SERVICE COMPUTATION 2012

The Fourth International Conferences on Advanced Service Computing

ISBN: 978-1-61208-215-8

July 22-27, 2012

Nice, France

SERVICE COMPUTATION 2012 Editors

Alfred Zimmermann, Reutlingen University, Germany

Pascal Lorenz, University of Haute Alsace, France

SERVICE COMPUTATION 2012

Foreword

The Fourth International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2012), held between July 22 and 27, 2012, in Nice, France, continued a series of events targeting service computation on different facets. It considered their ubiquity and pervasiveness, WEB services, and particular categories of day-to-day services, such as public, utility, entertainment and business.

The ubiquity and pervasiveness of services, as well as their capability to be context-aware with (self-) adaptive capacities posse challenging tasks for services orchestration, integration, and integration. Some services might require energy optimization, some might requires special QoS guarantee in a Web-environment, while other a certain level of trust. The advent of Web Services raised the issues of self-announcement, dynamic service composition, and third party recommenders. Society and business services rely more and more on a combination of ubiquitous and pervasive services under certain constraints and with particular environmental limitations that require dynamic computation of feasibility, deployment and exploitation.

We take here the opportunity to warmly thank all the members of the SERVICE COMPUTATION 2012 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 SERVICE COMPUTATION 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the SERVICE COMPUTATION 2012 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that SERVICE COMPUTATION 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of service computation.

We are convinced that the participants found the event useful and communications very open. We hope Côte d'Azur provided a pleasant environment during the conference and everyone saved some time for exploring the Mediterranean Coast.

SERVICE COMPUTATION 2012 Chairs:

SERVICE COMPUTATION Advisory Chairs

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Mihhail Matskin, KTH, Sweden
Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan
Bernhard Hollunder, Hochschule Furtwangen University – Furtwangen, Germany
Paul Humphreys, Ulster Business School/University of Ulster, UK
Arne Koschel, Fachhochschule Hannover, Germany
Michele Ruta, Politecnico di Bari, Italy

SERVICE COMPUTATION 2012 Industry/Research Chairs

Ali Beklen, IBM Turkey, Turkey
Mark Yampolskiy, LRZ, Germany
Steffen Fries, Siemens Corporate Technology - Munich, Germany
Emmanuel Bertin, Orange-ftgroup, France

SERVICE COMPUTATION 2012

Committee

SERVICE COMPUTATION Advisory Chairs

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain
Mihhail Matskin, KTH, Sweden
Hideyasu Sasaki, Ritsumeikan University - Kyoto, Japan
Bernhard Hollunder, Hochschule Furtwangen University – Furtwangen, Germany
Paul Humphreys, Ulster Business School/University of Ulster, UK
Arne Koschel, Fachhochschule Hannover, Germany
Michele Ruta, Politecnico di Bari, Italy

SERVICE COMPUTATION 2012 Industry/Research Chairs

Ali Beklen, IBM Turkey, Turkey
Mark Yampolskiy, LRZ, Germany
Steffen Fries, Siemens Corporate Technology - Munich, Germany
Emmanuel Bertin, Orange-ftgroup, France

SERVICE COMPUTATION 2012 Technical Program Committee

Julian Andrés Zúñiga, Ingeniero en Electrónica y Telecomunicaciones / Unicauca, Colombia
Ismailcem Budak Arpinar, University of Georgia, USA
Irina Astrova, Tallinn University of Technology, Estonia
Ali Beklen, IBM Turkey - Software Group, Turkey
Emmanuel Bertin, Orange-ftgroup, France
Juan Boubeta Puig, University of Cádiz, Spain
Radu Calinescu, Aston University-Birmingham, UK
Florian Daniel, University of Trento - Povo, Italy
Giuseppe De Pietro, Institute for High Performance Computing (ICAR) / National Research Council of Italy (CNR) - Napoli, Italy Manuel
Leandro Dias da Silva, Federal University of Alagoas, Brazil
Erdogan Dogdu, TOBB University of Economics and Technology - Ankara, Turkey
Massimo Ficco, Second University of Naples, Italy
Steffen Fries, Siemens Corporate Technology - Munich,, Germany
G. R. Gangadharan, Institute for Development & Research in Banking Technology [IDRBT] - Hyderabad, India
Luis Gomes, Universidade Nova de Lisboa / UNINOVA-CTS - Monte de Caparica, Portugal
Gustavo González, Mediapro Research - Barcelona, Spain
Andrzej M. Goscinski, Deakin University - Victoria, Australia
Victor Govindaswamy, Texas A&M University-Texarkana, USA
Mohamed Graiet, Institut Supérieur d'Informatique et de Mathématique de Monastir, Tunisie
Sven Hartmann, Clausthal University of Technology, Germany

Copyright Information

Table of Contents

PROYEVA: System to Evaluate the Projects Quality in Contests

System for the Technical Evaluation of Product Quality and Projects Participating in Invention Contests and Innovation Through the use of an External Metric and Quality in-use Plan

Laura Silvia Vargas-Pérez
ITCM Instituto Tecnológico de
Ciudad Madero, Tamaulipas, México
laura_silvia_vargas@hotmail.com

Agustín Francisco Gutiérrez-Tornés
ITESM Instituto Tecnológico de Estudios
Superiores de Monterrey-CCM, México
agustin.tornes@itesm.mx

Edgardo Manuel Felipe-Riverón
Centro de Investigación en
Computación, CIC, IPN, México
edgardo@cic.ipn.mx

Abstract: **PROYEVA is a system based upon international standards and Mexican models that allows for a comparative analysis of different projects and products involved in innovation, invention and creativity contests-based on the following characteristics: quality in-use, functionality and usability through an external metrics and quality in-use metrics plan in a visual environment. PROJEVA software is a practical application of the PROYEVA model. It allows a general quality evaluation of projects and products in technical competitions. It provides support for judges and facilitates a more objective and impartial evaluation. It also provides guidance on the ranking evaluation procedures and documentation. PROJEVA is directed to organizations, companies and end-users who need to easily select products or projects, with the highest quality among the contesters to be pronounced as winners.**

Keywords: contests; creativity; invention; innovation; system.

I. INTRODUCTION

Currently, it is not easy for an evaluator to render a judgment on projects which are outside the evaluator’s field of expertise. In many cases, he or she has to make hasty decisions and determine the quality in-use of a project based on subjective criteria, which does not allow for objective evaluation of the different aspects comprised.

There are several standard models that provide guidance for organizations in the measurement of the characteristics which give them access to a high quality level in their products and projects. In theory and in practice, it is necessary to adjust the models to obtain a qualimetric model with the aim of evaluating and measuring the quality characteristics. Often, these models are used for different purposes, such as buying, renting, using and adapting projects and products.

A project can be defined in terms of its distinctive characteristics: it is a temporary endeavor undertaken to create a unique product or service, developed at all levels of the organization, and can involve one person or thousands. They may involve just a single unit of an organization or the combined efforts of several. Projects are often critical components of the business strategy of the organization that has developed them; their durations are finite. Projects are not successive efforts; they involve doing something that has not been done before. Therefore, they are unique.

Because the product of each project is unique, the characteristics that distinguish the product or service should be developed progressively, which means "step-by-step procedures,” “continued progress by increments" while elaborated means "worked with attention to detail – “fully developed" [1].

In these cases, a comparative analysis of various products and projects helps one decide which to select as the best as far as its quality in-use.

As such, a methodology and quality in-use technical evaluation method for participating projects in creativity, invention and innovation contests through the implementation of Quality Metrics Models, external as well as quality in-use, and the use of supplementary software in support of the judges is proposed so said judges are able to issue a more accurate decision.

Thus, the PROYEVA model is introduced, based on international standards IEEE610 [2], IEEE1061 [3], ISO 9000-3 [4], ISO / IEC 9126 [5], ISO / IEC 14598 [6], ISO 9001 [7], Project SQUARE (ISO 25000) [8], SUMI [9] as well as in other Mexican models (MECHDAV [11-14], [16-18], MECRAD [19]).

The paper is distributed in the following way: Section II includes a short background; Section III explains in a general way the characteristics of the contests where the proposed system have been used; Section IV details the system’s methodology; Section V describes through figures the operation of the system; Section VI explains the final evaluation report, and finally Section VII includes the results, conclusions and future works.

II. BACKGROUND

The first project competitions have been geared mostly toward ``Experiments and Devices" projects organized year after year, for several decades by various academic bodies. These play a positive role by encouraging the contestants to demonstrate their ability and creative genius by submitting projects on the design of experiments, display or educational devices. In addition to inducing participants to research and learn, the presentation of projects helps in reaching practical objectives. It could be said that they are relevant to today's society which needs to motivate and stimulate the creative potential and capacity of professionals and students at all levels [10].

This type of event, on the other hand, also serves as a starting point to familiarize, the public, with the knowledge of what science generates, and is therefore an element in making this an integral part of popular culture.

To compete, the contestant must first make and take into account a systematic study on the feasibility of a particular project, and should take into consideration some of the following useful aspects.

Due to the importance of having high-level professionals, teachers, developers and individuals capable of providing technical scientific benefit to society, events to motivate their creativity are organized.

III. CONTESTS

Anyone who has an innovative idea to be made into a development project may participate. The idea should preferably be supported or based on technology and may be the result of the ingenuity of a person or group. There must be a Technical Committee, which reserves the right to evaluate and support the ideas presented and must not accept ideas that are not aligned with the specific objectives and spirit of the initiative of the contest in question. There are no restrictions on the participation of a group or one of the members nor the number of ideas presented, i.e. there may be several innovative ideas for a group or by some members of the group.

A. Current National Prototype Competitions

The technological, scientific and technical prototype exposition has been booming since the nineties. Creativity competitions are very important both for the institutions that choose an award, as major companies and entrepreneurs are looking for new ideas and services that provide added value to their production management.

B. State of the art

We conducted a thorough investigation on the possibility of the existence of systems (software) for the evaluation of projects in terms of quality concerns, focusing on this important issue in quality competition contests such as creativity, innovation and invention, or when assessing a technological, scientific, social, cultural, environmental project to be approved by and for society.

The investigation found that there is an issue about the contests, but the postulates are loosely based for example, there are degrees specializing in the field of project quality assessment, for determining the magnitude of the results of the evaluations, which are a fundamental element of cost-benefit analysis and cost-effective, widely used in project evaluation [10].

There were no courses that prepare and certify judges to evaluate projects involved in creative competitions to channel the above benefits and better prepare people as project quality evaluators.

Currently there is some software dedicated to the evaluation of projects. Among them are: evalAS [18] (Software for Investment Project Production Evaluation). The purpose of this software is to determine, in the best of cases, financial feasibility. It can also be used to determine profitability of industrial production projects, agriculture and forestry. In Intecplan [17], which only evaluates investment projects, both references have a totally different approach to evaluating projects in order to obtain a score to determine the best of its kind in creativity contests. The only previous software tool found, are the papers known as "Software for assessing quality in-use project with a plan for external quality metrics [15], which showed an initiation protocol of this investigation."

Vargas-Pérez et al. [20] describes an intermediate step in the project, and refers to the completion of the first stage of the project.

IV. METHODOLOGY

In order to evaluate projects - products participating in creativity, innovation and invention contests, the application of a metric plan within the framework of a methodology and a technical evaluation model of the quality of software products for visual environments, MECHDAV is required, which is derived from this proposal to evaluate products and projects participating in the contests mentioned in software in a visual environment.

The metrics program is reflected in a new model, with its methodology and evaluation software, PROYEVA Model Methodology and Quality Technical Assessment Project participant's creativity contests, which will guide the evaluation results obtained on quality in-use of a project, and propose actions to improve the process. In addition, it will control the process established for ensuring the quality of the evaluation of these projects to support the judges in the competitions for creativity, innovation and invention.

A. Metric Oriented to Quality Products Projects

It is important that product measurement (products) be done easily and economically, and that the measurement result is interpreted in the same way. The way in which quality characteristics have been defined does not permit them to be measured directly, so it is necessary to establish metrics that correlate these features in a product (project). Each internal and external quantifiable attribute interacts with its environment and is correlated with a feature that can be established as a metric. The basis on which the metrics are selected depends on the product, project priorities and needs of the evaluator.

A set of product metrics that can be applied to the quantitative assessment of the quality of projects is examined. In all the cases, the metrics represent indirect measures, and never really measure quality, but a manifestation of it. The complicating factor is the exact relationship between the variable measured and the quality of the product, which can be measured based on the classification of quality in use metrics. Quality in-use is the user's view of the quality of a system (project or product) and is measured in terms of the result of using it, instead of the properties of the product itself.

It is the combined effect of the characteristics of product quality as perceived by the user.

B. *Requirements analysis*

According to data collected by the potential users of products, different people involved, both as judges and competitors in creativity contests, have provided some of the requirements which when tested, refined and synthesized, provide components and parameters of the system to be implemented.

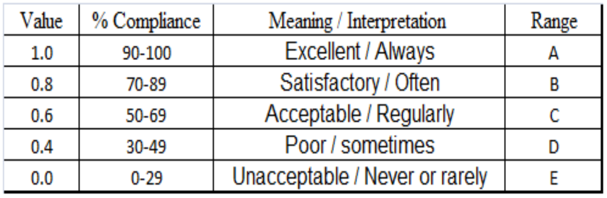
C. *Evaluation Process Applied*

To assess the quality of a product, the results of the evaluation of the different features need to be summarized. The evaluator must prepare a procedure for this which separates criteria for different quality characteristics, each of which may be in terms of individual sub-characteristics, or a combination of both. The procedure includes other aspects such as the specification's evaluation. In this part the scope of measurement is established, that is, the characteristics and sub-features set forth in the proposed quality model, which determine the starting point for the selection of attributes and metrics for evaluation.

Evaluation Metrics are grouped according to the corresponding sub-characteristics and attributes, and will serve to carry out the assessment.

Types of measurement are used to compare the quality in-use of the various products, and/or projects to be evaluated. They are represented by discrete evaluation variables of two types: binary discrete elemental evaluation variables and multilevel discrete evaluation variables. The numerical ranking scale for each of the metrics is presented in TablLe 1 [11-16], [20-26].

TABLE I. METRIC LEVELS RANGES



Value	% Compliance	Meaning / Interpretation	Range
1.0	90-100	Excellent / Always	A
0.8	70-89	Satisfactory / Often	B
0.6	50-69	Acceptable / Regularly	C
0.4	30-49	Poor / sometimes	D
0.0	0-29	Unacceptable / Never or rarely	E

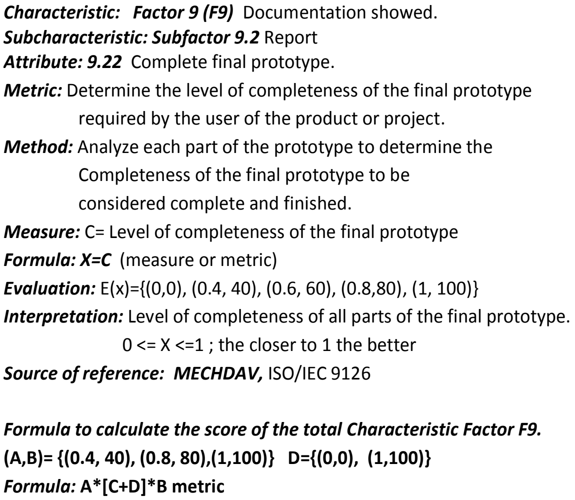
.

Translating the partial or total results of the evaluation of the quality of products projects is not an easy task, so a simple and understandable format to get a quick and reliable assessment of the quality of the different project representations should be selected. Checklists, **control matrix** and simple relationship tables are often chosen for this reason. Characteristic-Factor / Sub-Factor / Attribute / Metric. Figure 1 shows a documentation sample of one of the 42 combinations listed [20]-[26] and Table 2 shows the model PROYEVA arrayed in its 42 combinations [20-[26]

D. *Metrics Proposed for this Model*

Each component of the model requirements and methodology employed are divided into sub-components and parameters, which are represented by a metric, according to the application of the MECHDAV assessment model, which refers to this process. To calculate the metrics of each component and subcomponent mentioned, apply each of the formulas with their respective parameters described below:

1. Identify the area locating the project to be evaluated among the following four possibilities, corresponding to the most relevant project. The projects involved in creative competitions can be classified as follows: I. - Science - Technology. II. - Health and Environment. III. - Socio-economic, administrative and educational. IV. - Craft and Cultural.
2. Once the location of the project area is chosen, we suggest using the general procedure model proposed by PROYEVA (derived from MECHDAV) for 10 properties (factors), 26 sub-features (sub-factors), 42 attributes-metrics, which is fully represented by type I, then (somewhat fewer metrics) by type II, III and finally IV, which lacks several components of the model elements (attributes, metrics and sub-factors), in four levels of quality.
3. A score is assigned to each category or project type according to the PROYEVA compliance percentage for each combination of factors / sub-factors / attributes / metrics that apply, depending on the type of project. The first score assigned is the first metric that is calculated, which is given as follows for each of the types: I = 1.0, II = 0.9, III = 0.8, IV = 0.7.



Characteristic: Factor 9 (F9) Documentation showed.
Subcharacteristic: Subfactor 9.2 Report
Attribute: 9.22 Complete final prototype.
Metric: Determine the level of completeness of the final prototype
required by the user of the product or project.
Method: Analyze each part of the prototype to determine the
Completeness of the final prototype to be
considered complete and finished.
Measure: C= Level of completeness of the final prototype
Formula: X=C (measure or metric)
Evaluation: E(x)={(0,0), (0.4, 40), (0.6, 60), (0.8,80), (1, 100)}
Interpretation: Level of completeness of all parts of the final prototype.
0 <= X <=1 ; the closer to 1 the better
Source of reference: MECHDAV, ISO/IEC 9126

Formula to calculate the score of the total Characteristic Factor F9.
(A,B)= {(0.4, 40), (0.8, 80),(1,100)} D={(0,0), (1,100)}
Formula: **A*[C+D]*B metric**

Figure 1. Documentation about the 42 metrics used in PROYEVA

In the final grade for a project participant for each judge in any category, PROYEVA calculated metrics (equations) of each of the specified points, depending on the type of

ISBN: 978-1-61208-215-8

project that applies: the value assigned to each assessment, combined with the remaining fraction of each factor evaluated, accumulating the partial values, thereby calculating the result of each of the 10 factors. Finally, an equation is applied which represents the evaluation of all factors to be considered by the judges, for the project participant. The final score of a project is the combination of the recommendations given by all judges involved.

Finally, an equation is applied which represents the evaluation of all factors, enabling the judges to submit their opinion to the project. The final score of a project is the combination of the recommendations given by all of the judges involved.

TABLE II. MODEL PROYEVA

Characteristic / Factor	Subfeatures / Sub-Factor	Attribute / Attribute	Metric / Metric
1.1.1.1 F1	Project I	Science and Technology	/A
1.2.1.1 F1	Project II	Health and Environment	/B
1.3.1.1 F1	Project III	Social-Economic-Education	/C
1.4.1.1 F1	Project IV	Artisan-Cultural	/D
2.1.1.1 F2	Identification	Delimitation	/A1
2.1.2.1 F2	Identification	Hiposetis	/B1
2.2.1.1 F2	Objectives	General	/A2
2.2.2.1 F2	Objectives	Particles	/B2
2.3.1.1 F2	Scope	Techniques	/A3
2.3.2.1 F2	Scope	Socioeconomic	/B3
2.4.1.1 F2	Limitations	Techniques	/A4
2.4.2.1 F2	Limitations	Socioeconomic	/B4
3.1.1.1 F3	Originality	Invention	/A
3.2.1.1 F3	Originality	Innovation	/B
3.3.1.1 F3	Originality	Creativity	/C
4.1.1.1 F4	Feasibility	Financial	/A
4.2.1.1 F4	Feasibility	Tecnica	/B
5.1.1.1 F5	Justification	Socioeconomic	/A
5.2.1.1 F5	Justification	Tecnica	/B
6.1.1.1 F6	Formality	Level	/A
6.2.1.1 F6	Formality	Level of Complexity	/B
6.3.1.1 F6	Formality	Mathematical model	/C
6.4.1.1 F6	Formality	Graphic model	/D
7.1.1.1 F7	Registration	Pat	/A
7.2.1.1 F7	Registration	INDAUTOR	/B
7.3.1.1 F7	Registration	Utility model	/C
7.4.1.1 F7	Registration	Industrial Design	/D
7.5.1.1 F7	Registration	Integrated Circuit Layout	/E
8.1.1.1 F8	Level	Coverage	/A
8.2.1.1 F8	Level	Exhibition	/B
8.3.1.1 F8	Level	Contest	/C
8.4.1.1 F8	Level	Forum	/D
9.1.1.1 F9	Product	Over	/A
9.2.1.1 F9	Report	Full	/B
9.2.2.1 F9	Report	Prototype	/C
9.2.3.1 F9	Report	Manuals	/D
9.2.4.1 F9	Report	Models	/E
10.1.1.1 F10	Presentation	Item domain	/A

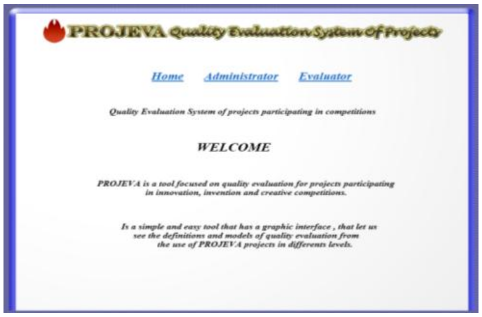
V. MAIN SCREENS OF THE PROTOTYPE PROYEVA

Figures 2, 3, 4, 5, 6 and 7 show some of the main screens that describe the operation of the system [20-26].

VI. FINAL EVALUATION REPORT

When the respective values of the selected project evaluation as well as the rate of quality compliance are obtained, a final evaluation report is generated in which the final results and the compliance percentages are given.

An outline is provided showing what the points are, where the product-producers stand out in quality as well as those which do not. It also dictates what level of quality is achieved according to the relevant points, and, if required, recommends changes so this draft is accepted as a draft-quality product.



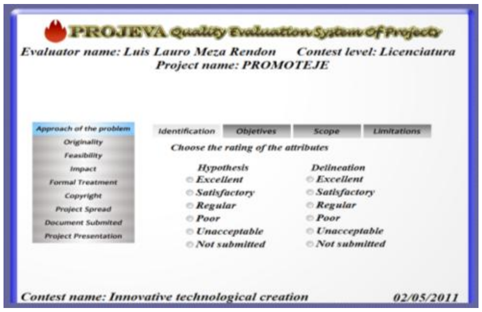
Quality Evaluation System of projects participating in competitions

WELCOME

PROJEVA is a tool focused on quality evaluation for projects participating in innovation, invention and creative competitions.

Is a simple and easy tool that has a graphic interface , that let us see the definitions and models of quality evaluation from the use of PROJEVA projects in differents levels.

Figure 2. Welcome Screen and Start at the PROJEVA System.



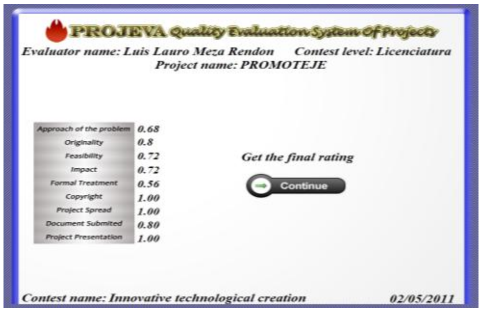
Evaluator name: Luis Lauro Meza Rendon Contest level: Licenciatura
Project name: PROMOTEJE

Choose the rating of the attributes

Hypothesis	Delineation
○ Excellent	○ Excellent
○ Satisfactory	○ Satisfactory
○ Regular	○ Regular
○ Poor	○ Poor
○ Unacceptable	○ Unacceptable
○ Not submitted	○ Not submitted

Contest name: Innovative technological creation 02/05/2011

Figure 3. View of a screen with points for evaluating the troubleshooting rubric.



Evaluator name: Luis Lauro Meza Rendon Contest level: Licenciatura
Project name: PROMOTEJE

Approach of the problem	0.68
Originality	0.8
Feasibility	0.72
Impact	0.72
Formal Treatment	0.56
Copyright	1.00
Project Spread	1.00
Document Submited	0.80
Project Presentation	1.00

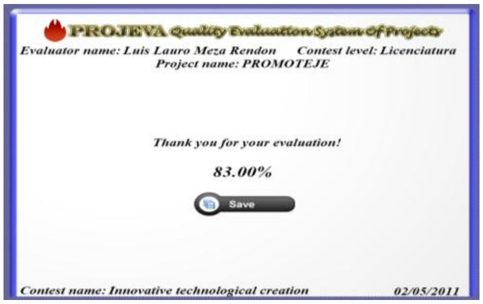
Get the final rating

Contest name: Innovative technological creation 02/05/2011

Figure 4. View of evaluation results of a project.

VII. RESULTS, CONCLUSIONS AND FUTURE WORKS

The preliminary phase of the PROYEVA project has been completed, covering the complete model and methodology for the technical evaluation of the quality of the projects participating in creativity, invention and innovation contests through the application of quality in-use metrics. In it, the first prototype of this type of software was developed, which is the proposed tool for a panel of judges to efficiently evaluate the quality in-use of the project participants in a particular creativity contest, [Copyright SEP INDAUTOR 03-2007-03201059300-01, and 03-2007-091813015000-01] (mathematical model and software) [23] [24] [25] [26] [27].



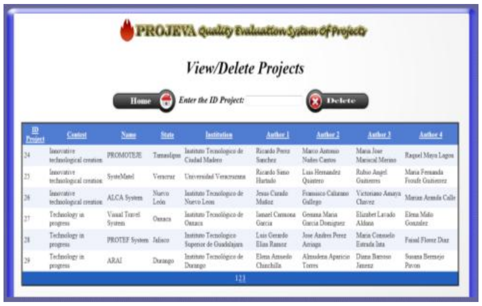
Evaluator name: Luis Lauro Meza Rendon   *Contest level: Licenciatura*
Project name: PROMOTEJE

Thank you for your evaluation!

83.00%

Contest name: Innovative technological creation   *02/05/2011*

Figure 5. View results of project evaluation.



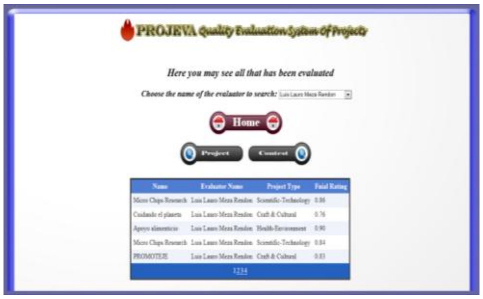
View/Delete Projects

Enter the ID Project:

ID Project	Contest	Name	State	Institution	Author 1	Author 2	Author 3	Author 4
24	Innovative technological creation	PROMOTEJE	Tamaulipas	Instituto Tecnologico de Ciudad Madero	Ricardo Perez Sanchez	Marco Antonio Nuñes Castro	Maria Jose Mariscal Merino	Raquel Maya Lagos
25	Innovative technological creation	SysteMatel	Veracruz	Universidad Veracruzana	Ricardo Sano Hurtado	Luis Hernandez Quintero	Rubio Angel Gutierres	Maria Fernanda Froufe Gutierrez
26	Innovative technological creation	ALCA System	Nuevo León	Instituto Tecnologico de Nuevo Leon	Jesus Curado Muñoz	Francisco Calixano Gallego	Victoriano Amaya Chavez	Marian Aranda Calle
27	Technology in progress	Visual Travel System	Oaxaca	Instituto Tecnológico de Oaxaca	Ismael Carmona Garcia	Gemma Maria Garcia Dominguez	Elizabet Lavado Aldana	Elena Malo Gonzalez
28	Technology in progress	PROTEF System	Jalisco	Instituto Tecnologico Superior de Guadalajara	Luis Gerardo Elias Ramos	Jose Andres Perez Arriaga	Maria Consuelo Estrada Iata	Faisal Florez Diaz
29	Technology in progress	ARAI	Durango	Instituto Tecnológico de Durango	Elena Amuedo Chinchilla	Almudena Aparicio Torres	Diana Barroso Jimenz	Susana Bermejo Pavon

123

Figure 6. Pproject list of participants in a contest.



Here you may see all that has been evaluated

Choose the name of the evaluator to search: Luis Lauro Meza Rendon

Name	Evaluator Name	Project Type	Final Rating
Micro Chips Research	Luis Lauro Meza Rendon	Scientific-Technology	0.86
Cuidando el planeta	Luis Lauro Meza Rendon	Craft & Cultural	0.76
Apoyo alimenticio	Luis Lauro Meza Rendon	Health-Environment	0.90
Micro Chips Research	Luis Lauro Meza Rendon	Scientific-Technology	0.84
PROMOTEJE	Luis Lauro Meza Rendon	Craft & Cultural	0.83

1234

Figure 7. List of projects evaluated by a jury in a contest.

There is also an English version PROJEVA for presentations abroad. The software will permit a very generic technical assessment, based on the quality in-use, creativity and project implementation. The assessment is very general, so it may issue an opinion on any project in any discipline and any level of competition: local, regional, state and national level, giving a reliable decision as a judge in creativity contests.

PROJEVA system is a service created for project quality in use evaluation, within innovation, invention and creativity contests, for different government agencies, industries and services that require an easy, fast and objective evaluation process which will help in the selection of a winning project in different categories.

This prototype is proposed for the creative competitions that take place in the National System of Higher Education Technology, for the state competitions organized by different universities, and national competitions organized by the National Institute for Women, National competitions of thesis, National Contests and Exhibition Projects Linking the different government sectors, among others. Additional formats are provided for manual evaluation of these contests. PROJEVA system can be adapted to various contests, for different juries as required. Projects may be installed in a multiuser environment for several judges, for various academic levels: primary, secondary, high school, undergraduate and graduate, and in a WEB environment. It will have the mobility to interact virtually any mobile device having WI-FI and it is in the range of broadband network provided by the host institute of the competition.

ACKNOWLEDGMENT

The authors of this paper wish to thank the Center for Computing Research (CIC) and National Polytechnic Institute (IPN), ITESM-CCM, ITCM-Tamaulipas, Mexico, for their support to this work.

REFERENCES

[1] R. Pressman, Ingeniería de Software. Un enfoque práctico, Mc.Graw Hill /Interamericana de España, S.A.U. 1998.

[2] IEEE Software Engineering Standards Collection, Standard Glossary of Software Engineering Terminology. IEEE, 1994. Std. 610.12-190.

[3] IEEE Std 1061, "IEEE Standard for a Software Quality Metrics Methodology", IEEE Computer Society Press, 1992.

[4] ISO 9000-3, 1991. ISO/IS 9000-3, "Quality Management and Quality Assurance", 1990.

[5] ISO/IEC 9126, Software Product Evaluation; Part 1: Quality, Characteristics and Guidelines for their Use; Part 2: External Metrics; Part 3: Internal Metrics. Part 4: Quality in Use.

[6] ISO/IEC 14598. Information Technology, Software Product Evaluation. (Part 1, 2, 3, 4, 5). 1998.

[7] ISO 9001, 1994, "Model for Quality Assurance in design, development, production, installation and servicing". 1994.

[8] ISO/IEC JTC C1/SC7 N2246. Software Quality Requirements and Evaluation Configuration. SQUARE2000. May 2000. Replaced by ISO/IEC 25000: 2005 Software Engineering -- Software product Quality Requirements and Evaluation (SQuaRE).

[9] SUMI: Software Usability Measurement Inventory. Human Factors Research Group, Ireland. European Directive on

Minimum Health and Safety Requirements for Work with Display Screen Equipment (90/270/EEC), 2000.

[10] Main site of research and development projects. Revista Espacios Vol.15 (1)1994. José Luis Solleiro. Evaluación de proyectos de investigación y desarrollo ¿alguna solución a este viejo problema? http://www.revistaespacios.com/a94v15n01/70941501.html . Retrieved on June 2012

[11] Laura S. Vargas-Pérez and Agustín F. Gutiérrez-Tornés. "Propuesta de un modelo de evaluación técnica de las herramientas de los ambientes visuales para el desarrollo de aplicaciones", ISBN: 84-688-783-1, Avances en Gestión de Proyectos y Calidad del Software pp.247-53. Oct. 2004. Universidad de Salamanca, España.

[12] Laura S. Vargas-Pérez and Agustín F. Gutiérrez-Tornés. "MECHDAV: a quality model for the technical evaluation of applications development tools in visual environments". 2nd Software Measurement European Forum, March 2005, pp.147-156, 2005. Rome, Italy.

[13] Laura S. Vargas-Pérez and Agustín F. Gutiérrez-Tornés. "MECHDAV: propuesta de un modelo sistematizado de evaluación técnica de la calidad del uso de las herramientas RAD para ambientes visuales". Revista de Procesos y Métricas. Editorial AEMES [Asociación Española de Métricas de Software], ISSN: 1698-2029. Volumen 3 Número 1, 2006. Madrid, España.

[14] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornés, and Edgardo M. Felipe-Riverón. "Propuesta de un modelo sistematizado de evaluación técnica de la calidad del uso de las herramientas RAD para ambientes visuales". ANDESCON2006 IEEE sección Andina. Noviembre 2006. Quito, Ecuador. CP126.

[15] Laura S. Vargas-Pérez and Jorge Peralta-Escobar "Software para la evaluación de la calidad en uso de proyectos mediante un plan de métricas externas de calidad". III Congreso Internacional Multidisciplinario de Investigación. Facultad de Comercio y Administración de Tampico Universidad Autónoma de Tamaulipas. Tampico, México. Noviembre 2006. ISBN: 968-9031-15-5.

[16] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornés, and Edgardo M. Felipe-Riverón. "MECHDAV: propuesta de un modelo sistematizado de evaluación técnica de la calidad del uso de las herramientas RAD para ambientes visuales". CSIC 2007. e-Revist@s Plataforma Open Access de Revistas Científicas y Electrónicas Españolas y Latinoamericanas. Centro de Información y Documentación Científica, 2007 www.fecyt..es, www.erevista.es. [January, 2008]. Retrieved on June 2012

[17] Inteligencia Tecnológica en Software S. de R. L. Mi., "Introducción a los Proyectos de Inversión". Intecplan® v1.0. 2004-2008. www.intecplan.com.mx. [March, 2009], Retrieved on June 2012

[18] Software para Evaluación de Proyectos de Inversión Productivos. Copyright 2000-2006 - Todos los derechos reservados Registro de la Propiedad Intelectual Nº 506866. 2000-2006, evalas@elsitioagricola.com. [December, 2009]

[19] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornes, and Edgardo M. Felipe-Riverón, "MECRAD: Model and Tool for the Technical Quality Evaluation of Software Products in Visual Environment". ICCGI-5.2 4th International Conference on Wireless and Mobile Communications (ICWMC 2008) and 3rd International Multi-Conference on Computing in the Global Information Technology (ICCGI 2008). July 2008. Product Number E3275. BMS Part Number CFP0840B-CDR. ISBN 978-0-7695-3275-2. Library of Congress Number 2008926137 pp. 107-112. IEEE Computer Society. IARIA. Athens, Greece, 2008.

[20] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornés, and Edgardo M. Felipe-Riverón, "Metodología y Software para la Evaluación Técnica de la Calidad de los Proyectos participantes en Concursos de Creatividad mediante un Plan de Métricas Externas." Congreso Iberoamericano ANDESCON2008 IEEE sección Andina. Octubre de 2008. CP 130. Cuzco, Perú. www.andescon.org/art_compu.htm.

[21] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornés, and Edgardo M. Felipe-Riverón, "PROYEVA: Sistema para evaluar técnicamente la calidad de proyectos y productos participantes en concursos de innovación e invención mediante un programa de métricas externas y de calidad en uso." VI Congreso Internacional de Innovación para la Competitividad. Consejo de Ciencia y Tecnología del Estado de Guanajuato. Universidad Iberoamericana SINNCO2011. León, Guanajuato. México. ISBN 978-607-8164-00-4.

[22] Laura S. Vargas-Pérez, Agustín F. Gutiérrez-Tornés, Edgardo M. Felipe-Riverón, and Jorge Peralta-Escobar. "Metodología Cuantitativa y Análisis Comparativo de la Calidad de Proyectos en Concursos Académicos." Memorias del XVI Congreso Internacional de Informática Educativa, TISE "Nuevas Ideas en Informática Educativa. Volumen 7, Universidad de Chile, Facultad de Ciencias Físicas y Matemáticas, Departamento de Ciencias de la Computación Edición Digital 2011. Santiago de Chile. Noviembre 2011. ISBN 978-956-345-770-5.

[23] Metodología y Modelo de Evaluación Técnica de la Calidad de Proyectos participantes en concursos de creatividad mediante un Plan de métricas: PROYEVA. Registrado el 20 de marzo de 2007 con número de registro público Copyright SEP INDAUTOR 03-2007-03201059300-01.

