was the foremost to include services beyond data transfer [3].

The current DICOM Standard is structured in parts, each layer being used to define different services and objects. An implementation of the DICOM Standard does not have to use all of its parts. Implementations may use the parts of DICOM, such as Service-Object Pair (SOP) classes, media storage profiles and attributes, necessary to support the designed architecture, as Conformance Statements refer to a specific implementation.

DICOM SOP classes and associated Information Object Definitions (IODs) are used to convey specific medical imaging information at the Data Format Layer. IODs are sets of Attributes that comprise a type of data element identified by tags [4]. Attributes are classified in three types: Type 1: Mandatory data elements. When classified as 1C, the mandatory character of data elements is dependent of the specified conditions; Type 2: Required data element but its value may be unknown. When classified as 2C, the data elements are required under the specified conditions; Type 3: Optional data elements.

A DICOM tag is a unique identifier of an Attribute, and corresponding data element, defined by a pair of numbers represented as (gggg,eeee), where gggg denotes the Group Number and eeee the Element Number. Group Numbers were given a meaning in the ACR-NEMA Standards Publication No. 300-1985 and ACR-NEMA Standards Publication No. 300-1988, known as version 1.0 and 2.0. For example, Group (0008,xxxx) was denoted by *Identifying Information*. In DICOM version 3.0 the Group's names are not mentioned, as new Attributes are now being assigned to Groups based on their similarity to the existing ones.

An architecture that implements the information flow aiming the optimization of the service provided to the patient was defined by the Integrating the Healthcare Enterprise (IHE). As the IHE architecture was developed for the specific healthcare environment based on communication standards, it provides a set of profiles to ensure communication between different systems, which use different communication protocols, allowing the information to flow through the hospital and become available where it is needed [5].

This paper presents a case-study of a diagnostic imaging department information flow analyses and optimization, using the DICOM standard embedded on most equipment, to improve data integrity, availability and structure.

II. Related Work

Data integrity is a known problem in healthcare, being the object of multiple studies in recent years. Arellano et al. [6] studied the problem of integrating multiple master person indexes (MPI) into a single enterprise person index (EPI). The authors evidenced the importance of standardization when filling in MPI files and the existence of a unique patient identifier. Cruz-Correia et al. [7] used three approaches, working simultaneously, to monitor data quality in a Portuguese public hospital where a Virtual Electronic Patient Record (VEPR) had been implemented. On the first approach a third party network monitoring application was implemented and configured to generate alerts on abnormal report retrieval and visualization rates. On a second approach patient identification inconsistencies were monitored by crossing data from the departmental IS and the administrative database. On a third stage data integrity was monitored through the verification of the physician digital signature on clinical records delivered by the VEPR. The first solution presented a large number of false alerts, and non-trigged alerts, due to the inactivity of some departments on the weekends. Nevertheless, achieved a significant reduction on the number of inconsistencies and increased data integrity. The data quality in healthcare was approached by Bates et al. [8] by inferring on the influence that information technologies (IT) have on data errors. The authors proposed a framework on recommendations using ITs to reduce errors in healthcare, divided in two categories: i) general recommendations; and ii) domain specific recommendations. The first category included recommendations such as the use of standards for data and systems and the communication between systems. The second category suggested the use of identification standards of consumables and the use of ITs to communicate asynchronous data.

III. Case Study Description

The case study described in this section is based on the data structure and information flow of a diagnostic imaging department; the name will not be mentioned for ethical reasons.

The studied HcP provides diagnostic imaging services on the modalities: computed tomography (1), mammography (1), conventional radiology (1), ortopantomography (1), densitometry (1), magnetic resonance (1), stereotaxy (1) and ultrasound scanning (5).

According to the language standard, the HcP information flow can be divided into three modules: i) Management, ii) RIS, iii) Imaging, as shown in Figure 1.

The Management module is the input point for the patient demographic and clinical data. Usually, the patient demographic data and exam information is inserted in the database upon an exam request made by the patient, the majority of the times made through a phone contact. This information is confirmed and updated on the front desk in the presence of the patient on the exam day. This module also manages the financial information taking into account the existing public and private conventions. Public conventions enclose entities such as the National Health Service (SNS, from the Portuguese Serviço Nacional de Saúde), the public officials assistance (ADSE, from the Portuguese Assistência na Doença aos Servidores do Estado), and the Public Security Police insurance, (PSP, from the Portuguese Polícia de Segurança Pública). Private conventions include the health insurances that have established a protocol with the clinic. Each one of these conventions has its own

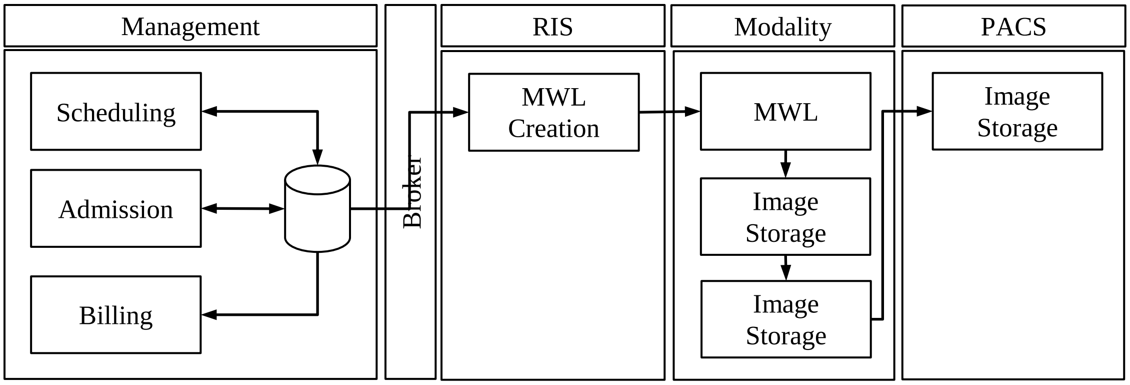


Figure 1 Illustration of the information flow at the studied HcP.

particularities concerning the invoicing exam description and rules.

The Broker receives the patient and exam information in HL7 from the Management Module, converting it to the DICOM standard before sending it to RIS.

RIS is in charge of creating a bridge between the Management and the Imaging modules. RIS receives the necessary patient and the exam information to create the Modality Work List (MWL) at the Modality request.

The Modality and the PACS together constitute the Imaging Module. The Modality issues the MWL requests to RIS, making the patient and the exam information available to the technicians at the exam room. After image acquisition and processing, at the demand of the technicians, these are sent for storage in PACS.

The aim of the work presented herein is to improve the information flow in between the described modules, in order to identify and resolve actions that can incur in information incoherence, to provide accurate information when and wherever it is needed.

IV. Methods

A holistic approach was used to determine the existing information flow. As the Management Module is a non-commercial system, during this phase the existing IS were identified and characterized according to the communication standard supported and data available. In a second phase the data consistency was verified, by crossing information from the ISs, outlined in the first phase, and the problems identified. The identified problems can be classified in two categories based upon the type of data in which they are related to: i) Patient data or ii) HcP internal data.

The identified problems related to patient data are as follows:

a) *Internal identifiers are attributed upon the request* – in case of patient no-shows, or late-cancellations; patient records with no data are sent to RIS for the MWL creation, causing empty entries on the modality;

b) *Absence of a unique patient identifier (UID)* – in the absence of an UID, when a patient returns to the clinic it is impossible to unequivocally identify him/her and multiple internal identifiers could be generated to the same patient. This patient misidentification leads to the existence of multiple records for the same patient being impossible to make his/her clinical history available to the clinical staff. Additionally, the propagation of this problem through the other systems could not be controlled or avoided;

c) *Creation of a new patient registry without cancelling the previous* **-** this action occurred when mistakes were found in the primary registry, causing not only multiple patient registries for the same exam but also a double patient billing registry that had to be filtered afterwards by the accounting staff.

d) *Consumables were registered as exams* – This occurred because the information generated by the Modality was not being returned to the billing system. As this information arrives at the Modality it reveals MWL inconsistencies, exams without data or associated images and, furthermore, inaccurate data passed forward to the billing system;

e) *Patients registered and their demographic data changed manually at the Modality* – this caused inconsistencies between the patient demographics and exam data existing in the Management Module and the real operational activities. Therefore, patients may not appear in the billing system or the existing information could be incorrect.

The identified problems relating to the HcP internal information comprises:

a) *Incorrect attribution of Accession Number* – Accession Number represents the request identifier for the Modality. As this identifier was attributed by Modality, different exams were attributed with the same Accession Number. In practice this means that if a patient is scheduled to perform more than one exam at the same Modality they would all be attributed the same Accession Number and, when sent to PACS they would be stored as a single exam;

b) *Patient request not matching the referring doctor prescription* – Different conventions have different exam descriptions and require to be billed according to different rules. Therefore, upon the patient request the information was inserted on the database according to the patient convention exam description and rules. Thus, the same patient had multiple requests for the same exam and the data sent to the modality was not correct;

c) *Inexistence of information on the MWL* – The exam description was not identified on the MWL message sent to the Modality by the RIS. This lack of information made it

ISBN: 978-1-61208-254-7

impossible for the technician to be acquainted of which exam was going to be performed by the patient until the paper copy of the referring doctor's prescription was delivered by the auxiliary staff.

To propose improvement changes to the information flow that would tackle the problems presented above with the minimum possible impact on the HcP workflow, it was essential to be acquainted with the existing workflows [9-11] and to know the Portuguese and International guidelines, protocols and best practices [5, 12-16]. This knowledge supported the information flow re-engineering and presents the fundamental issues, such as process interactions and information requirements and structure.

V. Results and Discussion

The patient data structure, shown in Figure 2, was defined in accordance with the DICOM Standard requirements and the existing database architecture, in order to minimize the changes to the last once the existing records could not be changed given the linkage to PACS and to avoid further inconsistencies. In this sense, was created a new identifier of the patient visit, *Episode ID*, which is attributed to all the exams performed on the same visit independently of the modality in which they ought to be performed. This identifier would replace the previous Accession Number. The *Accession Number* was kept in the structure but it is now attributed by exam requested and not by Modality. The *Accession Number* identifies a scheduled procedure, by HIS request, and it is duplicated to the *Study ID* by the Modality to identify the performed procedure. Each series, which denotes a set of acquired images, is identified by the Modality with a *Series ID* according to the nomenclature defined by the DICOM Standard.

The temporary registry data stores the patient requests. These patients have made a request and are scheduled for the day but, as they have not yet been admitted they are given no identifiers, unless they already exist in the database. This avoids the creation of empty patient records and registries from patients that have not yet been admitted are sent to the Modality through the MWL.

At the moment of admission it was made mandatory to fill the patient VAT number to tackle the patient misidentification. The VAT number is personal and allows the univocal identification of the patient. It was evidenced that, during admission, the creation of a new patient registry, when the original contained errors, is bad practice. However, such action cannot be controlled by the IS. To avoid several denominations for the same exam a look-up table was created based on the protocol lists existing in the modalities. Furthermore, and also in admission, the patients' requests are now registered according to the referring doctor's prescription and not to the patient convention rules. Thus, an interface was implemented to treat the patient request information according to the patient convention rules and create a list of provided services to be billed to the patient and stored in a financial information database, independent of the clinical information. In this sense, the consumables also no longer register as exams being managed by the Modality through the Modality Performed Procedure Step (MPPS) DICOM service. The MPPS is a complementary service to the MWL that enables the Modality to report on the performed exams. It is included in the MPPS message information relating: i) the patient demographics and IDs; ii) the exam performance, such as beginning and ending time of image acquisition, parameters used in the configuration of Modality protocols, number of series acquired, list of objects generated during acquisition; iii) the dose delivered to the patient; and iv) the consumables used during the exam such as contrast, anesthesia and number of film sheets used to print the exam [17]. The MPPS message is generated in three different study states: i) On patient registry at the Modality, taking the value IN-PROGRESS, acknowledging the moment in which the exam has started; ii) When the patient study is closed, taking the value COMPLETED, referring the moment when the exam has finished and reporting all the information regarding the performed exam; and iii) when, for some reason, the exam is discarded, taking the value CANCELLED. The activation of this DICOM service allows the modules prior to the Modality on the information flow, RIS and Management module, to return on the information sent to the Modality and be acquainted with the actual status of the operations.

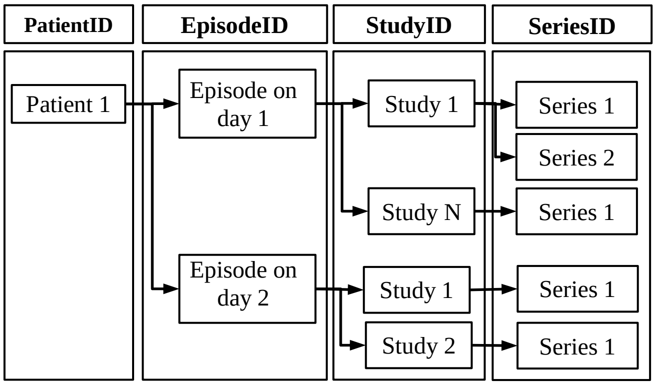


Figure 2 Illustration of the defined patient data structure.

These changes in the Management module facilitate better management of the MWL creation by RIS. The full information regarding the patient and the scheduled exam are now sent to RIS allowing both the patient and exam to be fully characterized by the Modality. This fact is of extreme importance as it not only avoids the information to be passed from the admission to the control room on paper, but also enables the Modality to create a DICOM structure making any previous exams of the same patient available to the technician. Furthermore, as RIS receives MPPS message from the Modality it is able to control the MWL according to the information sent.

Thus, patients are removed from the MWL if the status is COMPLETED or CANCELLED and maintained in queue if the status takes the value IN-PROGRESS.

The activation of the MPPS service implies some changes on the technicians' workflow. Consumables have now to be registered by the technician on the MPSS and studies have to be set to the COMPLETED status after image acquisition.

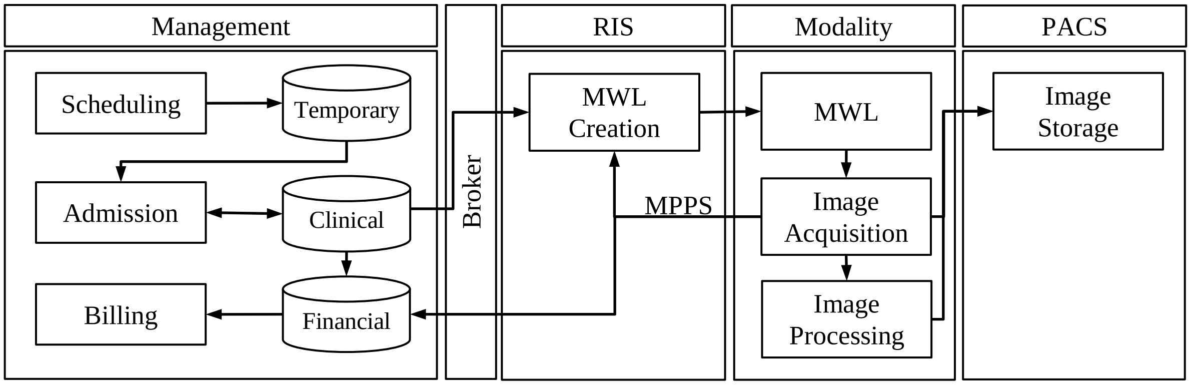


Figure 3 Illustration of the re-engineered information flow.

This action can either be performed automatically in the moment another patient is registered on the Modality, or manually by the technician when the acquisition is terminated. The activation of this service also requires that patients cannot be registered, or their demographic data changed manually on the Modality, as the MPPS message will only be sent for patients created by the MWL. The principle behind this fact is that the MPPS message will only be sent if there is a correspondence between the RIS and the Modality data, which does not happen with manual inputs. To avoid this from happening, the implementation of a management terminal on the control was suggested. As such, the technician can perform these actions without disturbing the information flow. The re-engineered information flow is shown in Figure 3.

VI. CONCLUSIONS AND FUTURE WORK

It was demonstrated that the implementation of information flows that apply the DICOM and the IHE guidelines can significantly reduce problems that affect data quality. The re-engineered information flow proposed herein significantly reduced the information problems at the studied HcP.

The implementation of the re-engineered workflow also facilitated an easier access to patients' previous examinations and minimized the number of duplicated records. The information flow has a clear and more efficient architecture, evolving towards the recommended paperless flow that minimizes human errors.

In the future it would be interesting to digitalize the referring doctor's prescription in order to make it possible to store it in PACS along with the study, as well as make it available on time to any clinical resource that has the need to consult it. Additionally, the creation of DICOM structured reports should be considered, that enable semantic queries to PACS based in fields such as the examined body region and evidenced pathologies.

REFERENCES

[1] M. Smits and G. Van der Pijl, "Developments in hospital management and information systems," in *Proceedings of the 32nd Annual Hawaii International Conference on System Sciences*, Maui, 1999, p. 9. [Online]. http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=773007&isnumber=16784

[2] W.D. Bidgood and S.C. Horii, "Introduction to the ACR-NEMA DICOM standard," *Radiographics*, vol. 12, no. 2, pp. 345-355, 1992.

[3] S. Khludov, L. Vorwerk, and C. Meinel, "Internet-orientated medical information system for DICOM-data transfer, visualization and revision," in *13th IEEE Symposium on Computer-Based Medical Systems (CBMS 2000)*, Houston, TX, 2000.

[4] "Digital Imaging and Communications in Medicine (DICOM) - Part 2: Conformance," Rosslyn, Virginia, 2011.

[5] Integrating the Healthcare Enterprise. [Online]. http://www.IHE.net

[6] M.G. Arellano and G.I. Weber, "Issues in identification and linkage of patient records across an integrated delivery system," *Journal of Healthcare Information Management*, vol. 12, no. 3, pp. 43-52, 1998.

[7] R. Cruz-Correia et al., "Monitoring the integration of hospital information systems: how it may ensure and improve the quality of data," *Studies In Health Technology And Informatics*, vol. 121, pp. 176-182, 2006.

[8] D. Bates et al., "Reducing the frequency of errors in medicine using information technology," *Journal of the American Medical Informatics Association*, vol. 8, no. 4, pp. 299-308, 2001.

[9] C. Granja, J. Mendes, F. Janela, J. Soares, and A. Mendes, "Optimisation-based on simulation: A diagnostic imaging department case-study," in *Second International Conference on Information, Process, and Knowledge Management (eKNOW)*, Saint-Marteen, 2010, pp. 32-36.

[10] J. Martins, C. Granja, A. Mendes, and P. Cruz, "Gestão do fluxo de trabalho em diagnóstico por imagem: Escalonamento baseado em simulação," *Informática de Saúde: Boas Prácticas e Novas Prespectivas*, vol. 1, pp. 85-96, 2007.

[11] C. Granja, P. Cruz, and A. Mendes, "Healthcare Decision Support System," in *10th International Chemical and Biological Engineering Conference*, 2008, pp. 1099-1104.

[12] American College of Radiology. [Online]. http://www.acr.org

[13] European Association of Radiology. [Online]. http://www.ear-online.org

[14] B.I. Reiner and E.L. Siegel, "The cutting edge: strategies to enhance radiologist workflow in a filmless/paperless imaging department," *J Digitl Imaging*, vol. 15, no. 3, pp. 178-1900, 2002.

[15] Eliot L. Siegel and Bruce Reiner, "Work Flow Redesign: The Key to Success When Using PACS ," *Journal of Digital Imaging* , vol. 16, no. 1, pp. 164-168, 2003.

[16] M.D. Ralston, R.M. Coleman, D.M. Beaulieu, K. Scrutchfield, and T. Perkins, "Progress toward paperless radiology in the digital environment: Planning, implementation, and benefits," *Journal of Digital Imaging*, vol. 17, no. 2, pp. 134-143, 2004.

[17] Digital Imaging and Communications in Medicine. [Online]. http://medical.nema.org

Risk as a Subjective Construct: Implications for Project Management Practice

Jose Irizar
Information Systems Department
TRW Automotive
Germany
Jose.Irizar@googlemail.com

Martin Wynn
School of Computing and Technology
University of Gloucestershire
Cheltenham, UK
MWynn@glos.ac.uk

***Abstract* - The management of risk is a key element of all mainstream project management methodologies. It has implications for the effectiveness of the project management process itself, and for the management and communication of knowledge that is an inherent part of that process. There are two main schools of thought regarding project risk management – 'risk as an objective fact' and 'risk as a subjective construct'. The former considers risk as epistemologically probabilistic, whilst risk in the subjective construct perspective allows multiple epistemological dimensions of risk. Here we review how 'risk as a subjective construct' features in existing risk management literature, and how these contributions can be classified or grouped together. The role of risk registers is then reviewed to determine whether this has any relationship with the 'risk as a subjective construct' concept. The paper then reflects upon the authors' future research programme and the possible implications for project management practice.**

Keywords - risk; risk analysis; subjective construct; project management; knowledge management; perception; stakeholders

I. INTRODUCTION

Project management is an established discipline in traditional industries such as engineering and construction, and other industry sectors such as education, IT, health, pharmacy and surgery have adopted project management in their organizations in recent years [1]. Project management has also grown from a tactical to a strategic discipline, with project managers playing an increasingly significant role in the execution of senior management business strategy [2]. Strategic project management, project performance tracking and systematic assessment of lessons learned may underpin strategy revisions and adjustments [3].

Despite the recognized criticality of project success for organizations, a considerable proportion of projects continue to either not meet their due dates, exceed budget, do not deliver the specifications, miss quality, underestimate risk or do not meet customer satisfaction. That is why project management failure remains an area of considerable interest in contemporary project management literature [4].

Formal risk management is a relevant part of project management. In fact, risk management has been identified as one of the major criteria for project success [4]. Hence, risk management has become a central component of some of the most deployed industry standard methodologies such as Project Management Body of Knowledge, PRINCE2, Systems Development Life Cycle, Capability Maturity Model Integrated, and Information Technology Infrastructure Library.

Comprehensive risk management implementation increases the probability of project success [5]. It is considered as the tool that limits the effect of unexpected events or prevents such events from happening. Therefore it is assumed, that risk management as part of project management contributes to overall project success [6]. Contemporary risk management literature can be assigned to two distinct schools of thought, risk as an objective fact and risk as a subjective construction. Both schools provide different definitions of risk, both are based on different ontological and epistemological principles, and both handle risk in a different manner [11].

Risk management is one of nine project management knowledge areas defined in the Project Management Body of Knowledge. Among project management practitioners it is one of the most critical activities for project success together with communication, resource planning and scheduling. Project management should always include risk management [7]. The Association of Project Management identifies and separates out a series of hard and soft benefits (see Table I) from deploying risk project management [8]. Bartlett [9] stresses individual benefits, and concludes that the major impact on deploying risk management resides in focusing the way the team members think, behave and work together. One further conclusion of the author is the contribution of project risk management to the organization as a means of identification of threats to the organization.

However, it is generally accepted that organizations tend to lack application of this knowledge. Bannerman [10] suggests the existence of a gap between the development of risk and risk management in the literature and the needs of the phenomenon in practice. Not only does there appear to be a disconnect between risk focused management research and the needs of project risk management in industry, but also the converse - the adoption of risk concepts and risk management methods in practice lags behind the new concepts and understandings found in the literature. Researchers and practitioners still have to learn from each other to reduce the level of project failure.

Different schools of risk analysis provide different risk definitions which may have significant implications and impact on the management of risk in the context of project management. Generally speaking two schools of thought have crystallized on project risk management, ‘risk as an objective fact’ and ‘risk as a subjective construct’. Risk as an objective fact considers risk as epistemologically probabilistic, while risk in the subjective construct perspective allows multiple epistemological dimensions of risk, encompassing experience, organization, culture and society which are to be taken into account to manage risk in the context of project management.

‘Risk as a subjective construct’ opens a new opportunity and approach to risk management, a new perspective on the creation and use of risk registers, and engenders a two-way communication process between stakeholders and project manager [11].

The research questions (RQs) addressed in this paper are:

RQ1: To what extent does 'risk as a subjective construct' feature in existing risk management literature, and how can these contributions be classified or grouped together?

RQ2: What is the nature of existing literature on risk registers and what relationship does it have with the 'risk as a subjective construct' concept?

RQ3: How could project risk management theory and practice be informed or improved by an assessment of the 'risk as a subjective construct' concept.

This introductory section is followed by a discussion of the research methodology, based on a detailed literature review. Findings and analysis are presented in section three and the final section draws together some conclusions from work completed to date and briefly outlines the authors' future research intentions.

II. Methodology

In the last two decades, qualitative research has found increasing recognition in the project management field [12]. A large number of empirical studies using qualitative data are available in academic literature and specialized journals. At the same time, management researchers and practitioners in particular rely on evidence-based policy [13]. In fact, most of the existing generally accepted standards in the field of project management are built around evidence-based policy and best practice.

The systematic review deployed in this research assumes that it is feasible and sensible to cumulate findings and generalize results to create new knowledge. The review attempts to identify, evaluate and interpret all available research relevant to the three research questions. The overarching aim is to synthesize existing evidence in a fair, rigorous, and open manner. This systematic approach is an aid to the grouping and structuring of findings, which will grow over time as new relevant materials are published, alerts collected and the search parameters or sources are adjusted.

An initial literature scoping exercise encompassed a range of disciplines that contribute to the discussion of the validity of the subjective construct of risk in the context of project management [11]. As a first step, evidence on 'risk as a subjective construct' was documented. The intention is to expand previous studies [11] by extending the period observed as well as the source of literature. Similar to other systematic researchers [14] who also accept the premise that

TABLE I. Hard and Soft Benefits of Project Risk Management [9]

	‘HARD’ BENEFITS		‘SOFT’ BENEFITS
H1	Enables better planning, scheduling and budgeting.	S1	Improves corporate experience and general communication.
H2	Increases the likelihood of a project adhering to its schedules and budgets.	S2	Leads to a common understanding and improved team spirit.
H3	Leads to the use of the most suitable type of contract.	S3	Helps distinguish between good and bad management (and good and bad luck!).
H4	Allows a more meaningful assessment of contingencies.	S4	Helps develop the ability of staff to assess risks.
H5	Discourages the acceptance of financially unsound projects.	S5	Focuses project management attention on the real and most important risks.
H6	Contributes to the build-up of statistical information for better decision- making.	S6	Facilitates greater risk-taking, thus increasing benefits gained.
H7	Enables a more objective comparison of alternatives.	S7	Demonstrates a responsible approach to clients.
H8	Identifies and allocates responsibility to the best Risk Owner.	S8	Provides a fresh view of the personnel issues on a project.

project management is under-represented in the leading management research journals, the current paper concentrates on the two flagship project management journals, Project Management Journal (PMJ) and International Journal of Project Management (IJPM), established in 1969 and 1983 respectively. The search which identified risk as subjective construct was broad, combining automated and manual searches. There were identified peer-reviewed articles published up to July 2012. Discovery service EBSCO search engines and indexing systems were used; in addition bibliographies of the initial papers were scanned for additional papers. The combined research strategies provided 90 articles for the RQ1 and 15 for RQ2.

Five areas were identified as interpretative contexts to understand 'risk as a subjective construct':

- Individual risk constructions
- Conflicts and contradictions
- Multiple rationalities
- Complexity - Size
- Perspective to project result / end product

For synthesizing the studies the technique chosen is ‘lines of arguments’. The articles selected examine different aspect of the same phenomenon. The interest for one author may be more focused on the disaster feature; the next may be stressing the uncertainty aspect; both relate to complex, big sized projects with stakeholders of disparate backgrounds. The ‘lines of arguments’ uses categories surfacing from the data. In the next section, as part of the analysis, categories are linked with personal interpretation to offer a holistic version of the risk analysis by the selected authors.

III. Findings and Analysis

As regards RQ1, the quality criteria applied to select the articles were:

1. A focus on project risk
2. Addresses real projects, case studies – not a theoretical discussion
3. Relates to risk, uncertainty or failure or risk analysis/risk register

The studies which clearly fulfill all of the three quality criteria were graded ‘A’ (Table II). An analysis of this literature suggests that one defining characteristic of 'risk as a subjective construct' is the way risks are identified. The identification of risk as a subjective phenomenon coincides with its creation – the risk exists only once the stakeholder has identified it. This is particularly noticeable for risks linked to an organization’s own qualities and deficiencies. Such risks show a significant limitation compared with traditional risk analysis based on external threats and probabilistic consequences. One further characteristic seems to be that risks apparently not identified by the existing project management systems are the ones originated by the organizational pathogens or organizational latent conditions. These are causes of failure, are created by actors, and often occur after a prolonged period, becoming evident or problematic after an adverse event occurs. Such conditions are the result of the individual’s subjective interpretation (example: ring-fencing of funds for particular task against other tasks, investment flexibility becomes limited, and is then followed by unforeseen calls for other tasks). One stakeholder’s pathogen is another stakeholder’s protection. These constructions may move from protection to pathogen in the project life cycle. Such different constructs engender discrepancies during project development. Failure may not affect all stakeholders. Organizational pathogens can be better treated as subjective interpretations [15].

Failure in complex systems provides evidence of competing and contradictory demands. The multi-nodality of complex systems shows conflicts and contradictions. These conflicts and contradictions are interpreted as deviations and misunderstandings by traditional project management. The practice of mixed top down/bottom-up, local empowerment and top down responsiveness does not easily fit with academic project management methodologies. Ivory and Alderman [16] suggest that predictive project management models cannot necessarily capture such complex models as evidenced with several case studies using qualitative data from three complex industrial projects. Projects are built as framed linear and non-linear interaction – the deconstruction of these interactions provides the opportunity to identify non-linear interactions, in which inputs lead to unexpected outputs and possible project failure. The analysis of NASA recurrent disasters shows the prevalence of different and opposing risk rationality within a project oriented organization, in which certain leaderships, with certain objectives, influence risk handling with fatal consequences [17]. This provides a paramount example of different perceptions and expectations (commercial vs. safety) with different risk constructions leading to collapse.

A knowledge based risk assessment template has been developed [18] to analyse the risks from both supply-side and demand-side perspectives. The author takes into account the fact that project participants may have perspectives that differ from those of the project manager. Both the definition and existence of risk phenomenon are considered as subjective in this model. Particularly interesting is Marrewijk et al’s contribution [19] to the concept of multiple rationalities in megaprojects. The authors oppose the assumption of projects having a single or shared rationality. Megaprojects offer - through their documentation and contract - a great example of the opportunities for ambiguities and multiple interpretations. Project participants with diverse cultures and rationalities will have different perceptions of uncertainty, ambiguity and risk. All this makes it hard to make “rational” and “consistent” decisions in such projects [20].

An area related to the megaprojects example is the build-operate-transfer (BOT) concession model. Yeo and Tiong [21] try to answer the question of how to positively manage such differences to achieve convergence of results. Actors in such a constellation are typically the representative authorities of the host governments, entrepreneurial promoters and banks. They represent different constructions of risk with different perceptions and expectations, values and motives. To sort out some of the conflicts that arise because of these different constructions of risk, the authors recommend an approach based on the Soft Systems Methodology [22].

Risk analysis could be subject to the impact of interest. The interest of the project owner is likely to be distinct from that of the project manager or a project team member. These different interests can be categorized as operational and strategic, and can lead to a different handling of operational and strategic risks. Krane et al [23] identify three major groups of stakeholders who will have different project objectives. They are:

1. The project team with a 'project internal' perspective, focusing on the project's deliverables, costs, and schedule. This is typically seen in a very short time perspective.
2. The customer or user, focusing on the benefits of the project or the project's direct effects. The time perspective will necessarily be somewhat longer than the project's perspective. They are the ones who will live with the end-product once the project is finished.
3. The project owner with a longer-term strategic perspective on the project.

TABLE II. GRADE 'A' ARTICLES USED IN LITERATURE ANALYSIS FOR RQ1

Author	Title of work	Region/Detail	Knowledge contribution	Practice implications
[15]	The pathogen construct in risk analysis	UK - Based on interviews to 22 project members	Risk origination in the way an organization sees the world. Pathogen link to practice and its subjective interpretation	Enhancement of risk identification by querying contradictory interpretations of the same entity.
[16]	Project Managements learns from complex systems failure	UK - Detailed examination 3 large size projects	Recognition of conflicts and contradictions as opposed to deviations and misunderstanding as consequence of diverse social positions, organizational responsibilities, values, and culture	Suggestion of mixed top-down/bottom-up approach to management.
[16]	Organizational behaviour and disaster: NASA	USA - Detailed examination of 2 large projects. Access to investigation board documentation and lessons learned.	Risk analysis impacted by interests, values and culture and objectives; risk phenomenon is subjective	Integrate devil's opinion, independent quality review, identification and elimination group behaviour; setup multiple groups under different leaderships to work on critical issues
[31]	Deviation, ambiguity , uncertainty project-organization	USA – The Millcorp case study	Knowledge sharing through interaction and communication between teams and contexts to enhance risk understanding	Interactive risk identification and analysis processes
[30]	Fall of firefly	USA – case study	Different risk constructions	Provides lessons learned / check list to identify different risk constructions for practitioners
[23]	Project Manager–Project Owner Interaction influences risk management	Norway - Analysis of 7 big complex projects	Project Owner – Project Managers' interaction results in lack of strategic risk attention, strong focus on operational risk	Recommendation to emphasize the identification of more short- and long-term strategic risks at all stages of projects
[18]	A knowledge-based risk assessment framework for evaluating web-enabled application outsourcing projects	Global	Practical tool to assess specific stakeholder risks	Template available
[19]	Managing public–private megaprojects: paradoxes, complexity, and project design	Netherlands and Australia	Multiple rationalities	Adequate project design to accommodate partners' culture to enable cooperation to achieve project objectives
[21]	Positive management of differences for risk reduction in BOT projects	Turkey, Thailand, Indonesia, Australia, Bangladesh, Malaysia, Canada and Hong Kong	Analysis of different constructions of risk	Proposed soft systems methodology to achieve convergence

These authors also use the term 'risk management systems' associated with the various stakeholder groups and point out that these distinct systems must communicate. The risk management systems communication is more a collaboration process to address the holistic view of the project and less a 'risk register' for dissemination to project participants.

The risk register review (RQ2) does not require an integrative synthesizing analysis, but Table III details the grade 'A' articles that were reviewed (using the same quality criteria as noted above). Risk registers are widely used as a tool or template. Integration and simplicity are common requirements. All of the selected studies described tools or analysis that could be adapted to incorporate and integrate several constructions. Project management applications supporting stakeholders' collaboration are primarily built according to specifications based on a project manager centric risk viewpoint. Although it may be too early to provide a definitive answer, it appears feasible to adapt the current systems and templates structure to incorporate several risk constructions. It is possibly not that much of a technical challenge to incorporate different risk constructs, but more of a challenge to the people involved to adopt and deploy the required processes.

Krane et al [23] [24] address the issue of how risks, once identified, are then distributed amongst different risk categories. This analysis provides a comprehensive general overview of how to deal with different risk items using a risk register. This comprehensive study proposes a basic categorization based on the levels of hierarchy of management objectives. They define the three categories as follows:

TABLE III. GRADE ‘A’ ARTICLES USED IN LITERATURE ANALYSIS FOR RQ2

Author	Title of work	Knowledge contribution	Practice implications
[32]	Categorizing risks in seven large projects	Analysis of 7 big complex projects – Categorization of 1450 risk elements as operational, short- term or long-term strategic.	Questions of identification and assessment of operational short-term strategic and long-term strategic guidelines
[8]	Knowledge based proactive risk management	Detailed description of mature interactive software application	Recommendations for adapting stakeholders interaction and collaboration
[33]	Integrated Methodology for Project Risk Management	Comprehensive practice example of risk management with risk register development including Delphi Analysis for final validation	Framework available for project owner - external consultant- collaboration that offers opportunity to adapt to other particular projects
[34]	Risk avoidance in bidding for software projects based on life cycle management theory	Integration of ‘bidding risk’ with project-life-cycle and risk response measures	Model suggested method for forecast, prevent, discover and reduce related risk completely and in a timely fashion, thus enhancing the probability of a successful bid.
[29]	Intervening conditions on the management of project risk: Dealing with uncertainty in information technology projects	Provides approaches to ensure risk is assessed; and to overcome practitioners risk management ineffectiveness perception	Application to improve risk management techniques when conflicting risk perceptions as opposed to deal with issue
[35]	Development of a Model for Risk Management at Corporate, Strategic Business, and Project Levels	Methodology to ensure risk management integration with internal and external stakeholders, in line with McGill risk knowledge database	Mainly theoretical, ensures consistency of risk register usage with corporate strategy
[36]	Comparing project management practices in new product development: a study in the automotive, aerospace and rail transport industry	Comparative approach, identification of best practices and suggestion for best practices transfer	Recommendations for best practice solutions exchange at inter-industry and intra-industry level
[37]	A Risk Register Database System to aid the management of project risk	Presentation of project risk assessment and project risk register in an automotive company through a project lifespan	Information about design and construction of risk registers in the automotive manufacturing industry
[38]	Project risk management practice: The case of a South African utility company	Risk management in practice involving stakeholders, adherence to very simple risk management processes	Insight, information for practitioners on how to integrate stakeholders using very simple risk management process. Ensures risk management is as part of mainstream business activities.
[39]	Risk management practices of leading UK cost consultants	Insight of usage of risk assessment techniques and risk registers	Outline of options for risk management approaches

1. Operational Risks—risks related to operational objectives of the project. This means risks related to the direct results from the project: its products.

2. Short-Term Strategic Risks—risks related to short-term strategic objectives of the project. The project owner will have a set of objectives related to his/her use of the project results. The short-term strategic risks are the risks related to those objectives, or the risks concerning first-order effects of the project— that is, risks pertaining to the effects that should be achieved for the target group or end users of the project.

3. Long-Term Strategic Risks—risks related to the long-term strategic objectives of the project. This means those risks related to the project purpose—the long-term objective that the project is meant to contribute to.

As regards RQ3, Macgill and Siu [25] stress the importance of a single architecture of risk knowledge as its epistemology. Their proposal is the establishment of a risk knowledge database for the promotion of knowledge dissemination. Macgill and Siu [26] also recognize the importance of risk analysis not as a philosophical or intellectual exercise, but as an applied discipline. The novelty of this proposal resides in the observation of the risk phenomena not only as a physical but also a social issue. People’s perception of risk results in acceptability of risk that does not necessarily correspond with the consensual body of peer group scientific knowledge. In this construct, risk based on people knowledge and the constructed social reality is not univocal. Risk appears to be contextual and the social actions adopted by individuals when facing a risk are related to their knowledge.

With this relatively new approach to risk, a new school of thought has been identified in the area of project risk management. Zhang [11] provides a systematic review of the position of 171 articles published between 1999 and 2009 regarding project risk in the two leading project management journals, Project Management Journal and the International Journal of Project Management. Only 12 out of the 171 articles belong to the 'risk as a subjective construct' category. The same author also suggests future exploration on identifying the epistemological dimensions of subjective risk for developing methods and tools to assess and evaluate subjective risk in the context of project management.

A significant blocker to developing and applying new risk management theory is its low uptake as a concept in industry [27]. A fuller understanding of the implications of risk as a subjective construct has the potential to significantly enhance the management of risk in projects and thus overall project outcomes. Some studies have already explored how mainstream project methodologies can be adapted to different company contexts [28], and recognition of the different origins and dimensions of risk can be seen in this context. The perceived effectiveness of different project risk management assessments has been analyzed to identify the root causes for manager's reluctance to deploy risk management processes and its consequences. This study [29] is a valuable input that could underpin an enhanced application of risk assessment in project management.

IV. Conclusion and Future Work

Of the authors that address 'risk as a subjective construct', no one presents a complete solution and none of them proposes a risk management system that could integrate several risk constructs. At the same time, literature on risk registers and current risk management is available but no relationship could be found between any of the existing proposals and the 'risk as a subjective construct' concept. This poses the question of how to adjust current approaches in order to integrate more than one risk construct.

What is clear is that different stakeholders see different realities - as Peter Drucker has put it, 'when intelligent, moral, and rational people make decisions that appear inexplicable, it's because they see a reality different to the one seen by others' [30]. This phenomenon, in the case of risk, has no unique or universally accepted interpretation and it thus requires further research and enhancement, which the authors are pursuing with regard to project management practice in the German automotive sector. In order to develop further knowledge in this field, interaction and communication between project teams and their contexts will be required. This knowledge - context dependent, situational, shared - will be created collectively [31]. If it can be successfully harnessed within project management methodologies and disciplines, it has the potential to significantly enhance eventual project outcomes.

References

[1] D. E. Hodgson, "Disciplining the professional: The case of project management," *Journal of management studies*, 39(6), . 2002, pp. 803-821.

[2] A. Brown, "Getting your projects to meet strategic goals," Paper presented at the PMI Global Congress Proceedings, Sydney, Australia and Denver, CO, USA, 2008.

[3] J. Tharp, "Align project management to organizational Strategy," Paper presented at the PMI Global Congress Proceedings, Hong Kong., 2007.

[4] D. McClure, *From the CIO trenches: Why some projects fail and others succeed*, Gartner Industry Research, 2007.

[5] R. Jen, Visual Ishikawa Risk Techique (VIRT) - An approach to risk management, *PMI Virtual Library*, 2009. Available: http://www.pmi.org/en/Knowledge-Center/Knowledge-Shelf/~/media/Members/Knowledge%20Shelf/Jen_2009.ashx. Retrieved: January, 2013.

[6] K. de Bakker, "Risk management affecting IS/IT project success through communicative action," *Project Management Journal*, 42(3), 2011, pp. 75-90.

[7] PMI., *A guide to the project management body of knowledge* (PMBOK®) (Fourth ed.) Project management institute, Inc., 2008.

[8] J. Arrow, "Knowledge-Based proactive project risk management," *AACE International Transactions*, June 2008, pp. 1-9. Available: http://ehis.ebscohost.com/eds/pdfviewer/pdfviewer?sid=7d282fa8-4eb3-4756-86ad-5f2d4aa0d29d%40sessionmgr15&vid=2&hid=1. Retrieved: January, 2013.

[9] J. Bartlett, *Project risk analysis and management guide*, Buckinghamshire: APM Publishing Limited., 2004, pp. 5-13.

[10] P. L. Bannerman, "Risk and risk management in software projects: A reassessment," *The journal of systems and software*, 81(12), 2008, pp. 2118-2133.

[11] H. Zhang, "Two schools of risk analysis: A review of past research on project risk," *Project Management Journal*, 42(4), 2011, pp. 5-18.

[12] T. Biedenbach and R. Müller, "Paradigms in project management research: examples from 15 years of IRNOP conferences," *International journal of managing projects in business*, 4(1), 2011, pp. 82-104.

[13] A. Bryman and E. Bell, *Business research methods*, Oxford : Oxford University Press, 2007, 2nd ed.

[14] B. Hanisch. and A. Wald, " A Bibliometric View on the Use of Contingency Theory in Project Management Research," *Project Management Journal*, 43(3), 2012, pp. 4-23.

[15] J. S. Busby and H. Zhang, "The pathogen construct in risk analysis," *Project Management Journal*, 39(3), 2008, pp. 86-96.

[16] C. Ivory and N. Alderman, "Can project management learn anything from studies of failure in complex systems?" *Project Management Journal*, 36(3), 2005, pp. 5-16.

[17] R. D. Dimitroff, L. A. Schmidt, and T. D. Bond, "Organisational behavior and disaster: a stydy of conflict at NASA," *Project Management Journal*, 36(2), 2005, pp. 28-38.

[18] W. L. Currie, "A knowledge-based risk assessment framework for evaluating web-enabled application

outsourcing projects," *International Journal of Project Management*, 21(3), 2003, pp. 207-217.

[19] A. van Marrewijk, S. R. Clegg, T. S. Pitsis, and M. Veenswijk, "Managing public–private megaprojects: Paradoxes, complexity, and project design," *International Journal of Project Management*, 26(6), 2008, pp. 591-600.

[20] M. Olzmann and M. Wynn, "How to Switch IT Service Providers: Recommendations for a successful transition," *International Journal On Advances in Intelligent Systems*, vol 5, no 1&2, 2012, pp. 209-219, IARIA journals.

[21] K. T. Yeo and R. L. K. Tiong, "Positive management of differences for risk reduction in BOT projects," *International Journal of Project Management*, 18(4), 2000, pp. 257-265.

[22] P. Checkland and J. Scholes, *Softt Systems Methodology in Practice*, J Wiley, Chichester, 1990.

[23] H. P. Krane, N.O.E. Olsson, and A. Rolstadas, "How project manager-project owner interaction can work within and influence project risk management," *Project Management Journal*, 43(2), 2012, pp. 54-67.

[24] H. P. Krane, A. Rolstadas, and N.O.E. Olsson, "Categorizing risks in seven large projects—Which risks do the projects focus on?" *Project Management Journal*, 41(1), 2010, pp. 81-86.

[25] S. M. Macgill and Y. L. Siu, "A new paradigm for risk analysis," *Futures*, 37, 2005, pp. 1105-1131.

[26] S. M. Macgill and Y. L. Siu, "The nature of risk," *Journal of Risk Research*, 7(3), 2004, pp. 315-352.

[27] R. Hynuk Sanchez, M. Bourgault Benoit, and R. Pellerin, "Risk management applied to projects, programs, and portfolios," *International journal of managing projects in business*, 2(1), 2009, pp. 14-35.

[28] M. Wynn. P. Turner, H. Abas, and R. Shen, "Employing knowledge transfer to support IS implementation in SMEs," *Journal of Industry and Higher Education*, Volume 23, No 2, April, 2009, pp. 111-125.

[29] E. Kutsch and M. Hall, "Intervening conditions on the management of project risk: Dealing with uncertainty in information technology projects," *International Journal of Project Management*, 23(8), 2005, pp. 591-599.

[30] Bud Baker, "The fall of the firefly: An assessment of a failed project strategy," *Project Management Journal*, 33(3), 2002, pp. 53-57.

[31] M. Hällgren and E .V. A. Maaninen-Olsson, "Deviations, ambiguity and uncertainty in a project-intensive organisation," *Project Management Journal*, 36(3), 2005, pp. 17-26.

[32] H. P. Krane, A. Rolstadas, and N.O.E. Olsson, "Categorizing risks in seven large projects—Which risks do the projects focus on?" *Project Management Journal*, 41(1), 2010, pp. 81-86.

[33] A. del Caño and M. P. de la Cruz, "Integrated methodology for project risk management," *Journal of Construction Engineering & Management*, 128(6), 2002 pp. 473-485.

[34] G. Xie, J. Zhang, and K. K. Lai,. "Risk avoidance in bidding for software projects based on life cycle management theory," *International Journal of Project Management*, 24(6), 2006, pp. 516-521.

[35] A. Merna and T. Merna, "Development of a model for risk management at corporate, strategic business, and project levels," *Journal of Structured & Project Finance*, 10(1), 2004 pp. 79-85.

[36] A. K. Müller, A. Wald, and A. Görner, "Comparing project management practices in new product development: a study in the automotive, aerospace and rail transport industry," *International Journal of Project Organisation and Management*, 4(3), 2012, pp. 203-217.

[37] F. D. Patterson and K. Neailey, "A risk register database system to aid the management of project risk," *International Journal of Project Management*, 20(5), 2002, pp. 365–374.

[38] R. van Wyk, P. Bowen, and A. Akintoye, "Project risk management practice: The case of a South African utility company," *International Journal of Project Management*, 26(2), 2008, pp. 149-163.

[39] G. D. Wood and R. C. T. Ellis, "Risk management practices of leading UK cost consultants," *Engineering Construction & Architectural Management*, 10(4), 2003, pp. 254-262.

Process Mining in Manufacturing Company

With Focus on Process Simulation and Prediction.

Milan Pospíšil
Department of Information Systems
BUT, Faculty of Information Technology
Brno, Czech Republic
ipospisil@fit.vutbr.cz

Vojtěch Mates
Department of Information Systems
BUT, Faculty of Information Technology
Brno, Czech Republic
imates@fit.vutbr.cz

Tomáš Hruška
IT4Innovations Centre of Excellence
BUT, Faculty of Information Technology
Brno, Czech Republic
hruska@fit.vutbr.cz

***Abstract*—Simulation can be used for analysis, prediction, and optimization of business processes. Nevertheless, process models often differ from reality. Data mining techniques can be used for improving these models based on observations of process and resource behavior from detailed event logs. More accurate process models can be used not only for analysis and optimization, but for prediction and recommendation as well. This paper analyses process model in manufacturing company and its historical performance data. Based on that observation, simulation model is automatically created and used for analysis, prediction and for dynamic optimization.**

Keywords-business process simulation; business process intelligence; data mining; process mining; prediction; optimization; recommendation

I. INTRODUCTION

Classic simulation can be used for the analysis of business processes. It is useful to try many variants of processes, measure the effects, and then decide on the optimal process settings. For example, the process can be redesigned, it is possible to change resource allocation, and search for the most optimal configuration with respect to context-based requirements (price, effectiveness, customer satisfaction, etc.). The current process configuration can be tested for how many cases it can handle over periods of time.

These models can be built manually, which is time consuming and error prone. The main disadvantage is that this approach cannot be used for predictions for operational decision, but only for strategic decisions. The operational decisions are important for internal logistics purposes. The casual models have some simplifications – for example probabilities of routing and naive execution time of task. These parameters are set based on long observation of processes, so they can work in long-term simulation for strategic decisions. Nevertheless, operational decisions need short-term simulation. These two simulation types differ significantly. Short-term simulation starts in current state of the process with allocated resources, cases in progress with known parameters and with waiting cases to handle. Routing probabilities and execution times can differ significantly for different case parameters, thus mine deeper dependencies is needed.

For example, assume repair process, there are two tasks – repair basic item and repair advanced item, repair basic item is executed in 90% of cases, repair advanced only in 10% of cases. Execution time of basic item is about one hour and execution time of advanced item is about eight hours. If our case has known attributes and it is usually available in runtime, based on data mining, which these attributes lead to advanced repair with 80% probability, classic routing probabilities are precise enough to be used. And there is another problem – execution time of task is also influenced with case attributes – some case attributes leads to longer execution time. Resources have to be also taken into account, e.g. some people work faster, some slower.

Predictions, recommendations, and dynamic optimizations could be accomplished by operational simulation. The system can warn us, that some cases will be probably late based on historic performance data. Then some different scenarios can be simulated and evaluated, then the system can recommend us actions and provide dynamic optimization of current running cases – for example; assign extra resources from non-critical case to critical, or use a different sub-process – when we have a slower / cheaper version or faster but more expensive.

This paper analyses processes of manufacturing company. Simulation model is built using process mining and used for predictions. Based on these predictions, managers can change priorities (reallocation of resources) or better plan their storage space, because working front is known, therefore they can better predict manufacturing time.

This work is based on our previous research and verifies our theory on process mining and simulation field [15].

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-254-7

II. RELATED WORK

Data mining techniques can be used in Business Process Management. This new area was called Process Mining [3, 6, 12, 13, 14]. It was based on analysis of information from event logs that were produced by business processes. Process discovery (figure 1) is one of the methods and it is able to find a process model from an unknown process using many sequence examples of tasks and case parameters.

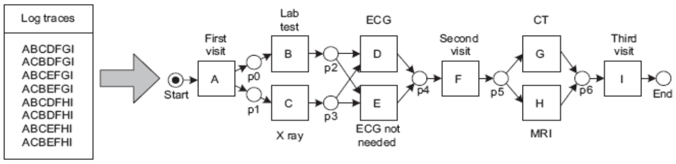


Figure 1. Process discovery (taken from [5]). It possible to discover a process model from log. The discovered process model must be able to replay most log traces.

Not only process model could be discovered, but also decision rules and social networks [5, 6, 10] and simulation models [5, 10, 11, 15]. Resource behaviour is also point of interest [8, 9]. Example of simulation model [5] is depicted in figure 2. It is possible to see routing probabilities and decision rules (decision rules are used when case attributes are known – that leads to better routing rules) and it is possible to see time distribution of tasks.

Some other research on process prediction was published in [1, 2, 4, 7, 10]. Wetzstein [4] used decision trees to analyse process performance in figure 3. As it can be seen, response time of banking service is higher than 210, KPI (key performance indicator) is always violated. If customer id is 123, manager can observe process bottlenecks, he can try to make banking service faster or find out why customer 1234 has problems.

Grigori [1, 2] uses similar approach, not for analysis, but predictions. Huge classifier is learned based on case attributes, start, and end time execution of tasks. Classifier can predict final time execution of case based on case parameters and time information from executed tasks. Evaluation of that approach compared to our approach is discussed in [15]. In addition, our work uses similar approach as [1, 2, 4] but it combines it with process mining.

Finally, when we mine deeper dependencies about routing rules and execution time of cases, we can use it for simulation for decision support [15]. Our previous work [15] is extension of papers [5, 10] and it adds some important features, some inspired by papers [1, 2, 4]. For example, execution time of cases is also predicted by classifier like decision rules.

This paper shows theory of [15] can be applied in real large manufactory company.

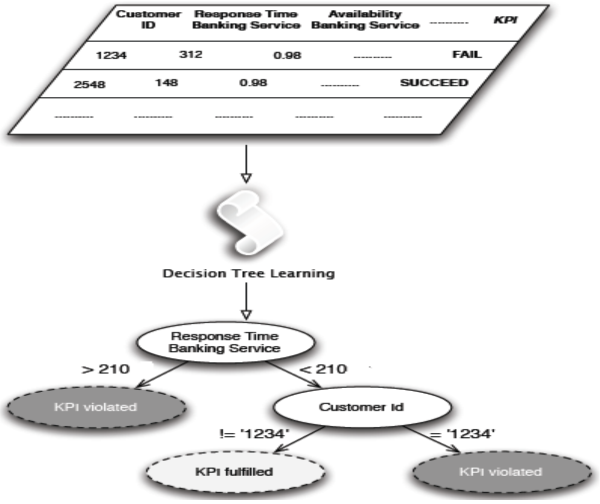


Figure 3. Process performance analysis (taken from 4). Decision tree is used for discovering factors that leads to KPI violation. We can see that KPI is violated when response time of banking service is larger than 210.

III. MORE PRECISE SIMULATION MODEL

As it is said in [15], there is need to build more precise model than the one described by papers [5, 10]. We will describe steps needed to accomplish that:

A. *Process Discovery*

If process is not known, it is possible to discover it using process discovery techniques. However, process discovery is not the most needed method for building simulation model. If explicit model is not present then it is possible to discover it, but the precision of the model will be lower than explicitly given real model. In some companies, discovered

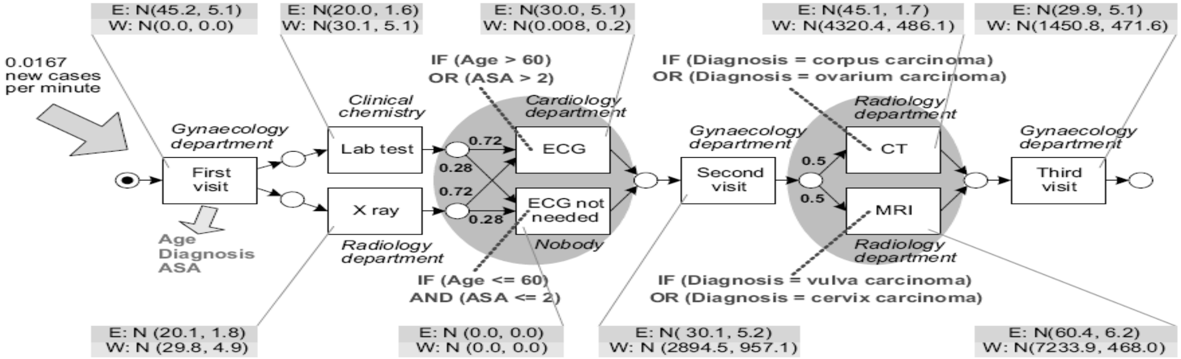


Figure 2. Simulation model [5]. Classic simulation model (taken from [5]) is enhanced by decision rules. Decision rules can make our routing probabilites more precise, because they depend on case attributes.

model could be more precise than official model but this is because these companies do not have their model formalized so well. This is not the case for manufacturing companies where prediction and usage of short time simulation is considered better.