[24] Software para Modelo de Evaluación Técnica de la Calidad de Proyectos participantes en Concursos de Creatividad mediante un Plan de métricas: PROYEVA. Registrado el 26 de septiembre de 2007 con número de registro público Copyright SEP INDAUTOR 03-2007-091813015000-01.

[25] Rediseño y adaptación de la metodología y modelo del sistema PROYEVA PLUS. Registrado el 6 de junio de 2011 con número de registro público Copyright SEP INDAUTOR 03-2011-060611393500-01.

[26] Sistema de Evaluación Técnica de la Calidad de Proyectos participantes en concursos de innovación, invención y creatividad mediante y un plan de métricas. Registrado el 18 de marzo de 2011 con número de registro público Copyright SEP INDAUTOR 03-2011-030711424100-01.

[27] PROJEVA: Diseño y Modelado de Interfaces para la implementación del sistema PROJEVA. Registrado el 05 de junio de 2012 con número de registro público Copyright SEP INDAUTOR 03-2012-060511055000-01.

Community-Commerce Brokering Arena for Opportunistic Cloud Services Offerings

Ethan Hadar
CA Technologies, Inc.
Herzelia, Israel
ethan.hadar@ca.com

Steven Greenspan
CA Technologies, Inc.
Ewing, New Jersey, USA
steven.greenspan@ca.com

Abstract—**Community clouds are formed when a cloud infrastructure is created to support the needs of multiple, independent service consumers that have shared concerns. Such infrastructures may allow member organizations to share resources on an opportunistic basis, i.e., one entity may provide services to the other members during off-peak times, and may in turn consume services from other members. Thus each member provides permanent or transient services according to their competencies and excess. This paper sets forth a conceptual architecture, examines the opportunities offered by such systems, and initiates a conversation on the business implications. Community-commerce brokering expands the choices for business from "where can I buy/use a service" to "whom do I want to sell my excess services to".**

Keywords - Cloud computing; community clouds; clearinghouse services; brokering services, community-commerce; service-oriented business ecosystems.

I. INTRODUCTION

Cloud Computing (Specifically, Infrastructure-as-a-Service and Software-as-a-Service) offers valuable delivery methods that allow large enterprises to provide specialized and non-specialized services to others. Specialized services might be: mortgages, human resources, healthcare and insurance, etc. Non-Specialized services might be – utility computing (storage and computing power), or commodity services such as e-mail. In addition, many enterprises have engineered their internal networks to accommodate peaks in service consumption for particular Information Technology (IT) services. Some of these companies have considered offering core competencies to other enterprises, for a fee, but they may also wish to provide, for a fee, underutilized storage and processing assets. Moreover, some group of independent companies may wish to provide services that work in combination with one another, each company providing a service appropriate to their competencies, establishing a Community-Commerce brokering system or "c-commerce" [3][6].

However, these companies may not want to offer services to competitors or the general public, for competitive and security reasons. Instead, they may wish to form consortia in which resources are shared on an opportunistic basis, within a community cloud framework. Such consortia have been called Service-Oriented Business Ecosystems (SOBE) [4][7]. For example, companies increasingly engage in open-innovation networks that require secure sharing of data and computing resources. Other companies may require sophisticated supply chain integration or the sharing data and domain-specific, proprietary software under highly specific policy and regulatory regimes [7]. Companies in a region that has high bandwidth within then region might form a SOBE because they cannot depend on global connectivity. Or, noncompeting companies may wish to pool resources so that the net cost of additional resources (as might happen in cloud bursting) is very low. Tai et al. [7] illustrate the value of SOBEs using the pharmaceutical discovery process.

This position paper discusses SOBEs that have agreed to share IT resources with appropriate constraints and Service Level Agreements (SLAs). However, this type of inter-enterprise or inter-agency cooperation has problems as well as benefits. The key benefits over and above the public cloud are that it leverages current assets and continues many of the roles and responsibilities of the current enterprise IT organization. Within a SOBE, trust among providers and consumers can be expected to be higher than in the public marketplace of cloud services, however trust must still be sustained though appropriate incentives and heuristics such as those found in [7]. Many additional problems present themselves in an ecosystem of interlocking internal clouds. Although collaborating companies want to increase IT utilization when they are not using these resources, yet they also want to ensure that they have the resources they need for critical business processes and for their peak usage. Many of the concepts developed for Peer-to-Peer (P2P) Grids are applicable here [1], except that they reformulated within a SOBE cloud with strong privacy guarantees, membership services, etc. Unlike [2], the ecosystem that we explore requires a brokering system; the incentives are not necessarily maximizing profits by providing more services within the SOBE to other SOBE members, but may also include the equitable sharing of resources to maximize profits from customers outside of the commerce-community. Hence, commerce-community members are exchanging services with one another, but are also selling services outside of the commerce-community.

Our position paper presents a conceptual brokering approach to c-commerce, where the system enables and facilitates new ad-hoc possibilities for providing and consuming IT services. Participating in the c-commerce arena, service providers may offer services opportunistically, as they become available. These services are not part of their core business, but shared among businesses within the c-commerce consortia on an opportunistic basis. For example,

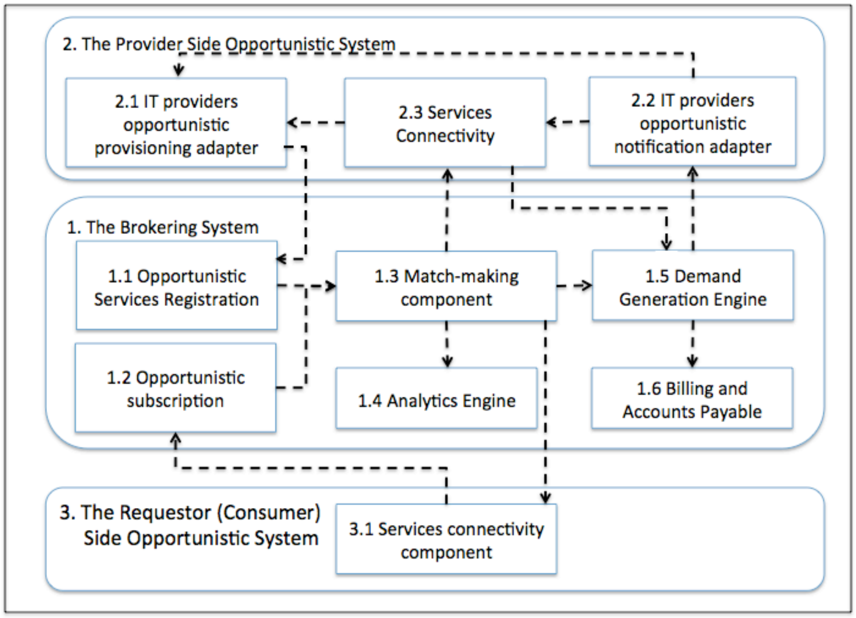

2. The Provider Side Opportunistic System
2.1 IT providers opportunistic provisioning adapter
2.3 Services Connectivity
2.2 IT providers opportunistic notification adapter
1. The Brokering System
1.1 Opportunistic Services Registration
1.3 Match-making component
1.5 Demand Generation Engine
1.2 Opportunistic subscription
1.4 Analytics Engine
1.6 Billing and Accounts Payable
3. The Requestor (Consumer) Side Opportunistic System
3.1 Services connectivity component

Figure 1: Conceptual architecture of a community-commerce brokering system

one company within a commerce-community might provide a payroll service they have developed, but consume the risk assessment service of another commerce-community member. To motivate future research in this area, this paper first describes a conceptual model for supporting c-commerce brokering, then discusses a prototypical usage pattern. This is followed by an examination of the advantages of the proposed system, and concludes with a discussion of the community-commerce concept and future research directions.

II. CONCEPTUAL ARCHITECTURE OF THE SYSTEM

The conceptual brokering system displayed in Figure 1, conducts matchmaking between potential service providers and requestors, and facilitates the negotiations between the sides. The central system has several conceptual components:

An **Opportunistic Services Registration componen**t enables IT service providers to offer active or dormant services, as well as acknowledge the provisioning of a specific service to a specific consumer. The registration component can block consumers that do not act according to agreed-upon commitments, e.g., not paying fees on time.

An **Opportunistic Subscription Component** allows consumers to view existing active services and inactive ones, as well as their underlying specifications and details. The component enables the requestors to register for the services, and even register for several alternative services according to preference and a set of subscription rules, such as time of day, signaling bandwidth, distance, etc. The subscription component can block providers who routinely fail to meet SLAs and other agreed upon commitments.

A **Match-making Component** balances the offered services and requests (requirements), and notifies providers about potential consumers (while collecting match-making fee).

The **Analytics Engine** generates historical demand and offerings lists, and identifies, according to classifications, trends in potential services that the broker is involved with.

The **Demand Generation Engine** is used in case the match-making component cannot accommodate a match according to the consumer's criteria. This component notifies the providers of a need (new requirements that may be addressed with other services that they offer)

The **Billing and Account Payable Component** bills both the providers and consumers, with the overall charges or revenue, in a monthly account. The costs of the services are collected by the broker, and transferred on a periodic basis from overall consumers to overall providers. If the provider is a consumer of other services as well, the broker will pay or collect the net difference. The payment schedule is determined by the contractual agreements between the broker and the subscribers (net-consumers or net-providers). Notably, a single broker may handle payment arrangements for many different consortia.

In addition to the main brokering system, there are interaction modules for service providers and requestors:

One interaction module is the **Provider Side Opportunistic** system that contains the **IT provider's Opportunistic Provisioning Adapter**. This adapter enables the IT provider to publish the potential services offerings (what), as well as their provisioning date (when) to potential consumers. When a potential service is not publically available (even for a limited time), it is denoted as dormant (or inactive) service and can be published without any availability options. When a dormant service is allocated for a consumer, the provider will send a token that will indicate the identity of the requested service, its duration, its availability time, and the consumer's identity.

IT providers Opportunistic Notification Adapter enables the IT providers to be notified that a request may match a potential offering that is not "Active". When the requestor of a service acknowledges the offering of a dormant service (tailored just for this specific consumer), the approval and handshake of the negotiated service is send via this adapter as well.

A second interaction module is the **Requestor (Consumer)** Side **opportunistic** system. This system enables the requestor of services to subscribe to services registered in the brokering system. The requestors can define criteria for the services, and the component presents an acceptance approval for offered potential alternatives for final selections, if exists.

The **Services connectivity component** is identical for both service providers and consumers. It provides activation codes to service requestors that indicate billing, or provides activation codes to providers that indicate payments.

A. Prototypical usage pattern

The service provider offers an active service (description, availability time, and cost, etc.), or offers a dormant one, and publishes the offering to the brokering system. The brokering system registers the offering in the match-making component.

A requestor (potential consumer) browses the offered services in the *Match-Making* component (active or dormant), and subscribes to the services via the *Opportunistic subscription* component, including an order of prioritization (in case several similar services are available).

The brokering *Match-Making* system locates a potential existing (active) offering, assigns an activation code, and informs the service provider and requestor via their respective connectivity components. The activities of the *Match-Making* component are logged in the *Analytics* component for trends analysis.

In case no active service exists, the *Match-Making* component defers the request to the *Demand* engine that searches dormant services. When the *Demand* engine locates a potential offering, it connects to the provider's *IT providers opportunistic notification* adapter, and reports the request. If the provider can provision such a service, the *Demand Engine* gets a commitment from the *Services connectivity* component of the requestor, and immediately notifies the provider via *the IT providers opportunistic notification* adapter to provision the service with the specific token issued by the *Providers' IT providers opportunistic provisioning* adapter.

The Demand *Engine* delegates monitoring of the service once it is provisioned to the *Billing and Account Payable* component for further monitoring.

The *Billing* component reports to the *Analytics* engine for further analysis of trends.

III. ADVANTAGES OF THE PROPOSED SYSTEM

There are several advantages that are consequents of the conceptual brokering system, all parts of a cloud c-commerce arena.

Maximization **of capacity utilization** allows enterprise IT departments to offer excess of IT services on a transient base (limited duration). For example, such excess capability may be offered on a regular schedule (e.g., between 3 am and 6 am), or on an ad hoc basis.

Proactive revenue generation notifies potential enterprise IT departments of an external need for IT service. As a result, a service provider may balance internal capacity and job scheduling, optionally accommodating the requests for an external service.

Triggering negotiations and trade enables consumers to be engaged with potential IT service providers, request a service, and negotiate its availability date, timeframe, and associated costs.

Business offering expansion enables service providers to examine a potential service. The offering can be a primary service that varies in scale, used as an alternative to existing service, or deployed when testing a brand new offering.

Business opportunities detection enables a consumer of services to offer services as well, in an opportunistic way as well, whether or not they wrap services of other parties, composite services, or the consumer's original services.

Arbitrage transactions profit enables the brokering service to generate revenue based on advertisements of the providers, match-making fee, payment and billing processing fee, survey and analytics of trends of usage (requests), or trends of offerings (what is available). Because the providers of these services are also consumers of other services, the brokering service can create a single bill that takes into account all of the negative and positive charges.

Simplification of transactions complexity reduces the amount of funds and transactions transferred between providers or consumers, and facilitates trade-offs between providers and consumers, while reducing the brokering fee.

IV. DISCUSSION

This position paper presented a conceptual brokering arena, aimed at exploiting opportunistic service offerings within a Services-Oriented Business Ecosystem (SOBE). The presented paper triggers discussions around the applicability of such systems. Existing prototypical brokering and commerce arenas revolve around established service providers offerings, in which the broker acts as sophisticated IT service catalog. The notions presented in this paper are focusing on the interactions, registration, financial, and commerce potential aspects of such collaboration. In addition, the restriction based on competitive advantages is imperative, in which the provider can restrict several consumers, and deny service based on the end-users identify, all according to the provider internal business or compliance policy.

The game change is not "where can I buy/use a service", rather of "who do I want to sell my excess services to".

This approach may form business alliances, without financial transactions, aimed at temporary treaty between prototypical competitors, in order to block a third one.

Brokering may be extended to any type of intervention in the middle ground of C-commerce (service-commerce): permanent, transient, or opportunistic. Brokering and auditing may be operated through a third party or though a consortia of SOBE members. Notably the agreements would not be peer to peer, but between the SOBE member and the broker.

Such an approach for opportunistic brokering and marginal changes facilitates services demand and requests according to business changes, enabling consumers to become providers. Thus, the distinction between requestors and consumers is introduced, separating negotiating parties and contracted ones.

"Service leverage" is introduced, tilting the balance between market forces. The dynamic nature of the cloud is no longer the playground of the Managed Service Provider, but may also be influenced by Telco companies, large IT enterprises, and other large-scale commercial organizations.

Cloud brokering concepts are just emerging, and have largely been the domain of Managed Service Providers (MSPs). The present paper extends these brokering concepts to community cloud infrastructures. Future research directions may include expansions into financial combinations and chains of brokering arenas.

ACKNOWLEDGMENT

The authors thank Debra Danielson, Donald Ferguson, Peter Matthews, Serge Mankovskii, and Gabriel Silberman for many discussions on cloud brokering and infrastructure.

REFERENCES

[1] Buyya, R, Yeo, C.S., Venugopal, S., Broberg,J., and Brandic, I. "Cloud Computing and Emerging IT Platforms: Vision, Hype, and Reality for Delivering Computing as the 5th Utility", Future Generation Computer Systems, Vol. 25, Number 6, Elsevier Science, Amsterdam, The Netherlands, June 2009, pp. 599-616.

[2] Coêlho, A. and Brasileiro, F. V. "Smarter Heuristics for Business-Driven Services Selection in Multi-services P2P Grids". 2010 IEEE International Conference on Services Computing, Miami, Florida, July 2010, pp. 417-424 .

[3] Grivas, S.G., Uttam Kumar, T., and Wache, H. "Cloud Broker: Bringing Intelligence into the Cloud", 2010 IEEE 3rd International Conference on Cloud Computing, Miami, FL, USA, July 2010, pp. 544 – 545.

[4] Lia. S., Fana, Y., and Lib, X. "A trust-based approach to selection of business services." International Journal of Computer Integrated Manufacturing, Vol. 24, Issue 8, 2011, pp. 769-784.

[5] Nair, S.K., Porwal, S., Dimitrakos, T., Ferrer, A.J., Tordsson, J., Sharif, T., Sheridan, C., Rajarajan, M., and Khan, A.U. "Towards Secure Cloud Bursting, Brokerage and Aggregation", IEEE 8th European Conference on Web Services, Ayia Napa, Dec. 2010, pp. 189 – 196.

[6] Satsiou, A. and Tassiulas, L. "Trust-based exchange of services to motivate cooperation in P2P networks", Peer-to-Peer Networking and Applications, April 2010, pp. 1-24.

[7] Tai, S., Desai, N., and Mazzoleni, P. "Service communities: applications and middleware". In Wohlstadter, E. (Ed.): Proceedings of the 6th International Workshop on Software Engineering and Middleware, SEM 2006, Portland, Oregon, USA, November 2006, pp. 17– 22.

An Approach to Find Integration and Monitoring Points for Container Logistics Business Processes

Tugkan Tuglular, Dilek Avcı, Şevket Çetin
Izmir Institute of Technology
Izmir, Turkey
{tugkantuglular,dilekavci,sevketcetin}
@iyte.edu.tr

Gökhan Dağhan, Murat Özemre
Bimar
Izmir, Turkey
{gokhan.daghan,murat.ozemre}
@bimar.com.tr

Tolgahan Oysal
Yekare Consulting
Izmir, Turkey
tolgahan.oysal
@yekare.com

***Abstract*— An end-to-end intermodal container transportation usually requires business processes of various enterprises to work together in a fast, accurate, reliable and secure way. This can be achieved with integrated business services, which are monitored constantly for the above mentioned quality objectives. Although there is academic research currently available for the general topic, to the best knowledge of the authors of this paper, no environment is defined and implemented for container logistics business that provides an integration and monitoring framework for business processes of the sector under consideration. One of the steps in the specification and development of such a framework is to seek points for integration and monitoring. This paper discusses an approach to find integration points for container logistics business processes as well as monitoring those points of integration.**

Keywords-business process; integration; monitoring; container logistic.

I. Introduction

Although container logistics business is composed of well-defined services, such as loading/unloading containers from ships/trailers/trains, storing containers in depots, etc., which are handled by various companies, due to competition the very same goods traveling the same route may be stored in different depots or carried by different shipping companies. With intermodal transportation, the number of possibilities for carrying a container from one point to another increases dramatically as shown in Figure 1. With intermodal container transportation getting more support from the European Union, and therefore, introducing reduced costs, the pressure on enterprises to use different transportation modes, such as sea, land, and rail, is increasing constantly. Under these circumstances, the necessity for integrated business services is unavoidable for companies in logistics business and for customer satisfaction, which is usually defined with service level agreements (SLAs), monitoring of integrated business services becomes crucial.

Currently, integrations shown in Figure 1 are real and running among heterogeneous applications, which means that different processes in different business units are supported by different applications [1]. This indicates that different semantics for the data to be exchanged exist in the container logistics environment. For instance, every country has its own customs regulations. Even though they stem from trade union regulations, which introduce some standardization, it is almost impossible to find two countries having exactly the same regulations. In spite of this fact, in our case, all transportations are of container type and there are EDIFACT [2] message standards for containers, such as CODECO (container gate-in/gate-out report message) [3], COPARN (container announcement message) [4], COPRAR (container discharge/loading order message) [5], etc. That means some standardization exists at data integration level for some of the container logistics business processes. However, our observations show that a similar standardization is not valid for Enterprise Resource Planning (ERP) systems or financial systems that are part of container logistics business.

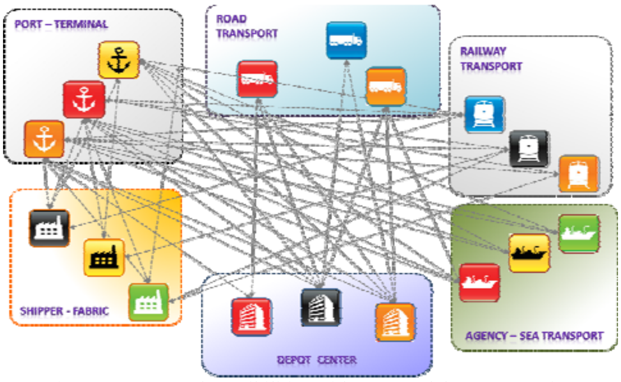


Figure 1. Increased Possibilities with Intermodal Transportation

With the realities explained above in mind, we propose a method based on "need to know" analysis and formal concept analysis supported by semantic similarity analysis along with ontology analysis to find integration points for container logistics business processes. After integration points are found and decision on integration implementation is made, we utilize audit, control and monitoring design patterns to achieve quality objectives set by SLAs. These two methods in action constitute the novelty of our approach. The next section explains our approach to find integration and monitoring points for container logistics business processes with a running example of port-agency integration of gate in-out, which coordinates container entry or exit to/from port.

II. Approach

In Bimar case, there are 180 integrations running on Microsoft's BizTalk server. These integrations are among different actors, such as COPARN integration between port and agency as well as between agency and depot. Business processes (BP) of these integrations are modeled using swimlane diagrams. These swimlane diagrams show the integration points for existing integrations, such as given in Figure 2, which shows integrations among Port, Agency, and Depot using EDIFACT message standards for containers. For instance, booking BP of agency (software) triggers and supplies necessary information in COPARN message format to booking BP of depot (software). This is an integration of agency-depot and booking is an integration point between agency and depot.

The “need to know” principle means that access to the information must be necessary for the conduct of one's official duties [6]. In other words, data or information should be entrusted to those who must have knowledge of it for its necessary usage. If you go backwards from usage to the need, then you may discover what is needed in order to start, continue or complete a business process. In case of the running example, port needs to know about each entering container, whether it is empty or not, whether it contains dangerous goods or not, whether it contains frozen goods or not. Depending on this information, which is supplied by the agency-port integration, port software will trigger corresponding business processes possibly through another integrations or web services.

The purpose of formal concept analysis (FCA) [7] is to support the user in analyzing and structuring a domain of interest. Such a method allows us to automatically obtain similarity scores without relying on human domain expertise [8]. Here, we plan to apply FCA to a domain knowledge, which includes inputs and outputs of each business process in that domain, to discover similarities of input and output, which will indicate an integration point. In case of the running example, imagine that containers with frozen goods are stacked in a special storage area at the port. The stacking BP and gate-entry BP as part of the port domain are represented in the port domain knowledge with help of corresponding ontologies. When FCA is applied to port domain knowledge, a high similarity score between “frozen goods” output in gate-entry BP and “frozen goods” input in stacking BP will be found an it may indicate a possible integration between to business processes. We call this operation as integration discovery.

We also plan to apply FCA to existing integrations to reduce number of integrations. Similarities of input among existing integrations will be searched. High similarity of inputs belonging to two or more integrations may mean that those integrations can be combined. In case of the running example, triggering stacking BP and inspection of frozen goods BP can be combined into a single integration with similar input. We call this operation as integration reduction.

It is assumed that an integration point is also a monitoring point using the anology of a local area network connected to Internet through a router (integration point) with a firewall on it (monitoring point). What to monitor, to what depth, and how frequently depends on the SLA defined for the integration. In case of the running example with agency-port integration, which informs the port about the entering trailers with containers, trailer stops for a while and this stop causes a trailer queue at the port entrance, which causes accumulation of empty containers at the port, which may also cause lack of space for unloading. Such an integration point should be monitored heavily, meaning that different views should be compiled from different sources and rules should be checking these views for anomalies.

The audit, control and monitoring design patterns (ACMDP) introduced by Trad and Trad [9] is based on flow-encoded switching design pattern and supported by Observer Design Pattern [10], Model-View-Controller Design Pattern [10], and the Decision Making Design Pattern [11]. ACMDP design patterns suggest factors and views. Factors determine characteristics and quality status of the phenomena under observation whereas a view is defined as a set of one or more factors. By using ACMDP approach and Microsoft’s Business Activity Monitoring (BAM) tool, we plan to add monitors and rules to the integrations, to observe integrations through BAM portal and to send e-mail and SMS notifications to the people on duty.

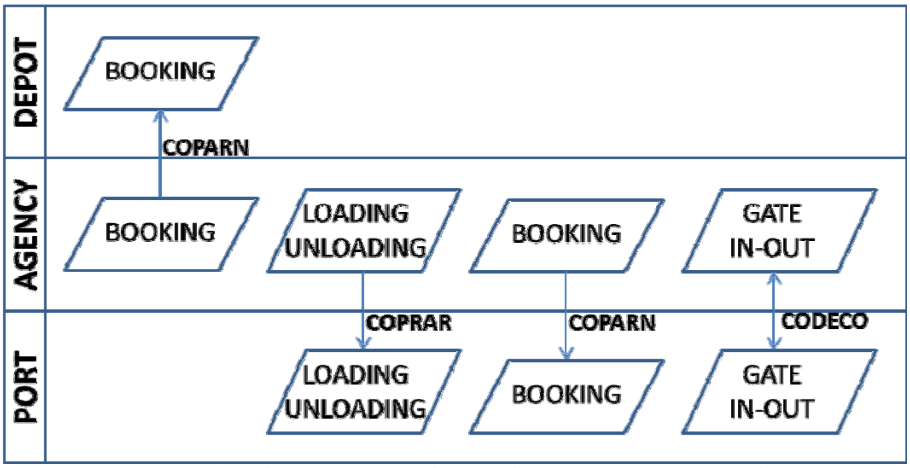


Figure 2. Example Swimlane Diagram with Integrations among Port, Agency, and Depot

There may be integration requests coming from customers, who know what is needed for the business goal and have a rough idea from which business process the necessary data can be taken. In this case, we plan to use the “need to know” analysis method proposed in this paper to validate the rough idea of the customer. In some other cases, the customer may not have an idea about what is necessary for the business goal and which business process can provide necessary data. For those cases, we plan to use “need to know” analysis and similarity analysis method as explained with a running example above. Those cases will show the benefit of the method.

III. PLANNED FUTURE WORK

At the moment, we apply this approach to the integrations developed and handled by Bimar to find integration points, and therefore, monitoring points. The next step will be finding similar integration points using FCA method, so that number of integrations can be reduced along with monitoring points. Furthermore, we plan to collect not only business data but also data exchanged among various levels of software, database, network, and operating systems, from monitoring points and analyze it to define largest possible set of monitors be used in the configuration of monitoring parameters as well as in the service level agreements.

Later, we plan to develop an integration and monitoring framework for container logistics business processes. Currently, all integration development and monitoring configuration are performed manually by drag and drops, connecting two points and writing expressions. The goal is to replace manual operation with code development, so that the process can be improved, for instance, by inclusion of unit testing. Since libraries, APIs, and frameworks are needed for code development, an integration and monitoring framework for container logistics business processes will be developed.

ACKNOWLEDGMENT

This work is supported by SAN-TEZ 00933-STZ.2011-1 Research Grant.

IV. CONCLUSION

A combined approach to find integration points and monitoring points for container logistics business processes is introduced. The proposed approach is composed of “need to know” analysis and formal concept analysis. After integration points are found and reduced respectively, the resulting set integration points are also accepted as monitoring points. Monitoring requirements will be determined by using the approach set forth by the audit, control and monitoring design patterns and agreed upon service level agreements. An integration and monitoring framework for container logistics business processes will be developed to enable coders to implement monitoring requirements efficiently.

REFERENCES

[1] T. Puschmann and R. Alt, “Enterprise Application Integration - The Case of the Robert Bosch Group”, Proceedings of the 34th Hawaii International Conference on System Sciences, 2001.

[2] UNECE, “The United Nations rules for Electronic Data Interchange for Administration, Commerce and Transport”, http://www.unece.org/cefact/edifact/welcome.html, last accessed on July 8th, 2012.

[3] UNECE, "UN/EDIFACT, Message Type: CODECO", http://www.unece.org/trade/untdid/d00a/trmd/codeco_c.htm, last accessed on July 8th, 2012.

[4] UNECE, "UN/EDIFACT, Message Type: COPARN", http://www.unece.org/trade/untdid/d00a/trmd/coparn_c.htm, last accessed on July 8th, 2012.

[5] UNECE, "UN/EDIFACT, Message Type: COPRAR", http://www.unece.org/trade/untdid/d00a/trmd/coprar_c.htm, last accessed on July 8th, 2012.

[6] Wikipedia, “Need to know”, http://en.wikipedia.org/wiki/Need_to_know, last accessed on March 21st, 2012.

[7] B. Ganter and R. Wille, “Formal Concept Analysis: Mathematical Foundations”, Springer, Berlin, 1999.

[8] A. Formica, “Concept similarity in Formal Concept Analysis: An information content approach”, Knowledge-Based Systems, 21(1), pp. 80–87, 2008.

[9] A. Trad and C. Trad, “Audit, Control and Monitoring Design Patterns (ACMDP) for Autonomous Robust Systems (ARS)”, International Journal of Advanced Robotic Systems, 2(1), pp. 25-38, 2005.

[10] E. Gamma, R. Helm, R. Johnson, and J. Vlissides. “Design Patterns: Elements of Reusable Object-Oriented Software”, Addison-Wesley, 1995.

[11] A. J. Ramirez and B. H.C. Cheng, “Design Patterns for Developing Dynamically Adaptive Systems”, Proceedings of SEAMS ’10, May 2-8, 2010, Cape Town, South Africa.

An API for Autonomous and Client-Side Service Substitution

Herman Mekontso Tchinda[*†], Julien Ponge[*], Yufang Dan[*‡], and Nicolas Stouls[*]
[*]*Université de Lyon, INRIA, INSA-Lyon, CITI, F-69621, France – Email: first.second@insa-lyon.fr*
[†]*UMMISCO, LIRIMA, Université de Yaoundé I, BP 812 Yaoundé Cameroun*
[‡]*Departement of Computer Science Chongqing University, Chongqing, China*

***Abstract*—The service oriented approach is a paradigm allowing the introduction of dynamicity in developments. If there are many advantages with this approach, there are also some new problems associated to service disappearance. The particular case of service substitution is often studied and many propositions exist. However, proposed solutions are mainly server-side and often in the context of web-services. In this paper, we propose a client side API-based approach to allow service substitution without any restart of the client and without any assumption on external services. Our proposition is based on a transactional approach, defined to automatically and dynamically substitute services, by preserving the current run and collected data.**

***Keywords*-OSGi; Stale References; Substitution; Self-Healing Software.**

I. Introduction

The service oriented approach is a paradigm introducing loose-coupling into software architectures. A developer can simply choose an Application Programming Interface(API) describing a requested service and develop its software without knowing which implementation will eventually be installed on the final client system. Currently, most popular uses of this approach are done by web services, Android systems and the OSGi framework [1]. Main studies are about Web services, but with the *server-side* point of view [2]. It means that the service provider can make any assumptions on provided services with the objective that a service substitution can be done without any consequence on the client, even if the service is state-full, State-full services are the one that maintain internal state across successive invocations from the same requester.

Dealing with dynamism issues of services in SOA is a real challenge today. Every model implementing this architecture faces the problem of deprecated references caused by the services mobility. The OSGi component framework is one of the several models implementing SOA and in which stale references can be very harmful. In this paper, we propose to study the dynamic substitution of a service in the context of the OSGi framework. In order to introduce the problem of service substitution in OSGi, we briefly describe in the following the OSGi component framework and the problem of stale references.

A. OSGi Component Framework and Stale References

The OSGi platform allows a remote loading and dynamic deployment of applications. We propose to study the client point of view. A service is a running java implementation, whose interface is available in an open repository. Using a reference of service instead of a service object itself. But, this reference also has a drawback: the referenced service can be stopped and its dependencies deprecated at the moment of its use, leading to a stale reference. We are focusing on the case of a mobile platform with OSGi that can discover or lose connection to some service providers. In such a case, a service requested by a client can be lost while in use.

B. Shielding From Stale References

As mentioned previously, bundles can be dynamically unloaded and a service may be stopped without a prior notification, leading to *stale references*. A stale reference is a reference to a service that is no longer available, either because of the bundle offering that service has been stopped or the service associated has been unregistered [1]. When a bundle becomes unavailable, all the references to objects it provided should be released to allow garbage collector to do its work correctly.

Writing safe code for handling OSGi service references boils down to properly listening to the OSGi service registry and tracking which services are in, and which services are out. This also requires that each call to a service in a client code makes extra steps to ensure that it is effectively going to invoke a method on a service whose reference is not stalled. This is not easy as it seems, as concurrency is involved. Indeed, a thread may be invoking a service while another one is unregistering it. This easily defeats guarded accesses to a service reference if no intrinsic locks or fine-grained reentrant read/write locks are being used. While solutions such as *BluePrint Services* help in handling service events, it does not shield from concurrency nor it enforces that references are properly discarded in client code when a service is unregistered [1].

Hence, we cannot make any hard hypothesis on services lifetime, but we can propose some good practices in client development in order to be resistant to the substitution of services. The substitution is well known as a self-healing software technique [3], [4].

C. Dealing with Dynamic Substitution in OSGi

The problem of dynamic substitution in OSGi is linked to the problem of stale references: assume that a client Bundle is using a service supplied by any Bundle server of the environment and at one point, the service disappears whereas the client Bundle has not finished with it. If there exists an available service to replace the disappeared one, a substitution can take place. But because of stale references, client bundles programmers may not be aware of the unregistering of the service and should not look at a new service. A good policy should then be implemented to deal with the problem of stale references.

So, in order to solve the problem of stale references, the main steps are: *(i)* to detect the service unloading, *(ii)* to choose a new service and *(iii)* to load the new service by preserving the internal state of the unloaded one. We do not want to focus on the service selection problem, since this step has been largely studied elsewhere in the literature, e.g., [2]. We will focus on the two other steps.

In this paper, we propose an API-based approach for the development of the client, inspired from the transactional approach of concurrent systems. We will consider the problem of detecting unload and the one of loading the new service. In our proposition we will first consider the easy case of using a single service, before introducing the substitution of a service while using a set of state-full services.

If a software is developed by using correctly this API, we guarantee that it can, according to its preferences: *(i)* be actively notified of the unload by a specific exception, or *(ii)* continue its execution with a new service that has been automatically substituted, even if the service is a state-full one, with a very light overhead of code to write. We also claim that the development cost is low in comparison with the development cost of a similar software with the same capabilities but developed without this API. Finally, we will show that our approach does not restrict the expressiveness of developed software, which means that every program using service can be rewritten to use the proposed API.

Section II cites some other works of the domain and shows the gap we will try to fill with this proposition. Section III describes the contribution of this article, about service substitution. In order to fix the global understanding of the reader, Section IV describes the tool developed to show the feasibility of the approach. Finally, Section V concludes this work.

II. Related Work

In this paper, we propose a solution to deal with dynamicity in OSGi. Our proposition is a *client-side* solution allowing state-full services substitution in OSGi. Client-side means that we do not make any assumption on services, potentially provided by some different providers. In [5], we have an example of a case in which the substitution process fails because of a mishandling of stale references. In this section, we present some related works on service substitution in general. The section is ended by a presentation of some existing approaches dealing with dynamicity and stale references in OSGi.

In [6], authors propose an algorithm for CORBA service reconfiguration, that involves a passive link to the unavailable service and an active link to the new service, while keeping the application consistency and with a few execution disruption. In the case of stateless services, it is straightforward. But for state-full services, it is more complex. One should restore the state of the substituted service. In SIROCO [2] framework, there is a registry system, where a service can register its current internal state and thus make a checkpoint. When a service fail, the framework try to manage the new service in order to set its internal state in the late one of the previous service. A synchronization mechanism has been presented in [7]. The configuration manager provides a run-time kernel which provides a message repository for messages that has been sent by components.

OSGi specification releases some advises to use ServiceFactory Interface or Indirection mechanism in service object implementation in order to limit, *"but not completely prevent"*, the consequences of stale references [1, Section 5.4: Stale References]. In [5], by using Aspect Oriented Programming techniques, the authors propose a tracking stale references tool named Service Coroner that helps to find stale references for developed or maintained OSGi applications, and apply it in two cases study. Others approaches such as using Service Binder [8] or IPOJO [9] suggest to separate functional and non-functional aspects, by describing the services dependencies management information in meta data XML files and merge both at run time. Each of these approaches tackles a particular case of the stale references problem, but a general solution is not yet provided. An alternative solution is the use of a proxy [10], instead of a service references. The proxy manages load/unload of services and the client services do not longer keep a reference to a likely disappeared service and the problem of stale reference is then avoided.

Almost all the aforementioned approaches are *server-side* and do not tackle state-full services. For state-full services substitution, one should implement a transaction mechanism to restore the state of the substituted service. Our approach is based on a proxy that make the substitution possible, but we manage state-full services by adding a transaction mechanism.