B. Decision Mining

Decision mining is based on discovering routing rules in OR split nodes. These rules could be available too but sometimes they are not applicable. Assume situation, routing rule is based on one parameter that is inserted into system just before the decision. Thus, our predictor will know next path only in the time of decision – useless prediction. In these situations, decision mining has to be used. Decision mining is described in [5, 10, 15]. Classifier is learned on training data where inputs are case attributes and output is next path in process. Our work [15] describes another problem and that is missing attributes or 100% precise attribute known in the time of decision inserted by human (described earlier). If some attributes are missing then classic classifiers will not work in proper way. If there is 100% precise attribute then classifier is based only on that attribute. Solution is the same for both problems – it is necessary to build several classifiers for several milestones of the process – from the start (only subset of case attributes are known) to the end (all attributes are known).

C. Execution Time of Tasks

Execution time of tasks is the most important issue in short-term simulation. Process model and routing rules are important as well. However, in companies with predictable business processes (especially manufacturing companies), control flow and routing rules are used to be formalized.

Execution time of tasks will be described precisely in Section V.

IV. MANUFACTURING COMPANY

Our manufacturing company produces doors. Doors have their attributes (about twenty) and based on attributes, different operations are executed. Doors have different material, size, weight, different corner and edge types, different handle and glasses, etc. Every door has its ID and it can be modeled as case. Doors are manufactured in machines (tasks). Some machines work in parallel; some machines are bound to several tasks, so these machines must be treated as resources, because machine could be busy or working. People are working with machines or in manual workplaces. Routing probabilities are 100% accurate, because doors with specific attributes must be manufactured only by specific machine and with specific settings.

Resources are quite predictable, because they work on shifts and they are always available and planned several days ahead. The only unknown parameter is execution time of tasks that depends on case attributes – every case is modeled as one door, so case attributes are door parameters. Door parameters are known at the beginning of the process and are constant, so there is no need to build several classifiers for several periods of case execution [15]. Execution time also depends on people work rate, work queue and error rate (especially in manual workplaces), but this is issue beyond the paper.

Context-based predicting execution time of tasks quite precisely can help with several issues. First, managers can decrease storage spaces, because they could plan execution order of cases in order to decrease waiting times. Our prediction decreases variances of execution time and logistics have methods to plan storage spaces when there is low variance. They will also know if some doors will be probably late and for example, they can respond to that changing priorities, resource allocation, etc. Another useful issue is the analysis. Managers could measure which door types takes long time to produce and better calculate their price. For logistics, execution time is not as much important as influence of variance of execution time. It is possible to measure which door types (based on parameters) have high variance. Managers can focus on that door types and try to find out the cause of high variance, or produce them only in situations (if it is possible to wait) when variance is not such important issue.

V. PREDICTION OF EXECUTION TIME OF TASKS

The time deviation is sometimes high, but it can be decreased by data mining techniques. Thus, it is useful to examine data and find relationships between case parameters and execution time for each task in process. This can be solved as a classification problem, where case parameters are input attributes and execution time is the target attribute.

A. Classification and Prediction Models

There are number of classification models, every model has its advantages and disadvantages based on data type used for classification. Our problem is rather prediction than classification, but both terms are similar and many models support both of them.

In our case, we have 18 case attributes and one numerical target attribute. All attributes are categorical. Yes, some of them are numerical (width, height), but they are standardized to only few distinct values, so they can be numerical or categorical depending on requirements of classification/prediction model. What is more difficult for prediction, it is also our case, is that target attribute varies even for cases with the same attributes. This is typical for execution time, because work is performed not only by machines, but also by people and people do not work in coherent speed.

Another problem is high variability of door types. In manufacturing company, it is possible to make several millions variants of doors. This causes problem in prediction, because it is difficult to obtain enough data for prediction, it needs many examples. Attributes can also pod up high number of distinct values, it corresponds with high variability of door type (this is problem for neural network classification).

In the next section, some prediction models will be described and discussed its applicability.

B. *Neural Network*

We tested Neural network approach, but results were not satisfied. Neural network was not able to learn. It was caused by high number of input neurons - 303. Every categorical column had to be transformed to new columns. Every distinct value of that column created new column, which holds 1 or 0. So for example, column corner has four distinct values – left, right, top, and bottom. It creates four new columns that can acquire value 1 only once for a row (for the columns that belong to one categorical column). That transformation was necessary, because neural network can handle only numerical attributes. Target attribute was divided into several intervals and every interval was modeled as a single output neuron.

We think that network was not able to learn because of high number of inputs compared to number of training examples and mainly because of variability of output, (even identical training examples had little different outputs). Thus, we think network is not sufficient for our problem because of high number of categorical attributes and variability of target attribute.

C. *K-Nearest Neighbour (KNN)*

The method is based on simple idea of finding several examples from training set closest to input pattern. We simply computed number of differences of attributes between training example and input pattern. These differences (0 or 1, equals and not equals) were weighted. Weight of every attribute was computed by the same method described below in regression tree. Higher weight means that attribute have higher influence of execution time and it is considered more important. Then twenty nearest examples were given and mean, min, max and deviation of time was computed (we measured only mean, but deviation is also important in simulation and it is good indicator of supposed reliability of prediction).

Results (figure 4) were quite satisfied (there is only subset of real workplaces). We have compared prediction to simple algorithm – prediction based on mean of all execution time. Simplest predictor is the predictor that assumes mean value for every example. Differences in table are mean of all differences between real value and predicted value for every tested example.

Result was computed as follows: prediction was compared to most simple predictor that supposes always-average value of task execution time for all records. So:

Mean diff = $\sum \lvert \text{mean} - \text{real value} \rvert$
Predictor diff = $\sum \lvert \text{predicted value} - \text{real value} \rvert$
Final Score = Predictor diff / Mean diff

Mean and Predictor difference is computed as sum of differences over all tested examples. Mean difference is absolute value of mean and real value and Predictor difference is computed from predicted value and real value. Final ratio equals ratio of predictor difference and mean difference.

We have run test with 600 examples and compared them to dataset that contained about 10-20 thousands records for every workplace. Rating was computed as a ratio between difference computed by algorithm and difference computed by mean. So, result 0.5 means that we have decreased the variance of execution time of task about 50 %.

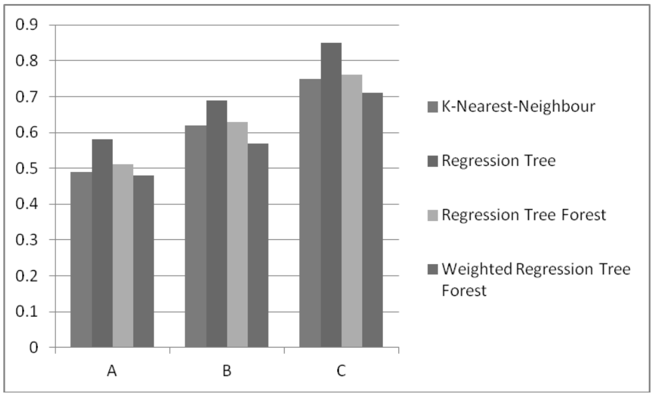


Figure 4. Experiments. Four methods were used on three workplaces. Note that a higher column means lower precision. Workplace A is machine does not depended on resource skills, workplace B is workplace with dependence on resource skills and workplace C is manual workplace (packaging) that does not depend so much on door type, but on resource performance.

Figure 4 shows that some results are satisfied, others not. For example, ratio of workplace A is good, Workplace C is not. Nevertheless, it is not the problem related to method, execution time is not based so much on attributes. It is because workplaces C perform packaging and that type of work is naturally quite independent of door types.

D. *Regression Tree*

Decision tree is popular model. It is simple, readable by human, and quite fast. Precision has not as satisfied results as K-Nearest Neighbour. However, the classification speed is several hundred times faster. Regression tree is decision tree with numerical target value. Nodes contain information about mean, min, max, and deviation of predicted value. Learning algorithm is similar to decision tree, but selection of split nodes differ. We have numerical target attribute so algorithm can be like that:

- For every column.
- For every distinct value of column.
- Take all target values of column grouped by current distinct value and compute deviation.
- 1 / Mean of all deviations is decision power for column.

This algorithm is similar to entropy computation, which is computed for categorical target value. The deviation is closed to entropy, because lower deviation points to better decision power. Computing of deviation can be also weighted by count of rows of groups divided by distinct values of column – distinct value with more rows should be more important. We have tried both approaches, but no significant precision difference was observed, even maybe precision was little lower. Algorithm described above works similar to ID3 algorithm. C4.5 algorithm has been also tried, but no significant difference has been found. Post-pruning was based on removing nodes with low row count (every node corresponds to subset rows of whole data set), because nodes with low row count are not representative.

Regression Tree had worse precision than K-Nearest-Neighbour (Ratio was about 1.2 – 1.3 times worse), but had also several advantages. It is more readable to human and it can be used to examine some properties of tasks – for example which combination of attributes positively or negatively affects execution time or which combinations of attributes have little ratio of prediction – that is represented by deviation of target values corresponding to some node of tree.

E. Regression Tree Forest

Regression Tree Forest is based on several Regression Trees. One extreme example is Random Forest. Random Forest creates many decision trees (more than one hundred) using classic (ID3 or C.45) algorithm with several differences:

- Every tree randomly selects subset of rows from training set (about 2/3).
- Every tree randomly selects subset of attribute columns (about 2/3)
- Every tree is not pruned and full-grown.
- Predictions made by voting of all trees by computing mean.

It is known that Random Forest is very precise model and still quite fast, because it is semantically similar to K-Nearest-Neighbour algorithm. Because learning time is quite long (it requires more than one hundred trees), we found it not suitable for real-time decision support. However, we have tried some trade-of between Random Forest and normal Decision Tree. We created several (about ten) trees and enforced different first splitting column for every tree. Enforced columns were ordered by their decision power. Thus, first tree root node begins with first (best) column; second tree root node begins with second column, etc. In addition, every tree randomly selects 70% of dataset and 70% decision attributes as it is said in Random Forest algorithm. Trees were pruned (opposite to Random Forest, which is not pruned) to about 10 min rows in a node.

It should be stress out that in normal Random Forest, result is computed by mean of all tree results. We selected best tree result by looking to the deviation of tree node. Best prediction could be measured by deviation of particular rows covered by tree node. Node with lowest deviation wins. This rule was necessary, because mean of all tree votes gave terrible results – mainly because we had only low count of trees compared to Random Forest.

Then we chose another improvement – tree result (mean and deviation) was not computed only by looking at the leaf node, but also taken into account the parent node. We computed mean by both node mean values weighted by their deviation (if child node had much better result – low deviation – than parent, its result will have bigger vote). That improvement was tested also in single ordinary Regression Tree and it increased precision too, but only slightly.

Our Tree Forest greatly improved accuracy of classifier, ratio was only about 1.05 times worse than K-Nearest-Neighbour and still order of magnitude faster that K-Nearest-Neighbour, which applicability could be problematic in real time monitoring for every tasks. Similar results can be explained, because random forest works similar to K-Nearest-Neighbour. It returns items that are close (by attributes) to predicted item, but it uses tree searching instead of searching in whole table.

VI. EXECUTION TIME AND RESOURCES

There is a little problem with resources. The resource information can be treated as normal case attribute, because it surely has impact on execution time of task, but there is a catch. For example, if we allow decision tree to build tree using resource attribute, final leaf will contain only records that have been executed only by that resource. This could enable problems, because sometimes, it is better to look for more examples, even from another resource. However, if we do not have such training examples and resource performance does not differ too much from other resources, it is good idea to look also to another resource records and consider them.

Second problem is related to dynamic changes. Even if the process is the same (e.g. technological process), workers performance could change over time. More experienced workers may be faster, so our algorithm could be prepared for that. We recommend following method, which little improved prediction in our manufacturing company.

Suppose K-Nearest-Neighbour or Regression Tree (or Forest) classifier. All that classifiers could be implemented to return set of records rather than final prediction (mean and deviation). The result (mean, deviation) could be implemented over those records, but with different weights. First, records that belong to the resource, which performance time is now predicting, should have bigger weight (for example two times higher) than other records. And second, these records (of our resource) should be considered in time plane. Newest records should have also bigger weights (for example two times bigger than oldest). Why is not possible to take into account time plan also to other records (another resources)? It is because we do not know about them so much in order to take into account their improvement and skills compared to our resource. This could be issue to another paper.

VII. Final evaluation

We have tested three tasks: One machine with little human interaction, second machine with manual work and third packaging with little dependence on door type, but with dependence on resource. As it was presented, Regression tree is always worse than other methods, while Regression Tree Forest is as good as K-Nearest-Neighbour, because it is optimized K-Nearest-Neighbour. Last method was weighted Regression Tree Forest. Weighting was described at Section VI. As it has been shown, weighting on workplace A did not improve result at all, because machine works independent on resources and time (it does not learn to work faster). In Workplace B and C, there was improvement. Workplace C had worst results, because packaging is not dependent on door type too much, but it is dependent on resource – we can see that weighting slightly improved performance.

VIII. Conclusion

We have tested our theory and our results were quite promising. It has been shown that the quality of results does not depend only on our methods, but mainly on manufactory itself. For example, if execution time cannot be predicted from case attributes in wanted precision, prediction will be useless. In our company, predictions helped lower execution time variance, which is very useful in internal logistics planning, but there is a question what precision is needed to implement some better planning techniques, which enables significant saving especially in space and time need for manufacturing production by improving input data for planning algorithms. We can also find subset of case parameters that have low time deviation and try to optimize their production. Other cases could be produced in another time or in other machines in parallel with another approach (slower but more robust). Some workplaces had bad time variance, but that were some manual workplaces like packaging, that were at the end of the process, so variance was such important issue.

Resources working speed was the biggest issue. There is not so much research in that very important area. In addition, dynamic aspect of process (new machines, resource improvement) is problem to solve. We believe these methods could reach maturity and could be used in some manufactories in future.

Acknowledgment

This research was supported by the grants of MPO Czech Republic TIP FR-TI3 039, the grant FIT-S-10-2, the research plan no. MSM0021630528 and the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

References

[1] D. Grigori, F. Casati, M. Castellanos, U. Dayal, M. Sayal, and M.C. Shan, “Business Process Intelligence”, Computers in Industry, Volume 53, Issue 3, Process / Workflow Mining, April 2004, Pages 321-343, ISSN 0166-3615, DOI: 10.1016/j.compind.2003.10.007.

[2] D. Grigori, F. Casati, U. Dayal, and M.C. Shan, “Improving Business Process Quality through Exception Understanding,Prediction, and Prevention”, Proceedings of the 27th VLDB Conference,Roma, Italy, 2001, 1-55860-804-4

[3] W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek, “Business process mining: An industrial application”, Information Systems, Volume 32, Issue 5, July 2007, Pages 713-732, ISSN 0306-4379, DOI: 10.1016/j.is.2006.05.003.

[4] B. Wetzstein, P. Leitner, F. Rosenberg, I. Brandic, S. Dustdar, F. Leymann, "Monitoring and Analyzing Influential Factors of Business Process Performance," Enterprise Distributed Object Computing Conference, 2009. EDOC '09. IEEE International, pp. 141-150, 1-4 Sept. 2009, doi: 10.1109/EDOC.2009.18

[5] A. Rozinat, R.S. Mans, M. Song, and W.M.P. van der Aalst, “Discovering simulation models”, Information Systems, Volume 34, Issue 3, May 2009, Pages 305-327, ISSN 030

[6] M Song and W.M.P. van der Aalst, “Towards comprehensive support for organizational mining”, Decision Support Systems, Volume 46, Issue 1, December 2008, Pages 300-317, ISSN 0167-9236, DOI: 10.1016/j.dss.2008.07.002.

[7] W.M.P. Van der Aalst, “Business Process Simulation Revisited”, 2010, ISSN: 1865-1348

[8] J. Nakatumba, A. Rozinat, and N. Russell, “Business Process Simulation: How to get it right”, 2010,Springer-Verlag,doi=10.1.1.151.834

[9] J. Nakatumba and W.M.P.V.D. Aalst, "Analyzing Resource Behavior Using Process Mining", in Proc. Business Process Management Workshops, 2009, pp. 69-80.

[10] A. Rozinat, M.T. Wynn, W.M.P. van der Aalst, A.H.M. ter Hofstede, and C.J. Fidge, “Workflow simulation for operational decision support”, Data & Knowledge Engineering, Volume 68, Issue 9, Sixth International Conference on Business Process Management (BPM 2008) - Five selected and extended papers, September 2009, Pages 834-850, ISSN 0169-023X, DOI: 10.1016/j.datak.2009.02.014.

[11] W.M.P. van der Aalst, M.H. Schonenberg, and M. Song, “Time prediction based on process mining”, Information Systems, Volume 36, Issue 2, Special Issue: Semantic Integration of Data, Multimedia, and Services, April 2011, Pages 450-475, ISSN 0306-4379, DOI: 10.1016/j.is.2010.09.001.

[12] W. M. P. van der Aalst, and A. J. M. M. Weijters, “Process mining: a research agenda”, Computers in Industry, Volume 53, Issue 3, Process / Workflow Mining, April 2004, Pages 231-244, ISSN 0166-3615, DOI: 10.1016/j.compind.2003.10.001.

[13] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. M. M. Weijters, “Workflow mining: A survey of issues and approaches”, Data & Knowledge Engineering, Volume 47, Issue 2, November 2003, Pages 237-267, ISSN 0169-023X, DOI: 10.1016/S0169-023X(03)00066-1.

[14] W. M. P. van der Aalst, “Process Mining”, Berlin, Heidelberg 2011, ISBN 978-3-642-19344-6

Pospisil, M., Hruška, T., “Business Process Simulation for Predictions” In: BUSTECH 2012 : The Second International Conference on Business Intelligence and Technology, Nice, FR, IARIA, 2012, s. 14-18, ISBN 978-1-61208-223-3

Analysing Resource Performance and its Application

Milan Pospíšil
Department of Information Systems
BUT, Faculty of Information Technology
Brno, Czech Republic
ipospisil@fit.vutbr.cz

Vojtěch Mates
Department of Information Systems
BUT, Faculty of Information Technology
Brno, Czech Republic
imates@fit.vutbr.cz

Tomáš Hruška
IT4Innovations Centre of Excellence
BUT, Faculty of Information Technology
Brno, Czech Republic
hruska@fit.vutbr.cz

***Abstract*—Performance analysis of resources is important part of process optimizing. It is useful to analyze performance (execution time) for particular tasks, time deviation, error rate and its improvement during time. It is also possible to use the analyzing task execution time by resource for different purposes. For example, it is possible to change process definition according to the results, make predictions using short-term simulation, or use it only as analysis of performance properties. This paper focuses on analyzing resource properties and then makes overview of its applications. Some of these methods will be evaluated on real data in manufacturing company.**

Keywords-resource performance analysis; business process simulation; business process intelligence; data mining; process mining; prediction; optimization; recommendation

I. INTRODUCTION

Current business process management system stores a lot of information about processes, data flow, resources, and execution time. The information is very valuable and it is possible to use for analysis [14, 15, 16]. Related work deals with process model discovery [14]. Another related work deals with mining decision rules or organizational models of companies [5, 6]. There is also work based on resource perspectives [8, 9]. Nevertheless, resources are an important part of business processes. This paper deals with resource performance analysis and its usage.

It is possible to analyze different performance perspectives. Basic properties can be execution time of task, its deviation, and error rate. However, other important properties could be analyzed – for example ability to raise performance during pressure or ability to quickly adapt to new task (set up time). However, simple look at execution time of task is not enough, because this execution time is not only based on resource productivity but also on another attributes. There is also an influence done by process change, e.g. new, faster machines or cooperation of multiple resources.

The information is useful to improve our process model. This paper describes several approaches related to mentioned topic. The first application is analyzing the current resource performance properties and improvement over time. Then manager can look at these data and make some assumptions – better resources will get more money, better resources can show other resources how to perform particular task much better.

Second application is about changing process definition according to resource properties. For example, more experienced resource does not need so many checkpoints as less experienced resource does. This can improve performance while technological logic process remains the same.

Third application deals with process prediction. This part is based on short-term prediction that uses simulation model built semi-automatically by process mining. These simulation models need more information about resources, because performance could differ significantly between best and slowest worker.

Next possible application is about allocation of resources. This can be static, e.g. manager can decide what resource should be assigned to specific task taking into account properties of resource. This method corresponds with second application – changing process definition. The other approach corresponds to short-term simulation and recommendations. System can simulate multiple scenarios of allocation of resources to task.

This paper is organized as follows: second Section concludes related work; the third Section is about analysis of resource properties. Forth Section is about its application, fifth Section evaluates some previous methods in real manufactory and last Section is conclusion and future challenges.

II. RELATED WORK

An interest in process mining raised in last decade. Process mining is based on several perspectives. First is process discovery [3, 12, 13, 14]. Process discovery is able to analyze process event log. Event log contains events that were executed during run of process. Every event corresponds to some task and some case. Events have also start and end execution time (they must be ordered at least). Using this information, process discovery is able to find process model of the task sequence (see figure 1).

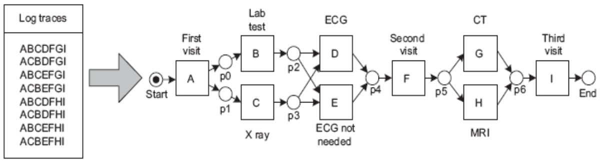
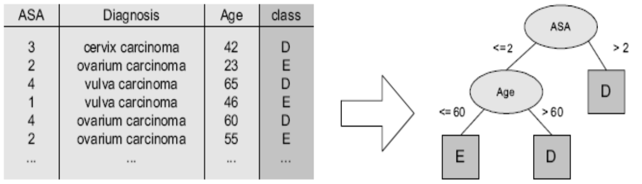


Figure 1. Process discovery [5]. It is possible to discover a process model from logs. The discovered process model must be able to replay most of log traces.

Another research deals with mining decision in routing points (OR split) [5, 7, 14, 15]. Using case attributes, decision rules could be discovered as classification problem.



ASA	Diagnosis	Age	class
3	cervix carcinoma	42	D
2	ovarium carcinoma	23	E
4	vulva carcinoma	65	D
1	vulva carcinoma	46	E
4	ovarium carcinoma	60	D
2	ovarium carcinoma	55	E
...	...	...	...

Figure 2. Decision rule discovery [5]. It is possible to discover decision rules from log. Target attribute is class, which corresponds to next task in process model.

Using previous methods, simulation model can be built [5, 7, 10, 15]. This simulation model (figure 3) can be used for either analysis, or short-term prediction and operational decision support [15].

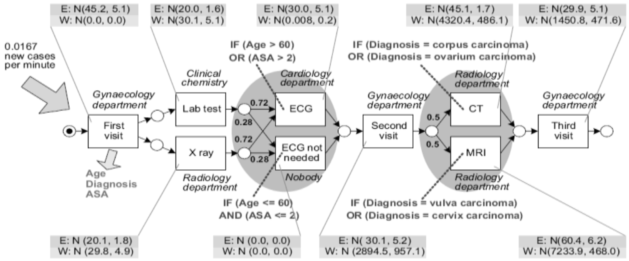


Figure 3. Simulation model [5]. Classic simulation model is enhanced by decision rules. Decision rules can make our routing probabilities more precise, because they depend on case attributes.

Resources are also point of interest. Mates [1] described Resource Dynamic Profiles, which he used for modification of process taking account resource attributes from their Dynamic Profiles. Process mining group also had some research in resources. For example, Song [6] discovered organizational model from process log. Nakatumba [8, 9] examined some resource properties as time availability and ability to increase performance when in pressure.

Another work in Business Process Prediction was from Grigori [2, 16]. She used classification over all case attributes for business process prediction. However, that work did not focus much on resource. Wetzstein [4] did related work where bottlenecks were identified using similar methods (classification). That work also covered resources (resources could be bottlenecks too).

Aalst [11] discovered simulation model using different methods (transition system), but he did not take into account the resource and case attributes at all.

Related work was also done in business process simulations for operational decisions [5, 7, 10], or another methods [2, 11, 16] but only our work [15] deals directly with execution time of tasks ([5, 7, 10, 16] did not examined it at all). However, execution time of task is also dependent on resource performance and thus it is the goal of this paper.

III. Analysing of Resource Performance

It is possible to analyze multiple resource properties:

A. *Execution Time Length of Task*

This property cannot be computed only by looking into the task execution time, because the execution time can be dependent on another attributes. The task with one attribute combination could be much easier than the same task with another attribute combination. For example, assume repair process. “Repair” is one single task, but there is difference between repairing of computer mouse and the notebook. Repairing mouse is easier task than repairing notebook in most cases. In classic workflow system, resource that repaired more mice than notebooks could be considered faster than resource that repaired more notebooks. However, this is false assumption, because second resource is maybe better. There can be the reason why he takes harder repairs.

One solution could be to divide one task into several task that are more similar, but this may not be necessary. The algorithm that computes real worker productivity looks like that:

- For every worker.
- For every worker record.
- Take worker time and predicted time (taken from classifier based on given attributes of record).
- Worker productivity = record time length / classifier result.
- Compute average productivity from these records.

The idea of the algorithm is quite simple. Every worker record of executed task is compared to predictor, that is able to predict execution time of task based on provided record attributes (these attributes must not contain resource id). Predictor should be some classifier like Decision Tree, Neural Network, or K-Nearest-Neighbour (that predictor has good results, but it is very slow). Predictor is able to predict execution time of task independently of resource, so resource id must not be part of its input attributes. That means the classifier learns its predictions from all workers and our worker time is compared to all workers times (for only similar task parameters). The productivity ratio is computed as worker time compared to all workers times that worked on task with similar attributes.

This algorithm needs some data cleaning. First, execution times that deviate from average too much are not considered and second, if prediction of classifier is based only on records, that belong to the worker himself, the result is not taken account, because that would mean comparison of worker to himself. Of course, this information should be taken from decision tree, because decision tree should return final leaf, which is based on particular records, whereas neural network does not provide this type of information.

B. *Execution Time Variation of Task*

It is important to analyze not only the resource performance, but also time variation. Time variation could be the same important information to manager as performance itself. Whole planning is based mainly on variation. Slow resource with low variation and error rate could be good for planning and stable process. Fast resource with high performance could be used in situation, where time is more important than stability.

Variation could be computed using the same method as performance. We will take every worker record and compare it to classifier that predicts variance of all workers according to the task attributes.

C. *Error Rate*

Error rate could be computed by looking for task execution that were marked as incorrect and then computed by the same method as two previous parameters, because some case attributes could lead to more errors. Error rate have similar usage as variance of time. It brings uncertainty into process that is not desired.

D. *Ability to Raise Performance in Pressure*

Nakatumba [9] described method how to compute this property using linear regression. We propose different solution. It could by usable to know performance, variability and error rate, when worker is in stress. Therefore, we will compute these parameters for records that were marked as high pressure. How to detect urgent records? It highly depends on context of business process. Usually, it is possible to check work queue, deadlines and available resources. Nevertheless, the question, if the task was in high pressure is beyond that paper.

E. *Resource Set up Time*

Set up time is important factor that have to be taken into account. When resource is changing task, its performance could be lower than the situation where the task is repeated. This can be computed in the same way as ability to raise performance in pressure. We have to compute performance, variability and error rate the same way but only for tasks, which were changed. If the performance (or variability and error rate) is significantly worse than average properties, we know that we have computed set up time that has to be taken into account in resource allocation. This could be useful information for resource allocation planning, because we can choose resources that have good set up time ability.

F. *Present, Historic and Actual properties*

Present parameters (performance, variability, error rate – average, in pressure, or set-up time) can be seen also in time plane. We consider present parameters as those, which are two months long. Historic parameters could be analyzed for several months', long periods of time. Using this, we can see if the resource productivity is growing or falling. Actual properties are those, which are valid for example one last week. This reflects actual performance of resource and this could be useful information for prediction – see Section IV.