III. Contribution

We propose to add a “safe service use” layer into a service framework such as OSGi, in order to make softwares being more fault tolerant. This layer is an API that can be used by clients to be aware of services unload.

To describe what have to do this API we first introduce usual approaches of fault tolerant systems. Next, we describe our solution for the simple case where the client use a single service. Finally, we extend solution to take into account the case where several services are in use at same time.

A. Fault Tolerant System

Usually, fault tolerant systems are systems whose execution can continue to deliver correct service even if a fault occurs. In such a system, the first problem consists in identifying that a fault occurs. In our proposition, we define precisely what is a fault: the unload of a used service.

There are usually three families of treatment to recover an error [4]:

- to mask the error;
- to roll-forward in the execution until a new stable state is reached;
- to roll-back to the previous stable state and restart the execution from it.

Usually, to mask an error consists in having redundant information. Since we can not have it, we will focus on the two other treatment. We propose some mechanisms associated to the last two treatment families. In order to implement the roll-forward mechanism when a service disappears, we propose to throw an exception that explicitly advice the client that the service is no more available. Finally, to implement the roll-back mechanism when a service disappears, we propose an automatic substitution of the service by another one equivalent[1]. This substitution will be state-full service resistant.

In the following, we present these solutions in the context of a single service use and a multiple services use.

B. Safe OSGi Service Reference – Single Service

When a service is unloaded, its instance is kept in memory until the garbage collector dispose it, then while there is at least one reference to it. However, both the Java language and the Java virtual machine specifications do not support a notion of *"volatile / dynamic"* references [11], [12]. References to object instances cannot be changed "under the hood" unless explicitly re-assigned as part of a program control flow. This means that encoding a thread-safe and dynamic-aware behavior of service references need to be captured as part of a proxy indirection.

1) Proxy Indirection: A very common pattern for transparently mediating interactions between client code and a component in object-oriented languages is the introduction of a *proxy object*. They are most often used to enrich existing classes with cross-cutting concerns code such as logging, security or remote object exposition. A good example are the *Enterprise Java Beans*, where developers write simple Java classes, and EJB containers enrich them with support for security, transactions and other useful features. In our context, we will try to transparently add some enrichment without that services know that they can be substituted.

2) Proxy Requirements and Functionalities: The requirements for an OSGi service proxy depends on the usages. Hence, two kinds of policy and then requirements can be defined: Roll-forward policy and Roll-back policy.

In a *Roll-forward policy*, method invocations must throw an unchecked exception if the underlying service reference is staled. The client itself just need to take account the possibility of such exception.

In a *Roll-back policy*, when a method invocation reached a stale reference problem, we will try to transparently replace the unloaded service by another service, and then to make the invocation on the new service. However, if the unloaded service is state-full, the substitution can be the source of many unexpected problems. We then need to replay a part of the last commands. For instance, if the service need to be logged in, when the service is substituted, the login method has to be invoked again before any other use of the service. The part of the code that the API need to re-execute is called a *transaction*.

Transactional systems have been widely studied for several classes of problems and applications. The type of problem that we are tackling is actually close to a transactional memory [13]. However, the service and the client are developed by knowing that if a transaction fails, then it can be executed again. Our proposition is an adaptation of these existing results in the context of OSGi, where services are developed without knowing that such a substitution can occur. The client is the only one knowing this.

Since the API need to know precisely which part of the code has to be execute again in case of substitution, then the designer of the client must declare a part of code as the transaction. However, this code can be executed many times, since many substitutions can occur. Hence, this code has to be pure. It means that no side effect has to be done in the client by the transaction.

Finally, here are the sufficient requirements in the case of using a single service:

- **Awaited Behavior:** When a method is invoked, if the underlying service reference is staled, then the awaited behaviors are the following, for each policy:
  - Roll-forward policy: unchecked exception is thrown.
  - Roll-back policy:
    - If no other service: unchecked exception is thrown.
    - Else: substitution of the service, restarting the invocation from start of the transaction method.
- **Proxy Requirements:** It depends on policy:
  - Roll-forward policy: the client would consider the

[1] In OSGi, two services are equivalent if and only if they provide the same service interface.

possibility of an exception for each service call.

- Roll-back policy: the client must provide a pure method making the transaction.

C. Generalizing to the Invocation of Multiple Services

While a proxy is sufficient at the granularity level of a method invocation on a single OSGi service, generalizing the approach to the coherent execution of multiple services is more involving. Indeed, consider a block of instructions where several services are being used, and having a strong requirement for that block to be executed with a stable set of non-stale OSGi references. Given that, we cannot make any assumption on concurrency and the possibility for service references to become stale in the middle of a block execution. We need to provide a more powerful transactional-like framework to execute such blocks.

1) Requirements and Assumptions: Coping with the traditional definition of a transaction, we assume that a *transacted block* is a portion of code invoking a set of services, and that the whole block shall be successfully executed as a coherent whole. However, by opposition with the case of using a single service, we can generate side effects in used services. Hence, the transaction is pure only by the client point of view. Hence, in the context of multiple services, executing a transacted block requires:

- a declaration of the service interfaces it operates on,
- methods implementations to:
  1) put the block into a coherent initial state before its execution,
  2) execute the actual block code,
  3) finalize work upon successful execution,
  4) compensate possible side-effects in other services, if a stale reference caused a failure in the block,
- a retrial policy to control how the block execution is attempted again when a stale reference caused a failure.

The context of an OSGi platform imposes very loose assumptions on the transacted block implementations. Especially, services in use are not aware of being used in a transactional context, unlike Java EE resources that implement transactional APIs. Consequently, the correctness of performing a compensation operation or the ability to retry a block execution greatly depends on such services suitability in such a context, and their public specifications.

2) Invocation Atomicity – a Correctness Hypothesis in a Multi-Processed System: The OSGi specification states that OSGi service event listeners is notified when a service is unregistered [1]. A service reference becomes staled when all event listeners have been notified from the OSGi framework notification loop. Finally, we can take advantage of making a proxy to a service event listener in order to keep atomicity. Indeed, we can make a lock on the proxy object when performing a method invocation or when receiving a service unregistration event. This ensures a safe method execution as a reference cannot become staled in the middle of a method invocation.

3) Discussion: The generalization of the transacted execution of a set of services relies on strong assumptions:

1) services offer APIs to compensate effects in case an execution is aborted,
2) transacted execution blocks properly call compensation APIs,
3) intended compensation APIs are honored in service implementations,
4) services taking part in a transacted execution do not have further side-effects, or compensate them if client make a compensation.

In more traditional approaches, a transaction API is designed for resources to be managed by a transaction monitor. In the case of OSGi services, this would be translated to service interfaces extending such an API, making it impossible to use other types of services even if they offered compensation capabilities. We instead opted for a more open approach even if the usage of bad services or incorrect transacted block implementations can easily defeat the intended purpose.

IV. IMPLEMENTATION

The contributions presented in the previous section apply not just to OSGi environments. Indeed, any service-oriented architecture is based on the assumption that client code has no control over the services, including their availability and upgrades. We now detail how we implemented those contributions in OSGi in 2 steps. First we propose a simple service for building safe proxies to OSGi services, then we offer a service and an API for executing and defining transacted blocks. The interested reader can download the whole API and some examples at:

https://bitbucket.org/jponge/osgi-substitution

A. Configurable Service Proxy References

1) Overview: Proxies can be created at runtime in Java by creating a class that implements the *java.lang.reflect.InvocationHandler* interface and passing it to *java.lang.reflect.Proxy* for obtaining a proxy that is a subtype of one of more interface types. What we propose here is a very simple and minimalist API for generating proxies to OSGi services. It is exposed as an OSGi service of its own with the following interface:

```
public interface ServiceProxyBuilder<T>{
  public T getService(Class<T> c,
                      ServiceReference sr,
                      ProxyMode pm);
  public T getService(Class<T> c, ProxyMode pm);
  public T getFirstServiceMatching(Class<T> c,
                                   String filter,
                                   ProxyMode pm)
        throws InvalidSyntaxException;
  public ServiceBroker<T> getServices(Class<T> clazz,
                                      String filter)
        throws InvalidSyntaxException;
}
```

The interface mimics the OSGi service reference retrieval. *ProxyMode* parameters allow to specify whether a service reference becomes disabled after its backing service has been unregistered, or if another available service can be used in place. This allows to cater for both stateless and state-full types of services:

```
public enum ProxyMode {
  DISABLED_AFTER_UNREGISTERED, RELOAD_AFTER_UNREGISTERED
}
```

A *ServiceBroker* is used when dealing with several services for the same interface.

```
public interface ServiceBroker<T> {
  public Set<T> currentServices()
          throws InvalidSyntaxException;
  public void discard();
}
```

It is really close to the OSGi service trackers, except that it has the following semantics:

- *currentServices()* returns a set of service proxies currently matching the service interface and filter specification,
- returned service proxies have the *DISABLED_AFTER_UNREGISTERED* proxy mode,
- *discard()* is equivalent to the *close()* method of an OSGi service tracker.

2) Usage: The following code, extracted from the tests suite that we defined along with our implementation, shows an idiomatic usage of the service proxy builder OSGi service, in order to be substitution resistant.

```
ServiceReference ref = bundleContext.getServiceReference(
    ServiceProxyBuilder.class.getName());
serviceProxyBuilder =
    (ServiceProxyBuilder) bundleContext.getService(ref);
EchoService service = serviceProxyBuilder.getService(
    EchoService.class, RELOAD_AFTER_UNREGISTERED);

for (int i=1 ; i<= 10000 ; i++) {
    assertThat(service.echo("plop"), is("plop"));
}
```

B. Transaction Block API and Execution Service

1) Overview: We propose an OSGi service to execute transacted blocks whose interface is as follows:

```
public interface TransactedServiceExecutor {
  public T executeInTransaction(
          TransactedExecution<T> execution,
          RetryPolicy retryPolicy)
        throws TransactedExecutionFailed;
}
```

A transacted execution is specified through the following interface:

```
public interface TransactedExecution<T> {
  public void prepare();
  public T execute();
  public void finish();
  public void rollback();
}
```

It uses a parametric type T which is the expected return value type of a transacted block successful execution. The retry policy is a simple interface which is notified of potential stale reference errors, and can in turn decide whether a further attempt can be performed. It can also be used to implement delays between retrials. An example would be an exponential back-off delay over at most 10 executions.

The interface is defined as follows:

```
public interface RetryPolicy {
  public void notifyOf(Throwable throwable);
  public boolean shouldContinue();
}
```

By the way, a possible “retry forever” policy can be implemented as follows:

```
public class RetryForeverPolicy implements RetryPolicy {
  @Override public void notifyOf(Throwable throwable) { }
  @Override public boolean shouldContinue() {return true;}
}
```

2) Usage: A definition of a transacted execution implements the *TransactedExecution* interface. Given some fictious service interfaces *SomeService* and *OtherService*, an implementation could be as follows:

```
private class SomeTransaction
    implements TransactedExecution<Void> {

  @ServiceInjection public SomeService someService;

  @ServiceInjection(type = OtherService.class,
                    proxyType = MULTIPLE)
  public Set<OtherService> otherReferences;

  @Override public void prepare() { }
  @Override public <Void> Void execute() {
    for (OtherService s : otherReferences) {
      s.doThis(someService.doThat());
    }
  }
  @Override public void finish() {
    someService.release();
  }
  @Override public void rollback() {
    someService.undoThat();
  }
}
```

A more complete example would take greater care in the *prepare()*, *rollback()* and *finish()* steps. Fields annotated with *@ServiceInjection* are injected with service proxies. The definition for this annotation is as follows:

```
@Retention(RUNTIME)
@Target(FIELD)
@Documented
public @interface ServiceInjection {
  Class<?> type() default ServiceInjection.class;
  String filter() default "";
  ProxyType proxyType() default SINGLE;
  ProxyMode proxyMode()
      default DISABLED_AFTER_UNREGISTERED;
  public static enum ProxyType {SINGLE, MULTIPLE}
}
```

It is used to configure how proxies shall be configured. Especially, they can have service reloading capabilities enabled, and they can support a single reference or a set of instances like it is the case for the *otherReferences* set in the previous example. An OSGi service filter can also be specified. The block can be passed to the transacted executor service, which is also an OSGi service:

```
ServiceReference reference =
    bundleContext.getServiceReference(
        TransactedServiceExecutor.class.getName());
TransactedServiceExecutor transactedServiceExecutor =
    (TransactedServiceExecutor)
        bundleContext.getService(reference);
transactedServiceExecutor.executeInTransaction(
      new SomeTransaction(), new RetryForeverPolicy());
```

We used an optimistic approach. Having service proxies being injected into transacted blocks, we could have taken advantage of them to perform a giant lock spanning for the transaction execution lifespan. Indeed, it is possible to block the thread notifying that a service is going to disappear, thus keeping the reference valid until all receivers have been notified. Such an approach would avoid the need for rollbacks at the greater cost of limiting parallelism and breaking the OSGi framework requirements that service event notification handlers shall not block [1].

V. Conclusion

In this paper, we proposed an approach and a tool[2] to make a service aware to the stale reference problem. If a software is developed by using correctly this API, we guarantee that it can, according to its preferences: (i) be actively noticed of the unload by a specific exception, or (ii) continue its execution with a new service automatically substituted, even if the service is a state-full one.

Main properties of this contribution are: (i) this solution is client side, (ii) it does not make any assumption on used services and (iii) it can be used even if used services are state-full services.

This contribution is based on the fact that the client designer knows how to use desired services. Hence, we do not try to compute which behaviors are authorized by a service. The client designer has just to make a normal use of the service and to propose a sequence to rollback in a stable state before to make another try with another service, in case of substitution.

However, if a rollback is done on the external service, a rollback would also be done on the client itself in order to hold in a consistent global state. Since such a development model is risky, we prefer to give the following guideline in the client development: *do not make any modification of the client state from its transactional part.*

We also claim that the development cost is low in comparison with the development cost of a similar software with the same capabilities but developed without our API. Indeed, as explained and illustrated in the OSGi core specification [1, Section 5.4: Stale References], to make a correct use of a service without stale reference can be a little bit tricky. Moreover, we do not introduce any restriction to the expressiveness of services. Indeed, in the worst case, we can include the whole program in one transaction. Hence, if a service is substituted, then the whole program is restarted and fully executed with the new service. Hence, this API do not add any restriction in the software development.

In future work, we will consider the use of a service call logger, such as the Logos tool [14], and of a specification of used services, in order to propose an autonomous solution. We would try to propose some heuristics to automatically compute a call sequence to roll-back services in order to restart an interrupted transaction.

References

[1] The OSGi Alliance, "OSGi Services Platform, Core Specification, Version 4.2," June 2009, http://www.osgi.org [retrieved: June, 2012].

[2] M. Fredj, N. Georgantas, V. Issarny, and A. Zarras, "Dynamic Service Substitution in Service-Oriented Architectures." IEEE Computer Society, 2008, pp. 101–104.

[3] D. Ghosh, R. Sharman, H. Raghavrao, and S. Upadhyaya, "Self-Healing Systems - Survey and Synthesis," *Decision Support Systems*, vol. 42, no. 4, pp. 2164–2185, 2007.

[4] R. de Lemos, P. A. de Castro Guerra, and C. M. F. Rubira, "A Fault-Tolerant Architectural Approach for Dependable Systems," *IEEE Software*, vol. 23, pp. 80–87, 2006.

[5] K. Gama and D. Donsez, "Service Coroner: A Diagnostic Tool for Locating OSGi Stale References," in *34th Euromicro Conference on Software Engineering and Advanced Applications, SEAA*. IEEE, 2008, pp. 108–115.

[6] C. Bidan, V. Issarny, T. Saridakis, and A. Zarras, "A Reconfiguration Service for CORBA," *International Conference of Configurable Distributed Systems*, 1998.

[7] I. Warren and I. Sommerville, "A model for Dynamic Configuration which Preserves Application Integrity," in *3rd ICCDS*. IEEE Computer Society, 1996.

[8] H. Cervantes and R. S. Hall, "Automating Service Dependency Management in a Service-Oriented Component Model," in *CBSE*, 2003.

[9] C. Escoffier, R. S. Hall, and P. Lalanda, "iPOJO: an Extensible Service-Oriented Component Framework," in *Services Computing, IEEE International Conference on*. IEEE Computer Society, 2007, pp. 474–481.

[10] H. Ahn, H. Oh, and J. Hong, "Towards Reliable OSGi Operating Framework and Applications," *Journal of Information Science and Engineering*, vol. 23, no. 5, p. 1379, 2007.

[11] J. Gosling, B. Joy, G. Steele, and G. Bracha, *Java(TM) Language Specification, The (3rd Edition) (Java (Addison-Wesley))*. Addison-Wesley Professional, 2005.

[12] "Sun Microsystems, Java(TM) Virtual Machine Specification, The (2nd Edition)," Paperback, Apr. 1999.

[13] M. Herlihy and J. E. B. Moss, "Transactional Memory: Architectural Support for Lock-Free Data Structures," *SIGARCH Comput. Archit. News*, vol. 21, pp. 289–300, 1993.

[14] S. Frénot, F. Le Mouël, J. Ponge, and G. Salagnac, "Various Extensions for the Ambient OSGi framework," in *Adamus Workshop in ICPS*, 2010.

[2] Freely downladable at https://bitbucket.org/jponge/osgi-substitution

A Monitoring Approach for Dynamic Service-Oriented Architecture Systems

Yufang Dan*‡, Nicolas Stouls*, Stéphane Frénot*, and Christian Colombo†
*Université de Lyon, INRIA, INSA-Lyon, CITI, F-69621, France – Email: first.second@insa-lyon.fr
†Department of Computer Science, University of Malta – Email: first.second@um.edu.mt
‡College of Computer Science Chongqing University, Chongqing, China

***Abstract*—In the context of Dynamic Service-oriented Architecture(SOA), where services may dynamically appear or disappear transparently to the user, classical monitoring approaches which inject monitors into services cannot be used. We argue that, since SOA services are loosely coupled, monitors must also be loosely coupled. In this paper, we describe an ongoing work proposing a monitoring approach dedicated to dynamic SOA systems. We defined two key properties of loosely coupled monitoring systems: *dynamicity resilience* and *comprehensiveness*. We propose a preliminary implementation targeted at the OSGi framework.**

***Keywords*-Monitoring; Dynamic SOA; OSGi; Larva.**

I. Introduction

Service oriented architectures (SOA) is one of the current approaches to develop well structured software. It is focused on loosely coupled client-server through interfaces. The client usually requests service access through a repository. Subsequently, the client is bound to the service and is allowed to invoke methods as long as the interface types match. Among SOA approaches, we will focus on dynamic SOA, such as OSGi [1], usually used in 24/7 systems, where system may not be restarted when a service appears or disappears. In dynamic SOA, each invocation (potentially with the same client) must be considered as a completely new context change since potentially new services may appear and others disappear. From a dynamic SOA point of view, binding a client to a service is a matter of interface matching, but, neither the client nor the service has a guarantee that the other part behaves as expected. For instance, each time a client makes a request to a server, a formally specified constraint can be checked to ensure whether the client is authorized to perform that call or not.

Existing runtime monitoring tools such as JavaMOP [2] or Larva [3] weave interception calls using aspect-oriented programming techniques. These approaches work fine in SOA since client-server bindings are usually generated upon the first invocation and preserved throughout the entire client life cycle. On the other hand, in dynamic SOA, bindings may be reconsidered at runtime. Hence, any monitoring state wove into the service implementation gets reset.

Fig. 1 illustrates an example needing a dynamicity resilient monitor. Let us consider a client embedded on a mobile device based on a dynamic SOA platform and needing to communicate with a distant system according to a particular protocol. If two services S_1 and S_2 provide a single dedicated interface for accessing this distant system, but through different medium (WiFi, 3G, etc.), then the system could need to substitute the use of S_1 by S_2 and conversely. We propose a monitoring approach allowing such substitution without impacting the communication and then the monitored property. Hence, if any behavioral property has to be respected in the use of the distant system (such as "after a call of A, a call of B must occur") then the property is exactly the same in the whole system without any modification, even if the intermediate service is substituted.

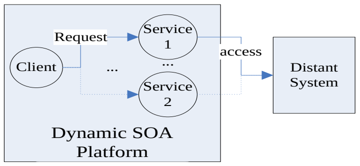


Figure 1. Example of Dynamic SOA System With Monitoring Restrictions

Our proposal is to bring a dynamic approach to runtime monitoring systems. We suggest that monitors must not be statically inserted into the observed system, but brought dynamically at binding time between the client and the service. Having a dynamic approach means that the service bindings and the behavioral monitoring bindings must be both considered as dynamic and loosely coupled. This article stresses the use for a dynamic runtime monitoring tool, that enables service substitution at runtime. This approach may preserve the current behavioral states and check that old and new service implementations are behaviorally compatible.

A dynamic runtime monitor must have two significant characteristics: *dynamicity resilience* and *comprehensiveness*. The former refers to the preservation of the behavior flow: in case of substitution of the monitored service, we want to keep alive the monitoring and the current state of the property. Hence, the property cannot be hard-linked to the code. The latter characteristic means that we cannot allow services to restrict what is observable by the monitor: if we want to check a property, we need to ensure that all the relevant events are monitored. We are not assuming that every service provides its authorized behaviors, but only that if an authorized service want to check the respect of a property on the framework, then no service can bypass

ISBN: 978-1-61208-215-8

this observation. The architecture relies on a generic event interception mechanism and a dynamic, loosely coupled, wiring mechanism for automaton verification. The verification automaton is extracted from Larva.

Section 2 of the article presents some runtime verification approaches and proposes a classification of them, showing the gap we propose to fill. Section 3 expresses the architecture model for a dynamic runtime verification tool. Section 4 illustrates our OSGi reference implementations. Section 5 shows our initial conclusions, and Section 6 our future works.

II. Related works

We can classify existing runtime verification approaches according to the monitor configuration with respect to the monitored service. Property may be: manually written inside the code (Hard-Coding), automatically injected inside the code (Soft-Coding) and kept out of the code (Agnostic-Coding). For each of these families, we will discuss the resilience to dynamicity and the monitoring comprehensiveness.

A. *Hard-coding*

In this category, where properties are manually injected at source time, we can cite all annotation techniques, like JML [4] or Spec# [5]. In both cases, the monitor is not *resilient to dynamic* code loading. If the monitored system is substituted, then its monitor is also substituted, since it is inlined code. However, this approach is interesting in terms of *comprehensiveness*, since we can observe anything in the program. A limitation of this approach is the dispersion of the monitor throughout the code, requiring significant intervention to write the property or to check its description is correct.

B. *Soft-Coding*

In this category, where properties are injected at compilation time, we can cite Larva [3] or JavaMOP [2]. These tools need a standalone description of a property and inject the synthesized monitor inside the code. Advantages are then the same as in the previous case, but specifying the monitor is easier, since the description of the property is centralized. However, this approach is still not *resilient to dynamicity*; at best, the tool may inject the property at first-time binding, but once injected, the property is hard-coded within the service.

C. *Agnostic-Coding*

In this last category, where the monitor is kept out of the code, we include any trace analyzes approach, such as intrusion detection systems [6] or liability by logs approaches [7]. The main advantage of the approach is the loose linking between the property and the monitored system. Hence, if a package is substituted, the monitor can observe it inside the logs and the monitored properties are still the same for the whole system. Moreover, the description of the property is located into a single location, which facilitates property management. However, such a system can be bypassed, since it can only observe what services accept to push or what it can pull from observation points. If a package provides a service without writing sufficient logs, then the monitor does not have sufficient information to check a particular property [8].

In this paper, we propose a *comprehensive* and *dynamicity resilient* monitoring approach for dynamic SOA. The kind of system we target, sits between the *Soft-Coded* and the *Agnostic-Coded* approaches.

III. An Architecture enabling Dynamic and Comprehensive Runtime Monitoring

In this section, we describe an abstract architecture of a monitoring system supporting specific features of dynamic SOA systems. After that, we will discuss about the *resilience to dynamicity* and the *comprehensiveness* of the proposed architecture.

Our proposition consists in dynamically inserting a monitoring proxy in front of each service, and externalizing monitors in some autonomous services. When an event occurs, a notification is sent to each monitor, which checks the event against its property.

Since services are treated as black boxes from the running environment's point of view, such architecture is designed to consider only interface properties. It corresponds to properties expressing the normal/authorized use of a service. We then address behavioral properties.

In this architecture, the scope of properties is not restricted to the use of a single service. Indeed, there is no restriction to add a monitor in front of several services, in order to observe a global property on the system.

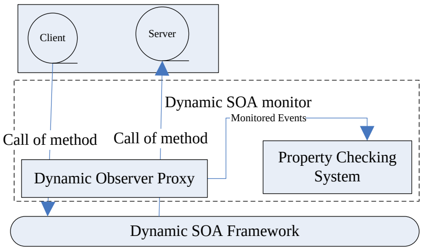


Figure 2. Monitor Abstract Architecture

Fig. 2 describes the whole abstract architecture, where we now detail the two main principles.

Resilience to Dynamicity: Since the monitoring system is not inlined inside the monitored service, but is externalized in an autonomous service, monitors are separated from the code. When changes occur in the framework, the observation mechanism and its properties may remain unaffected.

Comprehensive Monitoring: One of the main concepts of dynamic SOA is to have a framework which allows dynamic loading and unloading of loosely coupled services. Since the framework is in charge of providing an implementation to each service request, the framework adds a proxy between the client and the service to observe communications. This observation is comprehensive and no communication can bypass this proxy, since the client and the service do not know directly each other.

IV. OSGiLarva — A Monitoring Tool for OSGi

In this section, we present OSGiLarva, an implementation of the proposed abstract architecture, using the OSGi framework. Our implementation integrates two existing tools: Larva [3] and LogOS [9]. LogOS is a special logging tool based on the OSGi framework, developed at CITI Lab during the LISE project [7]. We will use it as a hooking mechanism to observe services' interactions. Larva is a compiler which generates and injects a verification monitor into Java code. We will use an adaptation of Larva to enable property verification. Fig. 3 describes the OSGiLarva implementation of the abstract architecture we proposed.

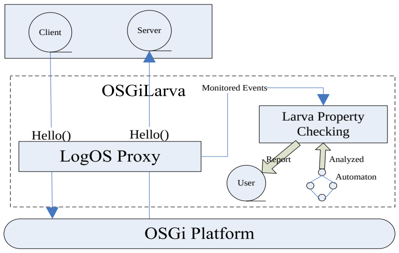


Figure 3. OSGiLarva Implementation

Currently, we use strictly the larva property description language. The addressed properties are the one that can be described by automaton.

We describe the monitor implementation with three key parts: we first present our adaptation of LogOS to intercept service interactions; Next, we give some details about our modifications of Larva and how it communicates with LogOS; Finally, we describe how the registration process of a service under OSGi will take into account an existing property monitor to insert it between the service consumer and the service itself.

A. LogOS: a Hook to Intercept Service Interactions

LogOS is a transparent toolkit for the OSGi architecture. As soon as the LogOS bundle is started, each registration of a service is observed by the system — thanks to the OSGi hooking mechanism — and a LogOS proxy is generated between the service and its consumer. Hence, every method call, including parameters and returned value, are automatically intercepted.

Each time such an event is captured, a corresponding LogOS event-description is forged and propagated to listeners. While the standard LogOS listener stores each event-description in a file, we have developed another listener that propagates events-description to the Larva monitor.

Finally, we have extended LogOS annotations, such that registered service interfaces may indicate at the deployment time their associated monitoring class. Interfaces and monitors can be provided separately from any implementation. In order to minimize the execution time of OSGiLarva, only registered events are notified to monitors.

B. Larva: a Property-Checking Monitor

Larva is a tool which injects monitoring code in a Java program to check a property described in a Larva script file. Upon compiling a script, the Larva Compiler generates two main outputs: (i) a Java class coding the property, and (ii) an aspect which links the monitoring code and source code. The aspect statically injects some calls to the monitor inside the Java software by using the AspectJ compiler. The Java code describing the property is then called each time an expected event occurs.

We keep the generation of the Java description of the property, which is the core of the monitor. However, we replace the static injection of the aspect into target services by an event-description propagation from LogOS to the monitor, based on dynamic proxy injection.

The new Larva provides a method accepting LogOS event-descriptions, which is dynamically called each time LogOS gets an event. Next, the Larva monitor starts analyzing this new event-description in the verified property.

C. Registration of a Service Providing Specification

In order to launch the monitoring of a service, we need to have a behavioral property to monitor. We propose to accept this property as a part of the OSGi bundle. An OSGi bundle provides three kinds of elements: a collection of *interfaces*, a collection of *services implementations* and a *bootstrap code* which is called when loading or unloading the bundle. Thanks to OSGi architecture, service interfaces, service implementations and bundles may have different lifecycles depending on the deployment scheme, since interfaces may be deployed with another bundle than service implementation.

As such, we keep the same philosophy, when providing properties that can be provided by the same bundle as implementation or by another one. Since interfaces are typing specifications of services, it makes sense to map the life cycle of properties to the one of interfaces. We then bind property monitor at the same time as interfaces and we keep monitoring until interfaces are removed.

In the current implementation of OSGiLarva, the property load is done by an explicit declaration referring to a, Larva-compiled, Java monitor in the manifest of the interfaces bundle. When this declaration is processed, LogOS is called. It loads the Larva property, generates a proxy between the new service and its consumer, and injects the given monitor inside the proxy.

V. Conclusion

In this paper, we have presented an approach to monitor dynamic SOA systems based on two main requirements: (i) *resilience to dynamicity* and (ii) *comprehensiveness*. The first one means that if a monitored service is substituted in the framework, the monitoring state is not reset. The comprehensiveness feature means that all services' interactions are monitored, i.e., it does not rely on the acceptance of the service designer or on the correct instrumentation of the service.

We have instantiated the approach in the context of the OSGi framework through a preliminary implementation, OSGiLarva, which integrates an adaptation of two existing tools: Larva and LogOS. Similar to Larva, OSGiLarva accepts the Larva property description language as input, hence inheriting all its features, including its expressiveness and its readability for non-expert users.

Our approach based on an OSGi hook to observe all occurring events seems to be inefficient when compared to injection-based monitoring tools, like Larva. However, this functionality is required to be resilient to the dynamicity of the system. Hence, we have to compare the OSGiLarva efficiency with that of other systems with the same features, such as classical log analyzes systems. In this context, our approach seems to have reasonable efficiency, since (i) we can configure filters in LogOS to push only the relevant events to Larva and (ii) we do not need to make hard drive accesses to read and write logs.

Finally, an interesting element of this approach is its non-intrusive aspect. Indeed, in contrast to the aspect oriented approach, we keep the original byte-code unchanged. This property can be interesting if we want to remove a monitor or always be able to check the binary signature of the code as an authentication credential [10].

VI. Future Work

At this point, we allow properties to take into consideration interface invocation events. In the future, we aim to introduce in the property description language the notion of loading/unloading of a bundle and the substitution of a service in order to express behaviors including such events.

Larva property description language allows timed properties, by the use of clocks. In the case of OSGiLarva, where a single service can be used by several consumers at the same time, we could want to introduce two levels of clocks: global to the service or associated to a consumer use. It seems that the foreach operator from Larva property description language could help us.

The current implementation of OSGiLarva is not finished according to our requirements. Indeed, the current declaration of events to log is done in the Java source file. In the future we aim to use the result of the Larva compilation to automatically define the list of events to observe. Finally, in a next version of the tool, we could make some propositions to reduce the OSGiLarva time cost. For instance, we could make OSGiLarva asynchronous, by exporting monitors in separated threads, or we can imagine that the use of a service need to be monitored only during a fixed time. If the property is respected during one week by a given consumer, we can consider that it will still respect it afterwards. In OSGiLarva, the removing a monitor is straightforward since it is non-intrusive.

References

[1] Open Service Gateway Initiative (OSGi), http://www.osgi.org/ [retrieved: June, 2012].

[2] P. O. Meredith, D. Jin, D. Griffith, F. Chen, and G. Roşu, “An Overview of the MOP Runtime Verification Framework,” *International Journal on Software Techniques for Technology Transfer*, 2011.

[3] C. Colombo, G. J. Pace, and G. Schneider, “Larva - safer monitoring of real-time java programs,” in *SEFM*, 2009.

[4] G. T. Leavens, K. R. M. Leino, E. Poll, C. Ruby, and B. Jacobs, “JML: notations and tools supporting detailed design in Java,” in *OOPSLA 2000 COMPANION*. ACM, 2000, pp. 105–106.

[5] M. Barnett, R. DeLine, M. Fähndrich, B. Jacobs, K. R. M. Leino, W. Schulte, and H. Venter, “The Spec# Programming System: Challenges and Directions,” in *VSTTE*, ser. LNCS, vol. 4171. Springer, 2005, pp. 144–152.

[6] C. Simache, M. Kaaniche, and A. Saidane, “Event log based dependability analysis of windows nt and 2k systems,” in *International Symposium on Dependable Computing*, 2002, pp. 311 – 315.

[7] D. Le Métayer, M. Maarek, E. Mazza, M.-L. Potet, S. Frénot, V. Viet Triem Tong, N. Craipeau, R. Hardouin, C. Alleaune, V.-L. Benabou, D. Beras, C. Bidan, G. Goessler, J. Le Clainche, L. Mé, and S. Steer, “Liability in Software Engineering Overview of the LISE Approach and Illustration on a Case Study,” in *ICSE'10*. ACM/IEEE, 2010, p. 135.

[8] H. R. M. Nezhad, R. Saint-Paul, F. Casati, and B. Benatallah, “Event correlation for process discovery from web service interaction logs,” *VLDB J.*, vol. 20, no. 3, pp. 417–444, 2011.

[9] S. Frénot and J. Ponge, “LogOS: an Automatic Logging Framework for Service-Oriented Architectures,” in *SEAAA*, 2012, p. to appear.

[10] P. England, “Practical Techniques for Operating System Attestation,” in *1st international conference on Trusted Computing and Trust in Information Technologies*, ser. Trust '08. Springer-Verlag, 2008, pp. 1–13.

Design of a Component-based Plant Factory Management Platform

Aekyung Moon·, Sangho Lee, Kyuhyung Kim

Embedded System Research Team, Electronics and Telecommunications Research Institute
161 Gajeong-dong, Yuseong-gu, Daejeon, 305-350, KOREA
{akmoon, comman35, jaykim}@etri.re.kr

Abstract—**In this study, we introduce convergence technologies of agriculture and IT that allow flexible development of greenhouse management services for different types of plant factory. There are various types of plant factory such as fully artificial light-type, sunlight-type and plastic house, etc. Currently, almost sunlight-type plant factory in Korea uses Priva as environment control system. However, it is expensive, but also difficult for user to change and integrate different sensors and actuators. In addition, most of the typical plant factory control systems have mostly been considered a specific type of plant factory. With this viewpoint, we propose a component-based methodology for providing various types of plant factory with flexible service development and deployment. A component-based platform developed by ETRI applies fully artificial light-type plant factory for monitoring and controlling microclimate of indoor environment. Finally, we develop a small plant factory prototype for a kind of fully artificial light-type plant factory in order to validate component-based service development methodology of plant factory. In the future, we implement sensor and actuators components for greenhouse of sunlight-type. We expect that the proposed service development methodology will be effective service deployment for various types of plant factory.**

Keywords-plant factory; greenhouse; environment control.

I. INTRODUCTION

The plant factory is a facility that aids the steady production of high-quality vegetables all year round by artificially controlling the cultivation environment (e.g., light, temperature, humidity, carbon dioxide concentration, and culture solution), allowing growers to plan production. There are various types of plant factory such as fully artificial light-type, sunlight-type and plastic house, etc. In the last several decades, greenhouse environment control systems for plant factory have been greatly developed. However, most of the typical systems have mostly been considered a specific plant factory types. Not only plant factory have many different types of sensors and actuators, but also environment control services are becoming more and more sophisticated considering correlation of microclimate information and autonomous control of actuator. Furthermore, each their services are often not reusable even in slightly different environment monitoring applications through sensors such as temperature and humidity.

In this paper, we design a component-based architecture of plant factory management platform. A component is a reusable and replaceable software module that enables complex functions to be developed easily. The main focus of component-based development is concerned with the assembly of pre-existing software components into larger pieces. Nevertheless, software reuse and component-based development are not yet state-of-the-art practice software development approaches in plant factory. We introduce a component-based methodology that can support to provide a variety of plant factory type with the flexible service deployment according to this platform. Component-based development is expected to shorten the development period, reduce maintenance costs, and improve program reusability and the interoperability of components.

The remainder of this paper is organized as follows. Section 2 presents the background of this research. Section 3 proposes system architecture and component-based service development methodology for plant factory. Section 4 describes case study of fully artificial light-type greenhouse. Finally, summary and concluding remarks are presented in Section 5.

II. BACKGROUND OF RESEARCH

There are various types of plant factory such as fully artificial light-type, sunlight-type and plastic house, etc. The fully artificial light-type plant factory also controls the closed indoor light environment using artificially light without the light of the sun as well as production cycle. The sunlight-type plant factory uses the light of the sun mainly, but supplements auxiliary lighting such as LED and fluorescent light.

The plant factory control system has a role to control the greenhouse environments; it gathers microclimate information around the crops from various sensors indoor parameters such as light, temperature, humidity, and CO_2 density as well as outdoor temperature, sun light, wind and rain. And it provides appropriate control functions. Figure 1 shows the example of sensors and actuators for plant factory. The most of the typical systems have mostly been considered a specific plant factory type. Not only plant factory have many different types of sensors and actuators, but also environment control services are becoming more and more sophisticated considering correlation of microclimate information and autonomous control of actuator.

TABLE I. EXAMPLE OF MAIN SESORS AND ACTUATORS AS PLANT FACTORY (GREENHOUSE) TYPES

Type	Example of Sensor/ Actuators Functions	
Fully artificial light Plant factory	Sensors	**Indoor**: Temperature, Humidity, Light, pH, EC, CO_2 density, Leaf Temperature, etc.
	Actuators	Cooling, Heating, Artificial light, CO_2 dosing, Water supply, Irrigation,

type		Circulation fan, Ventilation fan, etc.
Sunlight type	Sensors	**Indoor**: Temperature, Humidity, pH, EC, CO_2 density, Leaf Temperature, etc. **Outdoor**: Temperature, Sun light, Wind, Wind speed, Wind direction, Rain, Snow, etc.
	Actuators	Cooling, Heating, Supplemental lighting, CO_2 dosing, Water supply, Irrigation, Ventilation(Roof Vents, Side vents), Contains(Shading, Keeping warm), Misting, etc.
Plastic house	Sensors	**Indoor**: Temperature, Humidity **Outdoor**: Temperature, Sun light, Wind, Wind speed, Wind direction, Rain, Snow, etc.
	Actuators	Heating, Side Vents, Contains(Shading, Keeping warm), Irrigation**,** Supplemental lighting

Many research projects are tried to improve of existing plant factory (or greenhouse) systems. However, most of the typical systems have mostly been considered a specific plant factory types. Only exception is IntelliGrow [1] and BipsArch [2] by Aaslyng. IntelliGrow have functions as an addition to a generic Environmental Control Computer (ECC). The communication between PC and ECC was handled with a systems integration interface called BipsArch. But, this research has no consideration methodology for component and service development by composing.

III. COMPONET-BASED SERVICE DEVELOMENT

A. System Architecture

One of the benefits of growing crops in a plant factory is the ability to control all aspects of the production environment. This greenhouse management system has following considerations:

First, to optimize the environment for crop growth automatically, it can gather environment context information from various sensors and can provide appropriate control functions using various actuators. Second, it has limited easy installation and extension ability because most of the typical systems have mostly been developed based on the target specific plant factory. As a result, maintenance cost is increased. In particular, almost sunlight-type plant factory in Korea uses Priva [3] as environment management system. It is expensive, as well as difficult for user to change and integrate different sensors or actuators. Third, the microclimate control in a plant factory is a complicated procedure since the related variables are several and dependant on each other [4]. For example, the regulation of temperature can use heating, ventilation, cooling and water fogging. And ventilation control is calculated by wind speed and direction when weather station or the related sensors are in installed.

To meet above mentioned considerations, we design plant factory management platform that divided into three major systems, IMS (Integrated Management System for greenhouse), GOS (Greenhouse Operation System) and GC (Greenhouse Controller) as shown in Figure 1.

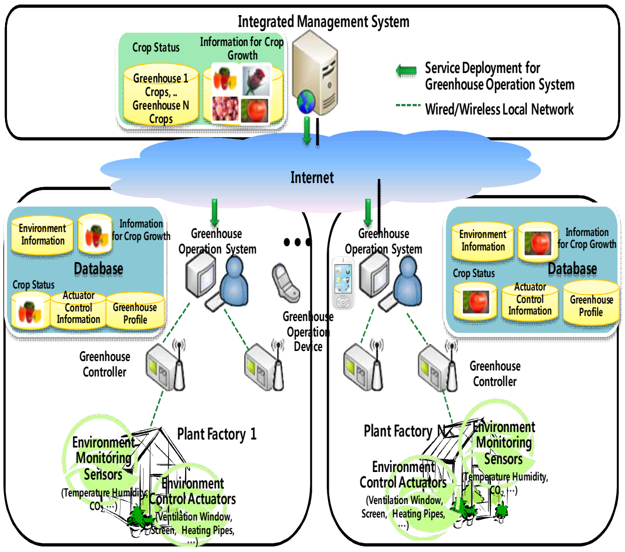

Integrated Management System
Crop Status
Information for Crop Growth
Greenhouse 1 Crops, .. Greenhouse N Crops
Service Deployment for Greenhouse Operation System
Wired/Wireless Local Network
Internet
Environment Information
Information for Crop Growth
Database
Crop Status
Actuator Control Information
Greenhouse Profile
Greenhouse Operation System
Greenhouse Operation Device
Greenhouse Controller
Plant Factory 1
Environment Monitoring Sensors
(Temperature Humidity, CO2,…)
Environment Control Actuators
(Ventilation Window, Screen, Heating Pipes, …)
Greenhouse Operation System
Environment Information
Information for Crop Growth
Crop Status
Database
Actuator Control Information
Greenhouse Profile
Greenhouse Controller
Plant Factory N
Environment Monitoring Sensors
(Temperature Humidity, CO2,…)
Environment Control Actuators
(Ventilation Window, Screen, Heating Pipes, …)

Figure 1. System Architecture

IMS deploys software components for them that sensors and actuators are installed in each actual plant factory. The communication of IMS and GOS uses TCP/IP protocol. IMS have databases to store crop growth environment information for a variety of crops. Database also stores crop status information received from GOS.

GOS includes the functions of internal and external environmental monitoring, life cycle management of sensor node and actuator node, and fault management and so on. For this purpose, GC gathers internal and external environmental contexts, sends control commands to the actuator. GOS installs services for agriculture and controlled horticulture received from IMS. In addition, it provides optimized environmental control services by feedback on the growth data of crops.

GC is consisted of embedded board with some control logics. GC supports the followings functions.

- Gathering contextual information from various sensors installed in plant factory environments. The communication of GC and sensors/actuators uses wireless protocols such as WiFi and Zigbee.
- Delivering microclimate information to GOS.
- Transferring control message from GOS to actuators
- Converting wireless protocol

B. Maintaining the Integrity of the Specifications

Plant factory management platform for a variety of operating systems has a role to install composed services according to sensors and actuators a installed on site. System provider / facility installation contractor consists of composed services using component registry of IMS as well as create new sensor / actuator components into component registry. Figure 2 shows the component-based service development methodology.

Components for new facilities (Sensor or Actuators) are required for the development and the produced components are registered to component repository by system developer or facility installation contractor. Obviously, component repository is helpful if we can utilize existing components when making a new component in that this reduces the development time and errors that might occur when creating the component from scratch.

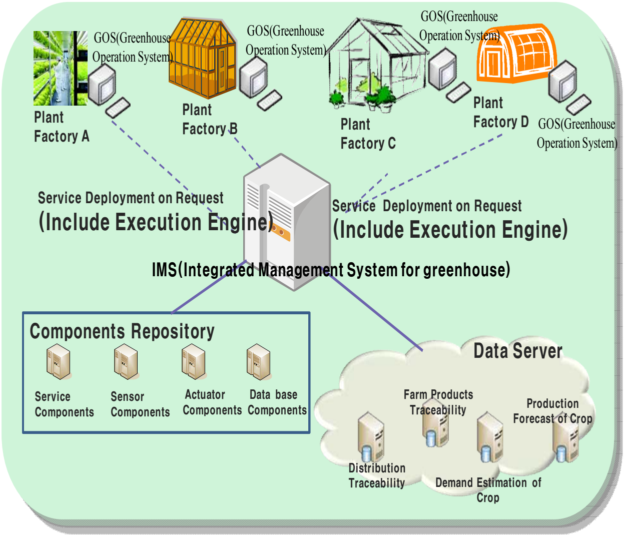

GOS(Greenhouse Operation System)
Plant Factory A
GOS(Greenhouse Operation System)
Plant Factory B
Plant Factory C
GOS(Greenhouse Operation System)
Plant Factory D
GOS(Greenhouse Operation System)
Service Deployment on Request
(Include Execution Engine)
Service Deployment on Request
(Include Execution Engine)
IMS(Integrated Management System for greenhouse)
Components Repository
Service Components
Sensor Components
Actuator Components
Data base Components
Data Server
Farm Products Traceability
Production Forecast of Crop
Distribution Traceability
Demand Estimation of Crop

Figure 2. Component-based Service Development

IV. Design of Components

We use OPRoS (open platform for robotic services) [5][6] in order to develop component. These tools run as plug-ins for the eclipse IDE style. OPRoS supports the full development lifecycle for robot software by providing a robot software component model, component execution engine, various middleware services, development tools, and a simulation environment. Not only do robots have different types of sensors and actuators, but also their services are becoming more and more sophisticated allowing them to run autonomously [7]. It is similar to plant factory. For this reason, we apply the OPRoS to develop the greenhouse management system.

A. Fully Artificial Light Type Plant Factorty

For verify the possibility of OPRoS to develop the greenhouse management system, we develop small prototype for a kind of fully artificial light-type plant factory. Fully artificially light plant factory for testing consists of two separate individual chambers. The specification of chamber is 2000 (W) x 1900 (D) (mm). Each chamber is installed temperature sensor, humidity sensor, light sensor, CO_2 sensor, pH sensor and EC sensor for collecting environment context. And for microclimate control of indoor environment, it is installed heater, cooler, ventilation fan, CO_2 generators, humidifiers and dehumidifiers. The light sources are LED and fluorescent light, respectively.

The entire system switches automatically according to the manual and automatic switch of control box. In the case of automatic, actuator can be controlled via the 485 communication protocol. The GOS is deployed software which is developed by component methodology and can control the two chambers; as well as a GUI made by component is installed to monitor microclimate of this chamber by several sensors.

B. OPRoS based Component Development

We use two tools provided by OPRoS that run as plug-ins for the eclipse IDE (Integrated Development Environment) style [5]. One is component authoring tool. The user needs to specify the port interfaces, callback functions, and a component profile when making an atomic component. The component authoring tool helps users to add implementations of callback functions and user-defined codes without any concern regarding various relationships between port interfaces and conformances defined in the component model, for example. The component authoring tool runs as a plug-in into the eclipse C/C++ development tools (CDT). It supports the GCC and Microsoft Visual C++ compilers. Figure 3 shows OPRoS based component that we designed. And we make components using component authoring tool.

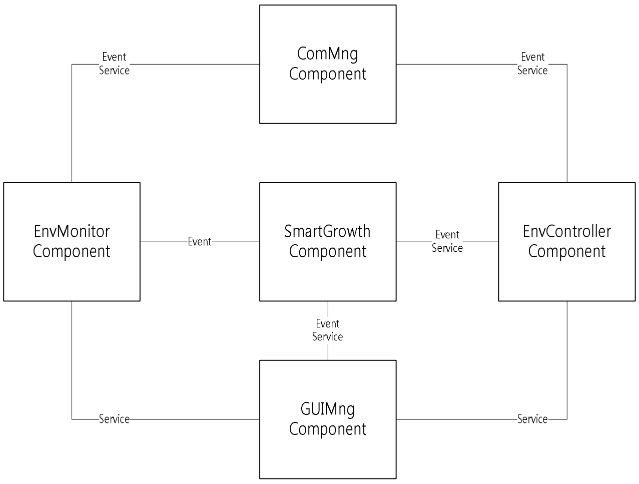


Figure 3. Design of OPRoS based Components

- SmartGrowth component refers to values of sensor node and controls actuators according to setting values for the regulation of the indoor environments in chambers. It is focus on creating the most appropriate microclimate for the maximization of crop growth.

- ComMng component has a role to communicate sensor node/actuate node with RS-485 protocol. We use wired communication protocol.
- EnvController component provides abstracted services of the actuator node to control a variety of environment.
- EnvMonitor component provides abstracted services of the sensor node to monitor a variety of environment.
- Finally, GUIMng component provides graphic user interfaces on monitoring services for indoor microclimate environment and outdoor environment as well as controlling services for regulation of indoor microclimate environment.

C. *Component Composer*

After developing components (SmartGrowth, ComMng, EnvController, EnvMonitor, GUIMng) for test plant factory using component authoring tool, the component composer is used building applications by composing components. It has a local repository to store components and imports component package for the component authoring tool [5]. A composite component can also be created by putting individual components into the composite component and connecting their ports to those of the composite component. The application developer drags and drops components onto the main diagram and connects components to build an application, as shown in Figure 4. Applications are composed with combinations of components according to message flow. Components communicate with each other via connections. A connection is established from a port of a sending component to a port of a receiving component. OPRoS provides inter-component communication for sending or receiving three types of information: method invocation, data, and events.

A service port allows other components to invoke its methods. It has an interface definition of a set of methods. A service port is either a provided or required type. A provided service port provides method services to other components. A data port is for exchanging data. It is either for input or output. An output data port sends data to input data ports of other components. Both the input and output should be of the same data type for a data exchange. An event port is for transmitting events. Although data ports and event ports are similar in that they transmit structured data, events are processed immediately. Figure 4 shows connection of components. The circle represents service ports and the rhombus represents event ports.

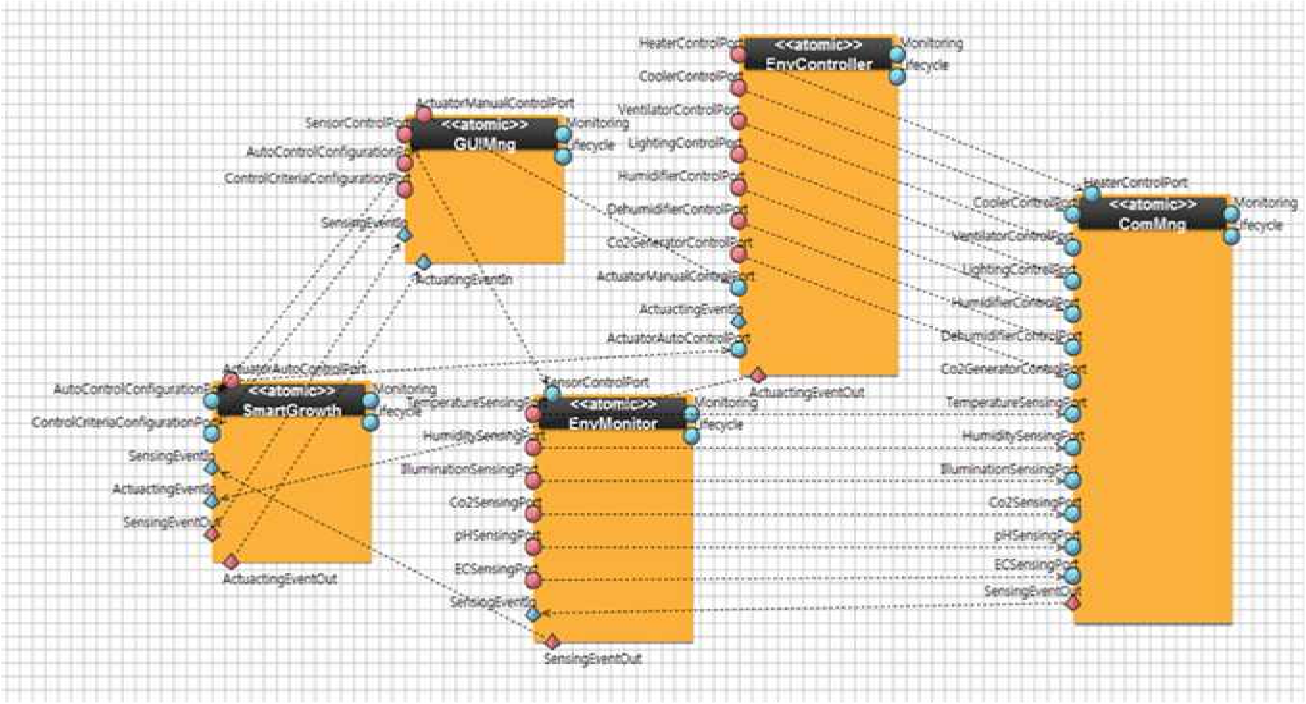


Figure 4. Componet Composer for Application

The composed application will deploy the target the operating system after build. Finally the application profile (.xml) and components (.dll) are packaged and deployed to component execution engine on target operating system of plant factory via a network.

D. *Experiment Results*

Developed applications using environmental monitoring and control components are installed in the fully artificially light plant factory previous mentioned.

EnvMonitor component gathers temperature, humidity, illumination, CO2 environmental information from sensor in test plant factory through ComMng communication component. And environmental information is stored in database. SmartGrowth component controls actuator based on information collected environment according to the service logic. Consequently, by adjusting temperature / humidity / photoperiod / CO_2, we grow lettuce (scientific namen : *Lactuca sativa).*

V. CONCLUSION REMARKS

We proposed a component-based methodology for providing various types of plant factory with flexible service development and deployment. And component-based platform developed by OPRoS applies fully controlled artificial light-type plant factory for monitoring and controlling of microclimate indoor environment. Finally, we develop small plant factory prototype for a kind of fully artificial light-type plant factory. Developed applications using environmental monitoring and control components are installed in the fully artificially light plant factory. It is shown that potential capability of OPRoS component-based software platform. In the future, we also implement sensor and actuators components for plant factory of sunlight-type. Eventually, we expect that the proposed service development methodology will be effective service deployment for various types of plant factory.

REFERENCES

[1] J. Aaslyng, J. Lund, N. Ehler, and E. Rosenqvist, “IntelliGrow: a greenhouse component-based climate control system,” *Environmental Modeling & Software*, 18, 2003.

[2] J. Aaslyng, N. Ehler, and L. Jakobsen, "Climate Control Software Integration with a Greenhouse Environmental Control Computer," *Environmental Modeling & Software*, 20, 2005.

[3] D. Kolokotsa, G. Saridakis, K. Dalamagkidis, S. Dolianitis, I. Kaliakatsos, “Development of an Intelligent Indoor Environment and Energy Management System for Greenhouses,*” Energy Conversion and Management*, pp. 155-168, 2010

[4] Priva, http://www.priva-international.com//en/, May, 2012

[5] C. Jang, S. Lee, S, Jung, B. Song, R. Kim, S. Kim, and C. Lee, “OPRoS: A New Component-Based Software Platform,” *ETRI Journal*, 32(5), 2010

[6] OPRoS, http://www.opros.or.kr, May, 2012

[7] C. Jang, S. Kim, M. Roh, and B. Seo, “Issues and Implementation of a URC Home Service Robot,” *16th IEEE Int. Conf. Robot Human Interactive Commun*., 2007, pp. 570-575

ISBN: 978-1-61208-215-8

Towards Context-Aware EHR-Based Healthcare Systems

Bogdan Niță, Dan Luca Şerbănaţi
Faculty of Engineering Taught in Foreign Languages
Politehnica University of Bucharest
Bucharest, Romania
nita_bogdan_07@yahoo.com, luca@serbanati.com

***Abstract*—In the medical world of today, healthcare professionals need to access the Electronic Health Record (EHR) of a patient not only on their desktop computer but also on mobile devices such as smart phones or tablets. Through the use of mobile electronic devices in everyday medical activities, the shortcomings of paper-based medical documents are limited removing the need of transcription or loss of documents and improving access to and search of patient's medical data. Mobility means awareness of a dynamic context whose richness can be used for infering new information and we believe that context-aware software approach is suitable for the medical world as healthcare professionals mobility is increasing. The proposed approach uses contextual information on the users of software applications involved in healthcare activities, patients and healthcare professionals to provide them with personalized, user-tailored views on the patients' EHR according to the scenario for which the application has been accessed. We have designed a system capable of perceiving context and reacting to ambiental changes to provide healthcare professionals and patients with EHR data dynamically generated according to the gathered information from the context. The user receives a customized view on the patient's EHR tailored upon her/his current context.**

Keywords-context aware computing; pervasive healthcare; mobile computing

I. Introduction

In the last years, attempts to implement national or regional healthcare systems supported by information systems that intensively use electronic devices have encountered many barriers: cultural, social and financial [1]. A major shift in healthcare will take place when care providers will realize that the critical element in work is the ability to exchange ideas, information and knowledge in a collaborative environment and that improved information *communication* benefits *caregivers and their patients*. From a social point of view, the main concerns are the privacy and the security of the stored data; it is well-known that a computer system is not 100% protected from data leaks or hackers' attacks. Regarding the financial issues, most of the professionals involved in the healthcare domain consider that investing in a system that would provide assistance in healing patients is not a must-have. This argument arises especially in countries or regions where capital resources are limited [1].

In the last years, the concept of mobility gained new dimensions by the multitude of devices emerging on market: personal device assistants (PDAs), smart-phones, tablet-PCs and notebooks. Nowadays, these devices are light, powerful and accessible and they provide mobility, ubiquitous access to information and enable easy interconnection of their users [9]. The growing market of mobile devices has encouraged the development of applications in medical fields. Several hospital-based prototypes have been proposed: "WardInHand" [6], "MobileWARD - Mobile Electronic Patient Record" [5], "Intelligent Hospital Software" or "Context-aware mobile communication". Each of these programs ended with questionnaires filled up by medical staff that interacted with it and the feedback is generally positive. Except "WardInHand", these prototypes make use of the idea of context-aware computing, a term first introduced by Schilit and Theimer and defined as "the ability of a mobile user's applications to discover and react to changes in the environment they are situated in" [8].

Studies [1, 7, 9] have shown that in last few years healthcare professionals have started using mobile devices as assistants in their daily work. However, ubiquitous access to EHR database does not improve the medical workflow. The guiding idea of this paper is that on the one hand, not all the information from a patient's EHR is needed at some point but only those relevant to the current situation, sometimes browsing a large amount of medical data and selecting the useful one may prove difficult and time-consuming. On the other hand context information can be used to better customize the presentation of the EHR's selected information to the final user. We explore the idea of endowment of an EHR system with the ability to perceive the contextual information of its users and to use it to sort out EHR data and consequently provide the users with it. We propose a system to react dynamically to situations for which it is accessed and provide its users with a customized view of the patient's EHR. A customized selection of EHR's data built upon contextual information will provide users with increased access and shorten the time needed to browse and search clinical data, increasing the efficiency of the medical staff. We also explore the idea of adapting EHR data to user's preferences and profile (e.g., font size, font color or background picture).

The paper is structured as follows: Section II depicts the contribution of the context-awareness paradigm in improving EHRs security and afterwards focuses on contextual attributes of healthcare domain's actors and emphasizes their particularities. Section III describes the overall structure of our model. In particular, we propose an architecture, make a thoroughly description of our approach and discuss its utility from a patient point of view. Section IV concludes the paper.

II. CONTEXT DATA IN HEALTHCARE ENVIRONMENTS

This section begins with a discussion regarding EHR systems security issues and improvements that may be brought by considering users' context. Afterwards, we outline how context-aware computing paradigm may add value to a conventional EHR system, and finally, we review various definitions of context-aware computing and present the contextual attributes of the actors involved in the healthcare process, healthcare professionals and patients, emphasizing their particularities.

A. Security and privacy

In the last few years, some of the reasons of reluctance in implementing EHR systems were the security and the privacy concerns regarding the data stored [11]. As many EHR systems are centralized (region- or country-wide) unauthorized access can have negative effects; once you are logged in to the system one can access and even modify private data of the patients causing possible permanent loss of data. Besides the integrity concerns, all the reads must be thoroughly authorized as the healthcare data distinguishes by its sensitivity. However, significant improvements were made towards security. For example, in the United States, Health Insurance Portability and Accountability Act (HIPAA) stipulates that patients can request a detailed list of all the accesses through their medical records [1]. One way to implement this facility is a user-centric auditing system, i.e., a portal where the user logs and sees a history of individuals who accessed their personal data.

Moreover, each patient having a network-accessible EHR can indicate which are the public information from her/his own record and which are private, each country implementing regional or countrywide EHR systems has imposed regulations regarding security and privacy of electronic stored data and backup systems have been proposed such that intentional or unintentional loss of data to be prevented.

As regarding security of EHR systems, studies have been made regarding healthcare domain actors' contextual attributes particularities and their influence in improving EHR's access control and security. Shetty and Loke proposed a dynamic context-aware access control model [12] and they suggested using subject, time and activity while replacing location with resource as location cannot stand on its own without resource or subject.

B. Mobility and context-awareness

Paper-based health records have shortcomings for both patients and healthcare providers, e.g., usually the health record of a patient is a single hard-copy kept by one medical institution rendering outside access to it difficult and cumbersome; manually search of data may be time-consuming; volatility of paper brings risks like partial loss of data or even irretrievable loss of data. Paper-based health records provide to healthcare providers an amount of data difficult to access, search or summarize and to the owner a risk of losing it.

Through the implementation of centralized EHR systems the mobility of medical documents has increased requiring to each healthcare institution only an Internet connection to be able to access the remote EHR. Studies [9, 11] have reported that implementation of EHR systems tend to reduce or even eliminate shortcomings of paper-based health records by increased security, ubiquitous access, reduced costs and the ability to predict and reduce the effect(s) of medical conditions. Implementation of different healthcare systems (e.g., WardInHand [6], MobileWard [5] or Clinical Trials Information System [7]) brought to the forefront of healthcare providers the mobiles devices, e.g., PDAs, tablets or smart phones moving the task of accessing, updating and changing medical data from desktop computers to mobile devices. This transition brought mobility to healthcare providers enabling access to EHR virtually anywhere, anytime being no longer constrained to use a desktop computer.

Powered with mobility enabled by mobile devices and the context-aware paradigm, the EHR systems may be seen from a new perspective, by adapting itself to the user's context in one specific situation, thus facilitating easy, quick and comprehensible access to situation-tailored EHR data. The transition of medical documents from desktop computers access to mobile devices involves moving from a relative static context to a dynamic one where not only the healthcare provider specialization involved is important but also the actors' location and the patients' circumstances, e.g., a routine or emergency situation or the device used for accessing data.

C. Particularities of contextual attributes in the healthcare domain

Interaction between the healthcare providers and an individual begins when the health state of the latter significantly worsens with respect to its regular state. From this point a team of healthcare providers (e.g., physicians, surgeons, nurses or radiologists) work together to cure the patient and to bring his health state to a normal level. During a medical treatment the team members provide each other with data according to their medical specialization and collaborate for improving the medical condition of the individual. Each member of the team has a specific role and given tasks which must be accomplished in a given period of time. The interaction with patients needs to be supported by accurate and detailed information regarding patients' past medical condition so the clinical history of the individuals should be structured and accessible in a meaningful way.

In [15], a context service is designed enabling the collection of contextual information which is further forwarded to interested clients. Henricksen, et al. [16] suggests that in pervasive computing applications the most relevant context information consist of the capabilities of the mobile devices, the characteristics of the network, user specific information and user preferences. Schilit, et al. [8] observed that contextual information goes beyond user's location because other things of interest are changing. Bettini, et al. describes a situation as being "a temporal state within context" [17].

Two of the most used context-aware computing definitions, that of Schilit, et al. [8] and that of Dey [18] place the space above other context categories. Individuals tend to associate a particular location with specific situations which they analyze and infer additional information upon [17]. The interaction of a healthcare provider with the EHR system begins when he logs in, at this point the system should gather the user's attributes (e.g., name, position inside medical institution or medical specialization), graphical interface preferences (e.g., font size or font type), location (e.g., inside hospital, inside ambulance or inside extrication car), the capabilities of her/his mobile device accessing the system (e.g., touch screen one, full keyboard one or size of the screen), etc. During a session, while some context information remain unchanged (e.g., user specific information or the capabilities of the mobile device accessing the system) others may evolve (e.g., user location or user preferences for information presentation).

Gathering patient-side context information is not a simple task to do as these data may be subject of uncertainty or/and inaccuracy of the source (e.g., human or devices). In certain circumstances like calling an ambulance or extrication service a second person reports about individual(s) requiring medical assistance. As in most cases the caller has little or no medical knowledge she/he may report vague information (e.g., the caller reports that an individual is bleeding but does not know the source of bleeding) or inaccurate data (e.g., an individual may have fainted but in fact he suffered a heart attack). Dey's definition of the context concept "any information that can be used to characterize the situation of an entity" [18] suggests that any information regarding the individual's needing medical assistance may be helpful to the system; of course, interpretation of possibly inaccurate data should be approached carefully. On the other hand, when the patient is hosted inside a healthcare organization medical observations are made by healthcare professionals, medical equipment is used for monitoring and all these data can be considered reliable, that is they accurately reflect the patient's health condition.

Our approach considers the context information emerging from a dynamic context analysis and uses it to compose a view on the patient's EHR that delivers to its users data tailored according to their current context, thus improving access and browsing of medical data.

III. Context-Aware Electronic Health Record

This section starts by presenting two scenarios and is followed by the description of our approach to an intelligent EHR system. We present how the system is able to react to the context inferred from the gathered contextual data of both the healthcare provider and patient and how it selects relevant data from patient's EHR and displays it according to the user's preferences. An architecture for the Context-Aware System is presented in Fig. 1. Throughout this section, short paragraphs expose current issues and based on these, we motivate choices to our approach. We conclude this section with a discussion on patients' benefit from using the system.

A. Scenarios

This section present two common medical scenarios and the main concerns of the roles playing these scenarios are identified. We then discuss how the Context-Aware EHR system can provide medical data to healthcare professionals and patients according to situation, place, time and available resources of both doctor(s) and patient(s).

Scenario 1

A paramedic working at SMURD Bucharest, an emergency rescue service established in Romania, during his shift is informed that he must take part in an extrication mission. On his way to the accident site he has 7 minutes to examine the EHR(s) of the victim(s) involved in the car accident and to gather personal information like age, previous diseases and chronic conditions of victim(s).

After providing first aid and after patient(s) stabilization the paramedic takes them to the hospital. Meanwhile, (s)he fulfills a mission report and enters it to the system providing all the details about the victim(s) and their health status to doctors from the emergency room.

Note: The 7 minutes interval is the average time of intervention of a SMURD ambulance in Bucharest [13].

Scenario 2

Andrew, a 16 years-old athlete is hospitalized for 2 days in Orthopedic Unit of the Hospital of St. John and St. Elizabeth from London. From the first medical tests doctors diagnosed the young boy with Osgood-Schlatter disease and before prescribing medications and sending him to a physical therapist doctors want to perform additional tests as his disease has relapsed. Meanwhile, during his stay in hospital, Andrew wants to know more on his disease and on possible causes of relapse.

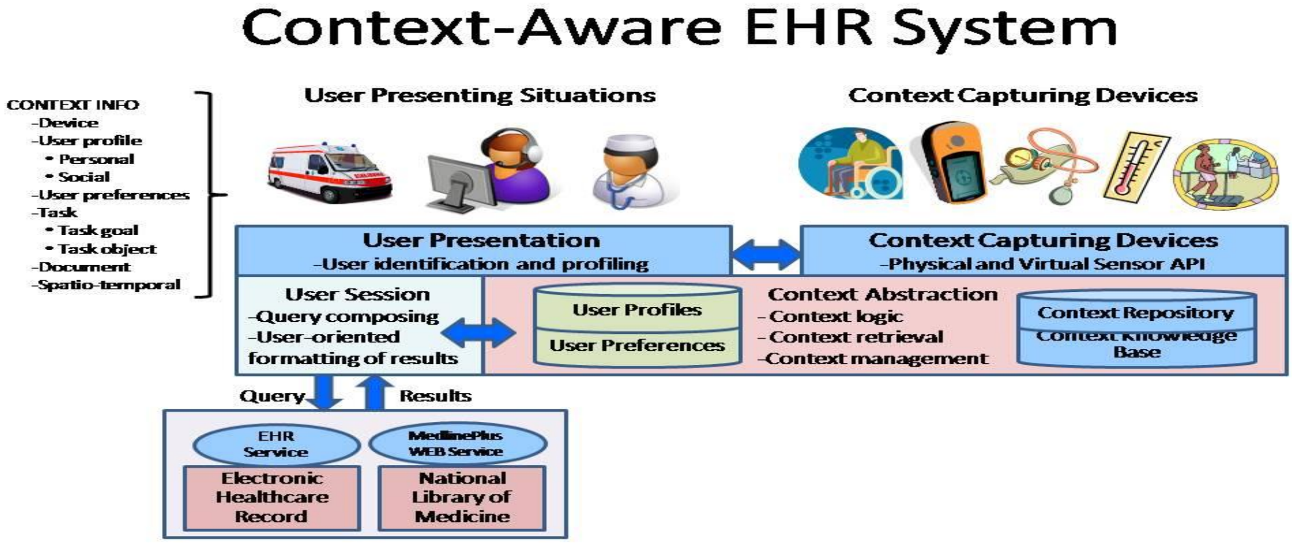


Figure 1: The architecture for the Context-Aware EHR System

In the first scenario, the time-critical situation requires an efficient management of both information and resources. On his way to the accident site the paramedic should be informed on the accident (e.g., number of victims, their health status, accident severity, weather conditions) and on the victim(s) (e.g., age, last known health status and previous diseases). When the paramedic accesses the EHR system (s)he should be provided with data according to both her/his and victim(s) context so that only the information needed for providing first aid (for example: age, last known health status, previous diseases) to be displayed upon his preferences which may include for example font type, font size, with/without images.

In the second scenario, while trying to inform about his disease, Andrew accesses his EHR and browses his past personal medical information. As his EHR stores a large amount of data, finding the desired information may prove laborious and also several medical terms specific to healthcare professionals may prove difficult to be understood. On one hand he needs to find out information on past symptoms and medical tests and on the other hand he needs to find out information on followed medical treatment and physical recovery.

When accessing the EHR, one way of making the EHR data comprehensible and easy to browse by the user is to perceive the user's context. For instance using Hyun-Yong Noh, et al. [19] indoor localization technique Andrew's location inside hospital can be determined with a high level of accuracy (Orthopedic Unit). Location along with information on device used for accessing the EHR (e.g., tablet PC or smartphone), the user's profile and preferences (e.g., font size or background picture) are used by the system for sorting out information and displaying it according to user's concerns. Moreover, for a non-professional, i.e., the user belongs to the patient profile, the system may provide more comprehensible synonyms of medical terms and web-site links related to current disease to provide information on treatments or physical recovery.

B. *Context discovering*

In order to react according to current needs, the system has to gather specific data about the current situation as well as data related to the main stakeholders: patients and healthcare professionals. As a registered user logs on to the system his profile and personal preferences are retrieved from the user database and used by the system to customize the user interface. This is particularly important as the user should feel comfortable with the way the data are displayed. According to her/his stored preferences, text will be formatted with the desired font size, font type and font color and the images will be properly sized.

As the user may access the system from a multitude of devices (e.g., desktop computer, tablet, smartphone or notebook) the next step is to gather the device capabilities (e.g., screen size, input mode or network connection speed). Sensing the device capabilities along with the available bandwidth avoids locking the connection until the transfer of a high resolution image (such as radiography) in which case it may be suitable to send a lower resolution image to the requesting device. Device's contextual information help improve the visual appearance of data displayed by smart scaling the images and the text to fit the screen size [14].

Once the user is authenticated, a session instance is created for her/him. As individuals have the right to expect their EHR data are kept confidential and they are the only to decide which medical information is public and which is private, our system comply with these regulations through inheritance of security, privacy and confidentiality levels regarding EHR data, owned by the EHR system. On each session instance an encrypted communication channel with the EHR database is established ensuring that the data cannot be intercepted by unauthorized persons.

During the user's session, the Context Service that runs in background continuously gathers contextual information from the context-associated sensors (e.g., GPS sensor, temperature sensor or light level sensor) and updates to the content of a Context Repository with it, in this way consistent and up-to-date contextual data is provided to the system.

Sensors can be both physical and "virtual". The physical sensors gather information such as:

- Environmental, e.g., temperature sensor or light level sensor;
- Location, e.g., by means of a GPS sensor;
- Medical data, e.g., heart rate or blood pressure.

"Virtual" sensors gather context data from software applications or from operating system. Regarding our approach, the "virtual" sensors may supplement certain unavailable sensors or communicate with third-party institutions (e.g., emergency call centers or police). For example, as the position inside a building cannot be determined using the GPS sensor, a "virtual" sensor invokes a localization service which determines the position of the user using indoor localization techniques proposed by Hyun-Yong Noh, et al. [19]. As emergency call centers have implemented an automatic detection system of the caller's context (e.g., GSM cell-based location, name or age) it is suitable to have a background application which automatically gathers these patient contextual data removing the need of manually forwarding these data to emergency medical institution.