G. *Triage*

Triage is term from business process reengineering. It means dividing one task into more special tasks. Our analysis could be more precise for tasks, which are more similar (in performance). This could be done semi-automatically (maybe fully automatically) by analysis techniques and manager. It is possible to use clustering methods to identify clusters (by execution time). If we found several clusters that are quite different, we can divide one task into that clusters (of course, only for analysis purpose). If those clusters are different (e.g. high distances between clusters) and have low variation (e.g. distances between items in cluster) then we could simplify the operation of analysis of performance (and variance), because we can then compute simple average of times.

IV. Getting Usage Information From Analysis

A. *Rewards and Experience*

Rewards are most obvious things when analyzing resource performance. It is sometimes difficult for managers to distinguish fast and slow resources only from simple analysis of execution times. Experience is also another valuable property of analysis, because when we found, that one resource has excellent performance in some combination of task attributes comparing to others, it could mean, he has a special approach that could improve performance to others. Similar method was described in [4].

B. *Modification of Process Model*

In some cases, it is useful to change process definition in runtime in order to adapt particular worker. This can be done by adding special rules using resource dynamic properties or switching variants of process based on results of previous observations. Monitoring and analyzing behavior of resources and products is one the most important source of process improvement.

C. *Prediction Based on Simulation*

Our previous work [15] proposed a method for operational predictions. These predictions were based on simulation model enhanced of process mining. For example, decision rules were discovered and execution time of tasks was predicted based on task attributes and particular resource. Using that (and process model, of course), the method is able to predict business process using simulation. Quality of this depends on predictability of process itself, complexity (complex process model full of communication with web services can be barely predicted) and quality of data. Nevertheless, our experience on data has shown, that this type of prediction could be usable is our manufacture.

What is the influence of resources and their attributes in this method? Significant, some tasks are heavily depended on resource that executes it. Predicted execution time could vary between two different resources (one slow, one fast). There are two basic approaches for this problem:

- Integration with classifier (predictor),
- post modification after prediction,
- prediction with resource attribute,
- prediction without resource attribute.

1) Integration with Classifier

First method is based on integration with instance classifier. Instance classifier is classifier that does not return final decision (in our case – time or variation) but rather set of examples that are near to current record we are predicting. This means we need to return set of records with similar attributes. It could be accomplished by several algorithms, for example K-Nearest-Neighbour, or Regression Tree (or Regression Tree Forest). Note that these classifiers must not contain resource as input attribute. In normal situation, result will be computed by simple average of results. However, this is not our case. There are several reasons for that.

First, we want to include both results from our resource and other resources. Second, people tend to change their performance over time, so later records are more important and third is almost the same as second, task performance could change over time due to some another reasons (better machines, ..) , so later records are twice more important. Also, if we are predicting task that is changed for resource (set up time), or resource is in stress (long working queue), we could give more weight to records, that are also changed or in hurry (and opposite – lower, if this is not the case). Result average and variance could be computed by weighting records.

How to compute these weights? It strongly depends on data. Sometimes, newer records are not such important as older records (resource improvement is not so important – for example some easy monotone work). We do not know how to set those weights, our experiments shows that those numbers ranges from 1.0 to 3.0 (another resource vs. this resource, old record vs. new record, etc.). These weights must be set manually in experiments. Of course, we can use some automatic optimization like evolutionary algorithm. Nevertheless, this is beyond this paper.

2) Post modification After Prediction

Previous method will work well for instance based classifier. However, there are situation, where instance based classification is not available – a lot of historic data that cannot be stored or another reason like performance. Some classifiers can learn from stream of data that are then forgotten. In that situation, we need to use post modification after prediction. Predictions have to be made without resource attribute. If it is, we do not have to modify its result.

Post modification is made by applying resource attributes (performance, variation) to classification result by multiplying it. Performance and variation of resource attributes is number about 1.0 that says how much better (or worse) resource is compare to other resources. Performance 0.5 means, that resource is two times faster than average resource. We deal with variation in the same way. Low variation means, that resource performance is stable, while high variation means unstable performance results.

3) Prediction With Resource Attribute

We could build predictor using resource attribute. There are two types of predictor: One returns set of instance records, second do not. The result of first one could be accomplished by similar method by weighting results of returned records. Second, one is final prediction and does not need any post modification. This approach has several limitations. These limitations depend on classifier we are using. In Regression tree, the result is based only on results by one resource (there is a way from root to leaf using resource attribute, so leaf will cover only records belongs to this resource), which could be not enough. There can be lot of experience in another record especially when there are many combinations of attributes and we do not have so many records for the same attributes. Note that in our case study, execution time varies even for the same attributes, if we want most probable result and variability, we should need much more than ten records (30-50 could be enough).

4) Prediction Without Resource Attribute

If work does not depend on resource, we can use simply prediction without resource attribute. This could be case for some workplaces with machines that are little dependent on resource. For this purposes, it could be easier to omit resource attribute at all.

D. Allocation of Resources

Based on simulation described in previous Section, we can predict future state of process. However, not only predict, we can also recommend some allocation rules. System knows (by analyzing resource performance) who is suitable for what work. Thus, there is space for system recommendation of resources allocation. For example, there are two work queues, one long queue with many same tasks and another with different tasks. We could choose resource that performs well in set-up time attribute. System can simulate those situations by using previous method and compare results, than recommend several good decisions. Static (long-term) allocation is also available.

V. Evaluation

We have tested some methods in real manufacturing company. This company produces doors. Company has multimple workplaces. These are machine workplaces or manual worklpaces. Execution time of one task (one door) varies too high. But it is dependent on attributes – door parameters (size, weight, material, type…) and also resource who serves the task. These attributes were all categorical. Yes, some of them were numeric (size, weight), but they were treated as categorical, because these parameters were standardized by company to only few values. Target attribute for prediction was execution length of task. There were about 19 attributes and hundered thousands of records (for one workplace). On these records, previous method were tested.

First, we have tried to make deeper analysis of workers performance and we have discovered that our analysis was closer (by opinion of manager) to real performance than simple average of execution times for particular tasks. Unfortunately, validation of this method cannot be made, because no one knows the real right result, but we believe that it is more precise than simple average of execution times of task.

Second test was more interesting. We have tested four methods from previous Section about simulation:

- Integration with classifier (predictor),
- post modification after prediction,
- prediction with resource attribute (classifier returned only rows that belong to one resource),
- prediction without resource attributes.

We have made some tests (figure 4) on workplaces divided into two sets – machine workplaces that are not so dependent on resource performance and hand workplaces, which are more dependent on performance. Result was computed as follows: prediction was compared to most simple predictor that supposes always-average value of task execution time for all records. So:

Mean diff = ∑│mean – real value│
Predictor diff = ∑│predicted value – real value│
Ratio = Predictor diff / Mean diff
Final score = 1 - Ratio

Mean and Predictor difference is computed as sum of differences over all tested examples. Mean difference is absolute value of mean and real value and Predictor difference is computed from predicted value and real value. Ratio equals ratio of predictor difference and mean difference. We turned over the Ratio, because it is more natural to see the better results as higher.

We can see (figure 4) that first method (integration with classifier) was best, as it was supposed. It has to be better than second method (post modification) because it takes account more deeper dependences (performance of resource is not so precise, because it is overall performance for task and resource could handle some task attributes better than other). Prediction without resource attribute worked well for machine workplace, because this workplace is not so much dependent on resource, but it did not work sufficient for hand workplace. Post modification after prediction is still best choice for resource-dependent workplaces if there is no historic data available and we have only predictor (neural network, regression tree with no leaf data, but only mean and variance information).

Triage (division of task into several tasks by clustering) was not tested, because there were about 18 attributes (high space dimension), high variance of execution time and one big cluster with hundreds thousands of overlapping records.

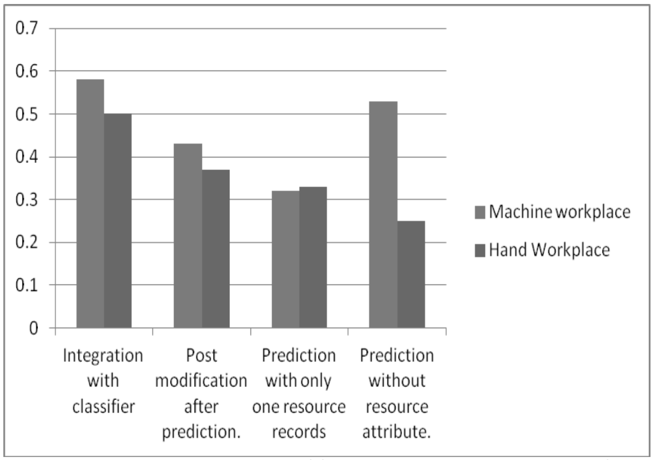


Figure 4. Experiment results of four method described in Section four. We tested two types of workplaces, machine, that are not so dependent on resource and hand workplace, which is heavily dependent of resource.

VI. Conclusion

The paper has been focused on internal context analysis, especially on resource performance analysis. Many different approaches were discussed. The analysis of internal context, especially combination task and resources, is very important for improving performance of process and it can be used for purposes presented in paper. Evaluation of methods were tested on data covered several million records.

Our future goal is adapting planning algorithms in such way they could increase the planning flexibility and precision, which is very important for logistics purposes, and current planning algorithms do not take so much individuality of product and resource into account. Other goal that has already almost been accomplished is to provide benchmarking methods for comparing performance of workers in company, because using only standards do not reflect individuality of particular tasks. Application of results of the benchmarking can improve overall performance of the process. The paper also describes suitability of particular data mining algorithm for area of research and evaluating it on real data created by manufactory.

Acknowledgment

This research was supported by the grants of MPO Czech Republic TIP FR-TI3 039, the grant FIT-S-10-2, the research plan no. MSM0021630528 and the European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

REFERENCES

[1] V. Mates: “Using Workflow Management System for Analysis Based on Properties of Resources”, In: DATAKON 2010 Proceedings (Ed. Petr Šaloun), Mikulov, CZ, Ostravská univerzita v Ostravě, 2010, s. 161-168, ISBN 978-80-7368-424-2

[2] D. Grigori, F. Casati, U. Dayal, and M.C. Shan, “Improving Business Process Quality through Exception Understanding, Prediction, and Prevention”, Proceedings of the 27th VLDB Conference,Roma, Italy, 2001, 1-55860-804-4

[3] W.M.P. van der Aalst, H.A. Reijers, A.J.M.M. Weijters, B.F. van Dongen, A.K. Alves de Medeiros, M. Song, and H.M.W. Verbeek, “Business process mining: An industrial application”, Information Systems, Volume 32, Issue 5, July 2007, Pages 713-732, ISSN 0306-4379, DOI: 10.1016/j.is.2006.05.003.

[4] B. Wetzstein, P. Leitner, F. Rosenberg, I. Brandic, S. Dustdar, F. Leymann, "Monitoring and Analyzing Influential Factors of Business Process Performance," *Enterprise Distributed Object Computing Conference, 2009. EDOC '09. IEEE International, pp. 141-150, 1-4 Sept. 2009* doi: 10.1109/EDOC.2009.18

[5] A. Rozinat, R.S. Mans, M. Song, and W.M.P. van der Aalst, Discovering simulation models, Information Systems, Volume 34, Issue 3, May 2009, Pages 305-327, ISSN 030

[6] M. Song and W.M.P. van der Aalst, “Towards comprehensive support for organizational mining”, Decision Support Systems, Volume 46, Issue 1, December 2008, Pages 300-317, ISSN 0167-9236, DOI: 10.1016/j.dss.2008.07.002.

[7] W.M.P. Van der Aalst, Business Process Simulation Revisited, 2010, ISSN: 1865-1348

[8] J. Nakatumba, A. Rozinat, and N. Russell, “Business Process Simulation: How to get it right”, 2010,Springer-Verlag,doi=10.1.1.151.834

[9] J. Nakatumba and W.M.P.V.D. Aalst, "Analyzing Resource Behavior Using Process Mining", in Proc. Business Process Management Workshops, 2009, pp. 69-80.

[10] A. Rozinat, M.T. Wynn, W.M.P. van der Aalst, A.H.M. ter Hofstede, and C.J. Fidge, “Workflow simulation for operational decision support”, Data & Knowledge Engineering, Volume 68, Issue 9, Sixth International Conference on Business Process Management (BPM 2008) - Five selected and extended papers, September 2009, Pages 834-850, ISSN 0169-023X, DOI: 10.1016/j.datak.2009.02.014.

[11] W.M.P. van der Aalst, M.H. Schonenberg, and M. Song, “Time prediction based on process mining”, Information Systems, Volume 36, Issue 2, Special Issue: Semantic Integration of Data, Multimedia, and Services, April 2011, Pages 450-475, ISSN 0306-4379, DOI: 10.1016/j.is.2010.09.001.

[12] W. M. P. van der Aalst, and A. J. M. M. Weijters, “Process mining: a research agenda”, Computers in Industry, Volume 53, Issue 3, Process / Workflow Mining, April 2004, Pages 231-244, ISSN 0166-3615, DOI: 10.1016/j.compind.2003.10.001.

[13] W. M. P. van der Aalst, B. F. van Dongen, J. Herbst, L. Maruster, G. Schimm, and A. J. M. M. Weijters, “Workflow mining: A survey of issues and approaches”, Data & Knowledge Engineering, Volume 47, Issue 2, November 2003, Pages 237-267, ISSN 0169-023X, DOI: 10.1016/S0169-023X(03)00066-1.

[14] W. M. P. van der Aalst, “Process Mining”, Berlin, Heidelberg 2011, ISBN 978-3-642-19344-6

[15] Pospisil, M., Hruška, T., “Business Process Simulation for Predictions” In: BUSTECH 2012 : The Second International Conference on Business Intelligence and Technology, Nice, FR, IARIA, 2012, s. 14-18, ISBN 978-1-61208-223-3

[16] D. Grigori, F. Casati, M. Castellanos, U. Dayal, M. Sayal, and M.C. Shan, “Business Process Intelligence”, Computers in Industry, Volume 53, Issue 3, Process / Workflow Mining, April 2004, Pages 321-343, ISSN 0166-3615, DOI: 10.1016/j.compind.2003.10.007.

A Model for Recommending Specialization Courses Based on the Professional Profile of Candidates

Antônio Eduardo Rodrigues de Souza
Electrical Engineering Post-graduation Program
Universidade Presbiteriana Mackenzie, UPM
São Paulo, Brazil
e-mail: aersouza@gmail.com

Sandra Maria Dotto Stump
Electrical Engineering Post-graduation Program
Universidade Presbiteriana Mackenzie, UPM
São Paulo, Brazil
e-mail: sstump@mackenzie.br

***Abstract*— The paper studies the candidates' professional profile on choosing a specialization course. A methodology based on the processes Knowledge Discovery in Databases (KDD) and CRoss-Industry Standard Process for Data Mining (CRISP-DM) is applied, and proposed a course recommendation model, using a technique of data mining based on decision trees for the discovery of relevant knowledge from database, which will identify the most suitable course to a candidate's profile. In this study, it is expected to be detected the specialization courses which best suits each candidate profile, giving support to academic institution to satisfy candidates needs and reduce the number of dropouts or changes.**

Keywords-recommender systems; data mining; data filtering techniques; academic counselling

I. INTRODUCTION

The last decade of the XX century was characterized by intense economic globalization, by the need for continuous and quick modernization of production systems, and by the extreme competitiveness in goods and services markets, requiring a better qualification of manpower. However, in the field of education there was the offer of courses strictly academic, which motivated the need of a more specific qualification for the exercise of certain professions, starting to be required master's degrees or doctorates, creating a growing demand by professionals with highly specialized skills profile and not focused on pure research. Such professionals could not be certainly formed as byproducts of courses targeted to the academic and scientific qualifications, but with a technical and scientific nature [1].

The need, by the companies, of increasingly well qualified professionals to meet the demands required, brings consequences as the imposition of education, specific training and qualifications. Moreover, the search for better job opportunities and salaries, have generated a strong influence on the demand for training programs. The professional has transited, increasingly, between the profession and the acquisition of knowledge. It is known that an undergraduate degree cannot over assure a successful career. Currently, professionals are obliged to seek more and more knowledge to achieve specific abilities and skills, meeting the existing shortage in the labor market. As the labor market dynamics change quickly, there is a demand for even more specialized professionals.

Specialization courses are sought after by professional options as a way to acquire and update knowledge [2]. Offered by institutions in various areas, targeted audiences with specific interests or general, diversification of courses or the lack of objective information, hinder the understanding of the purpose, the necessary prerequisites and other important factors in the decision to be made. Thus, a poorly chosen option may incur dissatisfaction, frustration and expectations need to change or even chosen option or even cancellation of the course.

In Brazil, the specialization is a post-graduation course. From Latin, *lato sensu* is an expression whose meaning is "broad sense."

Designed to be attended by persons performing other activities simultaneously, specialization consists of a course of professional qualification with a minimum duration of 360 hours. These courses are not evaluated by the Ministry of Education (MEC) and the Coordination of Improvement of Higher Education Personnel (CAPES), but have significant value to the labor market, especially those courses offered by renowned institutions [3].

It is considered that the topic is up to date and relevant to Higher Education Institutions (HEIs) that offer specialized courses, and presenting various types of courses, in order to give opportunity to continue the training of the candidate. It also shows the IES, the importance of offering specialized courses that are aligned both to the skills of the educational institution, as the interests of training and retraining of skilled manpower for the labor market.

In this context, concerns about the quality of information provided to prospective academic specialization courses at university, is an important aspect that should be considered and, moreover, contribute to the proper choice to meet the needs and profile.

The study becomes relevant since options of courses are offered to meet the expectations of candidates.

This paper is organized as follows: the first section presents the problem that is being studied. The background is presented in Section 2. Section 3 presents the proposed model.

ISBN: 978-1-61208-254-7

II. Background

Advances in science and technology have caused major changes in the global job market, making many professionals, in different areas, hoping to stay in or re-entering the job market, look for continuing education through specialization courses.

Nowadays, the new needs of qualified professionals have shown that the percentage of courses offered has increased to meet market demands. It also shows that quality is searched in the offered courses, which are mainly targeted at preparing more qualified professionals for the job market.

The present study proposes, from historical information of candidates for specialization courses at a private university in the city of São Paulo, to analyze and define behaviour profiles. The research is delimited to candidates who sought specialization courses during the last two semesters.

The database used on this study will serve as a source for data mining, from which will be extracted, in an accurate manner, information that will make courses suggestions for a future candidate. Thus, the objective is to develop a model based on candidate profiles, using artificial intelligence, filtering techniques and data mining, to customize offers of specialization courses. To achieve the objective of this research is intended to perform the following steps: a) Describe by means of literature the main techniques used in Recommender Systems and Data Mining; b) Show how Data Mining can help to identify knowledge in large data volumes; c) Produce an output interface for displaying candidates' data profiles and courses recommendations.

For the evaluation of recommendation model, tests will be performed with the database of candidates provided by the university. These candidates answered a questionnaire about goals and interests with respect to the desired course, the time in which they were enrolled.

III. Modeling the Professional Profile of Candidates

Will be held, initially, a documentary research, through a literature review, which aims to know the state of the art recommender systems, evaluate the filtering techniques and data mining available, and adapt them to the proposed model.

The work will be developed from the data collection of an exploratory, qualitative approach between those who were candidates for specialization courses at a private university, located in the city of São Paulo, between the years 2006 and 2011.

The university in study conducts semiannual selection of candidates for specialization courses. An amount of approximately three thousand candidates takes part of the process each term. The selection process consists of an electronic form, available on the website of the university, with questions which, besides personal information, allows depth knowledge of academic courses, objective, expectations regarding the content and reasons for choosing the course, for example, professional development, job promotion, etc. The responses of each candidate are recorded in Oracle database. With such information, the coordinator of each course verifies the information stored for each candidate and evaluates whether the chosen course is really best suited to the profile posted. Only at the time of publication of the results of the successful candidates who have knowledge of the course where it will be registered, since the electronic form allows the applicant to select three course options. Classes are formed with maximum students per classroom. If they are not offered various classes of the same course, the successful candidate will be relocated to one of the options selected by him, where there is still vague.

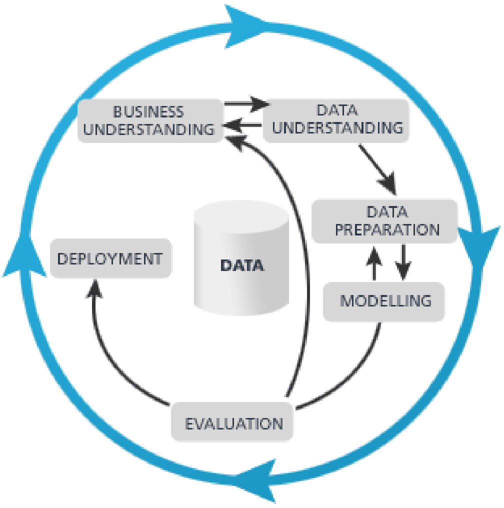


Figure 1. CRISP-DM process steps [4].

For the identification of knowledge it will be studied CRISP-DM [4], shown in Figure 1, and the KDD process [5], shown in Figure 2. Will be considered an approach based on collaboration, to best suit and adhere the needs of candidates from different areas.

A. CRISP-DM Process Steps

The first step comprehends the business understanding. All the understanding of the requirements of the candidate is elicited during this step during meetings with the responsible for the process of selections of candidates. The next step comprehends the data understanding, which involves exploratory analysis on the received data. After data comprehension, the next step is the data modelling, which means to create models based on algorithms, to find the

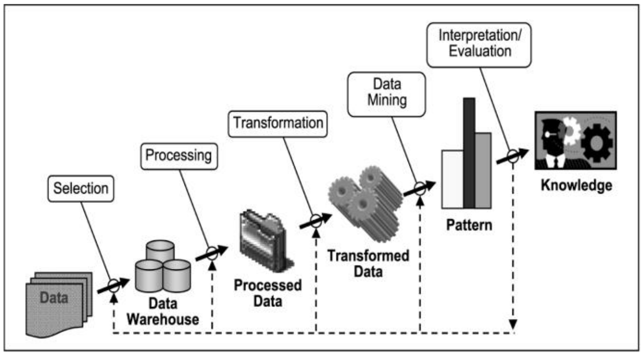


Figure 2. KDD process steps [5].

technique with greater effectiveness. Once the data modelling is finished, performance and gain with the proposed model are evaluated. The last step, installation, comprehends implementation of business rules validated by the pedagogical coordinator of each course in the database to be worked.

B. KDD Process Steps

First there will be the understanding and definition of a domain. Then will be selected within the domain established, the data on which the discovery is performed. Among the acquired data will be selected for analysis the factors that can be identified as potential influencers in choosing a course, such as interests and professional goals: Applicants may direct the choice in their training or opt for a more diversified; chance continuity of learning: candidates may tend to choose courses that enable continuous learning, knowledge and difficulty: where the complexity of the content and the selection criteria may influence the choice; Lack of information: applicants may not be aware of the content and target audience of the course; Location: a campus closest to the address of residence or work can contribute positively or negatively to choice. Besides these factors, the information will be considered professional profiles of the candidates listed on the registration form. The professional profile is a set of characteristics that need to be found in a candidate so that it can occupy a certain position, and can be divided into technical knowledge (education, training area, languages, work experience, computer) and behavioural profile (communication, interpersonal relationships, judgment, attitude, ethics) [6]. Due to the subjectivity and difficulty encountered in measuring values that express the behavioural profile will be considered in this work only the technical knowledge, as shown in the table below:

TABLE 1. ALL VARIABLES (ATTRIBUTES) USED ON THE STUDY

Variable	Description	Example
Graduation	Graduation course name	Administration, Architecture and Urban Planning, Computer Science
Degree	Received degree title	Bachelor, technologist, doctor
Post-graduation	Post-graduation course name	Administration, Architecture and Urban Planning, Computer Science
Professional activity	Activity name	Systems Analyst, Support Analyst, Architect, Controller, Journalist, Secretary
Position	Position name	Advisor, Analyst, Assistant Supervisor, Coordinator, Manager
Experience (years)	Experience time on the position	Less than 1 year, 1 to 3 years, over 3 years
English idiom	Knowledge level of English language	None, basic, intermediate, advanced, fluent

As a way of preparing for the next step, these data must be cleaned and processed. This cleanup includes removing noises, which are data errors or outliers, the adequacy of values that are out of context, the inclusion of missing values, selection and summary of variables to be used (see Table 2 and Table 3). Missing values, that is the absence of information contained in the records, are entered using a global constant, such as the average of each attribute or the average of all the variables of the same class. In case of nominal attributes a dominant subset will be used, whereas in case of non-nominal attributes will be used the mode of each class, which is the value which most frequently occurs in a data set. It will be selected the variables and eliminated some unnecessary variables after checking the information gain of each variable. Courses no longer offered by the university, and courses whose identification codes have changed, will be replaced by courses or equivalent codes in addition to eliminating redundant data, generating greater reliability.

TABLE 2. DATA CLEANSING SAMPLE

Received data	Cleaned and processed data
AdministraA?A?o	Administração
AutA?nomo	Autônomo
CiA?ncias	Ciências
SupervisA?o	Supervisão
TecnA?logo	Tecnólogo

TABLE 3. DATA TRANSFORMATION SAMPLE

Professional activities (Non transformed data)	Professional activities (Transformed data)
ADM, administrador de empresas, administração, administrador de empresas, administradora de empresas	Administração de Empresas
Adm./Financeiro, administrativo e financeiro, administrativo financeiro	Administração e Finanças
Advogada, advogado, advogada autônoma, advogado júnior, jurídico, jurídica, jurista	Advocacia
Analista de sistemas junior, analista de sistema, analista de sistemas informática, analista de sistemas, analista sistema, analista sistemas	Análise de Sistemas

Still in this step, it will be defined the techniques and data mining algorithms to be used, the domain selected to be then processed according to the technical characteristics of the algorithms. After this step, the data will be submitted to data mining itself.

To ensure a satisfactory number of elements of analysis it will be evaluated data from several previous semesters. The data will be received in Microsoft Excel spreadsheets, because of its easiness of being transformed into one of the formats accepted by the tool used in this step.

This stage also includes the use of classification rules. The classification will be used to identify candidate profiles, which also represent the choice of courses. The main mining technique and algorithm used in this step are: Decision Tree - hierarchical data, based on stages of decision (nodes) and the separation of classes and subsets. Major current algorithms: CART, CHAID, C5.0. For the proposal, it is intended to use association rules to classify data due to its better accuracy in recommending courses [7].

Will take part of the model the 56 courses offered semi-annually by the university, but the main focus of the analysis will be the courses that generate a greater number of dropouts or substitutions, and therefore, in addition to modification of various administrative procedures, can be the cause for discouragement, frustration and dissatisfaction from candidates.

For the information related to occupational profiles can be tested, it will be used professional profiles among exact, humanities and social areas, such as Business Administrator, Controller, Tax Attorney, Educator, Financial Analyst, Journalist, Human Resources Analyst, Psychologist, Analyst Systems, Advertising, Support Analyst, Secretary.

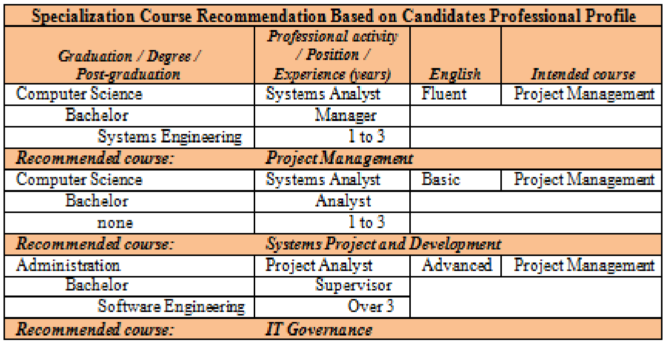
With the prepared database, it will be held the mining and machine learning steps. The tool that will be used is WEKA (Waikato Environment for Knowledge Analysis), developed by the University of Waikato, New Zealand. This tool was chosen because of its public domain, have been developed in the Java language, and working with various data mining techniques such as association rules, clustering, classification, and different algorithms.