However, sensors capabilities are limited. As there are situations in which the sensors cannot gather a particular contextual information we think it is important to endow the Context-Aware EHR System with the ability to manually input authoritative contextual data and thus enable explicit user actions (e.g., manual entry of the patient's symptoms) [20].

In the context discovery phase, the Context Service gathers all the available contextual information and stores them in the Context Repository and it is further updated as information changes.

The stored Context Repository data provide to the "Inference Engine" contextual information to be used in the inference phase which will further provide the user with a personalized view on the EHR's data and a quick and comprehensible access to medical data relevant for the concrete, particular state of affairs. Table 1 summarizes the contextual categories proposed for our system and offers a detailed view upon contextual attributes of each category.

C. *Making context information useful*

Field studies reported that the topic of searching information within EHRs is insufficiently explored and the lack of the search functionality often hampers healthcare professional's access to patient's data [21, 22]. However, as the patient's EHR comprises large amount of data, a keyword search utility [22] does not ease access to information nor facilitates browsing. During clinical workflows contextual attributes and patient's health state change and we envision that a system which proactively gathers these parameters will be capable of continuously provide situation tailored data.

The aim of a doctor is to heal the patient or at least to soothe his pain in case of an incurable disease. Doctor's medical knowledge is somehow insufficient as making a proper diagnostic relies not only on knowledge or aptness to predict patient's health state evolution but on patient's clinical history. Setting a proper diagnostic is one of the critical points of the workflow as it backs the patient's healing process. The clinical history offers a better understanding of how the patient's health state evolved up to the current one and how past medical conditions influenced the current health state.

The "Inference Engine" endows the Context-Aware EHR System with the ability of acting proactively. It accesses the Context Repository and based on the acquired contextual information it queries the EHR database and extracts situation relevant information. The search through EHR considers three of its major sections:

- Text-based information, e.g., name, age, medical history or allergies;
- The collection of clinical images, e.g., radiographs or photos;
- The collection of videos, e.g., oral examination videos made with an endo-oral camera or colon videos.

One of the challenges of the system is to extract and provide the user with medical data according to a given context and to filter irrelevant information. This is done by the "Inference Engine". Namely, it follows the idea of ranking text-based information according to its relevance in a given context. Using patterns, the raw contextual information is associated with a collection of related terms which are further searched within EHR database. As for example if the patient has a dislocated right ankle the system will search his medical history to search for medical data related to his

TABLE 1. SUMMARY OF CONSIDERED CONTEXTUAL ATTRIBUTES

Context category	Attributes	Source
Personal profile	Name, age, specialization...	Users' repository
Preferences	Font type, font size, font color...	Users' repository
Device capabilities	Screen size, input mode...	"Virtual" sensors
Environmental	Date, time, temperature, light level...	"Virtual" or physical sensors
Location	GPS coordinates, place inside building	"Virtual" or physical sensors
Medical data	Heart rate, blood pressure...	Physical sensors

right ankle. All the EHR data related to the patient's right ankle (e.g., diagnosis, medication prescribed or clinical decisions) will be high ranked as opposed to those which refer to his left ankle. A "Decision Frame" [23] is included to present the patient's relevant information. For the clinical images and videos the search is made through attached tags. Finally, gathered medical data are displayed according to the stored user's preferences and profile. In Section 4, we describe how specific medical terms and codes can be converted to general ones thus helping patients to understand their own medical data.

Additionally, a "Context Knowledge Base" is maintained and updated with information from the applications of healthcare professionals. It provides a collection of context-patterns extracted from previous contexts (e.g., the victim has suffered a frontal car accident so it may have caught his feet in the car body or the victim has suffered a heart attack and he may faint). Namely, each medical event and its contextual attributes is stored and in case of further events that match to a great extent it is shown. Previous medical situations and healthcare professionals' reports may offer a better insight of the actual situation as similar circumstances are probably to have similar effects. Each pattern is associated to the possible effects it may produce. Matching the current context with context-patterns in the knowledge base contextual information gain meaning and can be better used by the system to provide users with different point of views derived from previous similar cases. This inferred information will provide useful data for the current situation based on previous circumstances, thus predictions on patient's health state evolution can be made.

As noted above, the sensors, be they physical or "virtual" cannot gather certain information. Therefore, the context information may be vague and thus the system may retrieve insufficient EHR data. The Context-Aware EHR System does not restrict the user to one particular, scenario specific, patient's data view, if more EHR data are needed the system can be further interrogated by standard queries requesting the missing pieces of information. The retrieved data are further merged with the inferred one.

D. *The Context-Aware EHR System. A patient point of view*

The EHR is a collection of medical information which reflects patients' past medical history. It comprises highly structured data and its entries use standard codes and classifications for representing medical data. This standardization may seem fair for healthcare professionals since the EHR systems promote collaborative and goal-directed treatment planning [27, 26] but may appear cumbersome as patients don't usually have medical knowledge and thus they may encounter problems in accessing and understanding their own medical records. Recent works [25, 28] have explored the concept of "active patient" which no longer sits on the sideline waiting for the doctor to cure her/him but actively participate in his own care. Of course, to increase the efficiency of the "active patient" he must be endowed with tools.

MedlinePlus Connect is a health information resource freely provided by National Library of Medicine (NLM), National Institutes of Health (NIH), and the Department of Health and Human Services (HHS) aiming both patients and healthcare providers. It handles medical standard codes and retrieves related information (e.g., for a medication code the side-effects, dosage or special precautions for that medicine).

In order to obtain a patient-friendly EHR data view we make use of the above mentioned MedlinePlus Connect Web Service; a standard encoded message is converted to a non-professional user-friendly format and, consequently, lab tests, medications and diagnostic codes and other specific medical information gain meaning for the general public. Additionally, through "Context Knowledge Base" more context-specific information will be available for each context-pattern, i.e., patients can review similar previous medical situations thus be aware of the evolution of their disease and possible complications.

"Sensing user" role empowers patients to participate more in their healthcare workflow process and to be aware of what has been done and what is further planned in their care process.

IV. Implementation issues and prototypes

As the proposed system needs to proactively act according to a medical situation we first need a services platform to support context-aware application. In our future implementation we plan to use WASP platform [29], a robust and configurable services platform which gathers and provide contextual information to subscribed clients. We are aware that unavailability of certain contextual information may lead to incomplete data shown to users so we will implement and test the facility to refine the results through manually specifying certain contextual attributes and through standard queries.

One important issue regarding EHR systems is the unavailability of medical data due to its absence from the medical records or due to owner's desire to keep it private. We will develop several test scenarios and together with patients and healthcare providers we will try to enhance the quality of retrieved results. The patients are those to set their own EHR level of confidence and with respect to the proposed scenarios to have healthcare providers' opinion upon retrieved medical data degree of detail. This particular research is important on one side for patients as they can see the impact of different levels of confidentiality upon retrieved medical data and on the other side for healthcare professionals as unavailability of certain data leads to incorrect diagnostics or hinder him to have patient's clinical picture.

V. CONCLUSION AND FUTURE WORK

As sometimes search of information in large databases may be difficult, structuring EHR's information according to situations' specific needs enables a predictive medical information retrieval. Of the patient the system brings the opportunity of accessing the EHR medical information without the need of knowing the messaging standards (e.g., HL7 v3 or HL7 v2.x). The system uses the perceived contextual information and determines which medical data match the given circumstances and may be of interest.

The leading idea of this paper is to build a system to help EHR users retrieve medical data other than using standard ways which in certain cases proves to be cumbersome and time-consuming. To our knowledge, this study was the first to embed the conventional EHR database into a system which gathers user's contextual information and uses it to proactively retrieve patient's medical data tailored on user's device capabilities. Through a customized view of the EHR data, healthcare providers have they work supported by easy and comprehensible data access and also, patients are offered an interpretation of their medical data for a better insight. However, it is important to note that users are not restricted to this particular, scenario specific, patient's data view. If the Context-Aware EHR System retrieves exiguous information the user can query the EHR database typically in order to obtain the missing pieces of information.

We are currently planning a pilot deployment of our proposed approach followed by in-depth interviews and surveys with its users on application's usefulness. We also plan to build a complementary system to support entering to the EHR medical data along with of contextual information.

REFERENCES

[1] D. Gans, J. Kralewski, T. Hammons and B. Dowd, „Medical Groups' Adoption Of Electronic Health Records and Information Systems," Health Affairs, 24, no.5 (2005):1323-1333, doi: 10.1377/hlthaff.24.5.1323.

[2] http://ehealth-strategies.eu/database/documents/Estonia_CountryBrief_eHStrategies.pdf. Retrieved: February, 2012.

[3] http://www.sante.gouv.fr/acces-web-patient.html. Retrieved: February, 2012.

[4] http://ehealth-strategies.eu/database/documents/France_CountryBrief_eHStrategies.pdf. Retrieved: February, 2012.

[5] M. B. Skov and R. T. Høegh, "Supporting Information Access in a Hospital Ward by a Context-Aware Mobile Electronic Patient Record," in Crestani, F. et al. (Eds.) Personal and Ubiquitous Computing, 2006, Vol. 10, No. 4, pp. 205-214.

[6] M. Ancona, E. Coscia, G. Dodero, M. Earney, V. Gianuzzi, F. Minuto, and S. Virtuoso, „WardInHand: Wireless Access to Clinical Records for Mobile Healthcare Professionals," Proceedings of 1st Annual Conference on Mobile & Wireless Healthcare Applications, London, United Kingdom, 2001.

[7] M. A. Grasso, "Clinical Applications of Handheld Computers," Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems (CBMS '04), IEEE Computer Society, Washington, DC, USA, 2004, pp. 141-. DOI:10.1109/CBMS.2004.28.

[8] B. Schilit, N. Adams, and R. Want, "Context-Aware Computing Applications," Proceedings of the 1994 First Workshop on Mobile Computing Systems and Applications (WMCSA '94), IEEE Computer Society, Washington, DC, USA, pp. 85-90, DOI:10.1109/WMCSA.1994.16.

[9] N. Bricon-Souf and C. R. Newman, "Context Awareness in Health Care: a Review," International Journal of Medical Informatics 76(1), pp. 2–12 (2007).

[10] L. Barkhuus and A. K. Dey, "Is Context-Aware Computing Taking Control away from the User? Three Levels of Interactivity Examined," Proc. Ubicomp '03, Springer (2003), pp. 149-156.

[11] N. Menachemi, T. H. Collum, „Benefits and drawbacks of electronic health record systems," Risk Management and Healthcare Policy Journal, vol. IV, pp. 47-55 (2011).

[12] P. Shetty and S. W. Loke, "Modelling Context-Aware Security for Electronic Health Records," Encyclopedia of Information Ethics and Security (ed) M. Quigley, 2007, pp. 463-469, IGI Global.

[13] http://www.paginamedicala.ro/stiri-medicale/SMURD-Bucuresti-implineste-doi-ani_7437/. Retrieved March, 2011.

[14] D. Zhang, Z. Yu, and C. Chung-Yau, "Context-Aware Infrastructure for Personalized Healthcare," The International Workshop on Personalized Health (pHealth 2004), IOS Press, pp. 154-163, December 13-15, 2004, Belfast, Northern Ireland, United Kingdom.

[15] H. Lei, D. M. Sow, J. S. Davis, G. Banavar, and M. R. Ebling, "The design and applications of a context service," SIGMOBILE Mob. Comput. Commun. Rev. 6, 4 (October 2002), pp. 45-55, DOI:10.1145/643550.643554.

[16] K. Henricksen, J. Indulska, and A. Rakotonirainy, "Modeling Context Information in Pervasive Computing Systems,", Proceedings of the First International Conference on Pervasive Computing (Pervasive '02), Friedemann Mattern and Mahmoud Naghshineh (Eds.), Springer-Verlag, London, UK, pp. 167-180.

[17] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and Daniele Riboni, "A survey of context modelling and reasoning techniques," Pervasive Mob. Comput. 6, 2 (April 2010), pp. 161-180, DOI:10.1016/j.pmcj.2009.06.002.

[18] A. K. Dey, "Understanding and Using Context," Personal Ubiquitous Comput. 5, 1 (January 2001), pp. 4-7, DOI:10.1007/s007790170019.

[19] N. Hyun-Yong, L. Jin-Hyung, Sae-Won Oh, H. Keum-Sung, and Sung-Bae Cho, "Exploiting indoor location and mobile information for context-awareness service," Inf. Process. Manage. 48, 1 (January 2012), pp. 1-12, DOI:10.1016/j.ipm.2011.02.005.

[20] P. Coppola, V. Della Mea, L. Di Gaspero, S. Mizzaro, I. Scagnetto, A. Selva, L. Vassena, and P. Z. Rizio, „MoBe: Context-aware mobile applications on mobile devices for mobile users," Proc. of 1st int. workshop on exploiting context histories in smart environments (ECHISE 2005).

[21] T. Christensen, A. Grimsmo, "Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records," BMC medical informatics and decision making, Vol. 8 (28 March 2008), 12, doi:10.1186/1472-6947-8-12.

[22] K. Natarajan, D. Stein, S. Jain, and N. Elhadad, "An Analysis of Clinical Queries in an Electronic Health Record Search Utility," International Journal of Medical Informatics - July 2010, Vol. 79, Issue 7, pp. 515-522, DOI: 10.1016/j.ijmedinf.2010.03.004.

[23] E. Bayegan, Ø. Nytrø and A. Grimsmo, "Ranking of Information in the Computerized Problem-Oriented Patient Record," Medinfo 10 (Pt 1) (2001), pp. 594–598.

[24] K. Sadegh-Zadeh, „Fundamentals of clinical methodlogy 4. Diagnosis," Artificial Intelligence in Medicine, Vol. 20, Issue 3, pp. 227-241, November 2000.

[25] D. Vawdrey, L. Wilcox, S. Collins, S. Bakken, S. Feiner, A. Boyer, and S. Restaino, „A tablet computer application for patients to participate in their hospital care," Proc AMIA 2011, pp. 1428-1435.

[26] K. J. Leonard and W. J. Winkelman, „Overcoming Structural Constraints to Patient Utilization of Electronic Medical Records: A Critical Review and Proposal for an Evaluation Framework," Journal of the American Medical Informatics Association, 2004, March-April, 11(2), pp. 151-161.

[27] H. Stam van Ginneken, "Computer-based patient record with a cardiologic extension", Medinfo 1995, 8(Pt 2):1666.

[28] L. P. Vardoulakis, A. Karlson, D. Morris, G. Smith, J. Gatewood, and D. Tan, "Using mobile phones to present medical information to hospital patients," In Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems(CHI '12), ACM, New York, NY, USA, 2012, pp. 1411-1420, DOI=10.1145/2207676.2208601, http://doi.acm.org/10.1145/2207676.2208601.

[29] http://www.freeband.nl/kennisimpuls/projecten/wasp/. Retrieved: March 2012.

ISBN: 978-1-61208-215-8

Control-Flow Patterns in Converged Services

Andres Benavides, Gustavo Enriquez, Jesus David Ramirez, Cristhian Figueroa, Juan Carlos Corrales
Telematics Engineering Group
Universidad del Cauca
Popayán, Colombia
(andresbenavides, genriquez, jdramirez, cfigmart, jcorral)@unicauca.edu.co

***Abstract*—Nowadays, telecommunication-services providers have focused their interest in developing new converged services at low cost and fast time to market. For this purpose they have adopted composition approaches, which are based on the reuse of existing software components to create more complex services. However the composition of converged services has so far been done at code level which requires expert knowledge, makes the process tedious and complicated, and causes an increase in design and deployment time. This paper reviews the control-flow patterns supported in converged services, proposes a theoretical basis, and defines the most recurrent control-flow patterns present in these kinds of services.**

***Keywords*-Converged services; service composition; control-flow patterns.**

I. Introduction

Nowadays, telco service providers (hereinafter Telcos) are looking for new methods for service creation in order to react quickly and cost effectively to dynamic market conditions. Furthermore, convergence of networks, services and content is taking place at an increasing speed, and therefore Telcos are forced to reduce services time to market [1].

One of the methods adopted by Telcos to accelerate the time to market is called *composition*; which in Telco environments, is based on the reuse of existing software components in order to create converged services. These services integrate traditional telco services (e.g call forwarding, sms) with web services (e.g payment services, social services) and provide an added value to users willing to pay.

Therefore, Telcos have adopted telecommunications platforms like JAIN SLEE (JSLEE) [2], Sip Servlets [3], etc. to compose converged services keeping telco features (high throughput, low latency, high availability [4]) and additionally supporting web services invocation through specific adapters. These features have allowed the composition of a large amount of converged services; however, this process has been done at code level requiring an expert knowledge. Consequently, the composition becomes tedious and complicated, causing an increase in the service design and deployment time [5].

To address this issue, and in order to facilitate the creation and composition of services, some studies have focused on the separation of the service business logic and its implementation. For example [4], [5], [6] concentrated in the design of converged services using the Business Process Execution Language (BPEL) or the Java Bussines Process Management (jBPM) graphical tools. These services are then adapted to a JSLEE platform, and executed afterwards.

As can be seen, these works use well-known languages to create converged services and model features as execution order and data exchange, which defines the execution flow of these kinds of services. Moreover, these languages allow the use of recurrent structures known as patterns (e.g. split, merge, synchronization) which can be reused in the composition process. These patterns represent repetitive events within an execution flow and have been widely studied in Van der Aalst *et al.*'s [7] work. This study analyzes the execution flow from two perspectives: I) control-flow patterns (CFP) [8] related with services execution order; and II) data patterns [9] used in data flow to represent the information transfer among services. The previously named works have addressed service composition emphasizing the use of model languages and reusable structures, however these approaches have been found to present drawbacks due to the fact that patterns perspectives and typical composition languages, like BPEL and JPDL, were designed to model the execution flow of web processes, but not for converged services. Consequently telco-services execution features are left out.

Our work addresses the study of CFP support in converged services, in order to create a theoretical basis which allows the standardization of processes related with the creation and composition. The use of CFP as a means of categorizing recurrent solutions in the converged services field, brings advantages as: abstraction and description, in a concrete form, of the recurrent structures (e.g. splits, merges, loops) in specific contexts; fulfillment of composition requirements (e.g. define the correct order of services) and reduction of the modeling work as well as development time and cost, through the use of recurrent solutions. Therefore, it is a new insight within the processes related with converged services which allows understanding and systematizing the workflow perspective in this field; thus, we define a list of most recurrent CFP present in these kinds of services in order to take advantage of the impact of the workflow patterns [8].

This paper is organized as follows: Section II describes a method for detecting CFP in converged services; Section III shows the results obtained after patterns detection; and finally, Section IV presents the conclusions.

II. Methodology

The methodology used in this paper focuses on CFP detection in converged services and consists of two main parts: Service Catalog and CFP Detection. In the first one, we classify the services present in a convergent environment defining an architecture for services classification and a list of services which are then represented by a formal model. In the second one, we adapt an algoritm to detect substructures (patterns) in the entire service catalog, allowing to define a set of CFP applicable to the converged services domain. These main parts are described in more detail in the following sub-sections.

A. Service Catalog

We define a service catalog in order to classify a set of services present in a converged environment, the catalog is based on a literature review of service delivery platforms (Mobicents [10], Rhino [11]); some service providers (Vodafone, Telefonica, AT&T), and organizations (ITU, 3GPP, ETSI-TISPAN, OneApi [12]).

1) Service Architecture: The Services Architecture proposed in this paper (Fig. 1), is based on Al-Begain *et al.* [13], which provides an ideal environment for telco and web services composition. This study proposes an architecture divided in five layers: network, control, services, integration and implementation. For the scope of our work, we focuse only on the service layer in which we define the following modules: Telco Services (TS), Web Services (WS), Converged Services (CS), Data access Services (DAS) and Security Services (SS).

- TS module includes traditional telco services known as Basic Services (BS) [14], [15] and complementary functionalities which are called Supplementary Services (SPS) [16]. Additionally, this module includes composite services as result of BS and SPS combination (e.g Call + Call Hold) [17].
- WS module contains traditional web services (TWS) and Web 2.0 services (WS2). The TWS, are static services with simple request-response mechanism; while the WS2, allows a dynamic and interactive knowledge creation on the internet [18]. Additionally, WS module includes composite web services which combine TWS and WS2.
- CS module contains converged services which are defined as services, applications, and content, provided by different networks through different user-terminals [19]. Based on this definition, this module contains services using telco and web capabilities. It is important to bear in mind that a convergent service must include at least a TS and WS.
- DAS and SS modules are transversely located in the architecture as they can be used by TS, WS and CS modules. DAS module allows services to access to certain data using different standards (e.g. Java Data Base Conectivity JDBC); and SS module guarantees that services are safely consumed by users, using protocols such as: Internet Protocol security (IPsec) [20], Secure Sockets Layer (SSL) [21], Secure HyperText Transfer Protocol (S-HTTP) [22], etc.

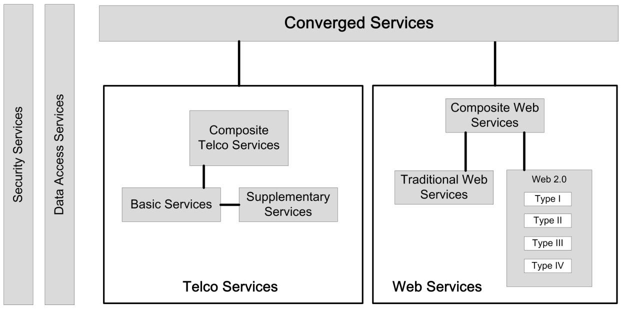


Figure 1. Service Architecture

2) List of Services: As presented at the beginning of this section, the list of services is based on the literature review of different organizations. In this list we show a set of 40 services (15 TC, 15 WS, 10 CS) which are part of a converged environment, however, in Table I due to the space restriction we only present some examples, which are listed according to the classification performed using the architecture shown in Fig. 1.

3) Formal Model: To describe the services contained in the catalog, we used two formal models: Petri Nets and Graphs. The first one, uses a set of places, transitions and edges to describe concurrent, asynchronous, distributed and nondeterministic systems [23]; and the second one, uses nodes and edges to represent and describe the existing relations between system objects [24]. In our work, Petri nets were used to provide a detailed and precise service description. This model allows the association between services of the catalog and CFP [8] which use the same representation. Additionally, we use the graph formal model to represent the services control-flow with a logical approach. With this model is possible to use gates like AND, OR or XOR to describe the structures that define the services control-flow. Fig. 2 and Fig. 3 (right) present the *Basic Call* service using Petri Nets and Graph model. In the first one, Petri Nets allow to represent states (ring, talking, busy recording) and possible events (invite, 200 Ok, busy, bye) of the service, while the second one, shows the service control-flow through logic gates: XOR-JOIN (XORJ) and XOR-SPLIT (XORS).

Table I
LIST OF SERVICES

Telco, Web and Converged Services		
Telco	BS	Call
		SMS
		Video Call
	SS	Call Hold
		Call Transfer
		Explicit Call Transfer
		Conference
		Advice of Charge
	Composite Telco Services	Call + Call-Hold
		Call + Call-Hold + SMS
		Call + Conference
Web	TRS	Payment
		Terminal Location
		Presence
		Terminal Status
		Device Capabilities
	WS2	Facebook Services
		Maps
		Skype Services
	Composite Web Services	Presence + Facebook
		Facebook + Call
		Travel Planning
Web+Telco	CS	Maps + Click to Dial
		Financial Instant Message
		Facebook Events Reminder

B. Control-Flow Patterns Detection

To detect CFP in converged services, a method composed of four main modules was used; this model is illustrated in Fig. 4.

- *Service Graph*: it contains a file that describes the service control-flow.
- *Query CFP*: they are the reference patterns used by the Pattern Detection module. These query patterns are stored in a database.
- *Pattern Detection*: it uses a patterns detection algorithm to define the number of patterns in the service graph.
- *List Of CFP*: represents the number of detected patterns in the service graph.

These modules are described in more detail on the following subsections.

1) Service Graph: Each service described in graph model is associated with a file description, in which the CFP detection is performed.

The file is divided into five sections, as shown in Fig. 3 (left). The first section of the file begins with an identifier (#graph0), the second section describes the number of graph nodes; the third section refers to the nodes labels, where *E* is an event and *T* is a task; the fourth section contains the number of edges (it should be at least one), and finally the fifth section shows the edge labels which represents the source and destination nodes, and defines the service control-flow [25].

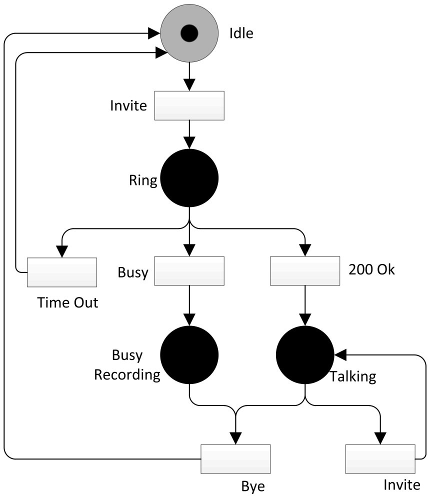
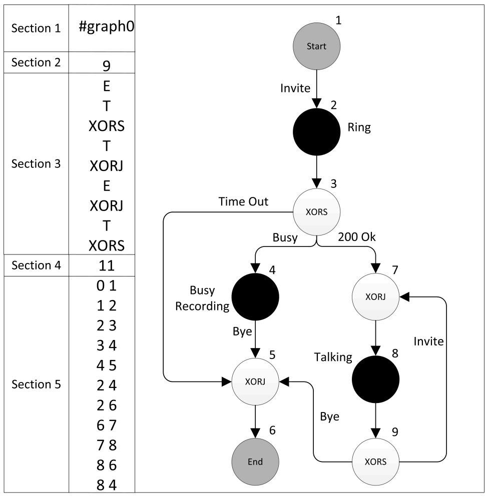


Figure 2. Basic Call Service using Petri Nets model



Section 1	#graph0
Section 2	9
Section 3	E T XORS T XORJ E XORJ T XORS
Section 4	11

Section 5	0 1
	1 2
	2 3
	3 4
	4 5
	2 4
	2 6
	6 7
	7 8
	8 6
	8 4

Figure 3. Basic Call Service using Graph model (right) and its description file (left)

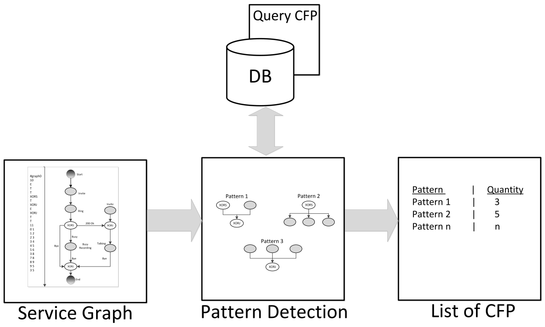


Figure 4. Method to detect CFP

2) Query CFP: Query CFP are the reference patterns used by the detection module to find recurrent structures in the graph service. In this paper we have defined a set of query CFP based on the evaluation performed in Wohed *et al.*'s [26] work, in which a study of graphical jBPM tool v. 3.14 and BPEL language v.1.1 was performed. This evaluation provided a set of 16 patterns directly supported by jBPM and 13 patterns directly supported by BPEL. In addition, we consider the composition and specializations relationships between those patterns [8]. The first one, exists when a pattern represents a more restricted form of another pattern, (e.g. *Multiple-Choice* pattern is a specialized form of *Parallel split* pattern) and the second one, combines two or more patterns (e.g. *Structured Loop* pattern is the *Exclusive Choice* and *Simple merge* patterns combination). Table II shows a subset of twenty-one query CFP, as result of the previous analysis. These set of query patterns are represented through logical gates. Fig. 5 shows some examples of query patterns taking as reference the Vander Aalst *et al.*'s work [8].

Due to the control-flow of some patterns can be similar, they may be associated to the same logical gate.

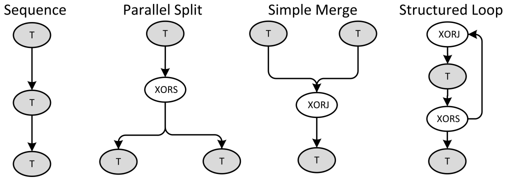


Figure 5. Query CFP represented through logical gates

Table II
QUERY CFP AND ITS ASSOCIATED LOGICAL GATES

CFP	Associated gate
Sequence	-
Simple merge	XORJ
Multi choice	ORS
Multi merge	ORJ
Parallel split	ANDS
Implicit termination	
Synchronization	ANDJ
Transient trigger	
Persistent trigger	
Exclusive Choice	XORS
Deferred choice	
Multiple instances without synchronization	
Structured synchronizing merge	ORS+ANDJ
Local syncrhonizing merge	
Generalized AND-join	ANDS + ANDJ
Critical Section	
Interleaved Routing	
Arbitrary Cycles	XORS + XORJ
Structured Loop	
Cancel Case	XORS or XORJ
Cancel Task	

3) Pattern Detection: CFP detection within the service graph (Section II-B1) was performed using the indexing graph method called GraphBlast [27] which allows searching for patterns within a graphs database. This method includes a subgraph isomorphism algorithm denominated VF2 [28] which executes comparisons between graphs, and determinate similarity among their nodes and edges structures (isomorphism), or whether a graph (sub-graph) is contained within another one.

4) List Of CFP: A list of CFP is the number of detected patterns in each service. For example, for the Basic Call service, as shown in Fig 3 (right) a set of two XORJ gates and two XORS gates were detected, which corresponds to the patterns: *simple merge, deferred choice, structured loop* and *cancel case.*

The result of the pattern detection is shown in Section III.

III. Results

In this section, we present the result obtained after the CFP detection. Table III shows the set of patterns found in the entire list of services (Section II-A2).

The CFP detection is based on the results provided by the proposed method (Section II). Each service is analyzed individually and afterwards a general analysis is performed.

For example, the Facebook Events Reminder service, represented by a graph model as shown in Fig. 6, is a converged service which alerts the user when a Facebook event takes place. This service invokes a web service called Facebook User Events, which has the basic information of the user events (date, time, place). Then a Presence service is used in order to check and determinate the user availability; in this way, if the user is available, the service call him to give the event information by means of a voice recording, but if the user is not available, a SMS with the necessary information is sent.

The patterns detected in Facebook Events Reminder service are: *secuence* (nodes 1, 2 and 3), *deferred choice* (nodes 4 to 5 and 6), *cancel case* (nodes 7 to end) and *simple merge* (nodes 5-6 to 7). As can be seen the *secuence* pattern represents the invocation of Facebook-, HTTP-Manager- and Presence- Service, while the *deferred choice* pattern provides the ability to defer the moment of choice in the service based on user status (unavailable or available), additionaly, *simple merge* pattern provides a means of merging call and SMS services without synchronizing them; finally *cancel case* pattern ends the service.

Another example is the Telco service composed by : Call + CallHold + Call Transfer + Conference wich has five patterns where the most recurrent pattern is *simple merge*. Furthermore, according to the services features, it also has patterns like: *deferred choice, cancel case, exclusive choice* and *structured loop* which are shown in the Fig. 7. The service graph model for this service is not present in this paper due to space restriction.

The analysis in the previous two examples was performed to the entire 40 services presented in the catalog (Section II-A) and it is illustrated in Fig. 8.

This result is based on a recurrence rate analysis performed to the CFP found in the entire amount of services.

As can be seen, the most recurrent patterns are *simple merge* (P5), *defered choice* (P8), *parallel split* (P2) and

Table III
FOUND CFP IN THE SERVICE CATALOG

ID	Pattern
P1	Sequence
P2	Parallel split
P3	Synchronization
P4	Exclusive choice
P5	Simple merge
P6	Generalized And-join
P7	Multi-Choice
P8	Deferred Choice
P9	Cancel task
P10	Cancel case
P11	Structured loop

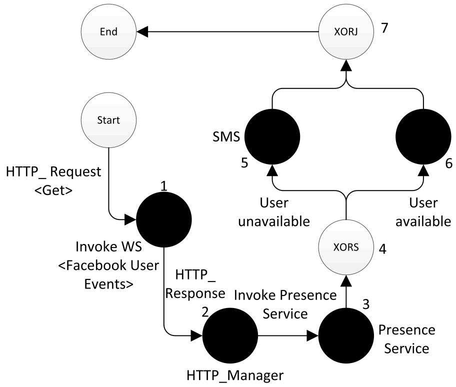


Figure 6. Graph model of Facebook Events Reminder service

cancel case (P9); while the the less recurrent patterns are *secuence* (P1) and *synchronization* (P3).

The result allows to identify the most recurrent CFP in converged services. In a practical environment, the CFP found in this paper can be implemented at code level once, and reuse them when needed; in this way it can be used as a code template in the processes of creation and composition of converged services. Thereby, the use of these recurrent structures (patterns) and a Service Logic Execution Environment like JSLEE, keeping telco features and supporting web services invocation, can provide a usable tool for developing new services in an easy way and with less time consumption.

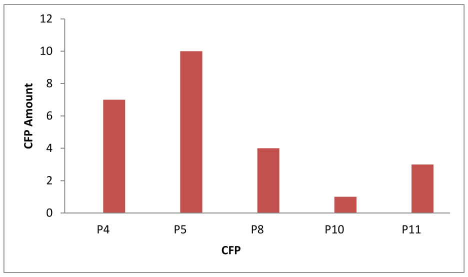


Figure 7. CFP in Call + Call Hold + Call Transfer + Conference service

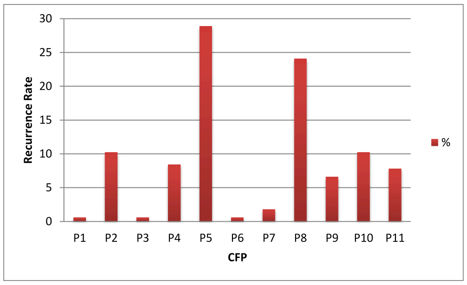


Figure 8. Recurrence rate of CFP

IV. CONCLUSION

This paper presented a study of the supported CFP in a converged services environment through the analysis of most recurrent CFP found in a set of services. The result obtained in this work serves as a theoretical basis to contribute to the standardization of practices related with creation and composition of converged services.

The CFP detection approach for converged services is based on two formal models: the first one is a graph model, which provides a clear description of the control-flow through logic gates; and the second one, is a Petri nets model which provides a more specific service description and allows the association between services in the catalog and CFP proposed in [8].

Additionally, a detection algorithm was adapted to find the number of CFP in each service, which results in a set of ten patterns applicable to the converged services domain. For this set of patterns, an analysis of recurrence rate was performed, concluding that the most common CFP within converged services are the *basic, stated based, cancellation* and *Structural* Patterns.

The next step in this work is a prototype development to experimentally evaluate the effectiveness of using CFP in the converged services environment in terms of time and ease. For this reason we are currently developing a graphical tool using GEF (Graphical Editing Framework) and JSLEE in order to create and compose converged services in an easier way. Fig. 9 shows the user interface within the Eclipse IDE; here services and patterns are illustrated as a blue building blocks, which can easily added to the workspace with a simple drag and drop functionality.

In addition, this study could be extended using data, resources and exception patterns defined in [7], in order to study other issues of converged services environment.

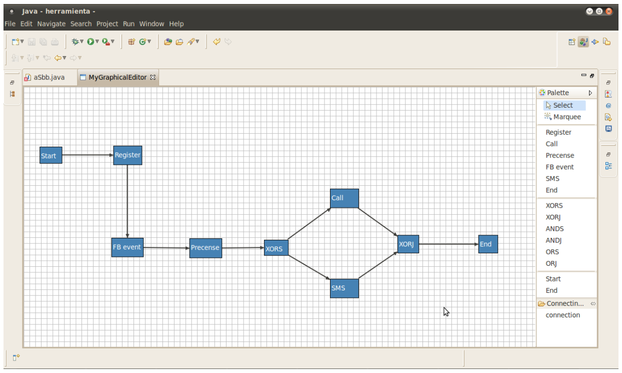


Figure 9. Graphical Composition Tool using GEF and JSLEE

V. Acknowledgment

The authors would like to thank University of Cauca and Telematics Engineering Group for supporting this research work.

References

[1] P. Falcarin and C. Venezia, “Communication Web Services and JAIN-SLEE Integration Challenges,” *International Journal of Web Services Research*, vol. 5, no. 4, pp. 59–78, 2008.

[2] Community Development of java Technology Specifications - Oracle, “Jain Slee (JSLEE) v1.1,” June 2012. [Online]. Available: http://jcp.org/en/jsr/detail?id=240

[3] J. Deruelle, V. Ralev, and I. Ivanov, *JBoss Communications Platform 5.1 - SIP Servlets Server User Guide*, 2011.

[4] M. Femminella, E. Maccherani, and G. Reali, “A software architecture for simplifying the JSLEE service design and creation,” in *Software, Telecommunications and Computer Networks (SoftCOM)*, vol. 22, 2010.