As a result generated, the knowledge gained will be examined by a specialist, such as the educational coordinator of each course, to improve understanding of the knowledge discovered by the mining algorithm. If mining results are not satisfactory, several process steps can be carried back.

The aim is to analyze comparatively the number of changes in previous years and those requested by implementing them in the proposed model. In sequence will be evaluated, by the number of candidates who changed course or dropout, the assertiveness of the recommendation by the proposed model, in comparison to data collected by the enrolment system.

The last phase includes the development of an online report, for easy viewing of results, where it intends to use the Java language to implement it, due to its high performance in a web environment. This information will allow the university staff to identify which professional profiles are best suited to specific courses, considering, also, possible relocations.

Based on different candidates' profile, Figure 3 shows part of the report that will be used to support the courses coordinators on course recommendation. It also shows which courses best suits to each candidate profile, according to the discovered model.



Specialization Course Recommendation Based on Candidates Professional Profile			
Graduation / Degree / Post-graduation	*Professional activity / Position / Experience (years)*	*English*	*Intended course*
Computer Science	Systems Analyst	Fluent	Project Management
Bachelor	Manager		
Systems Engineering	1 to 3		
Recommended course:	*Project Management*		
Computer Science	Systems Analyst	Basic	Project Management
Bachelor	Analyst		
none	1 to 3		
Recommended course:	*Systems Project and Development*		
Administration	Project Analyst	Advanced	Project Management
Bachelor	Supervisor		
Software Engineering	Over 3		
Recommended course:	*IT Governance*		

Figure 3. Output report prototype.

IV. CONCLUSION

This ongoing study has introduced another approach for recommending courses based on candidates' profile, taking into consideration the career that a candidate is following. The use of professional profiles for recommending specialization courses can provide a better qualitative recommendation, matching the professional career and academic objectives.

This paper reports, based on groups of courses that have a high number of dropouts or changes, that it is possible to extract relevant knowledge from the professional data of applicants for specialization courses in any educational institution, and use this knowledge for better planning of course offerings and classes' sizes, as well as provide support to academic advising.

With the contribution of artificial intelligence by means of algorithms and techniques of data mining, the studies developed on this work provided a better understanding of the techniques and concepts used in knowledge discovery, relevant to a selection process of candidates for specialized courses. With the knowledge gained was possible to develop a model of professional activities, with their specialization courses recommended by the applied algorithm. This model allowed us to identify the most relevant courses to professional profiles participants in the selection process.

The suggested model can help to define the table of courses, leading the university to rethink whether a particular course deserves to be offered in subsequent semesters. Still, the presented result, besides providing support to the coordinators of courses in the selection of candidates, also serves to alert the institution about the lack of understanding, from candidate point of view, about the contents presented by the offered course, so that the university could take action to improve disclosure of their courses.

The generated model can support decision-making of any academic institution which intends to improve their academic counselling and reduce the number of students dissatisfied with the course, avoiding dropouts or course changes.

ACKNOWLEDGMENT

The authors of this paper thanks Mackenzie Research Funding (MackPesquisa) for sponsoring this research.

REFERENCES

[1] A. C. Giuliani, A. F. N. Netto, M. C. Ponchio, M. S. Neto, and C. M. Batista, “MBAs, mestrados acadêmicos, mestrados profissionais e doutorados em administração: suas contribuições para o ensino e a pesquisa,” Revista de Administração da UNIMEP, vol. 5, Jan. 2007, pp. 52-73.

[2] S. M. C. Silva, “A pós-graduação no contexto atual: uma exigência do mercado de trabalho,” http://artigos.netsaber.com.br/resumo_artigo_23229/artigo_sobre_a_pos-graduacao_no_contexto_atual:_uma_exigencia_do_mercado_de_trabalho, [retrieved: 7 February, 2013].

[3] Brasil, Diário Oficial da União, Resolução CNE/CEN 1/2001, Apr. 9, 2001.

[4] P. Chapman et al., “CRISP-DM 1.0 - Step-by-step data mining guide,” The CRISP-DM Consortium, 2000.

[5] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, “From data mining to knowledge discovery: an overview,” AAAI Press, 1996, pp. 1-34.

[6] S. A. Araújo, D. Santos, and M. A. Bonaldo, “Redes neurais artificiais aplicadas em análise de perfis profissionais,” Proc. CBComp, Oct. 2004, pp. 280-284.

[7] N. Bendakir and E. Aïmeur. “Using association rules for course recommendation,” Proc. AAAI Workshop on Educational Data Mining, Jul. 2006.

Innovative Approach for Agile BPM

Marco Mevius
HTWG Konstanz
Constance, Germany
mmevius@htwg-konstanz.de

Rolf Stephan
Axon Active AG Schweiz
Munich, Germany
rolf.stephan@axonactive.com

Peter Wiedmann
HTWG Konstanz
Constance, Germany
pewiedma@htwg-konstanz.de

***Abstract*—Current challenges for companies require a high flexibility of business processes. The systematic combination of Cloud Computing and Business Process Management presents advantages, which are particularly useful for the modeling and automation of business processes. A targeted enhancement of Business Process Management with the help of these advantages generates significant new potentials for the optimization of business processes. Besides introducing topic-relevant fundamentals, this paper establishes an agile method with a corresponding support toolset, which allow the immediate capturing, fast implementation and high adaptability of business processes. This method can be used for handling the complexity of missing or insufficiently modeled business processes. In addition the paper presents a specific reference architecture. Based on this reference each phase – modeling, automation, monitoring - passes through the cycle efficiently and independently.**

Keywords – Agile Business Process Management, Cloud Computing, Service-oriented architecture

I. INTRODUCTION

For companies, fast and efficient business processes adaptability is increasingly becoming a critical competitive factor [1]. Furthermore, the belief that suitable tools for Business Process Management (BPM) are required is becoming more accepted [2]. In this context, Cloud Computing offers process participants properties which improve the conditions beyond conventional BPM. These advantages apply to both the development and the utilization of BPM tools. High scalability and possible cost reduction are two examples of these advantages [4]. BPM tools can be divided in tools for IT-BPM and Business-BPM. In IT-BPM, the tools can be categorized in the following classes: modeling, simulation, automation and monitoring. This categorization correlates with the BPM-cycle in general [5] and provides the foundation of this paper. Besides the constant enhancement of BPM and Cloud Computing, the topics of Service-oriented architectures and agile software development are also gaining importance. Together, these elements point to the need for a flexible and customer-specific composition of BPM tools. Therefore, the opportunity for an improved process orientation through the increased agility of supporting services is a topic of high relevance [6]. In section 2 the connection of BPM and Cloud Computing will be presented.

In section 3, it will be shown how the possibilities of a service-oriented combination of BPM and Cloud Computing can be used by applying BPM(N)^{Easy}. The acronym BPM(N) ^{Easy} paraphrases the combination of Business Process Management (BPM) and Business Process and Notation (BPMN) with the ambition of making BPM easier. The BPM(N)^{Easy} method is supported by an agile toolset for efficient and effective BPM.

II. BPM AND CLOUD COMPUTING

With the rapid development of IT in the context of launching and running cloud-based architectures, companies are faced with new problems. In particular, collaborative business processes in use across company borders offer essential optimization potential through the combination of BPM and Cloud Computing. An essential commonality of BPM and Cloud Computing is the flexible and agile approach [cf. 15]. The Cloud Computing paradigm can be called an “enabler” of an improved combination of service-oriented architectures and an agile proceeding regarding the management of business processes. But this potential depends on different framework conditions. These are outlined from an economic and technical perspective below and build a further major motivation for BPM(N)^{Easy}.

A. Technical view

From a technical view, three dimensions can be identified for a successful design, implementation and operation of (BPM) tools in cloud environments: *programming*, *integration* and *security* (according to [7, 8]).

- *Programming* - Complex and distributed systems are ubiquitous in business IT landscapes nowadays. In connection with the goal of reaching a higher usability and flexibility, this complexity translates into new requirements for the Software Engineering unit. To solve this issue, the adoption of new or alternative program languages is necessary. Relying on new innovative concepts and techniques, the effort invested in development has to be reduced to render the complexity of these new IT landscapes manageable.
- *Integration* - Integration can be split in data integration, function integration and process integration. In light of challenges involved, the

topic of integration plays a key role in different scenarios. For instance, a cloud-based workflow engine could control variable activities distributed across company borders. For a smooth running of several business process instances, there is a need of defined integration interfaces and structured methods.

- *Security* – (IT) security can be divided into three categories: functional security, information security and data security. All of these categories have a significant relevance for BPM, especially regarding complex business process grids. Functional security specifies how the current state corresponds with the target state of functionality. Information security is focused on the unauthorized changing or extracting of information. Data security takes care of the process-related data.

Furthermore, from a technical point of view, the question of which business processes are most appropriate for running on a cloud-based architecture has to be answered. Possible risks, for example insufficient integration options or application programming interfaces have to be taken into consideration.

B. Economic view

Two dimensions can be listed from the economic point of view.

- *Availability* - Services which are provided by a cloud infrastructure can be accessed any time. Based on a higher abstraction level, the customizing and the application setup become significant easier. In addition to the simplified procurement, the end user is able to work with the service immediately.
- *Investment risk* – In the context of variable billing models such as pay-per-transaction (pay in case of an actual use) the usage of cloud-based service results in certain charges. These charges contain all relevant costs (e.g. server costs, support costs, etc.). On this account, investment costs are significantly reduced, e.g. risk is minimalized in the procurement of a business process supporting application.

According to the user-oriented study by Northbridge [9], in which 417 companies of different sizes were surveyed, 75% of respondents estimate that two-thirds of the processing power will be obtained from the Cloud. The described promising tension between BPM and Cloud Computing generated the motivation for developing and testing the BPM(N)^{Easy} method in the context of the research project "BPM@Cloud" at KIPS (http://kips.htwg-konstanz.de) as presented in the following.

III. METHOD AND IMPLEMENTATION

For performing cloud- and service-oriented BPM, it is important to have a method which makes it possible to handle the complexity of technical and economic conditions.

A. BPM(N)^{Easy} method

The major intention of the method is to support the potential of BPM by a consistent use of cloud- and service-oriented infrastructures. Therefore, the method is based on the outlined (technical and economic) dimensions and connects aspects of agile software engineering with the conventional BPM cycle. Furthermore, the method is divided in two phases and three steps. The phases provide the time frame for performing the steps.

Figure 1 shows an overview of the BPM(N)^{Easy} method phases and steps:

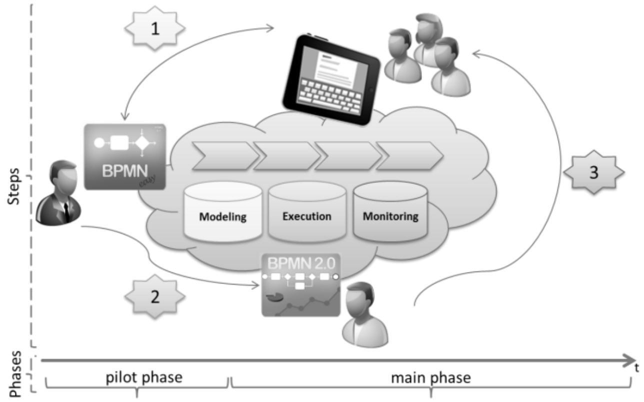


Figure 1. BPM(N)^{Easy} method

An additional intent of the method is to allow for a location independent modeling of business processes, supported by a continuous and integrated tool landscape. The number of unstructured business process repositories can be greatly reduced by using such a homogeneous system of resources to manage and store the business processes and information [cf. 10]. Furthermore, after running through a number of "Sprints" [cf. 11], the created (executable) business processes are provided in a cloud environment [cf. 12]. This release of business processes (or parts thereof) offers the chance to utilize these as service-oriented components in existing business process architectures. Within the steps of BPM(N)^{Easy} the "what must be done" is declared and the phases are used to specify "when and how it must be done".

1) Steps

The method steps are based on the conventional BPM cycle and describe the procedure into three categories:

- *Modeling* - The initial capture of a process, if there are no suitable business process models available, is modeled by using standardized interviews which are based on observations carried out on site with help of a mobile application. The BPM(N)$^{\text{Easy}}$ notation thereby reduces the BPMN 2.0 standard (http://www.omg.org/spec/BPMN/2.0/*)* on important control elements while allowing the collection of valuable information through various mediums e.g. by adding a video to a process activity. All business processes are saved in a Cloud repository with which they can be retrieved, analyzed, and changed at any time.

- *Enrichment* - New business process models are proposed automatically for enrichment. This enrichment includes the design of human-machine interaction by creating user interfaces, and the integration of needed services. For these actions, the latest concepts of software engineering such as library mechanisms and loose coupling or overwriting, are used. The result is a semi-automated or automated business process.
- *Monitoring* – The monitoring is separated into two categories. On the one hand, technical data are monitored, for instance to monitor the time required to access third party systems or application programming interfaces. On the other hand, the monitoring is based on specific indicators for measuring efficiency or effectiveness. A drill-down to the lowest-level information of an activity of the processes should be deliverable. Both categories are aligned with the previously discussed technical and economic considerations.

2) *Phases*

The typical flow through the various stages of the BPM often leads to the fact that important requirements at the beginning of a BPM project have not been considered or adequately described. Also, technical problems are often only identified during the implementation of the business process. Therefore the entire project costs can be significantly increased. In contrast, BPM(N)$^{\text{Easy}}$ is based on the assumption that complex BPM projects can’t be planned at the beginning. The approach of BPM(N)$^{\text{Easy}}$ follows an empirical, incremental and iterative concept to increase the predictability of the process quality and to reduce project risks. Within a predefined cycle, there is always the aim to generate an executable version of the business processes in order to get feedback as early as possible. Fundamental principles of the agile method are transparency, controlling and flexible adjustment.

The phases of the BPM(N)$^{\text{Easy}}$ method define in which intervals and with what accompanying activities the steps have to be carried out.

Figure 2 shows the different stages with the two phases schematically:

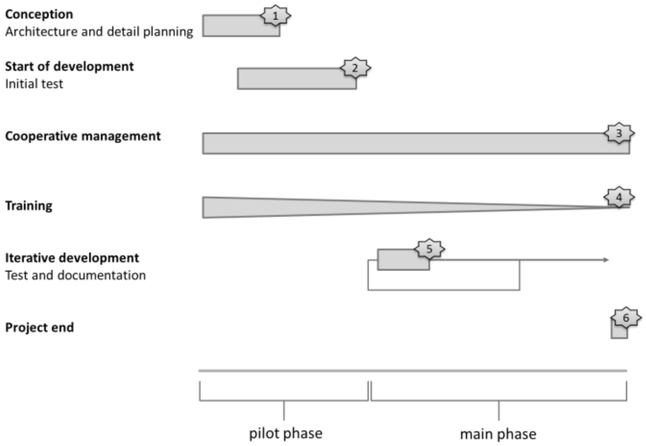


Figure 2. Method phases - pilot and main phase

The high parallelism of the tasks is meant to favor a consistent and systematic implementation of the BPM cycle from the very start. The tasks are derived with the execution of the steps, i.e. the phase sections (1-6) within the pilot and main phase are related to following methods steps:

(1) The conception task describes the modeling of the processes and the implementation of the system architecture.

(2) Shortly after the beginning and before the first sprint of the main phase, users should be able to find a runnable application, i.e. the cycle (steps) was already completed once. This increases the bilateral understanding of business and IT departments.

(3) The "Cooperative Management" is executed as a support activity over both phases. Examples for this task are the coordination of the project members or the writing of the product backlog – a list which contains all implementation requirements.

(4) Typically, new applications have to be provided to the users for reaching the goal of operating independently. This "training" supports the critical coordination process between processes developers and process users.

(5) During the main phase, each sprint is connected to a run through the cycle. If a task is completed already, for instance the modeling of a business process, the task has to be omitted.

(6) The end is defined by an end date and a final closing test. Conditions to closure include a completed cloud repository and an application without access restrictions.

- *Pilot phase* – In the pilot phase, the creation of a detailed specification is the focal point. Furthermore, the architecture for the underlying system has to be set up. Very soon afterwards, the recording of business processes begins and the implementation of the first pilot process application in a prototypical way is initiated. Accompanying this, various training sessions are conducted for the core team and the users to introduce them step by step to the new system. From this early first prototype, which can be used in the productive environment already, other design criteria and components are derived. In addition, all components are re-usable, and therefore an accelerated development is possible. The continuous contact through cooperative management also ensures the high acceptance of the users.
- *Main phase* - The main phase consists of sprints. A Sprint describes a timed interval in which the Sprint task list has to be worked through by the responsible team members. This task list has to be set up at the beginning of each Sprint. The task list is not bound to sequential development sections, but includes tasks for setting up the product – the business process – in general. Based on this iterative process, the possibility of parallel application testing is increased. Parallel to the goal of rapid, iterative development, the sprints of the main phase improve the communication between business and IT significantly and minimize typical barriers. Moreover, the final approval gets simplified by involving the end user in all steps of the method intensively. Through the direct coupling of the development team and users, the risk of implementation errors due to lack of consultation processes or misunderstandings can be reduced considerably.

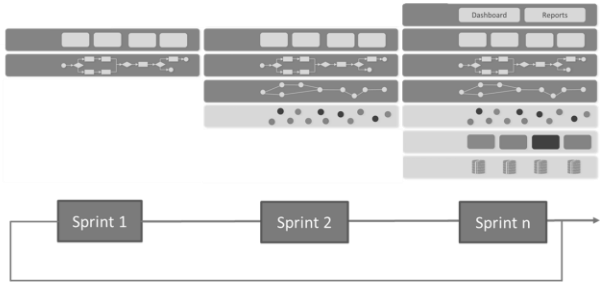


Figure 3. Illustration of the Sprint cycle

Figure 3 illustrates how the automated or semi-automated business processes in the context of the BPM(N)^{Easy} method are produced and how a system can be enriched through different levels. As previously introduced, an essential aim in each Sprint is the focus on the ultimate goal of the BPM project. As early as after the first sprint, it has to be possible to run through the business process from the "spring in the valley". Providing the full functionality of each process activity is neglected initially. At the end of Sprint n a system is available, which covers the defined requirements completely.

B. Reference architecture and implementation

As part of an operational application of the BPM(N)^{Easy} method a supporting reference architecture was developed and combined with various technical components.

1) Reference architecture

As an adequate basis for the method BPM(N)^{Easy} a generic reference architecture for a cloud-based BPM infrastructure was developed as shown in Figure 4 (in allusions to [13]).

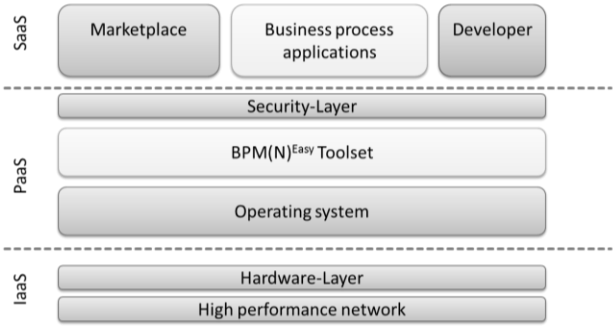


Figure 4. Reference architecture

The generic reference architecture is based on system-oriented services. The cloud typical layers, Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS) can be distinguished [14]. IaaS describes resources, such as the network connection or processing power which are provided as a service. In this case the infrastructure is flexible and can be dynamically controlled depending on the load. PaaS describes a standardized platform for running basic applications (e.g. an operating system). PaaS is needed to load the various SaaS products from the toolset (suitable to the current step of the method). The SaaS layer contains the software which is necessary to offer services and user interfaces for the administration, execution and monitoring. In general, all three levels of the discussed technical aspects - programming, integration and security - are considered. Examples are the security layer within the PaaS, which protects the core architecture, or the separate developer interface, which allows for an easy development access.

As a result of the reference architecture's modular design, a high degree of integration can be ensured. Therefore, the selection of the process application and the corresponding toolset are vendor-independent.

2) *Implementation*

Based on the generic reference architecture in the context of a specific application project within the BPM@Cloud Labs, various components were used to apply the method and to test and validate the usage on a selected business process.

- Xpert.ivy BPM Suite

The Xpert.ivy suite of Axon Active AG (http://axonactive.com) was used for the development of an executable process model. Connected to a cloud repository, it is possible to create a high degree of structuring reusable modules and execute the created process on the web, managing it by versions. The Xpert.ivy module "Monitor" provides functionality for real-time measurement of process parameters and status. On basis of this module service level monitoring and reporting are supported.

- Fujitsu Cloud

Fujitsu services were used both as Infrastructure and Platform-as-a-Service to create an own cloud repository for storing all relevant data. In addition, the services were the base for the Software-as-a-Service level which gives the option to bring release-ready processes immediately up to the Fujitsu Cloud Store. For instance, the SaaS layer contains components which enable the users to monitor or perform the business processes.

- Mobile Easy Tool

This mobile application is based on Android (http://developer.android.com) and has been developed at the KIPS. It can be used for modeling the selected business processes interactively. The application provides a few BPMN 2.0 elements which can be dragged & dropped easily for modeling the business processes intuitively. Moreover, all steps/activities can be enriched by adding metadata directly on site (videos of interviews, images, etc.). Tablets with camera function are used for this purpose. An impression of the very user-friendly ("Easy")- OnClick technology is shown in Figure 5. The Android native features facilitate the operation of the application. For example, all share options, such as e-mail, MMS or Google+ are available.

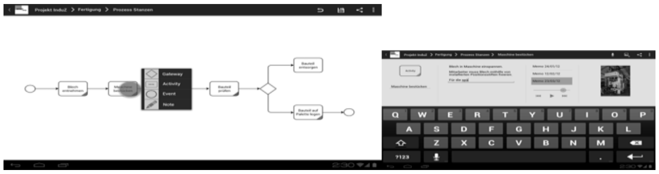


Figure 5. BPM(N)^{Easy} mobile application

An example of a typical approval process is shown in Figure 6.

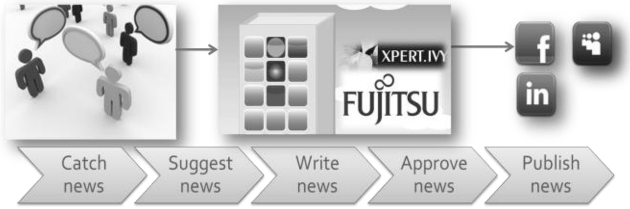


Figure 6. BPM(N)^{Easy} test process

The business process describes the steps involved in publishing company news on social media platforms. On the one hand, the exemplary process enables an overview of the complete phases and steps to go through. On the other hand, requirements emerge which have to be dealt with, such as the collaborative, mobile news writing and the integration of external systems. In sum, all presented technical challenges must be considered and dealt with.

The test scenario was started by modeling the business process with help of the simple and efficient BPM(N)^{Easy} notation. Using cloud repositories, the communication between business and IT was ensured at any time so that early user tests could be performed agilely. Within the Sprint cycles, the Ultimo, a defined project end, has always had the highest priority. From a technical point of view, it was also required to develop a specific, security-related client management which initiates a trigger automatically when the business process has been subscribed to over the cloud marketplace. This trigger receives various activities at the application level of the Ivy server, such as creating new users or the activation of other functions (in terms of Basic / Premium versions). The service-oriented usage of third-party systems (social media platforms) has been implemented by calling web services - therefore the system can easily be extended.

IV. SUMMARY AND OUTLOOK

The integration of BPM tools or entire business processes (BPaaS) in a cloud environment can be assigned a high potential. Cut costs and reduced complexity represent fundamental goals for BPM projects. Furthermore, the improved distribution of services or business processes increases the ubiquitous availability of business applications and provides a significant target value.

The transfer research project BPM@Cloud is currently run by the Constance Institute for Process Control (KIPS) in collaboration with the Axon Active AG. This paper introduced an agile method called BPM(N) ^{Easy}. Furthermore it has been presented a reference architecture which supports the agile method for efficient and effective BPM.

The phases and steps of the agile method do not require rigorous planning at the beginning of the BPM project. Therefore, a highly flexible and close cooperation with all participants is possible. As a result of this, the implementation of the business processes can be achieved significantly.

The test of the method was performed at the BPM laboratory of KIPS and reached a successful result regarding the implementation of new business processes. Related and further work is currently investigating to what extent BPM(N)^{Easy} can be used not only for the first implementation but for continuous business process improvement. From the authors' point of view, there is significant potential for optimization of business processes in this context. In addition, cloud-based services will be created to complement the methods-supporting toolset.

V. References

[1] Bekele, T.M.; Weihua Zhu; Towards collaborative business process management development current and future approaches, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on vol., no., pp.458-462, 2011.

[2] Chaudhry, S.; Moller, C.; Advances in Enterprise Information Systems II, The role of BPM in the IT value-chain: Exploring how manging business processes can decouple business and IT, pp. 56-58, CRC Press/Balkema, AK Leiden, 2012.

[3] Chee, B. J. S.; Franklin Jr., C.: Cloud Computing: Technologies and Strategies of the Ubiquitous Data Center. CRC Press Taylor & Francis Group, 2010.

[4] Voorsluys, W.; Broberg, J.; Buyya, R.: Introduction to Cloud Computing. In (Buyya, R.; Broberg, J.; Goscinski, A. Hrsg.): Cloud Computing Priniples and Paradigms. John Wiley & Sons, Inc., Hoboken New Jersey, 2011.

[5] Weske, M.: Business Process Management: Concepts, Languages, Architectures, Springer Verlag, Berlin New York, 2010.

[6] Abelein, U.; Becker, A.; Habryn, F.: Towards a Holistic Framework for Describing and Evaluating Business Benefits of a Service Oriented Architecture.13th IEEE Enterprise Distributed Object Computing Conference Workshops, Auckland, 2009.

[7] Briscoe, G.; Marinos, A.: Digital Ecosystems in the Clouds: Towards Community Cloud Computing. 3rd IEEE International Conference on Digital Ecosystems and Technologies, Istanbul, 2009.

[8] Fehling, C.; Konrad, R.; Leymann, F.; Mietzner, R.; Pauly, M.; Schumm, D.: Flexible Process-based Applications in Hybrid Clouds. 2011 IEEE 4th International Conference on Cloud Computing, Washington, 2011.

[9] Skok, M.: Future of Cloud leadership panel. Version 2011-Jun-20.2 Future of Cloud Computing. http://www.futurecloudcomputing.net/media-gallery/detail/91/286, accessed 19th October 2012.

[10] Kurniawan, T.; Ghose, A.; Le, L.; Dam, H.: On Formalizing Inter-process Relationships, BPM 2011 Workshops, Part II, Springer Verlag Berlin Heidelberg, 2012.

[11] Schwaber, K.; Sutherland, J.; The Scrum Guide, The Definitive Guide to Scrum: The Rules of the Game, 2011, http://www.scrum.org/Portals/0/Documents/Scrum%20Guides/Scrum_Guide.pdf, 19th October 2012.

[12] W. van der Aalst, Business Process Configuration in The Cloud:How to Support and Analyze Multi-Tenant Processes?, Ninth IEEE European Conference on Web Services, 2011.

[13] Jiang, J.; Le, J.; Wang, Y.; Sun, J.; He, F.: The BPM Architecture Based on Cloud Computing, Knowledge Acquistion and Modeling, 4th International Symposium, Sanya, 2011.

[14] Mell, P.; Grance, T.:The NIST Definition of Cloud Computing, Recommendations of the National Institute of Standards and Technology http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf, 2011, accessed 19th October 2012.

[15] Trifu, M.: SOA, BPM and Cloud Computing: Connected for Innovation in Higher Education. In (Jusoff, K.; Zhu, Z. Hrsg.): ICEMT 2010 International Conference on Education and Management Technology. IEEE Verlag, Cairo, 2010.