[5] T. Eichelmann, W. Fuhrmann, U. Trick, and B. Ghita, “Enhanced concept of the TeamCom SCE for automated generated services based on JSLEE,” in *Eighth International Network Conference (INC)*, 2010.

[6] D. Zhu, Y. Zhang, B. Cheng, B. Wu, and J. Chen, “HSCEE : A Highly Flexible Environment for Hybrid Service Creation and Execution in Converged Networks,” *Journal of Convergence Information Technology*, vol. 6, no. 3, pp. 264–276, 2011.

[7] “Workflow Patterns Initiative,” June 2012. [Online]. Available: http://www.workflowpatterns.com/

[8] W. van der Aalst, A. Hofstede, N. Russell, and N. Mulyar, “Workflow Control-Flow Patterns, A Revised View,” in *BPM Center Report BPM-06-22*, 2006.

[9] W. Van der Aalst, N. Russell, A. Hofstede, and D. Edmond, “Workflow Data Patterns,” *QUT Technical report, FIT-TR-2004-01*, 2004.

[10] Open Source Cloud Communications Platform, “Mobicents,” June 2012. [Online]. Available: http://www.mobicents.org/slee/docs.html

[11] Open Cloud, “Rhino’s SIP resource adaptor (RA), services and components,” May 2012. [Online]. Available: https://developer.opencloud.com

[12] GSMA, “OneAPI,” June 2012. [Online]. Available: http://oneapi.aepona.com/

[13] K. Al-Begain, C. Balakrishna, L. A. Galindo, and D. Moro, *IMS : A Development and Deployment Perspective*, Chichester, West Sussex, U.K, 2009.

[14] ITU, “Definition of Teleservices, ITU-T Recomendation I.240,” 1989.

[15] ITU-T, “Definition of Bearer Services, ITU-T Recomendation I.230,” 1988.

[16] ITU, “Definition of Supplementary Services, ITU-T Recomendation I.250,” 1988.

[17] 3GPP, “3GPP TS 27.173 - Multimedia Telephony Service and supplementary services,” Tech. Rep. Release 9, 2010.

[18] S. Shang, E. Li, Y. Wu, and O. Hou, “Understanding Web 2.0 service models: A knowledge-creating perspective,” *Information & Management*, vol. 48, no. 4-5, pp. 178–184, 2011.

[19] F. R. Mejia, *Evolución de los centros de acceso público a las TIC*, serie cepa ed., 2009.

[20] P. Asadoorian, “An Introduction to IPsec,” May 2012. [Online]. Available: http://www.pauldotcom.com/IPSEC.pdf

[21] GNU, “The GNU Transport Layer Security Library,” 2012. [Online]. Available: http://www.gnu.org/software/gnutls/documentation.html

[22] Network Working Group, “HTTP Over TLS - RFC 2818,” May 2012. [Online]. Available: http://tools.ietf.org/html/rfc2818

[23] J. L. Peterson, *Petri net theory and the modeling of systems*, prentice-h ed., 1981.

[24] H. Bunke, “Graph Matching: Theoretical Foundations, Algorithms, and Applications,” pp. 82–88, 2000.

[25] D. Rivas, D. Corchuelo, C. Figueroa, J. Corrales, and R. Giugno, “Business Process Model Retrieval based on Graph Indexing Method,” *Lecture Notes in Business Information Processing*, vol. 66 Part 3, pp. 238–250, 2011.

[26] P. Wohed, B. Andersson, A. H. Hofstede, N. Russell, and W. Van der Aalst, “Patterns-based Evaluation of Open Source BPM Systems : The Cases of jBPM , OpenWFE , and Enhydra Shark,” *Information and Software Technology*, vol. 51, no. 8, pp. 1187–1216, 2009.

[27] A. Ferro, R. Giugno, M. Mongiovi, A. Pulvirenti, D. Skripin, and D. Shasha, “GraphBlast: multi-feature graphs database searching,” in *NETTAB 2007 Workshop*, 2007.

[28] L. P. Cordella, C. Sansone, and M. Vento., “Graph Isomorphism Algorithm for Matching Large Graphs,” *IEEE Transactions on Pattern Analysis and Machine intelligence*, vol. 26, no. 10, pp. 1367–1372, 2004.

Towards an Integrated Service Rating and Ranking Methodology for Quality Based Service Selection in Automatic Service Composition

Alexander Jungmann and Bernd Kleinjohann
Cooperative Computing & Communication Laboratory (C-LAB)
University of Paderborn, Germany
global@c-lab.de, bernd@c-lab.de

Abstract—**The paradigm shift from purchasing monolithic software solutions to a dynamic composition of individual solutions entails many new possibilities yet great challenges, too. In order to satisfy user requirements, complex services have to be automatically composed of elementary services. Multiple possibilities of composing a complex service inevitably emerge. The problem of selecting the most appropriate services has to be solved by comparing the different service candidates with respect to their quality in terms of inherent non-functional properties while simultaneously taking the user requirements into account. We are aiming for an integrated service rating and ranking methodology in order to support the automation of the underlying decision-making process. The main contribution of this paper is a first decomposition of the quality-based service selection process, while emphasizing major issues and challenges, which we are addressing in the On-The-Fly Computing project.**

***Keywords*-Service Composition; Service Selection; Quality of Service (QoS); On-The-Fly Computing**

I. Introduction

Nowadays, software engineers have to increasingly face up to the paradigm shift from the 40 years old principle of purchasing software as monolithic and closed standard solutions to the principles of Service Oriented Architectures (SOA) and Service Oriented Computing (SOC) [1], which shall enable purchase and execution of services on demand. Individually requested services may have to be composed of elementary services in order to fulfill the demanded requirements [2]. In this context, the problem of automatic service composition is a major challenge, since appropriate services have to be identified and correctly interconnected. Dependent on the amount of services, different possibilities to compose a complex service inevitably emerge. For that reason, a convenient methodology for choosing between services and composite services, respectively, which provide the same functionality but may differ in their non-functional properties, is required. We refer to this process as *quality-based service selection*.

Different approaches for determining the quality of a service or composite service in order to select the ”best” one out of a set of alternatives can be found in literature (e.g., [3] and [4]). In this context, a wide range of different non-functional properties is considered under the general term of *Quality of Service* (*QoS*) by which service candidates are compared with each other in order to identify and select the most appropriate one.

In contrast, we are aiming for an integrated service rating and ranking methodology, which facilitates the automation of quality based service selection by providing an overall taxonomy that reflects significant sections of the entire process. Each of these sections is investigated on service property level in order to identify and classify the essential challenges. Based on this taxonomy, generic solutions shall be determined and provided in order to enable the automation of the entire process. By doing so, our methodology shall enable the incorporation of arbitrary properties such as non-functional service properties or user requirements. However, we do not want to develop the one and only solution for quality based service selection in automatic service composition. In fact, we want to drive the generalization of this problem forward by developing a holistic representation, instead of providing just another solution for a specific problem setup. For that reason, we first of all identified some major issues that have to be taken into account during our work. Furthermore, a basic taxonomy of the selection process was already determined.

The remainder of this paper is organized as follows. Section II encapsulates the problem from our point of view. Section III lists the relevant major issues we identified so far. Section IV depicts our very basic taxonomy of the entire selection process. Section V describes some existing approaches, which are briefly discussed in Section VI. Finally, the paper concludes with Section VII.

II. Problem Description from the On-The-Fly Computing Point of View

A major vision of the On-The-Fly (OTF) Computing project is the automated composition of individual services based on services that are freely traded on global markets and that can be flexibly interconnected with each other. In this context, users may formulate a request, which contains information such as user information, preferences, domain-specific information and constraints. A so-called OTF service provider has the task to automatically compose an appropriate complex service, which matches the requested functional as well as non-functional properties. Depending on the size and granularity of the available service pool, multiple alternatives will inevitably emerge. Consider, e.g., the following scenario:

A traveling user has to wait at a train station of a large city for changing to another train. Since there is still plenty of time left (e.g. 3 hours), he spontaneously decides to do some sightseeing. He is only interested in specific classes of points of interests. Furthermore, he may walk or use the public means of transport. He uses his mobile device to put a request to an OTF service provider to compose a service, that in turn produces a convenient sightseeing tour by taking all the available information into account. The generation of his personal tour, hence the execution of the composite service, should be as cheap as possible.

The requested composite service could, e.g., consist of a basic trip planning service that in turn requires additional information such as cartographic materials, local points of interest and local public means of transport. Regarding global markets, services that provide the same required information are usually offered by more then one provider. In order to select the most appropriate service out of the set of service candidates, non-functional properties such as performance and cost as well as the specific user requirements have to be incorporated.

III. Issues in Quality Based Service Selection

Until now, we have identified the following major issues with respect to quality based service selection. The list, however, is not exhaustive, since our work is still at the very beginning. It may be modified and extended during our future research.

Implicit and explicit properties: Properties that have to be considered during the selection process can be *implicitly* given, e.g., in terms of non-functional service properties or *explicitly*, e.g., in terms of context-sensitive properties such as user preferences or user information.

Level of information: In many of the current approaches (cf. Section V), service properties are assumed to inherit the same *level of information* (e.g. quantitative values). However, different service provider may describe the same non-functional properties based on different levels of measurement (different scales). Furthermore, service properties may not be only described on different levels of measurement, but can also be non-existent for particular services.

Different hierarchical levels: Considering a global market, composite services and elementary services may match the same functional properties. In order to decide for the most suitable one, services have to be compared on *different levels of hierarchy*. The topmost level corresponds to the individual service that has to be composed, while the lowest level is defined by the granularity of the available elementary services in the service pool. In general, the number of levels in between cannot be defined in advance.

Local selection vs. global selection: While *local selection* of service candidates may not appropriately consider the overall quality of the final composite service, it is computational very efficient. On the other hand, *global selection* may identify the best overall solution, but ends up in a combinational problem, which is proven to be NP-hard. Either way, the point of time of decision-making has to be taken into account, since it essentially affects the rating and ranking strategy.

IV. Significant Sections of the Quality Based Service Selection Process

The intended methodology for quality based service selection has to provide generic solutions for the issues mentioned in Section III. The very first step towards such a methodology is the investigation of the entire service selection process on service property level by systematically disassembling the entire process in order to identify sections that depend on inherent service properties as well as scenario specific user requirements. In this context, we identified the following taxonomy, which reflects significant sections during the selection process.

A. Acquisition:

For the acquisition of property values, different techniques can be used. Single values may be accurately acquired by measuring. Other property values in turn have to be acquired from a series of measurements, in which the measured values vary from each other. Still others are not based on any metric at all, but have to be estimated from previous observations or are arbitrarily defined. Independently, service properties may also change over time.

B. Representation:

After acquisition, a property value has to be appropriately represented, while the representation in turn depends on the type of acquisition. In this context, descriptive statistics provide convenient methods for describing data. Single values can be classified with respect to their level of measurement (qualitative vs. quantitative values) or with respect to their scale, namely nominal, ordinal and metrical scale. Series of measurements are usually accumulated and represented as a distribution, which in turn can be approximated by means of statistical quantities (e.g., measure of central tendency and measure of dispersion) or fuzzy sets. Furthermore, methods of multivariate statistics such as cluster analysis enable a reduction of acquired data by means of abstraction and generalization, respectively, if required.

C. Utility Functions:

A utility function usually assigns a single value to an elementary or composite service. This value expresses the service's quality with respect to the explicitly and implicitly given properties. Service candidates can be compared with each other and consequently ranked in order to support the decision-making process. However, utility functions may be based on different sets of service properties (e.g., due to incomplete service descriptions) or have to incorporate different hierarchical layers.

D. Aggregation and Decomposition:

In order to rate and rank a composite service, the property values of the underlying elementary services have to be aggregated, e.g. by means of addition or multiplication, while the aggregation functions generally depend on the composition structure (parallel, sequence, loop etc.). On the other hand, a decomposition of property values is also of interest, when breaking down the global selection problem to a set of local ones or when services have to be compared on different hierarchy levels. However, aggregation may not only take place on service property level, but also on utility value level.

E. Objective Functions and Optimization Objectives:

Service rating heavily depends on the particular optimization problems, that are usually defined in advance. User preferences preset objective functions such as costs and availability and a specific optimization objective like minimizing and maximizing, respectively. Apart from this, a user may also desire a specific range of satisfaction by defining relative boundaries. In this context, homeostatic approaches provide convenient methods to deal with these types of optimization goals. However, service properties may also depend on each other. For that reason, multiple goals have to be simultaneously considered since they can negatively affect each other, leading to multi-objective optimization.

V. Related Work

Zeng et al. [5] introduced an approach that bases on a multi-dimensional quality model for elementary services as well as composite services. In their work, five generic quality criteria (service attributes) are considered: execution price and duration, reputation, reliability and availability. Each attribute is assigned a specific aggregation function in order to determine the quality vector of a composite service. Based on this quality model, the selection of services is then formulated as a global optimization problem, which is solved by means of linear programming methods.

Alrifai and Risse [6] proposed a combination of global optimization and local selection in order to increase the efficiency of quality driven service composition. To combine local decision making strategies with global optimization, global QoS constraints are firstly decomposed into local ones. These local constraints are then used as upper bounds for the quality values of elementary services, so that services that violate the constraints can directly be discarded. The quality values of composite services are computed by means of pre-defined aggregation functions. A utility function finally maps the quality values of an elementary or composite service onto a single real value.

In [7], not only non-functional service properties (QoS attributes) but also behavioral service properties (transactional attributes) are considered. A local optimization with respect to common QoS attributes is combined with a global consideration of transactional attributes (e.g., compensability) in order to ensure a reliable execution of composite services. A set of non-functional properties such as execution price or execution duration is defined for elementary services and transfered to composite services by means of pre-defined aggregation functions. Furthermore, user preferences are expressed as weights over the non-functional attributes.

Ben Mabrouk [8] proposed an efficient service selection algorithm which is formed as a guided heuristic. First of all, a set of service candidates for each activity in a composite service is identified based on advertised QoS of services in order to perform a preliminary filtering. In a second step, a selection phase refines this first filtering and ensures the global compliance of user preferences. Services are grouped with respect to their QoS values into a set of so-called QoS levels, which in turn are used to determine the utility of the service candidates. Dependent on the type of composition (i.e. sequence, AND, XOR and loop) aggregation functions determine the quality of a composite service.

In [9], attributes (QoS properties) such as cost or response time of elementary as well as composite services are considered. In order to calculate the values for composite services, aggregation functions for each QoS property are defined in advance, dependent on user QoS constraints and the type of composition (parallel, sequence and combinations of both). Furthermore, a utility value models the user's priority with respect to QoS criteria. The utility value of an elementary service property is created from normalized QoS values and from weights, which reflect the priority of a QoS property. The overall utility value of an elementary service is obtained by summarizing the values of all particular QoS properties, while the overall utility value of composite services correspond to the sum of the utility values of all contained elementary services.

In many cases, the value of a service attribute may be difficult to be precisely defined. To overcome this problem, fuzzy sets and fuzzy logic [10] can be integrated to allow the representation of imprecise and vague information, respectively. Fuzzy sets can be interpreted as a generalization of crisp sets. The characteristic function of a crisp set assigns a value of either 1 or 0 to each element in the universal set, meaning nothing but an element either belongs to the crisp set under consideration or not. By generalizing this characteristic function such that the values assigned to an element fall within a specific interval, e.g. $[0, 1]$, the membership of these elements can be indicated in a more fine-grained way.

In [11], the selection of a single service is formulated as constraint satisfaction problem in the fuzzy domain. The user's preference to a specific service attribute (QoS criterion) is denoted by a fuzzy expression, which is composed of a group of fuzzy sets connected by the logical *and* operator. The user's overall preference to a service is subsequently formulated by connecting the fuzzy expressions of all service attributes with the logical *or* operator. After mapping the service composition problem into a fuzzy constraint satisfaction problem (FCSP), a depth-first branch-and-bound method is applied in order to find a solution.

Another approach that makes use of fuzzy logic was introduced in [12]. In comparison to [11], not only attributes of elementary services are considered, but also those of composite ones. For this purpose, the same non-functional properties of composite services (e.g., price or security) are aggregated according to the type of the composition. In this context, four different types of compositions are differentiated: sequence, parallel, choice and loop. Each combination of non-functional property and composition type is assigned a specific aggregation function. Furthermore, the user's preferences are modeled in the fuzzy domain in terms of fuzzy IF-THEN rules, which facilitates an efficient evaluation of good approximations of service attribute values [13]. The quantitative ranking of candidate services is finally achieved by an inferencing step for all n rules and a subsequent aggregation step of all inferred n values.

Pfeffer et al. [14] also proposed a fuzzy based approach for representing and evaluating service attributes. In their work, the values of service attributes may additionally change over time based on monitoring data of executed services. For that purpose, the monitored values for a particular service attribute are accumulated over time. The resulting distribution is approximated by fuzzy triangle functions and modified, whenever a new value was acquired. User preferences are likewise formulated in the fuzzy domain for each particular service attribute and superimposed with the associated service attribute value by multiplication. The ratio between the multiplication area and the original area is then interpreted as a service's fitness.

VI. Discussion

Apart from the inconsistent terminology, each of the approaches in Section V deals with some of the identified issues of ours mentioned in Section III, while covering parts of the taxonomy briefly described in Section IV.

All approaches are assuming a pre-defined set of non-functional properties, which are either precisely represented in terms of quantitative values or imprecisely represented in terms of fuzzy sets. However, none of the described approaches attends to the acquisition of property values and the resultant influences, except for Pfeffer et al. [14], who additionally consider monitored data for elementary services. Composite services, in turn, are not covered by the work of Pfeffer et al., while all other mentioned approaches incorporate at least an aggregation mechanism for determining non-functional attributes of composite services.

Furthermore, although not explicitly mentioned as such, different objective functions (e.g., costs or availability) are minimized or maximized with respect to specific user constraints. In this context, only single and independent optimization goals are considered, while the optimization problem itself is either solved locally, globally or by some sophisticated combination of both. However, neither ranges of satisfaction (homeostatic methods) nor dependencies between non-functional properties are currently considered in any of these approaches.

The application of utility functions (utility values, fitness values) is indeed a common principle. It can be found in all described papers, except for the work of Lin et al. [11]. However, the usage of utility functions differ among the approaches. For that reason, a clear definition and classification within the service composition context is still missing.

VII. Outlook

Our next step will be to assemble a detailed survey of important existing approaches with respect to the issues pointed out in Section III and to the taxonomy briefly sketched in Section IV. In this context, the taxonomy itself will be elaborated in more detail, and extended, if necessary. Finally, a concrete example for demonstrating our approach will be developed. By doing so, we want to establish a basis for investigating quality based service selection algorithms based on our rating and ranking methodology.

Acknowledgment

This work was partially supported by the German Research Foundation (DFG) within the Collaborative Research Center "On-The-Fly Computing" (SFB 901).

References

[1] W. Tsai, “Service-oriented system engineering: a new paradigm,” in *IEEE International Workshop Service-Oriented System Engineering (SOSE)*, 2005, pp. 3 – 6.

[2] N. Milanovic and M. Malek, “Current solutions for web service composition,” *IEEE Internet Computing*, vol. 8, no. 6, pp. 51 – 59, 2004.

[3] J. M. Ko, C. O. Kim, and I.-H. Kwon, “Quality-of-service oriented web service composition algorithm and planning architecture,” *Journal of Systems and Software*, vol. 81, no. 11, pp. 2079 – 2090, 2008.

[4] H. Chang and K. Lee, “Quality-driven web service composition for ubiquitous computing environment,” in *International Conference on New Trends in Information and Service Science (NISS)*, 2009, pp. 156 –161.

[5] L. Zeng, B. Benatallah, M. Dumas, J. Kalagnanam, and Q. Z. Sheng, “Quality driven web services composition,” in *Proceedings of the 12th International World Wide Web Conference (WWW)*. ACM, 2003, pp. 411–421.

[6] M. Alrifai and T. Risse, "Combining global optimization with local selection for efficient QoS-aware service composition," in *Proceedings of the 18th International World Wide Web Conference (WWW)*. ACM, 2009, pp. 881–890.

[7] J. El Hadad, M. Manouvrier, and M. Rukoz, "TQoS: Transactional and QoS-aware selection algorithm for automatic web service composition," *IEEE Transactions on Services Computing*, vol. 3, no. 1, pp. 73–85, 2010.

[8] N. Ben Mabrouk, S. Beauche, E. Kuznetsova, N. Georgantas, and V. Issarny, "QoS-aware service composition in dynamic service oriented environments," in *Proceedings of the 10th International Conference on Middleware*. Springer, 2009, pp. 123–142.

[9] N. Hiratsuka, F. Ishikawa, and S. Honiden, "Service selection with combinational use of functionally-equivalent services," in *Proceedings of the IEEE International Conference on Web Services (ICWS)*, 2011, pp. 97 –104.

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

[11] M. Lin, J. Xie, H. Guo, and H. Wang, "Solving QoS-driven web service dynamic composition as fuzzy constraint satisfaction," in *Proceedings of the IEEE International Conference on E-Technology, E-Commerce and E-Service (EEE)*, 2005, pp. 9–14.

[12] S. Agarwal and S. Lamparter, "User preference based automated selection of web service compositions," in *Proceedings of the International Workshop in Dynamic Web Processes (DWP)*. IBM, 2005, pp. 1–12.

[13] G. J. Klir and B. Yuan, *Fuzzy sets and fuzzy logic: Theory and applications*. Prentice Hall, 1995.

[14] H. Pfeffer, S. Krüssel, and S. Steglich, "A fuzzy logic based model for representing and evaluating service composition properties," in *Proceedings of the Third International Conference on Systems and Networks Communications (ICSNC)*, 2008, pp. 335–342.

Design Process for Decision Support Services to Support Multiple Agencies

Jari Leppäniemi
Software engineering
Tampere University of Technology, Pori unit
Pori, Finland
jari.leppaniemi@tut.fi

Abstract— **In this paper, we describe a design process for developing deployable information services in service oriented emergency management systems that support multiple agencies. These services aim to support decision-making based on situational awareness information. The service design process has been defined to support information needs in all phases of the lifecycle of emergency management. For this reason, we believe that the defined design process is general and thus it should be a useful base practice for similar tasks in many other operational environments. There are no type limits for the defined services; they could be independent services or compositions of other services. First, we give a short description of the stakeholders of the domain and our approach for useful emergency information. After that we describe the overall service defining process and give a supportive checklist for service designers and managers of different agencies.**

Keywords-emergency management;decision support; service design.

I. THE STAKEHOLDERS OF THE ECOSYSTEM OF EMERGENCY MANAGEMENT KNOWLEDGE

Many kinds of skills and a lot of knowledge are required by several kinds of participants in the management of major emergencies like natural and manmade disasters and catastrophes. These participants are thus vital organic parts of the ecosystem of emergency management knowledge and they are connected to the information and data networks either as consumers or producers – and in many cases as both. Some of the participants represent well-trained permanent solutions to frequent accidents but some of the parties are connected to a particular situation more occasionally. However, during the lifecycle of disaster management, the goals, overall picture and situational awareness must be maintained as well as possible [3]. So, who are the bodies interested in the knowledge of disaster management? Which groups are the stakeholders of the ecosystem? According to the Emergency Information Interoperability Framework workgroup [4], these stakeholders typically represent state-based authorities like rescue, firefighting, policing, health and emergency, safety, etc. Non-governmental organizations, international coordination agents, ICT solution providers, EM and NGO professional and academic communities may be in crucial positions during the emergency management lifecycle, as are often also the public and several types of volunteer organizations [4]. However, their list is not fully satisfactory. It may be completed by adding new groups like the general public, the victims of a disaster (people and industries), relatives of victims, the insurance and financial sector, and national airlines and other carriers. After some regrouping, refinement, and adding of these new parties, the list may be now defined as follows (examples are mainly US-based):

1) State-based emergency management agencies, for example:
   - operative authorities: Police, Fire and Rescue, Ambulance service, FEMA,
   - expert authorities: Weather service, EPA,
   - co-operative agencies: DHS, Army.
2) Domestic civil organizations, for example:
   - Red Cross, National Guard, Salvation Army,
   - Air National Guard.
3) Large international organizations and agencies, for example:
   - UN organizations: OCHA, WHO, WFP,
   - ICRC,
   - NATO EADRCC.
4) International non-governmental organizations, for example:
   - MSF, OXFAM,
   - MapAction.
5) International support services and projects, for example:
   - UNDAC,
   - EU MIC,
   - ReliefWeb, Sahana project.
6) Research and education organizations of emergency management, for example:
   - Universities and colleges,
   - ISCRAM, IAEM, and NetHope.
7) Information and communication solution providers, for example: EADS Ltd, Nokia Ltd.
8) Public: Non-affected citizens, Media.
9) Victims of the disaster, for example:
   - Peoples and their relatives,
   - Affected communities, companies, and industries.
10) Finance and insurance companies.
11) Airlines, carriers, energy companies, etc. needed for help during the incident.

The needs, available resources, and readiness to contribute to management and communication tasks vary between different groups. Fig. 1 describes an abstract view of the growing interoperability approach.

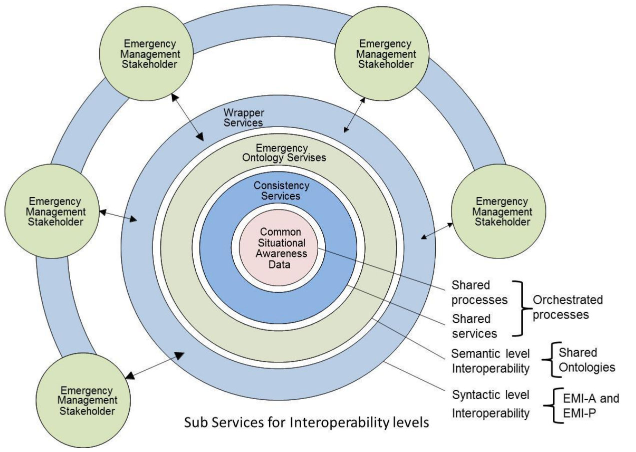

Emergency Management Stakeholder
Emergency Management Stakeholder
Emergency Management Stakeholder
Emergency Management Stakeholder
Emergency Management Stakeholder
Wrapper Services
Emergency Ontology Servises
Consistency Services
Common Situational Awareness Data
Shared processes
Shared services
Orchestrated processes
Semantic level Interoperability
Shared Ontologies
Syntactic level Interoperability
EMI-A and EMI-P
Sub Services for Interoperability levels

Figure 1. Interoperability services for Emergency Management Stakeholders.

The ability to interoperate via information and communication services could be classified using a five-level scale, starting from no interoperability and going up to business process level interoperability. The intermediate levels are syntactic level, semantic level, and service level interoperability. Different types of stakeholders are in the outmost circle in Fig. 1. Their ability to create a common situational awareness is enhanced via special interoperability services that are used to connect different parties at the appropriate level. Thus, business process level interoperability could be reached using all of these different types of sub-services. The essences of these sub-services are:

- Wrapper services for syntactic level interoperability,
- Ontology services for semantic level interoperability and
- Consistency services for service level interoperability.

Process level interoperability is the deepest level of information exchange and availability to communicate with other operating agencies. Cataldo and Rinaldi gave an example of knowledge sharing using the P2P approach and semantic web services [2] and another example of a service-based system meant to support medical information exchange in real time was given by Hauenstein et al. in [5].

In the next figure (Fig. 2), the different stakeholder groups are positioned based on their need for interoperability.

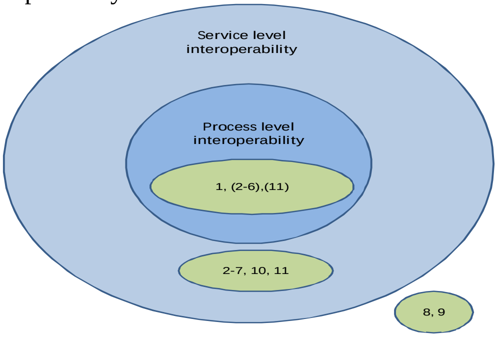

Service level interoperability
Process level interoperability
1, (2-6),(11)
2-7, 10, 11
8, 9

Figure 2. Interoperability needs of the stakeholder groups of emergency management ecosystem.

The key agencies like national authorities (group 1) are the focal point from the interoperability point of view. Their success is based on the practical suitability of the process level interoperability. This could be achieved by orchestrating the common/shared business processes. Domestic civil organizations (2), large international organizations (3), international NGOs (4), international support services (5), international projects (5), and research and education communities (6) could in some incidents also be positioned at the central point of interoperability, or in other words, in process level interoperability. However, normally their interoperability needs are fulfilled by a service level interoperability. This is based on usage of single services independently as a part of their normal operations. Probably most parts of the process level operations might be modeled and orchestrated beforehand and supervised based on the current needs of the incident.

The providers of information and communication technology (7) and finance and insurance companies (10) are positioned to be ordinary service consumers and providers in this model in a traditional way. The public (8), victims and their relatives, affected companies and communities (9) are positioned outside the interoperability service architecture. Their information needs are mainly covered by the dedicated services offered by groups (1) and (2) and also in some cases by groups (3) and (5). There are also auxiliary role stakeholders such as airline and carrier companies (11), whose fluent integration to support emergency management activities in different phases of the emergency lifecycle vary from critical to very important.

II. SYSTEM ARCHITECTURE FOR EMERGENCY MANAGEMENT KNOWLEDGE ECOSYSTEM

Some of the stakeholders are interested in emergency information only at certain times or when certain conditions and limits are met and then only temporarily and/or in an extreme hurry. For this reason, it should be possible to offer open interfaces to stakeholders and organizations that do not take part / operate daily in emergency management. The same applies to the wider public and third parties who might want to implement “ad-hoc” type mash-ups for information sharing and independent communication. Fig. 3 presents an overview of the system architecture for a service-based emergency management knowledge ecosystem [6]. To cite [6] “The main access to the services, i.e., the interfaces of the system are EMI-A and EMI-P. The former is an authorized interface for emergency management services and the latter is a public interface for the general public and non-authorized users of the system(s). Because the concept and function of a service bus is crucial for this ecosystem, there should not be a single point of failure in its functionality. Instead, it should be fault-tolerant and distributable.

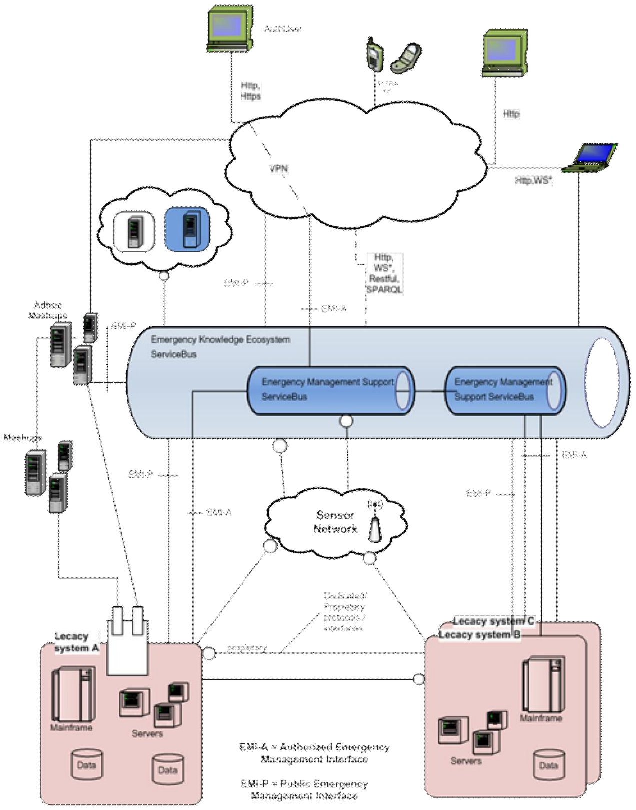


Figure 3. Service-oriented system architecture for Emergency Management Knowledge Ecosystem [6]

The legacy systems (authorities, research organizations etc.) located at the bottom of Fig. 3 interoperate with a large service-based architecture via open standards. They can preserve their proprietary interfaces and they can open new standard interfaces implemented by means of an appropriate technology (WS-*, RESTful, http, etc.) [10]. Sporadic and temporary users can be connected as an information service provider or to a normal user via a public and open service interface like EMI-P in the Fig. 3. A more permanent connection to the ecosystem is established using either the common emergency service bus (EMI-P) or a more controlled interface for authorized users (EMI-A), which supports the coordination and supervision of shared business processes." As mentioned above, an emergency management knowledge ecosystem consists of different types of stakeholders. Only some of them work in close relationship / interaction daily and have thus established routine ways of working and created shared concepts and mutual understanding. The coordination of the shared (and modeled) business processes could be based on the special control and service mediation layer described in Figure 4. Some of the services should be especially designed for supporting the modeled emergency management processes, in other words controlling the performance of the use cases, data and message translations, and routing messages to the correct endpoints.

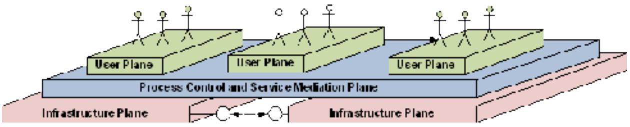


Figure 4. Shared business process supported by service bus and process control engine.

According to Fig. 3 and 4, we have the main user interfaces (the "views" in MVC model) in the user plane, the task-centric supervision ("control" in the MVC model) in the process control and a service mediation plane implemented by a service bus. Finally, the core data access ("model" in the MVC model), which is a composition of different types of entity and utility services and legacy systems, is positioned in the infrastructure plane in order to achieve the interoperability between the different stakeholders, as described in Figs. 1 and 2.

III. SERVICE DESIGN PROCESS FOR DECISION SUPPORT SERVICES

In this section, we will give a proposal for a service design process that is based on process models of the collaboration in major emergencies and disasters. The given proposal is comprised of an identifying process and a detailed checklist due to the various needs of the stakeholders on the field. We also believe that our approach might be valuable complement to domain analysis methods like presented in [8]. The main concern for the intended stakeholders is now a proper situational awareness for a decision maker. The decision points are not limited only to the hectic response phase of the emergency lifecycle. Instead, we believe that our approach might have a much wider applicability during the other emergency management phases (prevention & mitigation, preparation and restoration). The decision points (or moments in time) in different lifecycle phases [1] can be identified from accident investigating reports, by interviewing people and reviewing guideline documents, current practices, plans and process models. All processes must be analyzed and modeled for collaboration from the viewpoint of interoperability. The service design process for decision support services we propose is currently the following:

A. *Identify the main decision points:*

1) Choose some known disaster type.

2) Identify the main stakeholders, their tasks, and concerns.

3) Identify the main decision situations and decision makers at different levels e.g. strategic/operational/tactical (defined in [7]) and in corresponding phases of the lifecycle e.g. preparation, response, recovery, prevention and mitigation (defined in [1]).

4) Analyze whether it is possible for decisions to be also made by others (persons / roles) and/or at a different

moment of time – if the required information and decision-making power were available to them.

5) Document the options using B1 – B4.

B. For each decision point (A3, A4) also try to discover:

1) Which data, information and knowledge are most essential for this situation?

2) Is it possible to support the analysts and decision makers merely by presenting only that information?

3) Do the currently used decision data and information need prioritizing and/or reduction?

4) Is it possible to enhance and/or accelerate the decision-making process by filtering, composing, or otherwise preprocessing the data before showing it to the analysts and decision makers at a tactical, operational, or strategic level?

It should be crystal clear for the service designers that many of the earlier practices may have sub-optimal solutions that are caused by tight organizational limits and/or rigid information ownerships. In order to achieve optimal solutions, the design plans may need several iterations before the desired interoperability for enhanced collaboration and efficiency. Additionally, the cultural differences between the stakeholder groups, organizations, and agencies should be understood and elaborated in order to be able to introduce new services. The service design process could be further supported by detailed guidance and checklists. In the next section, we will give a preliminary checklist of issues that should be noted during the design and deployment processes of the services.

C. Checklist for decision support service designers and managers:

1) What sort of decision has to be made? (What sort of situation is to be followed or anticipated?)

Is the decision to be made tactical, operational or strategic?

Is the decision to be made a routine decision, kind of ad-hoc or is it somehow special?

Is the decision based on guidance or normal procedures or is it based on the experience or intuition of the decision maker?

Is the decision to be made classified or is it public?

Who has to be informed of that decision?

Who has the authority to make this decision?

2) What sort of data and information is needed for the decision? (Which data and information are useful, worthwhile, and wanted?)

Where is that data and information?

Who will have it now?

Who else needs it?

What are the types of that data and information?

How this information could be brought from its origin?

Is it time dependent?

Is it location dependent?

3) What information and data might be or need to be swapped between different agencies?

Who are the peers of the information exchange and data swapping?

Who has the authority to start the information exchange?

What are the reasons for the information exchange?