Is There Innovation or Deviation？

Analyzing Emergent Organizational Behaviors through an Agent Based Model and a Case Design

Tomomi Kobayashi, Satoshi Takahashi, Masaaki Kunigami, Atsushi Yoshikawa, Takao Terano
Tokyo Institute of Technology
Yokohama, Japan
kbys@triton.ocn.ne.jp

***Abstract*—This paper describes a new method for analyzing emergent organizational behaviors, which are causes of innovation and deviation phenomena, through an agent based model and a case design. Organizational deviation is inextricably linked to innovation, because their mechanisms are similar in terms of breaking operational standards. We have assumed that the former and the latter are different in external utilities, and under this assumption, we have developed a unified agent based model. The agent base simulations have been conducted based on the model, for analyzing the emergence process of innovation and deviation. The simulation results have been compared with case analysis in order to obtain an in-depth understanding of inextricably linked organizational phenomena, and to distinguish the similarities and differences between innovation and deviation.**

Keywords-Agent based modeling; organizational deviation; case design; organizational behavior

I. INTRODUCTION

Innovation is the act of producing something newly introduced. Organizational deviation is misconduct in organizational management. Companies tend to control organizational deviation strictly because they would get serious damage when it has been revealed. Direct control of deviation may, however, reduce the power of Innovation, because both deviation and innovation have similar mechanisms of breaking standards.

In sociology, deviation is classified into three categories [1]. First is criminality, second is violating conduct norms, and third is labeling. This paper is based on the concept of the second category, because it contains similar notions to innovation which is achieved from organizational improvement by breaking standards. Our model is built from the belief that organizational deviation and innovation have similar mechanisms, but they are different in the external utility or disutility.

Organizational deviation does not always occur according to immoral agents' wrongdoing [2]. It may emerge from unintentional behaviors of the agents with the bounded rationality, because they tend to act shortsightedly and to converge to local optima. It means that if agents have behaved aiming at Innovation, they would commit deviation unintentionally by producing disutility to the society. The shortsighted behavior is enhanced by the difficulty in recognizing the utility landscape. Therefore, we incorporate a hierarchical utility landscape into our model by expanding the landscape theory [3, 4]. The landscape theory explains the shortsighted behavior of agents by the limited range of view to landscape on which they behave.

The purpose of this paper is presenting a method for analyzing inextricably linked phenomena such as organizational innovation and deviation by combination of agent based simulation, manual simulation and case description.

The rest of the paper is organized as follows: Section II explains our unified model of deviation and innovation; Section III shows the result of computer simulation ; Section IV explains the case design and analysis approach ; Section V presents the results of the manual simulation ; Section VI describes the model based 'virtual' case ; and Section VII presents the conclusion and future work.

II. AGENT BASED MODEL

This section describes our unified model of deviation and innovation, which simplifies a real structure of an organization and the relation between an organization and a society. Agent based modeling method is applied in order to examine the bottom up emergence process of innovation and deviation. In this model, the implemented hierarchical utility landscape consists of three layers: individual, organizational, and social utility.

A. Structure of the Model

Figure 1 shows outline of hierarchical utility landscape in our model.

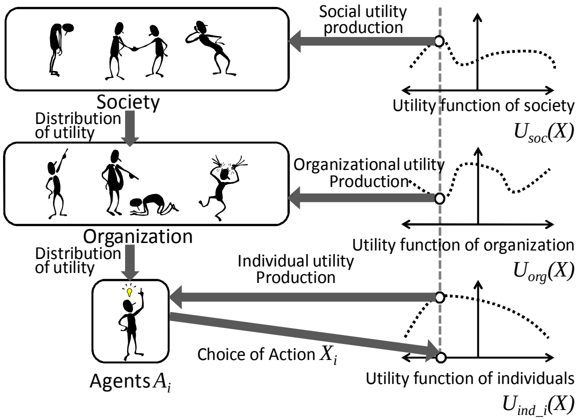

Social utility production
Society
Utility function of society
$U_{soc}(X)$
Distribution of utility
Organizational utility Production
Organization
Utility function of organization
$U_{org}(X)$
Distribution of utility
Individual utility Production
Choice of Action X_i
Agents A_i
Utility function of individuals
$U_{ind\_i}(X)$

Figure 1. Structure of the Agent Based Model.

In figure1, Hierarchical organizational structure which consists of three layers is brought into our model, because it is seen in many companies. Utility function of individuals means experience and values of each agent. Utility function of organization means business model of a company. Utility function of society means social norms.

In this model, agents choose their actions according to the rewards from organization and information from neighbors. As a result, their utility production which means contribution to an organization and a society is determined based on utility landscape. The accumulated organizational utility is distributed to all agents based on their amount of contribution through the system of rewards. The result-based reward is applied in this model Agents can recognize their own utilities, however, they cannot recognize organizational and social utility landscape completely. Therefore, both deviation and Innovation may emerge depending on experiment conditions, and this is the advantage of our model. Based on the model, we define two types of phenomena as shown in table 1 : a) Innovation is the increase of both organizational and social utility production, b) Organizational deviation is the decrease social utility production.

For example, in a Japanese pastry company case, the reduction of product disposals is consistent with their beliefs, in other words employees could recognize their individual utility. However they could neither recognize the social regulations, nor company's damages due to consideration of violating law. In other words, they could neither recognize social utility nor organizational utility landscape thoroughly. As a result, they conducted organizational deviation despite of aiming at innovation.

TABLE I. THE DEFINITION OF INNOVATION AND DEVIATION

Definition		Organizational utility production	Social utility production
a)	Innovation	increase	increase
b)	Deviation	increase/decrease	decrease

B. Utility Function

The Utility functions which are described in the previous section, are based on the NK fitness landscape model [5, 6]. NK model determines the values of N integer sequences, and utility landscape is defined by the combinations of K integers. Figure 2 shows a sample of integer combinations and their values, in case of N=6 and K=1.

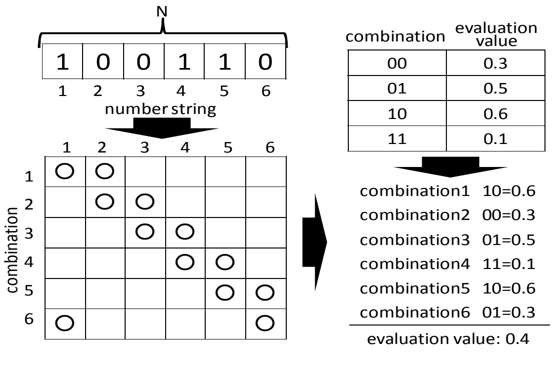


Figure 2. NK Model.

The variation of utility functions is described by number sequences and their evaluation values. Evaluation value is given between 0 to 1 depending on combinations of integers. The complexity of utility landscape depends on the number of integers and their combinations..

C. Choosing Actions of Agents

Each agent changes their action in order to increase their satisfaction according to the formula (1). In formula (1), the degree of satisfaction of agents increases along with the rising of their individual utilities: *Uind_i(X)*, rewards from organization: *Re*$_i$, and contributions for social utility: *Usoc(X)*. The index *i* means the number of agents.

$$S(U_{ind_i}(X), Re_i) = U_{ind_i}(X) + Re_i + U_{soc}(X) \qquad (1)$$

Agents imitate the actions of other agents whose actions are similar to them and receiving more rewards from organization according to the formula (2). P_j means the probability that agent$_i$ imitates the action of agent$_j$. *k* means the number of agent. *Lij* means the similarity of action between agent$_i$ and agent$_j$. Agents evaluate their satisfaction after imitation, and then return to original action when their degrees of satisfaction have been declined by the imitation.

$$P_j = \frac{Re_j \times L_{ij}}{\sum_{k \neq i} Re_k \times L_{ik}} \qquad (2)$$

The agents produce their own utility, and contribute to organizational and social utility as the result of their actions. The contributions of agents are accumulated in an organization and a society.

III. COMPUTER SIMULATION EXPERIMENT

Based on the descriptions of the model in previous section, we have developed the simulator according to agent based computational architecture [7] in Java language. This section describes settings and results of the agent based simulation experiment. Those results are confirmed by manual simulation and case description in following section.

In this experiment, the change in utility production amount of an organization and a society is analyzed by shifting the diversity of agents from 0% (uniform organization) to 100% (diversified organization). All agents have unique individual utility functions in the organization with 100% diversity, while they have common utility functions in the organization with 0% diversity. The other conditions are fixed.

Figure 3 shows the result that is emerged when improving diversification in agents. In figure 3, both social utility and organizational utility productions are increasing with improving diversification. This result suggests that the diversification in agents prompts Innovation type activities according to the definitions in Table1, and the result is corresponding to previous study [8].

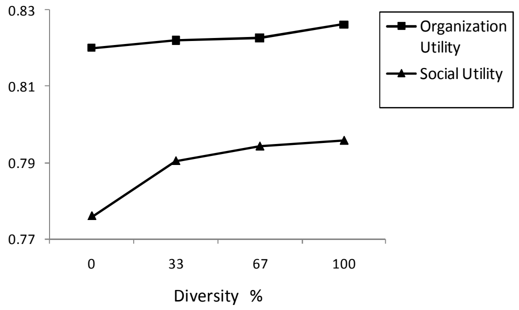


Figure 3. Utility production change that occurs with diversification.

This result means that mutual imitation in diversified organization makes individual utility production decline, because individual utility functions of agents are different from each other. As a result, agents tend to increase organizational utility and social utility production amount in order to complement the lowering of individual utility production, and to maintain their satisfaction which is determined by the formula (1).

IV. Case Design and Analysis Approach

Figure 4 presents the steps of case design and analysis. The simulation results in previous section are confirmed through these steps moreover innovation and deviation phenomena are analyzed from different point of view. The overview of case design and analysis approach is as follows.

The first part is Case Settings. In this part, the model elements are converted to the business management elements. Then case story template and utility landscape are developed according to those elements. The second part is Manual Simulation which is described in chapter V. In this part, manual simulation is conducted based on our model. The third part is Confirmation of Simulation Results. The result of agent based simulation which is described in chapter III is confirmed by the manual simulation in this section. The fourth part is Case Development. The could-be cases are developed in line with the case story template and the results of manual simulations in chapter VI.

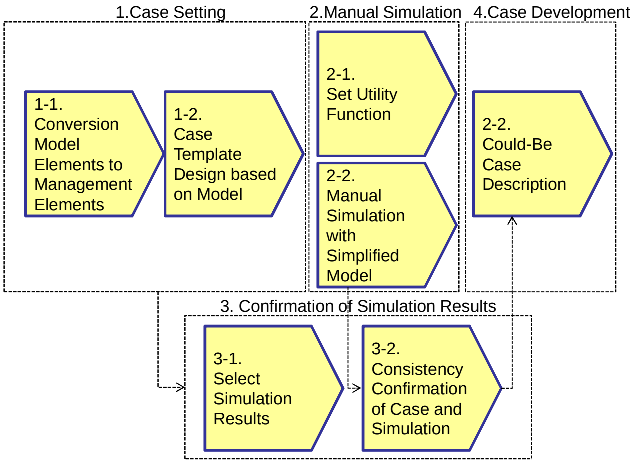

1.Case Setting
2.Manual Simulation
4.Case Development
1-1. Conversion Model Elements to Management Elements
1-2. Case Template Design based on Model
2-1. Set Utility Function
2-2. Manual Simulation with Simplified Model
2-2. Could-Be Case Description
3. Confirmation of Simulation Results
3-1. Select Simulation Results
3-2. Consistency Confirmation of Case and Simulation

Figure 4. Case Design and Analysis Approach.

The elements of the model described in chapter II are aligned with agent based model definition standard [9], and converted to the elements of business management in order to develop the template for case description. Table 2 shows the result of conversion.

The elements of business management listed in table 2 are organized and mutually interrelated in figure 5 as a template for case description. The common template shown in figure 5 is customized according to specific situations, and case stories are developed based on the template.

TABLE II. Comparison of Model Elements and Business Management Elements

Category	Elements of Agent Based Model	Elements of Business Management
Landscape	Hierarchical Landscape: Social Utility, Organizational Utility, Individual Utility	Social Norms, Business Model, Individual Value
	NK Model: Conflict between Social Utility and Organizational Utility	Conflict between Social Norms and Business Model
Characteristics of Organization	Utility Distribution	Personnel System
	Diversity of Individual Utility	Diversity of Organization
	Network Structure	Communication Environment in Organization
Behavior of Organization Members	Network Setting Up Rule	Encounter among Employees
	Behavior Status Change of Agents	Behavior Change of Employees
	Structure of Agent's Satisfaction	Source of Employee's Motivation
	Hill Climb and Imitation Algorithm	Learning Mechanism of Employees
	Incomplete Information Environment	Cognitive Limit
Result of Behavior	Variation in Social Utility, Organization Utility, Individual Utility Production	Variation in Legal Compliance, Corporate Earnings, Employee's Motivation

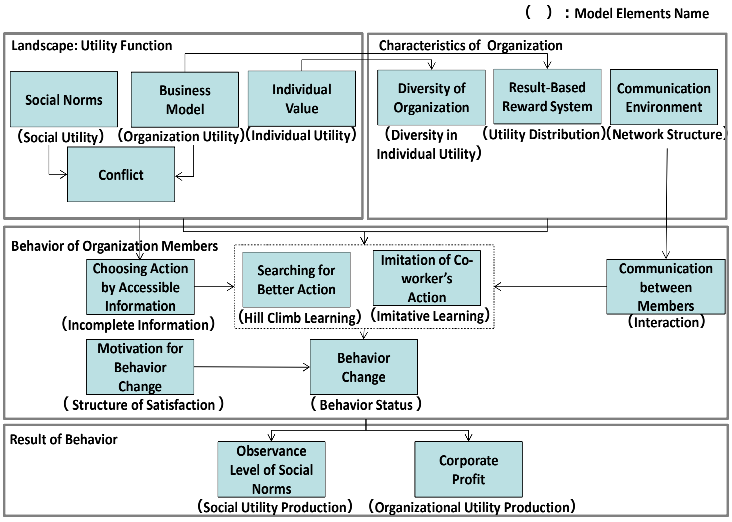

( ) : Model Elements Name
Landscape: Utility Function
Social Norms
(Social Utility)
Business Model
(Organization Utility)
Individual Value
(Individual Utility)
Conflict
Characteristics of Organization
Diversity of Organization
(Diversity in Individual Utility)
Result-Based Reward System
(Utility Distribution)
Communication Environment
(Network Structure)
Behavior of Organization Members
Choosing Action by Accessible Information
(Incomplete Information)
Searching for Better Action
(Hill Climb Learning)
Imitation of Co-worker's Action
(Imitative Learning)
Communication between Members
(Interaction)
Motivation for Behavior Change
(Structure of Satisfaction)
Behavior Change
(Behavior Status)
Result of Behavior
Observance Level of Social Norms
(Social Utility Production)
Corporate Profit
(Organizational Utility Production)

Figure 5. Case Story Template.

V. MANUAL SIMULATION AND CONFIRMATION OF COMUPUTER SIMULATION RESULT

This Before case development, utility landscapes are set in table 3 and table 4 using NK model which is described previously. We have set N=3 and K=1 for case settings and manual simulation. A fictional food maker is assumed in this paper. We assume that this food maker is required to reduce the production cost because of increasing competition however there are strict regulations in food industry.

TABLE III. CASE SITUATION SETTINGS ON NK MODEL

Options	Alternatives of manufacturing control (N=3)		
	1. Cost Reduction	2. Use-By Date Setting	3. Quality Control
0	Production Process Efficiency	Based on Guidelines	Bacteria Test by Devices
1	Waste Prevention of Raw Materials	Based on Case-by-CaseJudgments	Flavor Test by Human Work

TABLE IV. DEPENDENCE RELATIONSHIP BETWEEN ALTERNATIVES

Combination of Alternatives	00	01	10	11
Safety of Products	High	Medium	Medium	Low
Cost Reduction Effect	Low	Medium	Medium	High

Manual simulations using simple NK model are conducted based on table 3 and 4. Table 5 and 7 show the case settings which are social norms, the food maker's policy of manufacturing, and each assembly leader's policy. The situation of Case A is that all assembly leaders have same management policy, so that means uniform organization (Table 5). The Case B is that each assembly leader has different management policy, so that means diversified organization (Table 7). The manual simulation enables the confirmation of computer simulation results by tracing the behavior changing of each agent particularly.

A. *Manual Simulation: Case A*

In case A, all assembly leaders have same cost-conscious management policy as shown in table 5, and there is a certain degree of conflict between social norms and food maker's policy. Table 6 describes the process of manual simulation of case A, which is conducted according to landscape settings and behavior rules of agents. Each assembly line leader's behavior has been changed by searching for more satisfactory action, and also imitating of another leader's action according to NK model settings as shown in table 5.

Figure 6 shows the transition of utility production by assembly leaders as the results of behavior change. The social utility production means contribution to society by protecting of food safety. The organization utility means contribution to corporate objectives such as cost reduction. Assembly leaders receive rewards by their contributions. The individual utility means comfort of leaders which are achieved by the consistency with their management policies. As shown in figure 6, deviation type phenomenon emerges because social utility production amount is decreasing while organization utility production is increasing. It is because that all assembly leaders have cost-conscious management policy, and the food maker has also cost-conscious management policy while balancing with food safety.

TABLE V. UTILITY LANDSCAPE SETTING: CASE A

Combination (K=1)	Evaluation Value				
	Society	Company	Assembly Line Leaders		
	Safety-Conscious	Safety and Cost Balance	Leader1	Leader2	Leader3
			Cost-Conscious	Cost-Conscious	Cost-Conscious
00	0.4	0.1	0.1	0.1	0.1
01	0.3	0.3	0.2	0.2	0.2
10	0.2	0.4	0.3	0.3	0.3
11	0.1	0.2	0.4	0.4	0.4
000	0.40	0.10	0.10	0.10	0.10
001	0.30	0.27	0.20	0.20	0.20
010	0.30	0.27	0.20	0.20	0.20
100	0.30	0.27	0.20	0.20	0.20
011	0.20	0.30	0.30	0.30	0.30
110	0.20	0.30	0.30	0.30	0.30
101	0.20	0.30	0.30	0.30	0.30
111	0.10	0.20	0.40	0.40	0.40

TABLE VI. PROCESS OF MANUAL SIMULATION: CASE A

Step		1	2	3	4	5
Assembly Leader 1	Behavior Change	Initial Status	Imitate	Imitate	Stay	Stay
	Combination of actions	**000**	**001**	**011**	**011**	**011**
	Social Utility Production	0.40	0.30	0.20	0.20	0.20
	Organization Utility Produc	0.10	0.27	0.30	0.30	0.30
	Individual Utility Productio	0.10	0.20	0.30	0.30	0.30
	Reward	0.16	0.22	0.30	0.30	0.30
	Satisfaction	0.66	0.72	0.80	0.80	0.80
Assembly Leader 2	Behavior Change	Initial Status	Search	Stay	Stay	Stay
	Combination of actions	**001**	**011**	**011**	**011**	**011**
	Social Utility Production	0.30	0.20	0.20	0.20	0.20
	Organization Utility Produc	0.27	0.30	0.30	0.30	0.30
	Individual Utility Productio	0.20	0.30	0.30	0.30	0.30
	Reward	0.24	0.33	0.30	0.30	0.30
	Satisfaction	0.74	0.83	0.80	0.80	0.80
Assembly Leader 3	Behavior Change	Initial Status	Search	Imitate	Turn Back	Stay
	Combination of actions	**100**	**110**	**011**	**110**	**110**
	Social Utility Production	0.30	0.20	0.20	0.20	0.20
	Organization Utility Produc	0.27	0.30	0.30	0.30	0.30
	Individual Utility Productio	0.20	0.30	0.30	0.30	0.30
	Reward	0.24	0.33	0.30	0.30	0.30
	Satisfaction	0.74	0.83	0.80	0.80	0.80

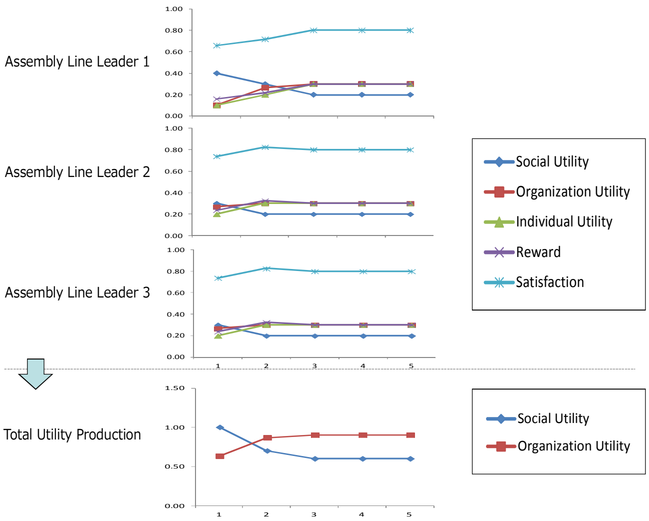


Figure 6. Manual Simulation Result: Case A.

B. *Manual Simulation: Case B*

In case B, each assembly leader has different management policy as shown in table 7, and there is a certain degree of conflict between social norms and food maker's policy. The management policy of leader 1 is safety conscious, leader 2 is balance of safety and cost, leader 3 is cost conscious.

Figure 7 shows the transition of utility production by assembly leaders as the results of behavior change. As shown in figure 7, innovation type phenomenon emerges because social utility production and organization utility production amount are increasing. It is because that leader 1 and 2 have found the appropriate action which enable the increasing of both social utility and organization utility by own searching. And in addition, leader 3 who has cost conscious policy, has imitated their actions although he could not find the appropriate action by himself.

The explanation of manual simulation process of case B is omitted.

TABLE VII. UTILITY LANDSCAPE SETTING: CASE B

Combination (K=1)	Evaluation Value				
	Society	Company	Assembly Line Leaders		
			Leader1	Leader2	Leader3
	Safety-Conscious	Safety and Cost Balance	Safety-Conscious	Balance	Cost-Conscious
00	0.4	0.1	0.4	0.1	0.1
01	0.3	0.4	0.1	0.4	0.1
10	0.2	0.3	0.1	0.4	0.1
11	0.1	0.2	0.1	0.1	0.4
000	0.40	0.10	0.40	0.10	0.10
001	0.30	0.27	0.20	0.30	0.10
010	0.30	0.27	0.20	0.30	0.10
100	0.30	0.27	0.20	0.30	0.10
011	0.20	0.30	0.10	0.30	0.20
110	0.20	0.30	0.10	0.30	0.20
101	0.20	0.30	0.10	0.30	0.20
111	0.10	0.20	0.10	0.10	0.40

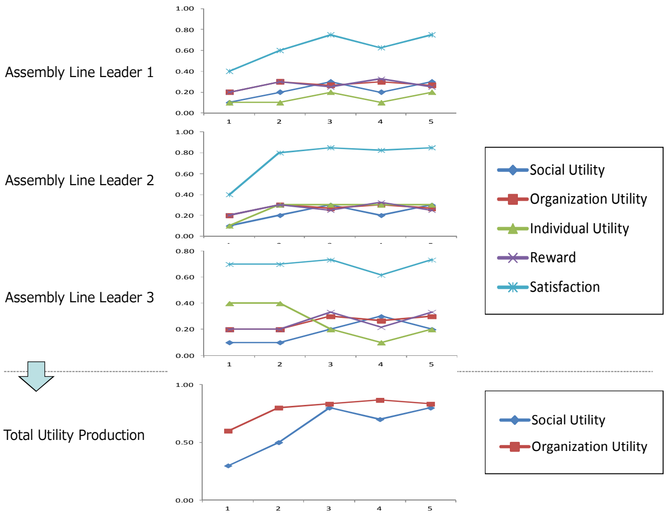


Figure 7. Manual Simulation Result: Case B.

C. *Confirmation of Computer Simulation*

The computer simulation result which is shown in figure 3 suggests that the diversification in agents prompts Innovation phenomena. This result is confirmed by manual simulation in Case A and B by observing the behavior change of each assembly leader. The results of manual simulation show that diversified organization tends to emerge innovation type phenomena in case B while uniform organization tend to emerge deviation type phenomena in case A.

VI. CASE DEVELOPMENT

A. *Purpose and Approach*

A model based fictional case is described in this section according to the case story template (figure 5) and the results of manual simulation. There are two purposes of model based case description. The first is to understand the emergence process of innovation and deviation at more detailed level than simulation. The second is to compare the description level of model based case with that of actual case.

The underlined portions are model elements and the italic words in parenthesis are the name of model elements. Only the story of case A is described in this paper. The stories of case B are omitted due to space limitation.

B. *Model based fictional case: Case A*

A food maker applied the product cost reduction policy because of severe competition in food industry. The company intended to balance cost reduction and product safety (*Organizational Utility Function*), however its policy was not completely fit to the requirements from consumers (*Conflict between Social Utility and Organizational Utility*).

In the food maker, the education programs were conducted for employees in order to strengthen their sense of cost reduction. As a result, most employees shared strong cost-consciousness(*Individual Utility Function*) in that company. Under the situation, cost reduction activities were executed in one of the factory in the food maker as follows.

There were three assembly line leaders in the factory whose cost reduction policies were different from each other at the beginning. The assembly line leader 1 applied the safest way in three leaders. He sought production process efficiency, set use-by date based on guidelines, and conducted bacteria test for quality control (*Behavior Status*). The leader 2 applied same method of cost reduction and use-by date setting as leader 1, however he conducted flavor test instead of bacteria test (*Behavior Status*). The leader 3 pursued waste prevention of raw materials, but ensured product safety by set use-by date based on guidelines, and conducted bacteria test (*Behavior Status*).

In that situation, the assembly line leader 1 who applied the safest policy received the least reward according to the result-based reward system (*Utility Distribution*). The leader 1 was frustrated at less reward and changed his way of quality control form bacteria test to flavor test by imitating the way of leader 2 (*Imitation Learning*). At the same time, the leader 2 and 3 applied the method of setting use-by date based on case-by-case judgments for more cost reduction through their trial and errors (*Hill Climb Learning*). This method had a risk of product safety decreasing, however it was consistent with their cost-conscious policy(*Individual Utility Function*). Therefore the leader 1 received less reward again because the leader 2 and 3 applied more effective cost reduction method(*Individual Utility Function*), even though he imitated the method of leader 2 previously. So that, the leader 1 imitated the method of leader 2 again (*Imitation Learning*), because leader 2 received more reward than him.

As described above, all three leaders applied more effective cost reduction method while sacrificing the safety of products. They recognized the methods which are effective for product safety, but they were not satisfied with those methods because of inconsistency with their cost-consciousness and less reward(*Structure of Agent's Satisfaction*). As a result of assembly line leader's behavior change, the factory achieved cost reduction target (*Organization Utility Production*), however its risk of reducing the product safety increased significantly (*Social Utility Production*)..

VII. CONCLUSIONS AND FUTURE WORK

This paper presented a method for analyzing inextricably linked phenomena such as organizational innovation and deviation by combination of agent based simulation, manual simulation and case description.

As described previously, a Japanese pastry company intended to conduct innovation in order to increase their organizational utilities, however they fell into deviation by unintentional decreasing of social utility because it falsified the expiration dates of products. According to the results of simulation and case description, it is detected that the emergence of deviation or innovation depends on the diversity of organization in this paper. It is also detected that how the diversity of organization impacts on behaviors and learning activities of agents.

The advantage of this method is that it enables to approach toward inextricably linked phenomena with unified model and the combination of multiple method of analysis. The unified model enables to observe that small changes of model parameters would cause both deviation and innovation phenomena. Thus it is the contribution to organizational innovation and deviation research area, because previous studies tend to approach from one side, such as only innovation side [8] or deviation side [10]. The model based 'virtual' case description is the novel method in terms of creating future scenarios compared to standard case study which is based on past phenomena.

In the further work, we would refine the method of manual simulation and case description with applying this method to another type of inextricably linked phenomena. And we would develop the novel method for creating future scenarios by integrating agent based simulation and model based case design. It is expected to reduce unexpected problems in organizational management.

REFERENCES

[1] Hougetsu, M. (2004). Sociology of Deviance and Control. Yuhikaku. (in Japanese)

[2] Baucus, M.S. (1994). Pressure, opportunity and predisposition: A multivariate model of corporate illegality. Journal of Management, 20 699-721.

[3] Axelrod, R. (1999). The Complexity of cooperation. Princeton Univ. Press.

[4] Kijima, K. (2001). Generalized Landscape Theory: Agent-based Approach to Alliance Formations in Civil Aviation Industry. Journal of System Science and Complexity, 14, 2 113-123.

[5] Kauffman, S. (1993). The Origins of Order: Self-Organization and Selection in Evolution, Oxford University Press.

[6] Kauffman, S. (1995). At Home in the Universe: The Search for Laws of Self-Organization and Complexity, Oxford University Press.

[7] Axtell, R., L. (2000). Why Agents? On The Varied Motivations for Agent Computing in the Social Sciences. Center on Social and Economic Dynamics Working Paper No. 17.

[8] Page, S.E. (2007). The Difference. Princeton University Press.

[9] Grimm V., Berger U., DeAngelis D., et al. (2010). The ODD protocol: A review and first update, ecological modeling, 221, 2760–2768.

[10] Reason, J. (1997). Managing The Risks of Organizational Accidents. Ashgate Publishing Limited.

Semantically Standardized and Transparent Process Model Collections via Process Building Blocks

Jörg Becker, Nico Clever, Justus Holler, Johannes Püster, Maria Shitkova
University of Muenster – ERCIS
Münster, Germany
firstname.lastname@ercis.uni-muenster.de

***Abstract*—Process model repositories management is a complex endeavor including modeling and publishing challenges. Existing modeling notations like BPMN or EPC are not able to cover the requirements induced by the volumes of such process model collections (PMC). The modeling technique proposed in this work addresses these requirements and enables organizations to efficiently manage their respective PMC. In fact, the proposed notation based on process building blocks allows for the efficient handling of PMC of any size. Its integrated structure of building block based process models combined with the concepts of layers, attributes, glossaries, reference models, and variants makes it a universal yet semantically standardized process modeling technique. The conceptual definition of the modeling technique and a prototypical instantiation and implementation are introduced. The practical applicability of the technique is justified through an evaluation in practice.**

Business Process Management, Process Modeling, Process Model Collections

I. MOTIVATION

Business process modeling (BPM) is a fundamental requirement in most management and IS projects [14]. A lot of companies have undertaken such an initiative for the purpose of business reorganization, certification, human resource planning or traditional software engineering [6]. The more complex the environment is, the more business processes models it contains [12]. In addition to process models, organizational charts, and a multitude of various additional documents related to the process models are created during BPM projects. All these artifacts form so-called process model collections (PMC), which according to [8] are being of great attention among researchers nowadays.

The most common modeling notations that are used for the task of process modeling are Flow Chart Diagrams, PetriNets, Integrated Definition for Function Modeling (IDEF0), Event-driven process chains (EPC), Unified Modeling Language, (UML), Business Process Model and Notation (BPMN). These existing modeling notations are subject to limitations, which have been criticized for different reasons by practitioners and researches in the field of BPM [3], [22], [1]. These limitations include a lack of standardization [24], which again imposes challenges on reusability, collection organization and variant management. The difficulties of managing PMC are therefore partly accounted for by the modeling language used to create the process models.

Besides the research endeavor conducted there are still open issues to be addressed [8], such as querying, mining, refactoring, re-use, similarity search, merging, variant management and collection organization. In the following, we will focus on the areas of reusability, because it is a fundamental idea of all modeling efforts, and follow up on ideas of collection organization as well as variant management. We will further discuss these areas from the perspective of a proposed process modeling technique, which is to be understood as a combination of a modeling notation including syntactical rules and a complementing modeling tool facilitating the application.

Therefore, the goal of this paper is to address the above-mentioned problems by answering the following research questions:

- *RQ1: How can the problem of organizing process model collections, including such aspects as variant management and storage of supporting model information, be resolved with the help of a modeling technique?*
- *RQ2: How can a modeling technique support the re-use of process models within a process model collection by semantic model standardization?*

The remainder of the paper is structured as follows. In the second section a literature review on the existing problems in two areas of BPM and PMC is carried out and their interrelations are highlighted. Our research method is presented in section three. In section four the conceptual model of the proposed modeling technique is introduced. Section five is devoted to the presentation of a prototypical implementation and evaluation of the modeling method. The paper is concluded within section six with the discussion of the findings and outline of open issues.

II. RELATED WORK

The problems of BPM can be classified into 3 groups in terms of their occurrence before, during and after the conduction of process modeling projects [24] as follows:

- *Before* process modeling: as most of the existing modeling methods are not intuitive, process modelers as well as process model users have to learn and understand the selected process modeling language before starting a BPM project [24].

- *During* process modeling: The most common techniques like Flow Charts, PetriNets, IDEF0, EPC, UML or BPMN allow for a high degree of freedom during modeling. They do not provide naming conventions, or standardized levels of modeling abstraction. As a result the created models differ greatly if several specialists are involved in the modeling process [3].
- *After* process modeling: Because of the high degree of freedom, the resulting process models are complex and their semantic is not standardized. They are therefore hard to analyze and re-use. In most cases, proceeding activities are a tedious manual task, in which expensive consultants have to be involved. [4]

In addition to this classification, a global Delphi study conducted by Indulska et al. (2011) identified the most influential current issues and future challenges in BPM. The following issues are among the most significant ones according to the opinion of practitioners and researchers [14]:

- lack of standardization of modeling notations, tools and methodologies;
- model management problems, i.e. publication, version, variant, or release management;
- absence of clear definition of modeling level of detail;
- lack of identification of abstraction levels;
- need for establishing process modeling expertise and collaborative modeling.

Other issues are related to the complexity management of process models as processes with a high number of elements can cause comprehension problems in such activities as model validation, maintenance and utilization [16], [17].

Regarding PMC, there are several more tasks to be considered such as querying of process models, similarity search, variant management, merging, mining, refactoring, re-use, collection organization, and repository technology [8]. Due to the BPM issues identified above, in particular huge number of process elements, high degree of freedom for modelers and absence of standardized semantics, most of these tasks and especially the model re-use and process model collection management are problematic.

III. RESEARCH METHOD

For developing a modeling technique as presented in this paper, design science research methodology (DSRM) is applied. In the area of design science (DS) research, there is a variety of concepts of how to conduct a DS endeavor. They mainly differ in the role, design theories take over in the understanding and definition of DS. On the one hand, there are notions in which design theory plays minor or even no role at all. For example, March & Smith [19], Hevner et al. [11] and Benbasat & Zmud [5] consider the IT artifact as the core research object of information systems research. On the other hand, there are notions in which the theory of how to design the IT artifact is considered as the main research objective along with the IT artifact itself. For example, Walls et al. [23], Gregor & Jones [9] and Iivari [13] propose a differentiation between the theoretical design of an IT artifact and its concrete implementation, as well as the relationship between these two aspects. Other authors like Kuechler & Vaishnavi [15] identify a mutual relationship between kernel theories and design theories saying that kernel theories can influence DS as well as they can be re-influenced by DS results.

For the purpose of this work, an overemphasizing of the role of theories is deemed not beneficial. Therefore, the notion of DS proposed by March & Smith [19] which is taken up and refined by Hevner et al. [11] is chosen as a research method in this work. According to Hevner et al. [11] the design science research process can be presented as a cycle with five main steps (Fig. 1). In order to specify the problem domain we have conducted a literature review, which revealed a number of deficiencies in the area of PMC and served as a basis for solution objectives definition. The design phase of DSRM was fulfilled by concept creation, presented in the following section of the paper. Demonstration of the result is achieved through prototypical implementation of the modeling tool. Evaluation is performed by applying the created artifact in two business modeling projects. Finally, as design is inherently iterative according to DSRM, new requirements were identified for the artifact improvement after the evaluation phase.

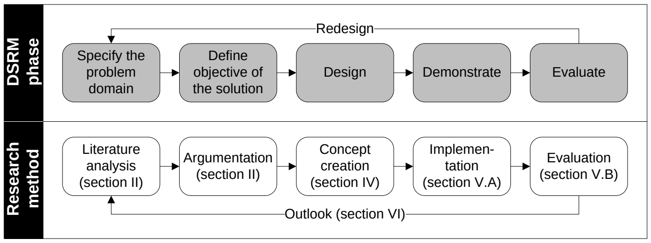

DSRM phase
Redesign
Specify the problem domain
Define objective of the solution
Design
Demonstrate
Evaluate
Research method
Literature analysis (section II)
Argumentation (section II)
Concept creation (section IV)
Implemen- tation (section V.A)
Evaluation (section V.B)
Outlook (section VI)

Figure 1. Research methodology

IV. CONCEPT OF THE MODELING TECHNIQUE

The main research artifact is a conceptual model based on several rationales, which are depicted in the following subsections.

A. Layers

In order to address the challenge of process complexity it is common practice to define layers of abstraction. The emerging question is how many layers are reasonable to support an adequate fit between necessary detailing of the process steps and constraining the amount of process information in one model with respect to usability and readability.

The most adequate amount of layers varies with respect to the modeling purpose of the modeling project. A workflow management system preparation project demands a higher level of (technical) detail in comparison to a management-oriented process modeling project. Hence, the challenge is to conceptualize a layer architecture which is able to meet the requirements for, e. g., both of the aforementioned scenarios.

B. Attributes

Despite the possibility to use the layers of abstraction, the here conceptualized technique proposes attribution as a mean to complement the process models with in-depth information on all process layers where applicable. By extending the process models with attributes, the challenge of complexity can be overcome more easily. Attribution reduces the need for sophisticated branching concepts for the control flow of the processes. Via the possibility to use different attributes on the distinct layers of abstraction, the aforementioned modeling purpose can be supported more easily. Hence, the concept of attribution fosters readability due to complexity reduction and expands the area of application due to the possibility to append attributes on any level.

Furthermore, manageable process attributes are a prerequisite for process analysis and reporting functionalities.

C. Glossary

Existing modeling techniques allow for a high degree of freedom in both syntax and semantics. These degrees of freedom also allow the modelers to arbitrarily label process elements, such as events and functions in EPC or BPMN.

Empirical studies verified that the terms used in modeling can vary heavily, especially, when developed timely, personally and regionally distributed [10]. On a word-based view, these problems are mainly caused by synonyms. As process element labels are normally composed of multiple words, the phrase structure of these words may also cause naming conflicts. It has been shown, that even when limiting the number of words to two, there are more than 20 different phrase structures being used by process modelers [7].

These issues, both on word and phrase structure base are called naming conflicts [2]. The re-use of models flawed in such a way is problematic, as they increase the complexity of the model and are thereby much harder to understand by the model users. Moreover, automated processing and analysis of the models is complicated or even impossible.

The key to prevent naming conflicts is standardizing the choice of words and the phrase structures to use before modeling and enforcing these standards during modeling [7]. Similar to the simple syntax our modeling technique strives for the simplest structures available to foster semantic standardization. Therefore, only a simple verb-object label is allowed for the phrase structures. These have proven to be understood better than all other phrase structures. [20] Within the process modeling context the verbs and objects can be interpreted as activities and business objects.

Standardization before modeling is achieved through a glossary, which is composed of several business objects. These business objects are again related to all activities, which results in a specific instantiation of the verb-noun phrase structure. The free definition of business objects and activities allows the modeling technique to be customized for any modeling scenario. This procedure is therefore chosen over the use of existing catalogues such as the MIT process handbook, although it requires more initial work [18].

The standardization is enforced during modeling, since all modeling processes have to be related to at least one glossary. Every process element is then labeled by linking the process element to one activity-business object combination specified in the glossary.

D. Reference Models

Besides the incorporation of the before mentioned rationales into the modeling technique, reference models are proposed to further enhance model creation. They allow simple and efficient model creation, since their reference character enables the modeler to easily adapt the model to his or her needs. Moreover, reference models foster models of high quality w. r. t. their best or common practice character. Furthermore, reference models facilitate storing, relating and finding the models by providing a frame which structures the process model collection in an enterprise.

E. Variants

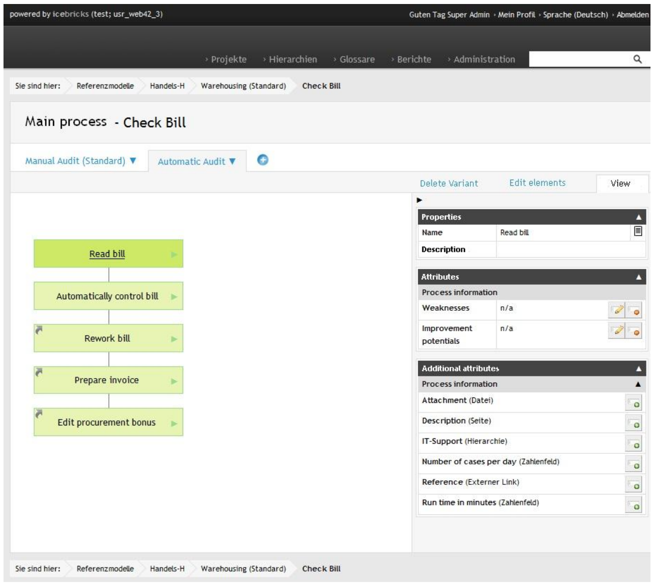
There are several scenarios where one outcome of a process is achieved by different process activities. This often leads to complex process models, since they take a range of possible circumstances into account in the sense of additional model components. A smart way to bypass this driver of complexity is to define several variants of one process. By this mean, the process model itself often remains simple with respect to branching and model elements but therefore the amount of simple model variants is increasing. It is a trade-off between complex models and several variants of one process model. Within the proposed concept, a new model variant is created, whenever the incoming and outgoing information of the process is the same, but at least one main process activity is different from the standard procedure.

V. PROTOTYPICAL IMPLEMENTATION AND EVALUATION

In this section, the prototypical implementation of the modeling technique proposed in the preceding section, is described. Moreover, an evaluation of the resulting tool in two medium sized enterprises is presented.

A. Implementation

Based on the conceptual model proposed in the preceding section, we have implemented a modeling tool prototype (Fig. 2) which fulfills all the aforementioned requirements. It is a web application based on the programming framework Ruby on Rails which follows the model-view-controller paradigm [21]. Therefore, it provides an elegant solution to separate the underlying data storage, the business logic and the presentation of the data. As the underlying database structure is easily exchangeable, the tool is able to be utilized in most scenarios and organizational IT infrastructures. Moreover, to facilitate an efficient and effective creation as well as utilization of the process models, the user interface for modeling as well as presentation of the models is designed to be highly intuitive also for non-process modeling experts. This is even enhanced by the use of JavaScript which is a client-side programming language fostering asynchronous handling of user input. Thus, irritating reloads of web pages are contained in the prototype.



Main process - Check Bill

Name: Read bill
Description:

Process information
Weaknesses: n/a
Improvement potentials: n/a

Process information
Attachment (Datei)
Description (Seite)
IT-Support (Hierarchie)
Number of cases per day (Zahlenfeld)
Reference (Externer Link)
Run time in minutes (Zahlenfeld)

Figure 2. Main process view in the prototype.

Within the conceptual model, a layer-architecture is described as a cornerstone for the modeling technique to be implemented. Within the prototype, this layer-architecture is realized as four-layer architecture. It consists of the layers process framework, main processes, detail processes and process building blocks (PBB) (Fig. 3). On the first layer, a *process framework* provides the modelers and model users alike with a process overview respectively process landscape comprising all relevant main processes within the depicted organization ordered by e. g. functional areas. The elements of the process framework are further specified on a more detailed level in the *main process layer*. Here, the main process steps are described in order to give a rough overview about the activities usually carried out during this process in the respective business area. As an example the main process "Check Bill" is shown in figure 2. To handle parallel steps, branching methods are supported by design on this layer. Each of the main process steps is further refined by a detail process on the *detail process layer*. Like in the superordinate layer, branching methods are provided on this layer to handle parallel activities. Every modeled element on this layer is represented by a so called process building block. These PBB are defined in detail on the fourth and most detailed layer. Here, the information about the atomic activities of the depicted processes can be provided. For example, attachments like videos, documents, hyperlinks, wiki pages, etc. are supported.

According to the description in the conceptual model, the prototype features attribution on each of the four model layers. Here, process-enhancing and additional information can be provided for each of the model elements on each layer. The attributes can be specified by the administrators of the tool and by providing administrators with the possibility to specify the concrete attributes themselves, the tool allows for utilization in any organization and business area. In the example in figure 2 two attributes were defined for "Read bill" activity of the main process: weaknesses and improvement potentials.

Corresponding to the utilization of a semantic modeling approach postulated in the conceptual model, a glossary is implemented in the prototype. With it, the aforementioned naming conflicts are contained. The concrete implementation in the tool allows for the creation of glossaries in which business objects and activities can be maintained. Moreover, an assignment of activities to business objects assures that only correct combinations can be assigned to process elements. The usage of the glossary and the abovementioned four layer architecture of the prototype are aligned as well. On the process framework layer, the elements – which are the main processes – can be assigned to a business object. On the subordinate layers – main processes and detail processes – the elements – detail processes respectively PBB – can be assigned a predefined phrase structure of a business object along with an activity. By this, modeling conventions are adhered and costly refinements or corrections are avoided.

Eventually, the usage of variants is facilitated within the prototype. On the main process and detail process layers different variants can be created whenever necessary. In figure 2 there were defined two variants of "Check Bill" main process: manual audit and automatic audit. The last three activities in the exemplary main process are the same for both variants and therefore are modeled only once in the standard variant ("Manual Audit") and afterwards inserted as references in "Automatic Audit" variant.

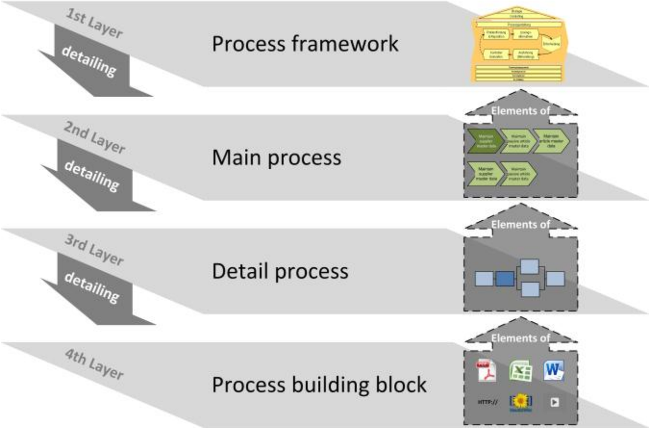


Figure 3. Layer concept of conceptual model.

B. Evaluation

The prototype has been evaluated in two process modeling projects. The characteristics of the companies, whose processes were modeled, are shown in Table 1. Although their characteristics differ, both companies are archetypical medium sized companies without a documented process landscape. As small and medium sized companies constitute the majority of all companies worldwide, the cases at hand are good examples for many process modeling projects to be conducted with the prototype in the future.

The first case is a business process reorganization project, which was conducted as the preparation for a consecutive ERP selection procedure. The project was structured in three steps, namely as-is process modeling to document the process landscape of the company, and two interrelated phases of to-be modeling and ERP selection to align both IT and infrastructure. All steps were conducted by a consulting company with the help of the prototype.

During the as-is modeling process, the consultants created the process models on the base of interviews with employees of the target company. As the company had not undertaken any modeling activities before the project, the processes were designed on the base of a reference model for processes in retail. Subsequently, the to-be processes were created in collaboration with all stakeholders. To support the to-be modeling phase, the prototype was customized with the attributes for detailed process element description, leading IT system, process owner, process executive, process relevance and importance of the process. Furthermore, examples of process related business documents like Excel files or scans of paper based documents were added to the processes. When engaging the as-is modeling phase, attributes for weaknesses, suggestions for improvement and optimization potential were added. In the last phase, the process models were annotated with attributes to document requirements for the new ERP software. The glossary for the process element labels was initially created before the to-be modeling phase and consecutively enhanced through the interviews. Re-use of the process models turned out to be simple, as small changes to the attributes were sufficient to re-use models created during different project phases.

TABLE I. OVERVIEW ON THE EVALUATION CASES

	Case 1	**Case 2**
Domain	Retail	Warehousing
Articles	~20.000, sports and fashion	~2.500, promotion material
Employees	>1.600	~250 worldwide
Customers	B2C	~6000, B2B
Modeling purpose	Process reorganization, ERP selection	Process documentation, preparation of software tests, knowledge base

The second case is a process documentation project that consists of an as-is process modeling phase. The project models will be used as a knowledge base for the company's employees and support the creation of test cases for an update of the ERP software. Like in case one no processes were documented by the company in advance and the processes were developed on the base of retail reference processes and interviews. The prototype was customized with the attributes for detailed process element description, process owner, process executive, SAP transaction codes for the test cases and links to the company-Wiki for knowledge management purposes. Analogue to case one, the glossary was enhanced consecutively during the interviews.

Although the two companies and the reasons behind the modeling projects differ significantly, the prototype and the proposed modeling technique could adapt well to both scenarios. The process modelers deemed the four layer modeling architecture well suited for the task at hand. They especially favored the simple syntax of the modeling notation and the alignment of modeling purpose and modeling technique through attributes. Glossary creation, in contrast, was carried out with reservations. The reservations were however dissolved for the most part during later stages of the project, when major renaming could be executed centrally in the glossary.

The results of the evaluation cases are promising, as they attest the modeling technique to be applicable in practice. All project stakeholders directly involved with the process models judged the collection organization support of the tool to be sufficient. Re-use of the process models turned out to be simple, as the tool could be adapted to different modeling purposes by small changes to attributes and glossary. The two aims of supporting collection organization and re-use of process models have therefore been reached.

VI. CONCLUSION AND OUTLOOK

A solution to the above stated research questions is presented in this work in form of a modeling technique which is based on a multi-layered process structure and the idea of semantic process building blocks. The proposed technique is generic with respect to the management of arbitrary PMC by allowing the technique to be tailored to the specific modeling purpose via attribution. On the one hand, the multi-layered structure allows modeling of business processes for companies of any size and from any business area. It enables efficient management of the resulting process model collection and includes features for variant management. On the other hand, the semantic building blocks in combination with a glossary allow the technique to create strongly standardized models. By adding extensive attribute support, flexibility regarding the modeling purpose is preserved.

With regard to an outlook, two aspects are in the focus of future research concerning the proposed modeling technique. On the one hand, further utilization of the attribution incorporated by the modeling technique is desirable. Here, especially the conceptual development and later implementation of elaborate analysis techniques and functionality is on the agenda. On the other hand, an in-depth evaluation of the prototype has to be carried out. While the two exemplary case studies have initially shown that the rationales seem to be valid, further validation will have to prove that to an even higher extent.

All in all, the concept of the modeling technique proposed in this paper and its prototypical instantiation have been validated by the evaluation of the prototype in two modeling undertakings. Thus, the requirements and research questions imposed by the challenges addressed in this work are met.

REFERENCES

[1] Aguilar-Saven, R.S.: Business process modeling: Review and framework. International Journal of Production Economics 90, 129-149 (2004)

[2] Batini, C., Lenzerini, M., Navathe, S.B.: A Comparative Analysis of Methodologies for Database Schema Integration. ACM Computing Surveys 18, 323-364 (1986)

[3] Becker, J., Algermissen, L., Pfeiffer, D., Räckers, M.: Bausteinbasierte Modellierung von Prozesslandschaften mit der PICTURE-Methode am Beispiel der Universitätsverwaltung Münster. Wirtschaftsinformatik 49, 267-279 (2007)

[4] Becker, J., Breuker, D., Pfeiffer, D., Räckers, M.: Constructing comparable business process models with domain specific languages - An empirical evaluation. Information Systems Journal, 1-13 (2009)

[5] Benbasat, I., Zmud, R.W.: The Identity Crisis within the Is Discipline: Defining and Communicating the Discipline's Core Properties. MIS Quaterly 27(2), 183-194 (2003)

[6] Becker, J., Rosemann, M., Uthmann, C. V.: Guidelines of Business Process Modeling. Business Process Management 1806, 30-49 (2000)

[7] Delfmann, P., Herwig, S., Lis, L.: Unified Enterprise Knowledge Representation with Conceptual Models - Capturing Corporate Language in Naming Conventions. In: Proceedings of the 30th International Conference on Information Systems (ICIS 2009). Phoenix, Arizona, USA (2009)

[8] Dijkman, R., La Rosa, M., and Reijers, H. A.: Managing large collections of business process models - Current techniques and challenges. Computers in Industry 63(2), 91-97 (2012)

[9] Gregor, S., Jones, D.: The Anatomy of a Design Theory. Journal of the Association for Information Systems 8(5), 312-335 (2007)

[10] Hadar, I., Soffer, P.: Variations in conceptual modeling: classification and ontological analysis. Journal of the AIS 7, 568-592 (2006)

[11] Hevner, A. R., March, S. T., Park, J., Ram, S.: Design Science in Information Systems Research. Management Information Systems, 28(1), 75-105 (2004)

[12] Hipp, M., Mutschler, B., Reichert, M.: Navigating in Process Model Collections: A new Approach Inspired by Google Earth. In: Aalst, W.M.P., Mylopoulos, J., Rosemann, M., Shaw, M.J., Szyperski, C. (eds.) BPM Workshops 2011. LNBIP, vol. 100, pp. 87-98. Springer, Heidelberg (2012)

[13] Iivari, J.: A Paradigmatic Analysis of Information Systems As a Design Science. Scandinavian Journal of Information Systems 19(2), 39-64 (2007)

[14] Indulska, M., Recker, J., Rosemann, M., Green, P.: Business Process Modeling : Current Issues and Future Challenges. In: van Eck, P., Gordijn, J., Wieringa, R. (eds.) Advanced Information Systems Engineering 2009. LNCS, vol. 5565, pp. 501-514. Springer, Heidelberg (2009)

[15] Kuechler, B., Vaishnavi, V.: On theory development in design science research: anatomy of a research project. European Journal of Information Systems 17(5), 489-504 (2008)

[16] La Rosa, M., Wohed, P., Mendling, J., Ter Hofstede, A. H. M., Reijers, H. A., van der Aalst, W. M. P.: Managing Process Model Complexity Via Abstract Syntax Modifications. IEEE Transactions On Industrial Informatics 7, 614-629 (2011)

[17] La Rosa, M., Hofstede, A.H.M.T., Wohed, P., Reijers, H.A., Mendling, J., Aalst, W.M.P.V.D.: Managing Process Model Complexity via Concrete Syntax Modifications. IEEE Transactions On Industrial Informatics 7, 255-265 (2011)

[18] Malone, T.W., Crowston, K., Herman, G.A. (eds.) Organizing Business Knowledge: The MIT Process Handbook. The MIT Press, 2003.