What are the specific impacts that are tried to be achieved?

4) What are the exact situations and the moments in time when certain data and information are exchanged and needed?

How often does it happen?

Who will be the initiator?

Who owns that data and information?

Who will be responsible for the maintenance of the data sources?

Who will be the overall coordinator of the data and information?

Who will be responsible for the reliability of the data?

Are the access rights of the different user groups to the data clearly defined?

5) What are the main obstacles and barriers to each type of information?

What languages and character sets are available?

To what extent is the data and information confidential?

What are the actual information packages and containers?

What are the technical formats and protocols?

What is the availability of the needed codecs and transformers?

6) What are the types of messages are required for interoperability?

Are the messages:

- Alert messages?
- Summons?
- Request for more resources messages?
- Letters of request?
- Notification messages?
- Announcement and information messages?
- Bulletins and press releases?

What is the confidentiality of the messages?

7) What are the available information and data networks and communication protocols at the decision time?

What are the alternatives for the primary networks (POTS, LA, SMS, CDMA, GSM, UMTS, etc.)?

Are the available networks secure enough for each type of message?

Is it possible to ensure that classified information is also secure on the end users' desktop or handheld?

The checklist is still preliminary and has not yet been tested in a practical situation. However, the list is composed based on the author's experience and research on interoperability issues in emergency management. It is practically impossible to give any references to the items of the list. The questions: What? (Data), How? (Function), Where? (Network), Who? (People), When? (Time) and Why? (Motivation) are elaborated from the different viewpoints that were presented in Zachman Framework by Sowa and Zachman [9]. The framework [9, 11 and 12] has been one of the inspiration sources when formulating the checklist.

IV. SUMMARY

The main stakeholders of the emergency management knowledge ecosystem were described as different groups in a service-based interoperability domain. The main types of subservices supporting different levels of interoperability between the various stakeholders were introduced. The service-based system architecture for decision support services aiming to enhance disaster and emergency management was also briefly presented. Finally, a proposal for a general service design process for emergency management decision support was described followed with a checklist for designers and managers.

REFERENCES

[1] D. Alexander, “Principles of Emergency Planning and Management”. US: Terra Publishing, Oxford University Press, 2002. (ISBN 9780195218381)

[2] A. Cataldo and A. Rinaldi, “Sharing Ontology in Complex Scenario using a Peer-To-Peer Approach”, International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), Vol. 1, pp. 92– 109, 2009.

[3] L. Carver and M. Turoff, “Human Computer Interaction: The Human and Computers as a Team in Emergency Management Information Systems”. Communications of ACM 50, 3, 2007, pp. 33–38.

[4] EIIF (W3C) Incubator Group Report 6 August 2009. Retrieved 15/02/2012. Word Wide Web, http://www.w3.org/2005/Incubator/eiif/XGR-EIIF-20090806/.

[5] L. Hauenstein, T. Gao, TW. Sze, D. Crawford, A. Alm and D. White, “A Cross-Functional Service-Oriented Architecture to Support Real-Time Information Exchange in Emergency Medical Response,” Proceedings of the 28th IEEE EMBS Annual International Conference New York City, USA, Aug 30–Sept 3. 2006, pp. 6478–6481.

[6] J. Leppäniemi, “Domain Specific Service Oriented Reference Architecture (Case: Distributed Disasters and Emergency Management),” International Journal of Computer Information Systems and Industrial Management Applications (IJCISIM), Vol. 4, 2012, pp. 043–054.

[7] OASIS EU-Framework Project 6 (FP6). Retrieved 15/02/12, Word Wide Web, http://www.oasis-fp6.org/

[8] I. Pansa, M. Reichle, C. Leist and S. Abeck, “A Domain Ontology for Designing Management Services,” Service Computation 2011: The Third International Conferences on Advanced Service Computing. September 25-30, 2011 - Rome, Italy

[9] J.F. Sowa and J.A. Zachman, “Extending and formalizing the framework for information systems architecture,” IBM Systems Journal, 31, 3. 1992, pp. 590–616.

[10] List of Web service specifications. In Wikipedia, The Free Encyclopedia. Retrieved 09:31, 04/05/2010, http://en.wikipedia.org/w/index.php?title=List_of_Web_service_specifications&oldid=354335607

[11] J.A. Zachman, “A framework for information systems architecture,” IBM Systems Journal, 26, 3. 1987, pp. 276–292.

[12] Zachman Framework. (2010, April 26). In Wikipedia, The Free Encyclopedia. Retrieved 09:24, 04/05/2010, http://en.wikipedia.org/w/index.php?title=Zachman_Framework&oldid=358410820 or http://www.zachmaninternational.com/index.php/the-zachman-framework

Learning as a Service – A Cloud-based Approach for SMEs

Ileana Hamburg
Institut Arbeit und Technik
Westfälische Hochschule
Gelsenkirchen, Germany
hamburg@iat.eu

Marius Marian
Department of Computers and Information Technology
University of Craiova
Craiova, Romania
marius.marian@cs.ucv.ro

***Abstract*: The paper deals with e-learning within small and medium enterprises (SME). Previous experience of authors within EU-funded projects showed that SMEs need to adopt some organizational and technological measures in order to improve e-learning readiness. In the same time, these measures correspond to the stages in building a lifelong learning (LLL) strategy by using the novel trans-theoretical model presented. The paper then proposes a new approach for LLL within European SMEs, based on cloud computing meant to both facilitate and reduce costs of accessing and management of e-learning strategies, technologies and content. This is a work in progress related with our intent to develop a set of guidelines for SMEs willing to incorporate cloud computing services in correlation with the trans-theoretical model.**

Keywords-e-learning; cloud computing; SME.

I. Introduction

E-learning refers to the support of modern communications and computer-based applications for two fundamental human-development processes, learning and teaching. In our paper, we want to emphasize the benefits of e-learning within small and medium enterprises (SMEs) from the novel context of cloud computing.

It is known [1] that due to the recent economic recession, many SMEs are facing the challenge of shortage in new skilled labor, and in combination with this the inability to realize innovative technological developments. SMEs can tackle this and gain competitive advantages by improving their performance through the use of new e-learning methods. The problem is that most e-learning technologies, methods and strategies have been developed for the needs of large companies, and are not flexible enough to be adapted for the specific learning needs of SMEs operating in diverse economic sectors. Limited capital and know-how resources in SMEs, difficulties to precisely delineate what competences are lacking, and the need for having flexible and efficient learning strategies hinder their employees to achieve a better qualification that helps them to cope with the increased marked competition and client requirements. What SMEs need are flexible learning models enhanced by technology to "reduce training costs per learner in order to be able to train a greater number of employees without increasing spending on training" [2], supporting the development of creativity, facilitating adaptive learning on the job, the deepening of linkage with other knowledge resources within company. The learning processes in SMEs differ from those in schools and higher education because they have to be integrated in working processes, and learning systems have to be implemented into the SME's workflow.

"Today companies are looking for services that provide what they need while giving them the convenience of and time to concentrate more on their business. Not only does cloud computing offer more flexibility than traditional methods, but also gives a business the luxury of letting their employees gain access to information while they are mobile as well as at their desks" said Shuveb Hussain (Head of Cloud Computing and Virtualization Research at K7 Computing Co.) [3]. Cloud computing will evolve from a futuristic technology into a viable alternative, not only for business but also for LLL strategies that have to be integrated with the business ones.

This paper will present work in progress within a cooperation of the LLL study groups of the Institut Arbeit und Technik Gelsenkirchen and University of Craiova.

Section two of the paper presents some learning strategies within European SMEs; section three describes benefits of cloud computing services in connection with social media to improve/substitute these strategies supporting personal, interactive and collaborative learning, and the last section enounces the future work steps.

II. Learning Strategies within European SMEs

E-learning within SMEs was the focus of two previous projects undertaken by the authors: ARIEL [4], [5] and SIMPEL (*SMEs: Improving E-Learning Practices*) [6]. ARIEL was an observatory EU e-learning project; it uncovered the widespread lack of successful take up of e-learning by European SMEs. In SIMPEL, an "optimal model" for the introduction of e-learning in an SME was developed and guidelines for all involved published. The SIMPEL findings have been used for the development of a framework for an LLL strategy in SMEs.

This framework suggests measures to improve LLL readiness and steps to develop LLL strategies. It uses a combination of the trans-theoretical model which is a model for behavior change, and recommendations from ARIEL, SIMPEL and eCASME (*eCApture of SME's training needs and specification*) [7] projects. It uses a top-down and bottom-up approach targeting both the individual and the organization. It aims first to raise awareness of the potential benefit of LLL to the individual and the organization. It aims to change the attitude and behavior of individuals and companies towards LLL. Last but not least the framework should help companies to implement sustainable LLL strategies.

All three projects mentioned earlier – ARIEL, SIMPEL, eCASME, required SMEs to adopt some organizational and technological measures. These measures correspond to the planning, action and maintenance stages in building a LLL strategy by using the trans-theoretical model for organizational behavior change. This is illustrated in Figure 7. At each stage we mention the electronic tools that may contribute to master the stage effectively.

A. Company Situation and Necessary Qualifications

The first step for an SME is to analyze its own business goals, the company situation, and also the difficulties encountered by the company in achieving these goals.

Once the analysis has reached a conclusion, the SME will be able to determine what qualifications are needed by the staff to overpass the identified difficulties. Some methods used to achieve such qualifications include LLL strategy, e-learning, short term qualifications, etc.

The electronic tool useful for gathering documents in various versions and making them available throughout the company (either for everyone or in base of differentiated access rights) is a Wiki portal (either on its own or as part of a Learning Management System – LMS, such as the open-source projects Moodle or Sakai). Additionally, a forum for discussions may found itself useful, again either stand-alone or as part of an LMS.

B. Concept

The next step is finding suitable offers and services for the qualification needs required by the work tasks. This implies determination of learning contents, forms and media used for the LLL strategy, and also identification of relevant knowledge and data flows.

For the internal communication and gathering of information, the wiki and/or the forum mentioned previously may still be used. Feedback sheets and/or databases such as provided in LMS (e.g. Moodle, Sakai, etc.) help gather the information even more precisely.

To find suitable offers, SMEs may use web searching, and particularly, social networks such as Xing, Facebook probably also being useful.

C. Planning

This step implies specifying LLL measures as well as the time, the actors, the technological and organizational infrastructure, and the tools needed for an efficient realization of these measures. This is followed by the preparation of a financial (business) part of the LLL model providing a framework for the economical dimension of the LLL strategy in the company, linking the planning with the process level of the implementation.

Here, an excellent planning instrument for SMEs is for example MindManager, linking mind maps with basic project management features. For the financial planning it is necessary to draw on the data of business or enterprise management software (depending on the size of the company, this may range from simple spreadsheets up to very specific enterprise planning resource planning packages, which vary greatly according to size, branch and needs of the companies concerned.

D. Implementation

In this stage, LLL solutions that correspond to the learning culture of the company will be produced (or purchased and customized) and put in use. This may cover the whole range: from buying standard learning software packages to subscribing to podcasts and other web-based offerings, to running a CoP (*community of practices*), and using/running an LMS with self-developed learning contents. For SMEs, it may be useful to gang together or to make use of offerings through professional associations in order to minimize costs.

A further step may involve tests and certifications. In all likelihood, SMEs will not go further than running online quizzes for testing knowledge. Certification will most likely be taken out of offerings by craft chambers and other officially recognized certification agencies (including universities).

E. Evaluation and improvement

SMEs will certainly want to estimate how effective and financially efficient the training was. A complete evaluation concerns human and financial resources, developed measures, participation, improved knowledge, behavior, competences and expectations of the participants to the LLL trainings, and the observable changes in the company.

This raises the issues of quality control of e-learning and return on investment (ROI). Here it is not possible to point to one or two tools that do it all. Many different parameters may play a role into this [8]. It is important not to follow a narrow, purely economic frame in this evaluation.

III. LLL Strategies For SMEs and Cloud Computing

In the following, we will briefly present some of the LLL strategies used for example by German SMEs ([9], [10], [11]), and we will then propose a new approach in which cloud computing can be employed by SMEs in their LLL processes.

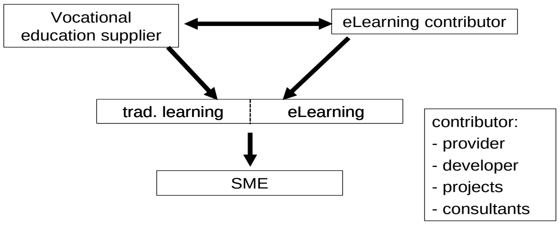


Figure 1. Cooperation with vocational education suppliers

In Figure 1, SMEs which are consumers of learning related products and services including e-learning, cooperate with vocational education suppliers to disseminate these, and to achieve their qualification needs.

Figure 2 exemplifies another strategy in which distributors of learning programs offer their services as a subscription. The point to be emphasized here is the distribution manner.

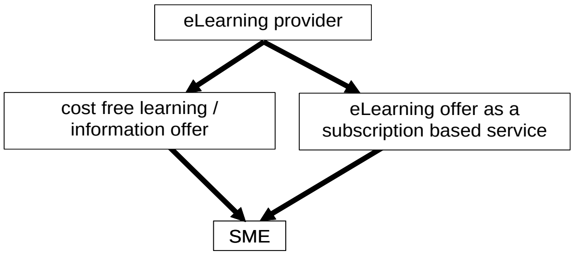

eLearning provider
cost free learning / information offer
eLearning offer as a subscription based service
SME

Figure 2. Subscription based services

The distributor in this business model gives first a trial offer or a free of charge basic information or learning offer. So clients know the services or products and request the suitable services for a fee.

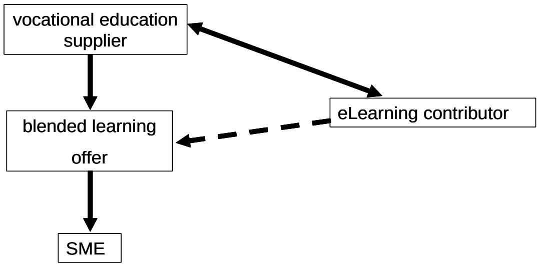


Figure 3. Refining of face-to-face courses

In Figure 3, the providers of learning services offer their services to the providers of traditional vocational education helping thus SMEs to build an integrated LLL strategy.

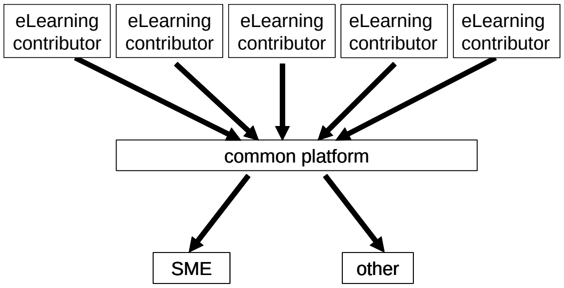


Figure 4. Content syndication model

It is a common practice to have different e-learning distributors pooling their contents on a common platform as depicted in Figure 4. This enables a common marketing and common standards (for certifications for example). The common platform can be established as a brand, after a while. The clients SMEs have then the possibility to access one platform for all courses. The content syndication model is designed mainly for smaller and niche/topic-oriented providers of learning content.

Another LLL strategy is that illustrated in Figure 5, where a franchisor offers a complete package of services to his franchisees. SMEs from one sector work together, in informal or formal ways, e.g. in associations, to develop and to use e-learning applications, contents, platforms or courses together (as seen in Figure 6).

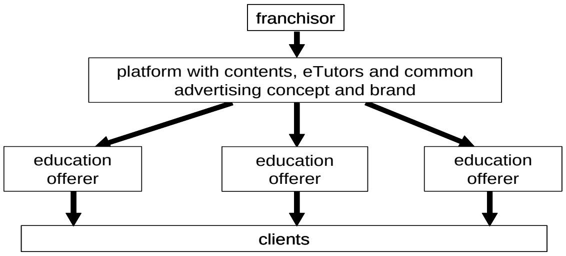

franchisor
platform with contents, eTutors and common advertising concept and brand
education offerer
education offerer
education offerer
clients

Figure 5. Franchising of e-learning

Among the LLL strategies presented so far, the two most used in the German SMEs community are *subscription-based services* and *content syndication model*.

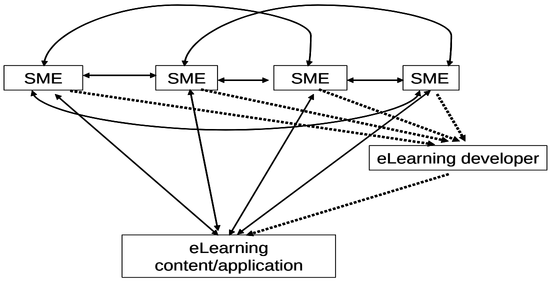


Figure 6. Sector courses developed by collaboration between SME

Cloud computing is a novel interpretation of sharing resources over Internet, on-demand and on a pay-per-use paradigm. These resources are diverse, ranging from software applications to data to computing and internetworking infrastructures. The cloud is seen as a natural evolution since it does not alter fundamentally the existing technologies; rather, it is growing on a successful collaboration/combination of them. Computing as a service and not as a product represents in fact the next public utility. SMEs could drastically reduce the costs pertaining to their LLL strategies and processes by adopting the cloud.

A first application of cloud computing is in the trans-theoretical model (see Figure 7). Some IT support or the entire IT infrastructure within the stages of the model can be moved into the cloud. Each stage as we have mentioned is using a certain number of software tools and these tools are better off in the cloud (due to reduced budget costs and delegated administration). In the traditional approach SMEs have to provide for the e-learning infrastructure with both hardware and software components. This means investing in the e-learning system: capital, human resources, etc. Traditionally, SMEs have developed their own intranets in which web-based e-learning systems were deployed. The economic crisis has redirected SMEs in finding alternatives to this paradigm. One of them is to shift towards incorporating cloud services in the e-learning process, and occasionally, aggregate in communities of sector-orientated employers. This could be the case for SMEs playing in the same sector of activity, having a consolidated economic position on external markets that agree to participate into a win-win approach with similar co-national actors for improving their employees' skills and qualifications. Similar actors are those that have interlocked business relationships and shareholdings. Common business interests and common needs could lead to a common e-learning strategy. Obviously, the main reason of moving the individual learning management system into the cloud remains the reduction of set-up, maintenance and evolution costs. One interesting consequence of this shift into the cloud is if we consider an association of employers and the planning step of the trans-theoretical model. The various companies' statuses could drive a joint effort for establishing a common list of needs of qualifications for that business sector, and then in the successive steps of the model, to prepare, purchase or share knowledge and competence-building processes within a guild of SMEs. Additionally, migration of the traditional learning system into the cloud demands for a common set of good practices. SMEs, knowledge providers and CSPs all benefit if such an informative guide would be available. The NetKnowing project is intending to develop such a set of best practices for European SMEs willing to move their learning management systems into the cloud.

The cloud is seen in literature in three major classes of services: infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS). IaaS is a complete virtual machine running a specific operating system, in most of the cases acting as a server. For an e-learning environment, the cloud approach means that the cloud service provider (CSP) is in charge with delivering the infrastructure of the learning system and its operational management. The CSP is also in charge with the customization of the learning solution for the particular constraints of the individual SMEs. These may concern primarily scalability (variable volumes of managed data) and high-speed computing. In practice, large CSPs such as Google Cloud Platform, Amazon Elastic Compute Cloud, and Microsoft Azure are already offering support for hosting e-learning solutions. Top open-source learning / content management systems allow configuration and operation into the cloud (e.g. Moodle directly on Amazon Web Services via BitNami, Blackboard Moodlerooms solution, SCORM Cloud for Sakai). But these commercial solutions need to be compared and evaluated taking into account various budgeting and business constraints of SMEs. A guideline for either individual or community of SMEs seems thus necessary.

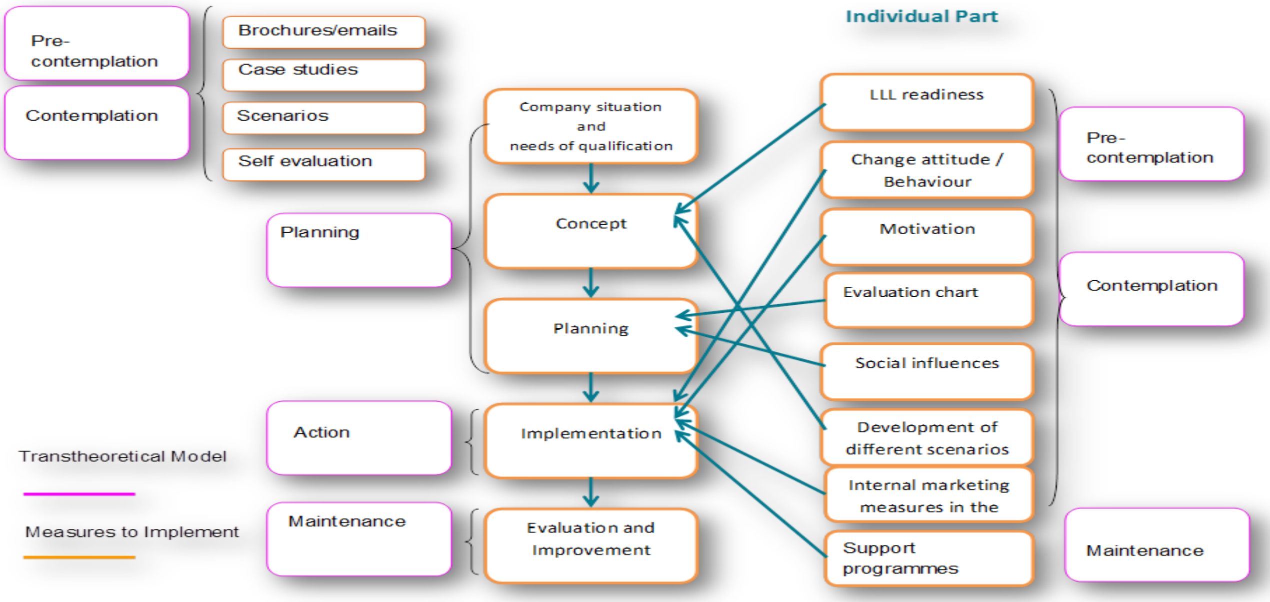


Figure 7. The trans-theoretical model

PaaS is is an extension of the IaaS to accommodate the middleware and to improve the performance in using it. It may be for example a web-based e-learning development platform containing the web/application server, the integrated development environment, the associated database and all additional utilities for development and testing. PaaS offers SMEs the possibility of acquiring on-demand usage-time for different types of software services. This includes a wide range of applications: office tools, graphic utilities, data storage facilities, etc. SaaS is dynamically scalable, device independent, and most of the applications are collaborative, allowing thus multiple users to share documents and work on them concurrently. Adding social media services through SaaS can only enhance this collaboration. A trivial example here, concerning office tools for SMEs, is Google Docs.

The deployment models into the cloud range from private cloud (an extension of the enterprise Intranet), to community-based cloud (the participants to this model are various organizations with a common mission or the same business goals), to public cloud (the CSP makes available its services to any customer via Internet), up to the hybrid cloud (actually an intersection between the previous deployment variants).

For our investigation, it is clear that SaaS and PaaS are the most suited categories for SMEs since the supporting IT instruments are out-sourced and need no longer be managed in-house. Taking for example the financial aspect of the software licenses this is no longer the task of the SMEs (instead CSPs will install, configure, update, fix and administrate them).

Security risks associated with the sensitive e-learning content and the related access control will be handled by the CSPs. Having the data stored into the cloud is a disadvantage for attackers since sensitive data (such as evaluation tests, exam answers, quizzes, etc.) is not as easy to locate on a particular hardware resource as is in the traditional intranet approach. Also, disaster recovery is no longer a problem of the SMEs instead this is delegated to the CSP. When client hardware terminals fail, the CSP in conformity with the service level agreement will ensure that the restoring time will be minimal and whether data restoring will be complete. Correlated with this, the SME's might get interested in aspects such as business continuity and data availability when moving the e-learning process into the cloud. Therefore, SMEs must establish and understand well the procedures necessary to have the data and the e-learning system available under any critical circumstance. Our current work concerns development of cloud adoption guidelines for SMEs addressing such sensitive topics.

For the subscription-based services, SMEs and e-learning providers benefit financially by moving the learning offer into a SaaS-like cloud service instead of buying a product (such as for example, the supporting software applications that need to be hosted and administered on the SMEs hardware/network). The same applies for the shared platform within the content-syndication model.

IV. Conclusions and Future Work

This paper describes a work in progress related with our intent to incorporate cloud computing services for the referred two most used LLL strategies in two on-going projects, NetKnowing 2.0 [12] and ReadiSME [13]. This means changing the technological paradigm for the current e-learning platforms and LLL strategies used by SMEs.

The goal is to investigate how much the cloud and social media are beneficial for both SMEs and e-learning providers, mainly in reducing costs and improving knowledge transfer. The next step is to describe and discuss scenarios with SMEs representatives about the proposed application in this paper, and to test them in 1 – 2 SMEs from project partner countries.

A guideline will be developed referring LLL in SMEs and cloud computing with precise indication how to use the cloud in different stages of a LLL strategy. Related security and privacy risks will also be addressed.

Acknowledgment

This work was supported by the strategic grant POSDRU/89/1.5/S/61968, Project ID61968 (2009), co-financed by the European Social Fund within the Sectorial Operational Program Human Resources Development 2007 – 2013.

The studies have been carried out within the innovation-transfer projects ARIEL, SIMPEL and eCASME supported by the EU LLP Leonardo da Vinci sub-programme.

References

[1] European Commission Report, "Assessing the Performance of European SMEs", September 2010, available on-line at http://ec.europa.eu/ enterprise/magazine/articles/smes-entrepreneurship/ article_10581_en.htm

[2] Steve Fiehl, Michel Diaz, and Antoine Solom, "The first e-learning barometer in Europe", October 2011, available on-line at http://www.crossknowledge.com/en_GB/elearning/media-center/publications.html?mictrl=cklgWhitepaper/show&sid=survey-e-learning-europe

[3] Michael Byrne, "Planning Gowth", January 2011, available on-line at http://www.smeadvisor.com/2011/01/planning-growth/

[4] ARIEL project, "Analyzing and Reporting the Implementation of E-Learning in Europe", March 2012, available on-line at http://www.ariel-eu.net.

[5] D. Beer, T. Busse, I. Hamburg, U. Mill, and H. Paul (eds.), "eLearning in European SMEs: observations, analyses and forecasting", 2006, pp. 2 – 6, Münster, Waxmann.

[6] D. Beer, T. Busse, I. Hamburg, and C. Oehler (eds.), "Improving e-learning practices in SMEs", Proceedings of the SIMPEL final conference, 2008, pp. 3 – 7, Brussels.

[7] eCASME project, "eCApture of SME's Training Needs and Specification", March 2012, available on-line at http://ecasme.amt.ul.ie.

[8] C. Porco, "Measuring the success of e-learning initiatives", available at http://www.prescientdigital.com/articles/learning/measuring-the-success-of-e-learning-initiatives/

[9] T. Hall and I. Hamburg, "Learning in social networks and Web 2.0 in SMEs' continuing vocational education", International Journal of Web Based Communities 5 (4), 2009, pp. 593-607.

[10] St. Engert, I. Hamburg, and J. Terstriep, "Promoting online education for new working environments in companies", in U. Demiray, S. Sever (eds.): Marketing online education programs: frameworks for promotion and communication. Hershey, 2011, PA: Information Science Reference, chapter 23, pp. 337-358.

[11] I. Hamburg, "Learning solutions and social media based environments for companies", in Life long learning for competitiveness, employability and social inclusion, International conference, 11-13 November 2011, Craiova, Romania, Editura Universitaria, pp. 31-37.

[12] NetKnowing 2.0 project, March 2012, webpage available on-line at http://www.netknowing.com.

[13] ReadiSME project, "Lifelong Learning readiness for SMEs", March 2012, webpage available on-line at http://www.readisme.com.

Analysis and design in providing a robotised cleaning and validation system for hospital environment

Alessandro Carlini, Giacomo Saibene, Valerio Turri, Giuseppina Gini
ASP, Politecnico di Milano
Milano, Italy
alessandro.carlini@u-bourgogne.fr
giacomo.saibene@asp-poli.it
valerio.turri@asp-poli.it
gini@elet.polimi.it

Alessandro Barardi, Alessandra Lo Moro, Michael Boris Mandirola, Bartolomeo Montrucchio
ASP, Politecnico di Torino
Torino, Italy
alessandro.barardi@asp-poli.it
alessandra.lomoro@asp-poli.it
michael.mandirola@asp-poli.it
bartolomeo.montrucchio@polito.it

Abstract— **Health care and hospitals services could greatly benefit from technological innovations in many fields beyond the disease treatment itself. For instance the cleaning process deserves a significant role among the services that a hospital must deliver. In this study we explore the fully automated and the traditional (human-based) cleaning protocol, we define its main components and discuss steps and benefits in the introduction of a partially automated solution based on a new robotic system.**

Keywords- Robotics, Cleaning services, Protocol Optimisation, Hospital

I. INTRODUCTION

Activities performed by humans have always been affected by innovations, such as the introduction of automation processes. Health care and hospitals greatly benefited from all these innovations, leading to a far greater awareness of the medical possibilities in treating diseases. However, the quality of health services is definitely more than just the medical aspects and the cleaning process deserves a significant role among the services that a hospital must deliver. This service, being still mostly manual, shows some criticalities that could be overcome by introducing a system with a higher level of automation.

Cleaning is a process composed by different tasks; essentially we identify two sub-processes: the cleaning activity itself and the verification of its effectiveness (cleanliness verification task). Given the current available technologies, introducing an automated cleaning system is still an uneconomical option; the inefficiency of the robots leads to a condition where costs are higher than benefits. On the other hand, designing and introducing an automated cleanliness verification system is a viable option.

Therefore, we argue that the cleanliness verification system can be introduced in a short term perspective, while the cleaning system itself only in a long term one. Accordingly, we focused on two core issues:

(i) the cleaning task in the long term, which focuses on the organisation and the sizing of a swarm group of robots;

(ii) the verification task in the short term, which focuses on the issues related to navigation and measurement of the cleanliness.

The cleaning task requires a system composed by simple automated units cooperating together, whose control system represents the critical issue. The cleanliness verification task may be performed by single automated units, through the adoption of simple positioning and moving methods and of basic sampling systems.

In the following Section 2 we start exploring the cleaning problems in a hospital. This chapter is based on a scenario developed with the San Raffaele Hospital (HSR) in Milan.

In Section 3 we recall the available automation and robotics technologies and propose our scenarios for solutions. Two solutions are devised: a short term one, where only the certification of cleaning is robotised and a long term one, where the cleaning itself is robotised.

In Section 4 we technically devise methods that could exploit robots to ensure almost full, verified coverage and methods to verify the sanitisation condition and to support the periodic control of the cleanliness.

In the Conclusion section we discuss about how to assess the sustainability of the solution and the consequent practical benefits.

II. APPROACHING THE CLEANING PROBLEM IN A HOSPITAL

The hospital is a complex institution. There are many problems and criticalities that are not exclusively related to medical aspects, such as the handling of objects (e.g. drugs and meals), the moving of patients, the transmission of data and the cleaning process. Among these, in accordance to the needs of our main stakeholders, we chose to innovate the cleaning process by studying the introduction of an automated system for the floor cleaning with a multidisciplinary approach.

Every day many people (patients, relatives, doctors and other workers) enter and exit hospitals. Patients under medical treatment and often with weakened immunity or contagious diseases, could be prone to infections. Hence, cleanliness deserves a special attention in hospital environments and any discussion involving it must deal with many different dimensions: economical, environmental, social, health and quality of service.

Most cleaning tools are driven by the operator who has complete control and is personally responsible for the result.

The most common machines are the vacuum cleaner, the washer-drier, the applicator machine and the wet vacuum cleaner, usually all of them operated by humans. There is not a big difference between devices used in hospitals or in other contexts, besides the products that are chosen as detergents and disinfectants.

There are many actions that could be taken to improve the situation from the points of view:

- economical: saving on materials, chemicals and on working time required by humans;
- environmental: optimising the dosage of chemicals, favouring green solutions;
- social: making dangerous and low-qualification jobs unnecessary, such as the janitorial services, in order to promote the creation of high-qualification jobs;
- quality: improving the cleaning system contributes to quality improvement.

Our analysis in a real hospital was possible trough a project of Alta Scuola Politecnica and the HSR. HSR has a prominent position at national and international level and it is qualified as a high specialisation hospital for the most important diseases. Some issues about the problems of cleaning and the most interesting topics emerged from meetings with HSR. It was clear from them that the entire cleaning system (route, dispensing of detergents and disinfectants, cleaning frequency) depends on the staff who is also responsible for potentially neglected tasks. There is a detailed planning of the cleaning procedures that is the guideline for operators in order to reduce processing errors. The hospital has been subdivided into areas that correspond to different levels of criticality and each of these has its own colour. This feature reflects a different cleaning protocol and the frequency in which the task must be executed. The size and the complexity of the system make the testing and the control more difficult. Specific operators perform the quality control also by visual inspection to establish the effectiveness of cleaning.

When considering automated systems, we observe that there are no examples of a completely automated cleaning system in structured environments, such as industrial buildings. Moreover we need to distinguish between two parts of the cleaning process; the cleaning task itself and the ex-post verification of its effectiveness. Accordingly, the implementation of an automated system involves two areas, in which significant improvements may be achieved: removing dirt and measure and localise the dirt.

III. THE SOLUTION DEVISED

Looking on the market, the cleaning machines fall in two categories: human operated or automatic. Today most of the machines for profound cleaning are industrial machines very heavy and manually driven. The automatic solutions are for small systems for home. Moreover, secondary functions as the removal of the dirt from the containers are still manual.

The analysis outcome of the described situation lead us to conclude that an automatic cleaning system cannot be proposed today in a hospital environment, both for the high costs and for the unfulfilling effectiveness. Even the use of the available automatic systems to cooperate in floor cleaning is still unfeasible and can lead to costs higher than the benefits.

On the other hand, the design and the introduction of a semi-automated cleanliness verification system is a viable option, since it needs a less complex technological platform.

Therefore, we concluded that the introduction of the cleanliness verification system can be done in the short term, while the introduction of the full cleaning system is to be done in the long term. Accordingly, we analysed these two phases of the cleaning process from two complementary perspectives: the short-term and the long-term. Our long-term solution focuses on the organisation and the sizing of a swarm of robots, while our short-term solution focuses on the issues related to navigation and measurement of the cleanliness. In devising a solution, we may say that part of the complexity and of the time of cleaning can be reduced if we exploit a team of robot for performing the floors' cleaning in hospitals.

Our vision regards an automated cleaning and validation system that is economically and socially sustainable within hospital environments that are commercially available. Hence, we do not analyse the co-building of an integrated hospital to such cleaning solutions, but we assume the feasibility constrains of current hospital structures as given and we analyse the case in which the cleaning system is introduced into an already existing building.

IV. LONG TERM SOLUTION

Our long-term solution can propose a team of robots that are able to clean and at the same time verify the effectiveness and efficiency of cleaning carried out.

From this long-term perspective we developed an optimisation-based approach inspired by the work of Altshuler et al. [1], in which a multi-agent system, the swarm described by the authors, plays a central role. The swarm is defined as a decentralised group of multiple autonomous agents, that are simple and have limited capabilities. A key principle in the notion of swarms, or multi-agent robotics, is the simplicity of the agents in the aspects [3] of memory resources, sensing capabilities, computational resources. Each robot of the team does not know the global system.

Regardless of the improvement in performance, such swarm systems are usually much more adaptive, scalable and robust than those based on a single, highly capable agent, if they are properly sized.

The cleaning problem assumes a grid, partly dirt, where the dirty part is a connected region of the grid. On this dirty grid region several agents move, each having the ability to ‘clean’ the place (‘tile’, ‘pixel’ or ‘square’) it is located in, while the goal of the agents is to clean all the dirty tiles in the shortest time possible. These agents work in a dynamic environment in which a deterministic contamination spread is simulated every d time steps.

A. Defining the number of robots

A way to decide whether k agents can successfully carry out their cleaning task is to provide a lower bound valid for each cleaning protocol [2]. We want to find the minimal number of agents necessary in order to carry out their task with a specific initial dirty area and a certain contamination spread step *d*. Due to the dynamic nature of the problem, we introduce a shape factor which takes into account that the contaminated region can change during the cleaning process. Our analysis is carried out from a mathematical and analytic point of view, starting as in [2, 3, 15, 16] with agents moving in a dirty region, each having the ability to clean the place it is located in.

The cooperative cleaners' problem has been studied by Wagner et al. [16] assuming as world model a regular grid of connected rooms (pixels), a parts of which are dirty. The dirty pixels form a connected region within the grid (matrix). Agents are able to move on this region and to clean the "dirty pixels". Due to the dynamic nature of the problem, a deterministic evolution of the environment is adopted. The goal of the agents is to clean the spreading contamination in as little time as possible.