[19] March, S. T., Smith, G. F.: Design and natural science research on information technology. Decision Support Systems 15(4), 251-266 (1995)

[20] Mendling J., Reijers, H.A., Recker, J.: Activity labeling in process modeling: Empirical insights and recommendations. Information Systems (IS). Special Issue on Vocabularies, Ontologies and Rules for Enterprise and Business Process Modeling and Management 35(4), 467-482 (2010)

[21] Morsy, H., Otto, T.: Ruby on Rails 3.1. 2nd, Updated and Extended Edition, Galileo Press, Bonn (2012)

[22] Vergidis, K., Turner, C.J., Tiwari, A.: Business Process Perspectives: Theoretical Development vs. Real-World Practice. International Journal of Production Economics 114(1), 91-104 (2008)

[23] Walls, J.G., Widmeyer, G.R., El Sawy, O.A.: Building an Information System Design Theory for Vigilant EIS. Information Systems Research 3(1), 36-59 (1992)

[24] Weiß, B. Process Modeling and Analysis in Banks: Leveraging Business Process Optimisation in the Financial Sector, Doctoral Thesis, University of Münster, Münster (Germany), (2011)

An Ontology-Aided Process Constraint Modeling Framework for Workflow Systems

Shasha Liu, Manuel Correa, Krys J. Kochut
Department of Computer Science
The University of Georgia
Athens, GA, USA
{shasha, correa, kochut}@cs.uga.edu

***Abstract* – Specification of non-functional and domain-specific constraints in workflow processes and incorporating them within workflow applications have posed persistent problems for workflow designers. In order to address these problems, we propose a constraint handling framework consisting of a Process Constraint Ontology and a Process Constraint Language. The extensible ontology allows workflow designers to specify constraint knowledge and vocabulary specific to their domain of interest. Subsequently, process constraints are formulated in the constraint language by utilizing the constraint concepts from the ontology. The constraints are connected to the affected process elements (activities, data, and performers), deployed along with the process definition, and enforced and handled at runtime by the workflow enactment system. Based on the proposed framework, we have implemented a prototype of a constraint-enabled workflow management system and used it to incorporate and enforce geospatial constraints for an emergency management workflow process.**

***Keywords* – process modeling, constraints, non-functional requirements, ontology, process constraint language**

I. INTRODUCTION

As defined by the Workflow Management Coalition (WfMC), a workflow is "the computerized facilitation or automation of a business process, in whole or part" [1]. Over the past few decades, workflow systems have been successfully applied in numerous areas of industries, including banking, manufacturing and scientific research. A workflow is often specified by a process definition language. However, one common disadvantage of current process definition languages is the lack of the capability of describing additional process constraints and non-functional requirements (NFRs). For example, Business Process Model and Notation (BPMN) does not include support for NFRs [2]. Nevertheless, functional and non-functional constraints are vital to process definitions of virtually all workflow applications during their design and development [3].

Constraints and NFRs have been a focus in workflow research since the introduction of workflow management systems due to their high impact on the overall success of workflow applications. Quality of service (QoS) is an important subset of NFRs [4]. Other process constraints may involve such factors as geographic or network locations and properties of system resources. For example, an emergency handling workflow process may require some of its tasks to be executed in close geographic proximity. Similarly, a scientific workflow may prohibit transfers and analysis of the generated experimental data by external workflows due to privacy or security concerns. In order to express such application-specific constraints and other NFRs with a process definition, workflow designers need an intuitive and clear method to specify the workflow constraints and NFRs, and these workflow constraints and NFRs should be enforced at runtime by the workflow engine to meet users' needs. Mapping these high-level requirement specifications to the low-level workflow execution remains a big challenge for the researchers and developers of workflow systems.

We summarize the high-level objectives of a constraint-enabled workflow system as follows: (i) constraint specifications should be expressed using a commonly agreed-upon vocabulary; (ii) constraint specifications should be reusable, extensible and intuitive to create; (iii) constraint specifications should be attached to any process elements, and their validation and enforcement should be supported by workflow runtime.

The main contribution of our work is twofold: (i) we have created a process constraint ontology, named ProContO, which enables process designers to express and share their knowledge of process NFRs and domain-specific constraints; (ii) we have developed a process constraint language, named PCL, which can be used to specify process constraints and NFRs in terms of process elements (such as BPMN tasks and data objects) and the constraint vocabulary defined in the ontology. Using our approach, a workflow process designer can define a workflow process and clearly specify a variety of constraints and NFRs that go beyond the expressiveness of typical process definition languages. An important aspect of PCL is that the constraint expressions can be deployed as part of the workflow application, and then evaluated and handled at runtime under a constraint-enabled workflow management system.

The rest of the paper is organized as follows. Section II presents three motivating workflow examples, while Section III contains a review of the related work. A process constraint ontology for constraints modeling is introduced in Section IV. Section V introduces process constraint language (PCL). Section VI proposes a general architecture for a constraint-enabled workflow enactment system and presents our prototype implementation, which is capable of handling geospatial constraints in workflow processes. Conclusions and future work are discussed in Section VII.

ISBN: 978-1-61208-254-7

II. Motivating Examples

In this section, three motivating examples are discussed to illustrate the importance of specifying constraints and non-functional requirements in workflow processes.

The first example is a simplified purchase approval process: a manager requests a new computer purchase, and this request is approved by another manager. In order to avoid fraudulent activities, managers who request and approve should be different. Although BPMN can specify an actor or a role, such a constraint cannot be specified.

GlycoQuant IDAWG^{TM} workflow is used by scientists at the Complex Carbohydrate Research Center at the University of Georgia to perform quantitative glycomics analysis. One part of the workflow can be represented as four sequentially connected tasks shown in Fig. 1. The raw data are produced by the mass spectrometer experiment task. Transferring raw data directly over the open internet may not be feasible due to the large data size (e.g., in gigabytes) and security concerns. Instead, it is preferable to transfer the data pre-processing task to the computer storing the raw data, and then only transfer the segmented and encoded data back to more powerful servers for further computational analysis.

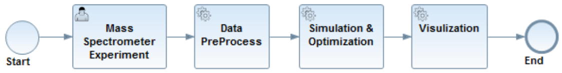


Figure 1: GlycoQuant IDAWG Workflow

The final example illustrates domain-specific geospatial constraints in an emergency management workflow that go beyond those simple constraints like task performers or the input/output data. Fig. 2 shows a fragement of the workflow process dealing with tornado emergencies. Once a tornado warning has been issued, schools within the tornado path need to be evacuated while shelters and hospitals that will not be affected but are near the area need to get prepared. It is apparent that geospatial constraints, such as the distance between a school and tornado path, are impacting the task execution.

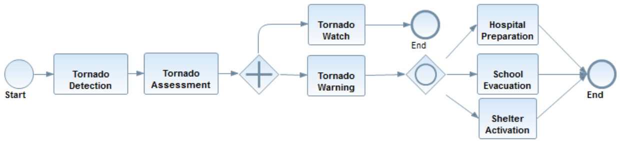


Figure 2: Tornado Emergency Workflow

III. Related Work

The work on representing constraints and NFRs within the general software engineering models has received a lot of consideration. Formalized language and graphical representation have been applied to define the constraints. NoFun [5] is a formalized language aiming to facilitate quantitative analysis of NFRs. A framework consisting of two languages, the process-NFL and the product-NFL, was proposed in [6] for building non-functional software architecture both in software developing phase and for the final software products. In [7], two additional artifacts, named Operating Condition and Control Case, have been introduced to BPMN to better discover and represent NFRs at an early phase of the business development life-cycle. In [8], flow model is applied to connect conceptual activity diagrams in UML and technical activity diagrams in BPMN for the purpose of design process continuity.

In summary, the formalized language representations help developers to easily express and document constraints, while the graphical notations provide an intuitive way for eliciting and visualizing them. A given constraint or an NFR can be stated by different vocabulary, which may lead to imprecise and ambiguous specifications [5]. This is a difficult problem for the reported modeling approaches. In [9], the authors addressed this problem by embedding specific keywords in their modeling framework to control the concepts and vocabulary used by developers. However, it is next to impossible to reach a consensus on a good set of constraint concepts expressive enough to cover a wide variety of application domains.

Recently, a lot of work has been focused on building ontologies for QoS, NFRs and domain-specific requirements in Web services, business and scientific workflows to promote such consensus regarding the constraints concepts and relationship among them. In [10], Dobson, et al. developed an ontology to model non-functional aspects in service-centric systems, based on which, he and his colleagues later presented a domain-independent ontology for NFRs and illustrated its application in a business trip service [11]. DAML-QoS [12] is another example of using ontology to model QoS for web services. However, based on a rapidly developing interest in scientific workflows, there is an increasing need for a general-purpose constraint specification framework that can model both non-functional and other domain-specific requirements. Our work addresses these issues by introducing an extensible process constraint ontology and a process constraint language. We also propose a software framework suitable for the development and execution of workflows incorporating constraints and NFRs.

IV. Process Constraint Ontology

Specification and handling of process constraints should be an integral part of a well-designed workflow application. Our motivating examples presented a few types of constraints that may be found in many other processes with similar types of requirements. We believe that ontologies offer the requisite expressive power to define the knowledge about process constraints, their classification and relationships, as well as suitable relationships connecting them to process components.

The high-level classes of our Process Constraint Ontology, (ProContO), are shown in Fig. 3. The *ProcessElement* class represents components in process definitions (activities, data objects, and performers). A *ProcessElement* may have a number of *Constraint-Attributes*, which represent simple constraint properties, such as the execution time of a task, its geospatial position, the size of an input data or a host's network location. However, many constraint attributes may have to be computed by suitable operations (for example the distance between locations of two tasks). Such operations are represented by the *Constraint-Operation* class. A *ConstraintOperation* takes *Constraint-Attributes* as its parameters and produces a *Constraint-Attribute* as its output. The three classes and relationships among them serve as the backbone of our process constraint ontology, which will be explained in greater detail in the rest of this section. An important aspect of our approach is that the ontology is meant to be extensible and the three classes are regarded as the roots of their respective hierarchies.

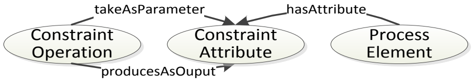


Figure 3: Backbone of the Process Constraint Ontology

A. Process Elements

Existing process definition languages, such as BPMN and XPDL, use different names for the components in a workflow process, but their functions and relationships are similar. Following the BPMN specification, ProContO includes the *Activity* class, used to represent a task within a workflow process, as shown Fig. 4. An *Activity* may input and output *Data*. Each *Activity* is executed by a processing entity, represented by the *Performer* class. *Process-Elements* can be described by attributes, such as the execution time of an *Activity* or the size of a *Data* object. Such properties can later be used in defining process constraints. For example, *SizeAttribute* and *Temporal-Attribute* are types of *ConstraintAttributes* in Fig. 4.

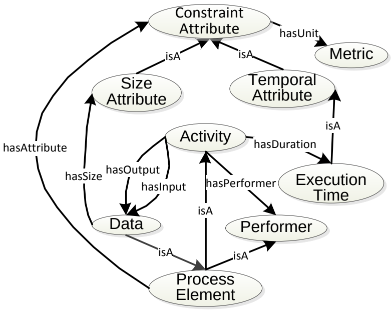


Figure 4: *ProcessElement* and *ProcessConstraint*

B. Constraint Attributes

The *ExecutionTime* is a subclass of the *Temporal-Attribute* and can be used to describe the properties of an *Activity*. Similarly, the *SizeAttribute* can be used to describe properties of *Data* elements. As shown in Fig. 4, *Activity* within the *ProcessElement* module is related to *ExecutionTime*, a subclass of *ConstraintAttribute*, by the object property *hasDuration*, and *Data* to *SizeAttribute* by *hasSize*. Both *hasDuration* and *hasSize* are defined as sub-properties of the *hasAttribute* relationship in the ontology (not depicted in the figure). Process designers may extend the ontology by adding additional *Constraint-Attributes* suitable for their application domain.

Another important part of the constraint ontology is the *Metric* module. It is meaningless to define a numerical constraint without giving its unit. For example, the *ExecutionTime* of an *Activity* may be specified in milliseconds or hours. Defining various units of measure is important but goes beyond the scope of this paper. An example of a metric ontology called QoSOnto in [10].

C. Constraint Operations

Some process constraint attributes cannot be expressed as simple properties attached to process elements and must be calculated via particular operations. The *Constraint-Operation* class is introduced to handle such constraints.

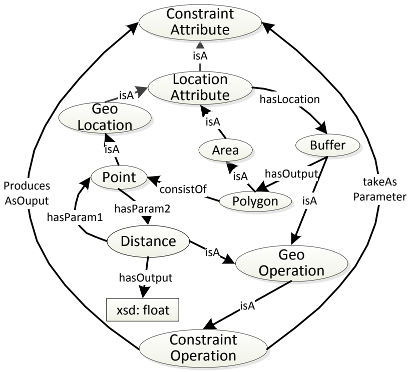


Figure 5: *ProcessConstraint* and *ConstraintOperation*

Considering the domain specific constraints in the Tornado Emergency Workflow depicted in Fig. 2 as an example, locations of hospitals and schools in the emergency system are viewed as geographic locations specified by their longitude and latitude. The distance between them can be calculated at runtime. The *Distance* operation, as shown in Fig. 5, takes two *Points* as parameters. The computed distance to the projected tornado path can later help to determine which nearby hospitals should be alerted to the tornado. For example, the hospitals closer than 30 miles to the tornado path should be prepared for evacuation, while those 30 to 50 miles away should be prepared to accept injured patients.

The *Buffer* operation is an example of an operation that produces a *ConstraintAttribute* as the output rather than a numeric value. In the context of the Tornado Emergency Workflow, the *Buffer* operation can be used to calculate the buffer area (a zone around a map feature), which is a *Polygon* area outlining the tornado path. As explained later, the operations defined here will have corresponding executable functions available for the runtime system.

Our process constraint ontology can facilitate the modeling of process constraints since it (i) serves as a concept vocabulary enabling process designers to use common language when specifying constraints, (ii) allows the specification and correctness validation of process constraints, (iii) can be easily extended by process designer to represent a variety of application domains.

V. Process Constraint Language

During process design of control and data flows, additional constraints are elicited and added as classes and instances in the constraint ontology. The next step involves specifying constraint expressions and connecting them to suitable elements in the process definition. To enable this, we have created an ontology-aided process constraint language (PCL). It serves as a declarative specification language for formulating and documenting additional process requirements. PCL constraint expressions are attached to the designed process and ultimately deployed to the enactment service for execution. Syntax of PCL is similar to that of the Object Constraint Language (OCL). Although the constraints discussed in motivating examples are difficult to specify via current process definition languages (e.g., BPMN), they can be defined in PCL constraints as described in the following subsections.

A. *Expressions*

A PCL expression is a logical assertion of a constraint, which evaluates to a Boolean value (true or false**)**. Table I shows the outline of the syntax of PCL expressions (defined using the Extended Backus-Naur Form). A literal is the smallest expression in PCL, which can be a string, a number, or a name. For brevity, we don't precisely define strings and numbers. A name is an identifier referring to a concept in the constraint ontology or a name of an activity in the process definition. It is also used to identify a constraint. Larger expressions are formed with the use of unary and binary operators (Table II), which include arithmetic, logical operators and navigation operators. The two navigation operators are used to traverse relationships in the ontology. The difference between "." and "→" is that the "." operator navigates the ontology by class names on the other side of associations, while the "→"operator uses the name of a specific relationship.

TABLE I. PCL EXPRESSIONS

expression	::=	logical_expr
logical_expr	::=	relational_expr {logical_op relational_expr }
relational_expr	::=	arithmetic_expr [relational_op arithmetic_expr]
arithmetic_expr	::=	unary_expr {arithmetic_op unary_expr }
unary_expr	::=	[unary_op] navigation _expr
navigation_expr	::=	primary_expr [navigation_op name]
primary_expr	::=	"(" expression ")" | if_expr | constraint_call | literal
if_expr	::=	"**if**" expression "**then**" expression "**else**" expression "**end if**"
constraint_call	::=	name "(" [constraint_parameters] ")"
constraint_params	::=	expression {"," expression }
literal	::=	string | number | name | "**true**" | "**false**"

A constraint call is an invocation of a constraint operation defined in the constraint ontology (an instance of a class in the *ConstraintOperation* hierarchy). The name should always starts with a lower case letter. As an example, consider a call *buffer(tornado.geoLocation)*, where *Buffer* is the name of a *ConstraintOperation,* while *tornado* is an alias of the Tornado Assessment activity in the process shown in Fig. 2. The *LocationAttribute* of a tornado is accessed through the "." navigation operator. An alias of an activity can be declared in the context declaration of a process constraint, which will be explained in the next subsection. Due to space limitations, additional PCL elements, such as quantifiers to deal with constraints on sets of process elements or their attributes, are not discussed here.

TABLE II. PCL OPERATORS

Operator			Associativity
unary_op	::=	"**not**"	right-to-left
logical_op	::=	"**and**" | "**or**" | "**xor**"	left-to-right
relational_op	::=	"="|">"|"<"|">=" |"<=" | "<>"	
arithmetic_op	::=	"+" | "-" | "*" | "/"	
navigation_op	::=	"." | "→"	

B. *Constraint Declarations*

Connections between constraints and processes are specified in the context definition part, which is used to list the process activities involved in the constraint. As the name of an activity can be long, activity aliases can be introduced at the same time. A constraint is identified by its name and includes one or more conditions, which are either invariants, pre-, or post-conditions. The syntax of constraint definitions is shown in Table III.

TABLE III. PCL CONSTRAINT DECLARATION

constraint_declaration	::=	"**constraint**" name context_definition condition { condition }
context_definition	::=	"**context**" [alias ":"] name {"," [alias ":"] name}
condition	::=	constraint_type [name] expression {"," expression}
constraint_type	::=	"**inv**" | "**pre**" | "**post**"

An invariant condition (**inv**) must hold during a workflow instance execution. More specifically, it is checked *before* and *after* the execution of all activities listed in the **context** definition. In case not all of the constraint attributes used in the expression are available (have already been established) due to the relative ordering of activities determined by the process control flow definition, the assertion is considered true. Consider an example shown in Table IV. Before and after the execution of the PurchaseRequest activity, the constraint is true, as the Performer of PurchaseApproval is not available yet. The actual verification of such a constraint can only be performed once the activity of PurchaseApproval starts and its *Performer* has been determined.

Pre-conditions (**pre**) define the required status of process elements *before* they start to execute. If more than one activity is declared in the context definition, the constraint expression specified within the **pre** clause needs to be verified at the starting point of each activity instance. Again, the constraint is trivially asserted as true if some attributes are not available yet (task has not executed yet).

Examples of pre-condition definitions are shown in Table V. Similarly, post-conditions (**post**), are evaluated *after* the execution of each involved activity.

The name given to a constraint not only facilitates the documentation and serialization of constraints along with the process definition, but also makes it easier to connect constraints with suitable exception handling methods within the process. If a constraint fails during a process instance execution, an exception is thrown and made available to the workflow engine and handled in a proper way, according to the defined exception handler. PCL is a declarative language used to specify constraints, while the process definition language, such as BPMN, is the proper place to provide detailed logic for handling of failed constraints. This issue will be further discussed next.

TABLE IV. CONSTRAINT DEFINITION OF THE PURCHASING WORKFLOW

```
constraint  ApprovalPermission
context     t1: PurchaseRequest,  t2: PurchaseApproval
inv         t1→hasPerformer <> t2→hasPerformer
```

TABLE V. CONSTRAINT DEFINITION OF GIS AND IDAWG$^{\text{TM}}$ WORKFLOW

```
constraint  GeoLocationProximity
context     tornado: TornadoDetection,
            shelter: ShelterActivation
pre         not  withIn( shelter.geoLocation,
                      buffer(tornado.geoLocation) )
            and distance(tornado.geoLocaion,
                       shelter.geoLocation) ) < 50 mile
--------------------------------------------------------
constraint  DataTaskCoLocation
context     t1: MassSpectrometerExperiment,
            t2: DataPreProcess
pre         if  t1→hasOutput > 1GB
            then t2.networkLocation =
                    t1→hasOutput.networkLocation
            else  true endif
```

VI. PROTOTYPE IMPLEMENTATION

In this section, we introduce a general architecture for a constraint-enabled workflow system and describe our prototype implementation based on the jBPM 5 workflow management system. jBPM 5's process definition is based on BPMN 2.0 and our prototype implementation adds the constraint specification and handling to a BPMN process definition by linking the PCL constraints to BPMN process elements and enforcing them during execution.

A. *System Architecture*

As shown in Fig. 6, our architecture has three layers: (i) the User Interface Layer, (ii) the Process Constraint Engine Layer, and (iii) the Context Awareness Layer. The User Interface Layer manages the interactions between a process designer and the underlying process-constraint engine. It not only provides intuitive graphical notation to define processes, but also provides a way of uploading the related constraint ontologies, as well as defining constraints using PCL. The Context Monitor module in the Context Awareness Layer is responsible for runtime context information, such as the network status, workload balancing statistics of the system and the geospatial information of a newly formed tornado in an emergency reaction workflow during the execution. The information is further processed by the Context Handler and fed to the Constraint Validator.

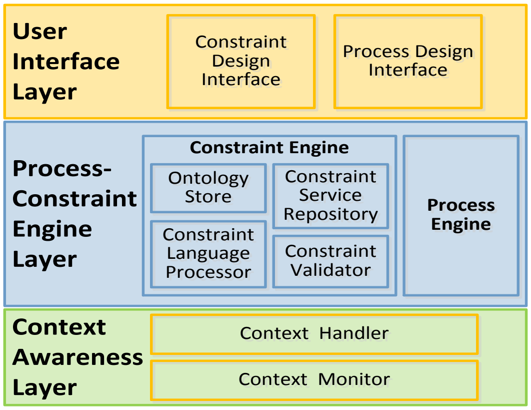


Figure 6: Prototype Architecture

The Process-Constraint Engine Layer is the core component of the constraint-enabled workflow system. It consists of the Constraint Engine and Process Engine.

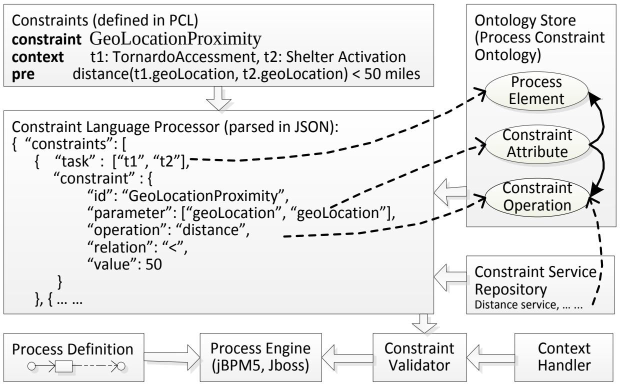
Within the Constraint Engine module, there are four sub-components. The Ontology Store contains the created ProContO ontology, including process elements, constraint attributes and constraint operations, while the Constraint Service Repository serves as a registry to manage the services related to the constraint operations and provide references to the Constraint Language Processor and Constraint Validator about how to interact with these services. The Constraint Language Processor parses the constraints specified in PCL and validates the syntax and semantics using the constraint ontology. It translates the constraints into executable constraint objects stored in the Constraint Service Repository. The translation keeps track of the mapping between the constraint operations defined in the ontology and the actual implementation of such operations (e.g., the code for calculating the distance between two points). This mapping enables process designers to focus on the constraint design without worrying about the underlying implementation.

The Constraint Validator is responsible for validation of constraints against the context information. It determines whether the constraint is satisfied and provides the validation results to the process engine.

One part of the Process Engine's functionality is to accept a process definition from the user interface layer, deploy it within the engine and execute process instances. In addition, since it is constraint-enriched, the Process Engine also receives input from the Constraint Engine concerning the constraints validation, and adapts its behavior accordingly, if needed. To be more specific, if the Constraint Validator does not detect any failed constraints, the Process Engine continues the normal sequence flow defined in the process. However, if one of the constraints fails, the Constraint Validator throws an exception corresponding to the name of the constraint. The Process Engine catches the exception and invokes the corresponding exception handler, if one has been defined in the process. If no suitable exception handler has been defined for the exception thrown by the failed constraint, a default action, such as suspend or terminate the execution of the current process instance is triggered.

B. *A Prototype Implementation*

Based on the general architecture, we have implemented a constraint-enabled workflow system prototype, focusing on enforcing geospatial constraints for emergency response processes [13]. It utilizes the existing jBPM 5 (from Redhat's jBoss) as the Process Engine. Domain specific constraint operations are implemented as Web services coded in Python. User defined geospatial constraints defined in PCL, are translated into JSON strings and mapped to the corresponding constraint operations. Every time a process instance is created by the process engine, the Constraint Validator generates a validator instance based on the associated constraint specification and evaluates them based the given runtime context information. The whole procedure is illustrated in Fig. 7. BPMN exception handling mechanism to is used to signal and handle runtime exceptions.



```
Constraints (defined in PCL)
constraint  GeoLocationProximity
context     t1: TornardoAccessment, t2: Shelter Activation
pre         distance(t1.geoLocation, t2.geoLocation) < 50 miles
```

```
Constraint Language Processor (parsed in JSON):
{  "constraints": [
    {   "task" :  ["t1", "t2"],
        "constraint" : {
               "id": "GeoLocationProximity",
               "parameter": ["geoLocation", "geoLocation"],
               "operation": "distance",
               "relation": "<",
               "value": 50
        }
    }, { ... ...
```

Ontology Store (Process Constraint Ontology)
Process Element
Constraint Attribute
Constraint Operation
Constraint Service Repository
Distance service, ...
Process Definition
Process Engine (jBPM5, Jboss)
Constraint Validator
Context Handler

Figure 7: Procedure of constraint-enabled workflow system

VII. Conclusion And Future Work

In this paper, we have addressed the problem of modeling and specifying domain-specific constraints and NFRs in workflow processes and incorporating them within workflow applications. We introduced a process constraint handling framework consisting of a process constraint language (PCL) and an extensible process constraint ontology (ProContO). ProContO allows workflow designers to specify constraint knowledge and vocabulary specific to their domain of interest, which we illustrated with examples. Process constraints are formulated in PCL, utilizing the constraint concepts defined in ProContO. PCL constraints are connected to process elements and deployed along with the process definition for execution. We have implemented a prototype of a constraint-enabled workflow management system and used it to create an emergency management workflow incorporating to handle geospatial constraints.

In the near future, we plan to integrate user interface of extending the ontology and creating PCL constraints with existing process definition tool. We also intend to integrate the framework with Web service composition methods.

Acknowledgment

This work was funded in part by the National Institutes of Health grant RR018502 for the Integrated Technology Resource for Biomedical Glycomics.

References

1 WfMC: 'Terminology & Glossary' (Winchester, 1999, 3rd edn. 1999)

2 Gorton, S., and Reiff-Marganiec, S.: 'Towards a task-oriented, policy-driven business requirements specification for web services', Proc. The 4th International Conference on Business Process Management (BPM 2006), 2006, pp. 465-470

3 Chung, L., and do Prado Leite, J.: 'On Non-Functional Requirements in Software Engineering': 'Conceptual Modeling: Foundations and Applications' (Springer-Verlag, 2009), pp. 363-379

4 Cardoso, J., Sheth, A., Miller, J., Arnold, J., and Kochut, K.: 'Quality of service for workflows and web service processes', Web Semantics: Science, Services and Agents on the World Wide Web, 2004, 1, (3), pp. 281-308

5 Franch, X., and Botella, P.: 'Putting non-functional requirements into software architecture', Proc. The 9th International Workshop on Software Specification And Design, 1998, pp. 60-67

6 Rosa, N.S., Justo, G.R.R., and Cunha, P.R.F.: 'A framework for building non-functional software architectures', Proc. The 2001 ACM Symposium on Applied Computing, 2001, pp. 141-147

7 Pavlovski, C.J., and Zou, J.: 'Non-functional requirements in business process modeling', Proc. The 5th Asia-Pacific Conference on Conceptual Modelling (APCCM 2008), 2008, 79, pp. 103-112

8 Al-Fedaghi, S.: 'BPMN Requirements Specification as Narrative', Proc. 3rd International Conference on Information, Process, and Knowledge Management (eKNOW 2011), 2011, pp. 68-75

9 Cysneiros, L.M., and do Prado Leite, J.C.S.: 'Nonfunctional Requirements: From Elicitation to Conceptual Models', IEEE Trans. Softw. Eng., 2004, 30, (5), pp. 328-350

10 Dobson, G., Lock, R., and Sommerville, I.: 'QoSOnt: a QoS Ontology for Service-Centric Systems', Proc. The 31st EUROMICRO Conference on Software Engineering and Advanced Applications (EUROMICRO-SEAA 2005), 2005, pp. 80-87

11 Dobson, G., Hall, S., and Kotonya, G.: 'A Domain-Independent Ontology for Non-Functional Requirements', Proc. The IEEE International Conference on e-Business Engineering (ICEBE 2007), 2007, pp. 563-566

12 Zhou, C., Chia, L.T., and Lee, B.S.: 'DAML-QoS Ontology for Web Services', Proc. The IEEE International Conference on Web Services (ICWS 2004), 2004, pp. 472-479

13 Correa, M.: 'Geospatial context awareness in business pocess modeling', The University of Georgia, 2012