The identification of the contamination status of the tile where the agent is located is carried out through a sensor installed on the platform. Moreover each agent is able to know the condition of tiles of the 8-Neighbours.

According to [16] we have developed a method to compute the lower bound depending on a number of parameters. This bound is plotted in Fig. 1 over the size of the contaminated region as a function of time, given an initial contaminated area (tiles), S_0 = 3000, dirty contamination spreading latency d = 3; 4; 5; 6 and different various number of cleaning agents k.

The lower bound over the cleaning time highlights the minimal time necessary for the team of robots to accomplish their job. The threshold, k_T , on the number of the agents depends on d: it is the minimum k that allows the system to complete the task in a finite time. If k is higher than k_T the objective can be reached. If k is lower than k_T the result cannot be reached even the time is infinite.

There is also an upper bound due to the fact that when the number of robots increases the probability of interference among them increases too, making it more difficult to reach the goal.

B. *Defining the spreading of dirt*

Those results are based on a deterministic approach regards to the contamination spread. It would be interesting to introduce a stochastic variant to enhance the dirt diffusion expression in order to make the model more realistic.

C. *Defining the control policy*

We can think about three different controllers for the robot cleaning system.

Firstly, a system that is activated when a human responsible decides that the hospital needs to be cleaned; the system works until the whole area is cleaned and then stops.

A second possibility is to integrate the probabilistic calculus of the diffusion of dirt into the cleaning plan. Namely, knowing when a spot has last been cleaned allows to calculate when it cannot be considered clean anymore, with a probability higher than a given threshold. In this approach we have a constantly active cleaning system and a guaranteed cleanliness level.

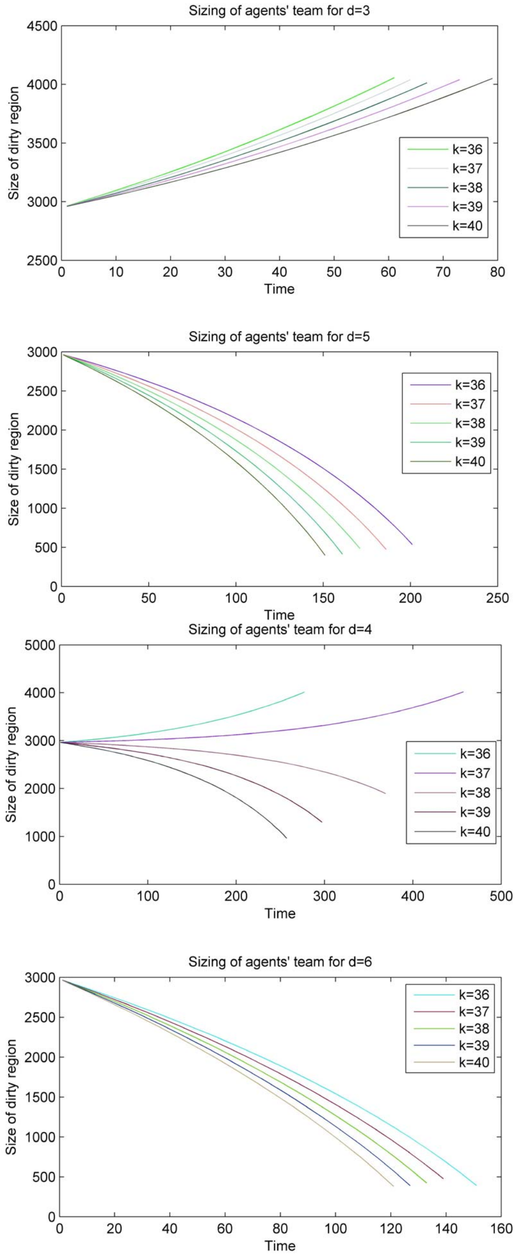


Figure 1. Evolution of the dirty region size (y axis) over the time (x axis) in function of the contamination spread step (d) and the number of cleaning agents (k). From top to bottom: given the same initial condition, each diagram presents the dynamic evolution respectively for d = 3, 4, 5 and 6; as evident, different values of the number of cleaning agents (k), in each condition, lead to accomplish the task, or to diverge from it, with different velocities. Logically the number of agent cannot grow infinitely ignoring the interference due to size and shape of the surface and agents.

A third possible approach is the integration of the cleaning system and the verification system, with verification agents that would be active at the same time as cleaning agents and that would provide them with cleanliness information: when a section is found dirty, the information is changed in real time in all the agents.

The implementation of those controllers is in the state of the art of robotics, not the construction of the real cleaning agents.

V. SHORT TERM SOLUTION

We outlined a solution for the verification system. We used a more practical approach, testing the positioning and navigation strategies as well as the methods for sampling, measuring and stocking the dirt on the floor.

The input of the navigation system is a complete map of the room and the fixed obstacles. The navigation algorithm guarantees that the robot collects dust samples in points randomly and homogeneously spread over the environment. At the same time the positioning system ensures the correctness of the sample coordinates.

A. Methods to measure cleanliness

We can divide the cleanliness measurement methods in two categories: the direct methods, applied directly to the part we are examining and the indirect methods, that require an analysis of the collected contaminants.

Different technologies support the dust investigation. The main traditional method is just visual inspection, that relies completely on human observation. Other methods measure the dirtiness in a scale using Bacharach scale (a paper that has different grey levels), or observe the movement of a drop of water (Water Break Test or Contact Angle test), or use UV spectroscopy. A simple method is the Scotch Tape Test: a strip of transparent tape is firmly pressed upon the surface, then it is displayed on an appropriate contrasting background.

On the other side we have to mention some semi-precise or precise methods that are mostly based on optical or electrochemical processes and generally require more laboratory equipment, time and cost. They are not necessary in our intended task of measuring the cleanliness of a floor. What is important is understanding whether the floor needs to be cleaned again and whether the cleaning process is performed correctly or not.

Therefore in the present work our choice is restricted to the gross verification methods. We consider the possibility to perform a deeper analysis for further certification needs. So our verification protocol is based on three main actions:

(i) take a sample of the contaminants
(ii) make the analysis
(iii) store it into the robot for further analysis.

The first analysis, more trivial but giving fast results is performed on the robot. In this case it is necessary to carefully consider the use of solvents, used in most of the indirect methods. A further problem lies in the fact that the floor should be still clean after the process. Other difficulties are related to the storage of liquids that would require a different recipient for each sample and more careful management if on board. Finally an important factor is the protocol proposed to the human resources to manage the system. Obviously one of the objectives of the proposed automation is to reduce the need of human labour compared to the current system (visual inspection, a simple method) while increasing the quality. So the automated method is expected to be more repeatable and robust than a method only dependent on human factors.

The question whether to evolve an integrated device (as a moving base) or design it ex-novo is a secondary choice. Results of our analysis show that today does not exist a device totally fulfilling our requirements, so it would be preferable considering the design of a new dedicated system.

B. Hardware equipment for the robot

Between the existing dirt investigation methods, only the principles of the Bacharach scale and the tape test (applied in an automated way) seem to be the solutions that comply with all the identified necessities.

Summarizing, the robot base has to carry an equipment composed of.

- the sampling device for the acquisition of the dirt level
- the measuring device to measure of the dirt level
- the stocking device to preserve the samples
- a battery
- motors to roll the tape

A small robot (about 50 cm in size) can move better near the obstacles. The robotic base should be equipped with sensors to detect obstacles (sonar, bumper) and encoders on the motors in order to use odometry.

The measurement is based on the tape test. An adhesive tape is rolled on two supports: on one roll there is the unused part, on the other there is the tape sector that has already passed through the measurement part. Inside each roll, there is a motor that moves it provoking the rolling/unrolling of the tape. The sampling is done by the contact of the tape with the floor. The measurement system is composed by a photodiode and a laser light source. The output consists of a series of zeros and ones: a zero means that the laser light was not able to achieve the photodiode whilst a one represents that the tape was transparent enough to allow the light to polarise the photodiode. The number of “1s” is the cleanliness level.

The sample, after the measurement, is stored, keeping a record of where the sample was taken. A tape of polytetrauoroethylene (commonly known as teflon) could be put aside the adhesive tape in the destination roll and the two tapes could be wound together. The robot keeps trace of its movement on the map and of the places where a sample has been collected.

C. Positioning and navigation

Now we focus on the choice of the positioning system. Most of the commercial robots used to clean the floor are not equipped of a positioning system, but typically use heuristic algorithms. Unfortunately this approach cannot be applied to our project. The dirt detector robot should know where it picks up each sample. From the different technologies [5] to implement a positioning system, we take odometry, computed from encoders on the wheels and the recognition of visual markers.

The navigation algorithm [5, 6] can be developed from a sequence of target points where the robot has to collect samples, using a modified version of the covering algorithms. The navigation algorithm of election for this robot is map based. It is able to manage critical situations such as fixed and mobile obstacles using on board sensors.

D. Experimental proof of concept

We have developed a proof of concept of the complete navigation method using the iRobot Create [12] and the Matlab environment [14]. iRobot Create (see Fig.2) is the research-orientated version of the more popular iRobot Roomba: it has the same robotic base of it, but it is not equipped with the cleaning system. It is possible to control it with a PC using a serial communication and the Create Toolbox Interface that allows to use a Matlab real-time script to control it. There is a Simulator to test the algorithms. On the hardware side, it is equipped with different sensors (bumpers, encoders) and through the PC it is possible to connect additional external devices (as cameras).

Figure 2. Virtual rendering of the robot at Politecnico di Torino

We have added a webcam pointing to the ceiling, where we put black and white markers. The images are analysed by RoboRealm [http://www.roborealm.com]. Given the position and orientation of the markers, the information coming from RoboRealm is used to compute the position of the robot using matrix calculus. This procedure theoretically guarantees a localization error < 1 cm; also small imperfections in the camera orientation could produce an error of 5 cm in positioning.

VI. Discussion and Conclusion

The quality of health services is definitely more than just the medical aspects and the highest cleaning condition deserves an essential role among the services that a hospital must deliver.

A preliminary analysis shows that cleaning condition is depending on many factors, both environmental and cleaning-task-related. The environmental factors are several in number and extremely variable; this condition emphasizes the impossibility for an optimization with a “time-scheduled” managing system and leads to consider an “on–condition” decision system to be the more appropriate to manage the cleaning activity. The cleaning-task-related factors are depending on the human factor in task accomplishment, on the protocol type and on the equipment management. The *traditional* cleaning process is still completely “human-based” (i.e. manually performed by cleaning staff). This solution implies relevant costs for the service and requires attention on protocols and verification of results to ensure health conditions both for the users and for the medical and cleaning staffs (not only dirt, but also detergents could be dangerous). The risk factors in the context of cleaning are present in all stages of the process. By the social point of view a robotised solution would improve the quality of the jobs making low-qualification ones unnecessary and improving the safety at work.

Market analysis and available technology suggest that a completely robotic cleaning system still remains an uneconomical option: the inefficiency of the robots leads costs be higher than benefits. Although robotising cleaning tasks appears to be less efficient than the *traditional* process, our analysis shows the automation of the *verification task* shall both improve the whole traditional protocol and lay the foundation for new future solutions.

A system based on a swarm of robotic agents was simulated and analysed, to reveal the most relevant characteristics and to evaluate possible novelties in cleaning protocol. Simulations show the existence of an optimised number of agents, depending on the surface characteristics and the accomplishment capabilities of each unit.

Successively the system was enhanced introducing a verification robotic unit.

This new element first implies a changing in the cleaning task management, now based on *“on–condition”* philosophy. Practically the availability of cleaning conditions data supports a new optimisation concept both for the intervention of the cleaning agents and in detergents handling and supplies. This reduced usage of the equipment leads important economic and ecological benefits.

A subsequent analysis highlighted that we obtain the same benefits by introducing the verification robotic unit also in an existent *traditional* cleaning protocol.

Today some cleaning protocols have been expressly developed to minimise the errors dues to the human factor (usually for the industrial field and somewhat dedicated to the hospital environment) but no optimisation is possible in this type of time-scheduled protocol.

Practically speaking, the cleaning process is a stand-alone task, performed far from the controller’s watch. Generally a person is charged of the occasional verification task, introducing besides a further incertitude factor in the system; obviously this person isn’t a cleaning-staff and (generally) he is an employ or a company manager, with a hourly rate higher than the cleaning-staff (i.e. each time-units dedicated to the verification task is more expensive than each one dedicated to the cleaning action).

In short, the introduction of the automated verification in the *traditional* process improves the system awareness and makes possible the “on-condition” optimisation.

The increased simplicity and the task automation leads to carry out more frequents checks and therefore to have more accurate data and a better knowledge of the *ambient system* and its dynamical conditions. The awareness of the real and detailed conditions allows a more incisive intervention and supports a better scheduling of the activities. The availability of new on-going data allows a continuing improvement of the protocol.

Finally, the faster discovering of anomalies and inconveniences and the improved effectiveness related to the more precise intervention, bring not only economical and ecological benefits, but also a significant enhancement of the cleaning condition.

We conclude this section with some economic considerations and a challenge for the future developments. The main outcome from our economic analyses is about the ineffectiveness and the expensiveness of the solution composed by cleaning robot agents. The immature present technology and the market conditions are the main limits preventing the success of a full-robotic cleaning stuff. Some keywords to describe these limits are: slowness in task performing, cost of energy, maintenance and productivity. Despite the estimated hourly costs are sensibly more favourable for the robotic solution (1,69 €/h) than the human-based solution (12.5 €/h), today the productivity represents an insurmountable obstacle (30 m^2/h and 400 m^2/h, respectively) and limits the interest towards this innovative solution.

ACKNOWLEDGMENT

The present work was developed within the framework of the ASP - Alta Scuola Politecnica. We thank the ASP school and faculty for the guidance and the support in project developing.

We also address a special thank to Alberto Sanna and the e-Service of HSR hospital, for the valuable introduction within the reality and the problems of the hospital environment.

REFERENCES

[1] Y. Altshuler, A.M. Bruckstein, and I.A. Wagner. “Swarm robotics for a dynamic cleaning problem”. IEEE Swarm Intelligence Symposium 2005 (SIS05), pages 209-216, 2005.

[2] Y. Altshuler, A.M. Bruckstein, and I.A. Wagner. “Shape factor's e_ect on a dynamic cleaners swarm”. Second Workshop on Multi Agents Robotics Systems (MARS) at the Third International Conference on Informatics in Control, Automation and Robotics (ICINCO), pages 13-21, 2006.

[3] Y. Altshuler, V. Yanovsky, I.A. Wagner, and A.M. Bruckstein. “Swarm Intelligence Systems” (Swarm Intelligence - Searchers, Cleaners and Hunters), volume 26 of Studies in Computational Intelligence, chapter II. Springer Verlag, 2006.

[4] Melissa J Perry Anila Bello, Margaret M Quinn and Donald K Milton. “Characterization of occupational exposures to cleaning products used for common cleaning tasks-a pilot study of hospital cleaners”. Environmental Health, pages 8-11, 2009.

[5] J. Borenstein, H. R. Everett, L. Feng, and D. Wehe. ”Mobile robot positioning: Sensors and techniques”. Journal of Robotic Systems, 1997.

[6] M. Hernando E. Gambao. “Control system for a semi-automatic faade cleaning robot”. International Symposium on Automation and Robotics in Construction, 2006.

[7] Joel M. Esposito, Owen Barton, Josh Koehler, and David Lim. “Matlab toolbox for the create robot”. www.usna.edu/Users/weapsys/esposito/roomba.matlab/, 2011.

[8] European Agency for Safety and Health at Work. The occupational safety and health of cleaning workers. 2009.

[9] J. Wang G. Beni. “Theoretical problems for the realization of distributed robotic systems”. IEEE Internal Conference on Robotics and Automation, 1991.

[10] A.M. Bruckstein I.A. Wagner. “Cooperative cleaners: A case of distributed ant-robotics”. Communications, Computation, Control, and Signal Processing, pages 289-308, 1997.

[11] A.M. Bruckstein I.A.Wagner, M. Lindenbaum. „Efficiently searching a graph by a smell-oriented vertex process”. Annals of Mathematics and Arti_cial Intelligence, 24: 221-223, 1998.

[12] iRobot. Roomba owner's manual, 2006.

[13] Lynne E Parker. Alliance: “An architecture for fault tolerant multirobot cooperation”. IEEE Transactions On Robotics And Automation, 14(2), April 1998.

[14] Cameron Salzberger. MATLAB Simulator for the iRobot Create. 1990.

[15] Israel A. Wagner, Yaniv Altshuler, Vladimir Yanovski, and Alfred M. Bruckstein. ”Cooperative cleaners: A study in ant robotics”. The International Journal of Robotics Research, 27(1):127-151, 2008.

[16] Israel A Wagner Yaniv Altshuler, Vladimir Yanovski and Alfred M Bruckstein. ”Multi-agent cooperative cleaning of expanding domains”. The International Journal of Robotics Research, 00(000):1-33, August 2010.

Enterprise Architecture Ontology for Services Computing

Alfred Zimmermann
Reutlingen University, Faculty of Informatics
Architecture Reference Lab of the
SOA Innovation Lab, Germany
alfred.zimmermann@reutlingen-university.de

Gertrud Zimmermann
ZIMMERMANN UND PARTNER
Enterprise Architecture Management Research
Pfullingen, Germany
gertrud.zimmermann@online.de

Abstract – **Enterprise services computing is the current trend for powerful large-scale information systems, which increasingly converge with cloud, computing environments. In this paper, we propose an original ontology-based Architecture Classification Framework for supporting cyclic architecture evaluations and optimizations of enterprise systems based on service-oriented architectures: ESARC - Enterprise Services Architecture Reference Cube. ESARC provides a standardized and normative classification framework for important architecture artifacts of service-oriented enterprise systems. Current approaches for assessing architecture quality and maturity of service-oriented enterprise software architectures are rarely validated and were intuitively developed, having sparse reference model, pattern, metamodel, or ontology foundation. Cyclic assessments of complex service-oriented systems and architectures should produce comparable evaluation results. Today architecture evaluation findings are hardly comparable. Our current idea and contribution is to extend the basic architecture classification framework of ESARC from our previous research by developing specialized metamodels and ontologies for a coherent set of reference architectures, to be able to support machine-based architecture diagnostics and optimizations in enterprise services computing.**

Keywords – service-oriented architecture; enterprise services; enterprise architecture; ESARC; reference model; refernce architecture; ontology; classification framework; diagnostics.

I. INTRODUCTION

Since recent years, innovation oriented companies have introduced service-oriented computing paradigms and combine them with traditional information systems. As the architecture of service-oriented enterprise systems becomes more and more complex, and we are going rapidly into cloud computing scenarios, we need a new and improved set of methodological well-supported instruments and tools for managing, diagnosing and optimizing complex enterprise service-oriented information systems. Service-oriented systems close the business – information technology (IT) gap by delivering efficiently appropriate business functionality and integrating legacy systems with standard application platforms.

Our research work and current innovation practice is about new methods for architecture assessments, and architecture diagnostics, monitoring and optimization. We intend to provide a unified and consistent methodology for enterprise architecture management for service-oriented and cloud computing systems. Our research results are currently validated and extended, and are further to be used for assessments and for integral monitoring of heterogeneous business processes and complex integrated information systems in commercial use [1] by members of the SOA Innovation Lab in Germany and Europe.

Our new introduced approach of an enterprise architecture ontology for services computing is a work in progress research, which is ongoing extended to cover the integral scope of the existing, but still evaluating, classification framework of our ESARC–Enterprise Services Architecture Reference Cube. In assessing the quality of implemented SOA vendor platforms and the integral architecture of service-oriented enterprise systems, we face the problem of not having real comparable evaluation findings from consecutive (cyclic) assessments. Only an architecture classification framework, which sets a relative standard of comparison, makes it possible to track the improvement path of different enterprise services and systems, their architectures and related technologies.

The current state of art research in enterprise services and cloud computing research lacks an integral understanding of architecture classification and semantic representation of service-oriented and cloud computing enterprise systems. Our previous assessment findings were done without an architecture reference model. As a result multiple evaluations of enterprise systems with service-oriented architectures were blurry and hardly comparable within a series of consecutive architectural tests and therefore have produced less meaningful assessment results.

The aim of our research is to enhance analytical instruments for cyclic evaluations of business and system capabilities of different service-oriented platforms and enterprise systems for real business enterprise system environments. In this paper, we disclose our ontology-based approach toward a unified classification framework for enterprise architectures of services and cloud computing systems.

The novelty in our current research paper for the ESARC comprises new aspects and extended ideas for Enterprise Architecture Management (EAM) for Services & Cloud

Computing (SCC). As worked out in this paper, metamodels and related ontologies for ESARC - Enterprise Services Architecture Reference Cube - are the useful extension and integration aid for a holistic set of reference architectures, which we have derived from the ESARC classification framework and the Open Group's standard on Service-oriented Architecture Ontology. Our architecture ontology should provide a base for semantic-supported navigation and automatic inference in architecture diagnostics.

In the following Section II, we present the main view of our original developed ESARC architecture classification framework. We define interrelating reference architecture domains of service-oriented enterprise systems, as part of an architecture layer model, which we built from integrated standards. In Section III, we introduce correlated architecture metamodels and our developed architecture ontology in the context of standards. In Section IV, we set the base for our study from the state of art and other related work. Finally, Section V summarizes our conclusion and gives ideas from current research and for future work.

II. Enterprise Architecture Reference Model

ESARC – the Enterprise Services Architecture Reference Cube – is an integral and continually growing ontology-supported architecture classification framework [2] to be used by enterprise and software architects, to define, structure, verify, and improve service-oriented enterprise and software architectures in a standard way. In order to specify our innovative enterprise and software architecture assessment method, we used a metamodel-based approach [3] for capability evaluations of architecture elements and their main relationships. For this purpose, we have extended, integrated and adapted elements from convergent architecture methods, architecture patterns [4] and [5], related standards and reference models from the state of art.

ESARC is an abstract architecture classification framework [3], which defines an integral view for main interweaved architecture types. ESARC was derived primarily from state of art research and standards [6] and [7], and from architecture frameworks like TOGAF [8], essential [9], the service model of ITIL, and from resources for service-oriented computing [10], [11], and [12]. The aim of the ESARC architecture classification framework is to be universally applicable in cyclic, repeatable and comparable architecture evaluations and structural optimizations of enterprise and software architectures for services and cloud computing. ESARC abstracts from a concrete business scenario or from specific technologies. The main focus of our present paper is to provide exemplarily a detailed view for the three main interdependent reference architecture views of ESARC: Business & Information Reference Architecture, Information Systems Reference Architecture, and the Technology Reference Architecture.

The Open Group Architecture Framework (TOGAF) [8] is the current standard for enterprise architecture and provides the basic blueprint and structure for the service-oriented enterprise software architecture domains. ESARC follows the main architecture domains of TOGAF and extends them substantially and in a unique way to a unified architecture classification framework. ESARC sets a standardization framework for cyclic diagnostics and optimizations of the following interrelated views of the reference architecture: Architecture Governance, Architecture Management, Business & Information Reference Architecture, Information Systems Architecture, Technology Architecture, Operation Architecture, Security Architecture, and Cloud Services Architecture.

The Architecture Governance and Management framework organizes main architecture types, like the Business & Information Architecture, the Information Systems Architecture, and the Technology Architecture. The architecture governance [12] cycle sets the abstract governance frame for concrete architecture activities within the enterprise software and product line development. The architecture governance cycle specifies constitutive management activities: plan, define, enable, measure, and control. The second aim of architecture governance is to set rules for architecture to comply with internal and external standards. Policies for governance and decision definition are set, to allow a standardized and efficient process for architecture decisions inside the enterprise architecture organization. Because enterprise and software architects are acting on a sophisticated connection path (from business and IT strategy to the realization of an architecture landscape of interrelated business domains, applications and technologies), architecture governance has to set rules for the empowerment of software architecture staff, defining structures and procedures of an architecture governance board, and setting rules for communication.

The ESARC - Business & Information Reference Architecture, as set in [3], extends the Business Architecture from TOGAF [8] and defines the link between the enterprise business strategy and the integral business and information design for supporting strategic initiatives. The Business & Information Reference Architecture provides a single source and comprehensive repository of knowledge from which corporate initiatives will evolve and link. This knowledge is model-based and is an integrated enterprise model of the business, which includes the organization and the business processes. The Business & Information Reference Architecture opens a connection to IT infrastructures, systems, as well as to software and security architectures. It provides integration capabilities for IT management, software engineering, service & operations management, and process improvement initiatives. The Business & Information Reference Architecture defines and models the business and information strategy, the organization, and main business requirements for information systems, like key business processes, business rules, business products, and business control information.

The ESARC – Information Systems Reference Architecture from [3] provides an abstract blueprint for the individual service-oriented application architecture to be deployed. It adds specific interactions and specifies relationships to the core business processes of the organization. The OASIS Reference Model for Service Oriented Architecture [13] is an abstract framework, which defines significant relationships among a small set of unifying architectural concepts for services computing. The reference model guides our correlating ESARC reference architectures, as in [14] and [15]. ESARC defines the abstract model for specific application architectures and implementations, which are in conformity with [7] and the Open Group's architecture standards [13], [14], and [15].

In ESARC – Information Systems Reference Architecture we have differentiated layered service types, inspired from [16]. The information services for enterprise data can be thought of as data centric components, providing access to the persistent entities of the business process. The capabilities of information services combine both elementary access to CRUD (create, read, update, delete) operations and complex functionality for finding/searching of data or complex data structures, like data composites or other complex-typed information. Close to the access of enterprise data are context management capabilities, provided by the technology architecture: error compensation or exception handling, seeking for alternative information, transaction processing of both atomic and long running and prevalent distributed transactions.

Process services [16] are long running services, which compose task services and information services into workflows, to implement the procedural logic of business processes. Process services can activate rule services, to swap out a part of the potentially unstable gateway-related causal decision logic. Process services are frontend by interaction services or by specific diagnostic service and process monitoring services. Often process services manage distributed data and application state indirectly, by activating task and information services.

The ESARC – Technology Reference Architecture [3] describes the abstract software and hardware capabilities that are required to support the deployment of business, data, and application services. This includes IT infrastructure, middleware, networks, communications, processing, and standards. The layers of the ESARC – Technology Reference Architecture and the layers of the ESARC – Information Systems Reference Architecture correspond to each other. Security services are part of an integral framework-based security system of standards and components and are impacted by mentioned services and distributed service technologies.

III. Enterprise Architecture Ontology

We have developed exemplarily metamodels and related ontologies seeded by a student research project [17] for the following main architecture domains from ESARC, as a starting and extendable set of work results: Business & Information Reference Architecture, Information Systems Reference Architecture, and the Technology Reference Architecture. Metamodels are used, as standardized in [18], to define architecture model elements and their relationships for the reference architectures of ESARC. Metamodels define models of models. In our approach for architectural modeling, as in [2], [3], and [5], we use metamodels as an abstraction for architectural elements and relate them to architecture ontologies.

The Reference Model for Service Oriented Architecture of OASIS [13] is an abstract framework, which defines basic generic elements and their relationships of a service-oriented architecture. This reference model is not a standard, but provides a common semantic for different specialized implementations. Reference models are, as in [13], abstract conceptual models of a functional decomposition of model elements together with the data flows between them.

Reference architectures, in [14] and [15], are specialized models of a reference model. It is a composition of related architectural elements, which are build from typed building blocks as the result of a pattern-based mapping of reference models to software elements. Architecture patterns, in [4], [7] and [10], are human readable abstractions for known architecture quality attributes, and represent standardized solutions, considering architectural constraints for given recurring problems.

The technical standard of Service-oriented Architecture Ontology from [6] defines core concepts, terminology, and semantics of a service-oriented architecture in order to improve the alignment between the business and IT communities. Following stakeholders are potential users of the SOA ontology, related architecture metamodels, as well as concrete architectural building blocks: business people and business architects, architects for the information systems and software architecture, architects for the technological infrastructure, cloud services architects and security architects.

In our understanding architecture ontologies represent a common vocabulary for enterprise architects who need to share their information based on explicitly defined concepts. Ontologies include the ability to infer automatically transitive knowledge. Our developed ontology for ESARC has some practical reasons: share the common understanding of the ESARC Architecture domains and their structures, reuse of the architectural knowledge, make architectural requirements, structures, building blocks explicit and promote reusability of architectural artifacts, separate the architectural knowledge according orthogonal architectural domains, classify, analyze, diagnose enterprise systems according to the service-oriented reference architecture od ESARC.

For our purpose, an ontology is, as in [19], a formal and explicit description of shareable and automatically navigable concepts of our architectural domain. For modeling purposes we are using UML class diagrams to represent concepts, and we are describing the attributes as properties (sometimes called roles) and role or property restrictions as facets. This structure of an ontology constitutes together with the instances of these concepts the knowledge base. Practically the knowledge base is a growing structure, which starts with the basic concept structures and is enlarged by a more or less amount of growing number of instances.

The SOA Ontology in [6] is represented in the Web Ontology Language (OWL) [20]. The ontology models the core concepts of SOA as classes and properties. The SOA ontology includes in addition natural language description of main concepts and relationships UML diagrams, which show graphically the semantic concepts as classes and the properties as UML associations. The intent of the UML diagrams are for explanations only, but are helpful constructs for understanding the modeled domain of SOA architecture and more concise than the more spacious formal descriptions in OWL. The SOA ontology defines the relations between semantic concepts, without mentioning the exact usage of these architecture concepts. To illustrate the SOA ontology the standard uses examples and descriptions of these in natural language.

The two core concepts of the SOA ontology in Figure 1, as in [6], are: System and Element. These two core concepts are generic and often used concepts to define a composite structure of systems that have elements. These abstract meaning of systems and elements is used in different specific architecture modeling situations. An example of an architectural element is ESB – the Enterprise Service Bus, which is an integration infrastructure for cooperating services. With the concept of *Element* the technical standard associates following core properties: *uses* and *usedBy* as well as the properties *representedBy* and *represents*. The technical standard of SOA Ontology defines additionally other concepts of the SOA Ontology like *HumanActor*, *Task*, *Service*, *ServiceContract*, *Effect*, *ServiceInterface*, *InformationType*, *Composition*, *ServiceComposition*, *Process*, *Policy*, and *Event*.

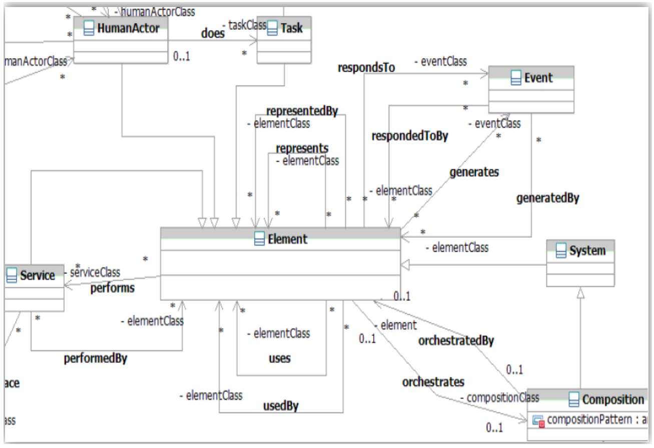


Figure 1. Open Group – Service-Oriented Architecture Ontology.

Element is the central generic service concept on which specialized model elements of ESARC – Enterprise Services Architecture Reference Cube – are constructed.

The metamodel of the ESARC-Business & Information Reference Architecture consists of specialized concepts, which are represented with associations and are generically linked using "is-a"-relationships with the generic concepts like Element and Composition from the Open Group's SOA Ontology [6].

To validate the developed metamodel from Figure 2, we modeled an instantiation scenario of the usage domain of a virtual travel agency. So the ontology was applied as an example to a particular domain by adding class instances of things in our test domain. A particular application, which is based on the ESARC Ontology and the generic SOA Ontology, can add new application-specific classes and properties. The result of these model concepts is a formal representation of the ESARC architecture-types and can be used as an aid to automatically navigate and infer architectural knowledge.

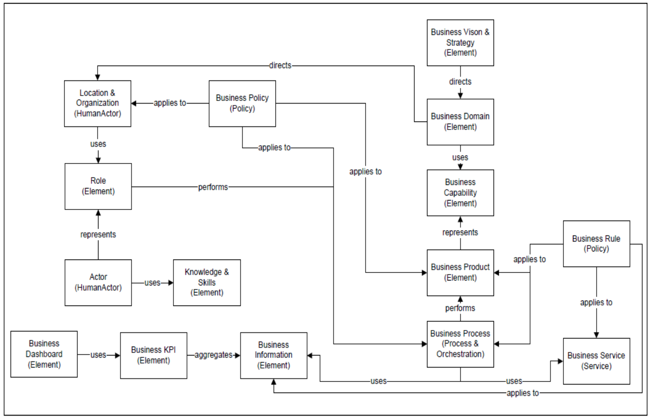

Business Vison & Strategy (Element)
directs
directs
Location & Organization (HumanActor)
applies to
Business Policy (Policy)
Business Domain (Element)
uses
applies to
uses
applies to
Role (Element)
performs
Business Capability (Element)
represents
represents
Business Rule (Policy)
applies to
Business Product (Element)
Actor (HumanActor)
uses
Knowledge & Skills (Element)
performs
applies to
Business Dashboard (Element)
uses
Business KPI (Element)
aggregates
Business Information (Element)
Business Process (Process & Orchestration)
Business Service (Service)
uses
uses
applies to

Figure 2. ESARC – Metamodel of Business & Information Reference Architecture.

We have developed the ESARC Ontology as in [20] and defined ontology concepts for ESARC using the ontology editor Protégé [21]. We have merged our specialized ESARC Ontology, as in [17], with the more generic SOA Ontology from [6]. The so-called *Asserted View* from Protégé in Figure 3 shows the *is-a-relationship* between specific concepts of the Business & Information Reference Architecture and the Open Group's generic SOA Ontology Reference Architecture.

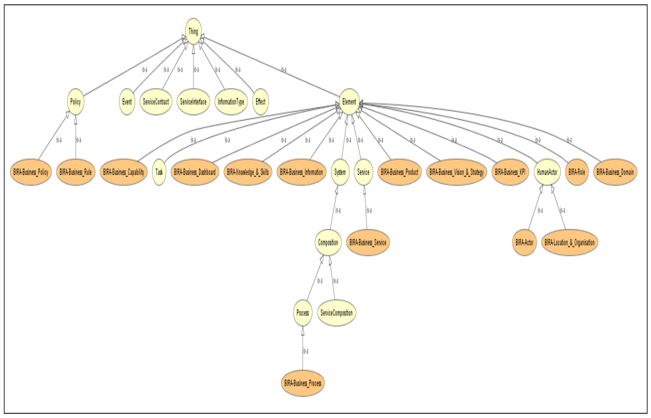


Figure 3. ESARC – Ontology of Business & Information Reference Architecture.

The terminal concepts are specific concepts of ESARC. In contrast we are representing the linked generic concepts of the SOA Ontology on the top of the diagram in Figure 3. Additionally, we determined knowledge properties for the modeled ontology concepts of ESARC. Using the developed ESARC Ontology, we can navigate in the multidimensional space of enterprise architecture management structures and enable a future research effort semantic-supported navigation for architects as well as a base for intelligent inference along specific inference chains. In addition, we have planned to add visualizations for these ontology concepts, as part of a sematic-supported architecture management cockpit.

IV. Related Work

Our research is based on formal architecture concepts from [7] and their relationships: software architecture, reference architecture, reference model, and architecture patterns. A reference model for SOA [13] is a generic fundamental model that embodies the basic idea and provides a decomposition of functionality of a given problem, together with the data flow between elements. The reference model contains an abstract technology agnostic representation of the elements and their relationships, showing the interactions between basic concepts. The concept of reference architecture [7] and [14], [15] is the result of a mapping of an architecture reference model to software elements and contains the related fundamental relationships between them.

Architecture ontologies are quite new study objects. Related and fundamental work on ontologies with their development processes and tools, as in [19], [20], and [21], will allow a better understanding of the modeled domain of enterprise services computing, can help to organize complexity in categories of interrelated concepts, and are an efficient and machine-understandable representation for the modeled classifications of concepts. Ontologies provide an aid both for software architects as well as for automatic inference procedures, to enable diagnostics and improvements within a predefined classification framework, which is defined by formal represented ontology concepts.

The Open Group's SOA Ontology in [6] is the fundamental work for of our ESARC Ontology. This basic ontology was the seminal work for our research. The ontology contains concepts and properties from the domain of service-oriented architectures. Formal OWL definitions are supplemented by text explanations and by UML class diagrams for the related models of the ontology. These diagrams and models are intended only for explanations of the formal OWL representations. We have done additional experimental work in a long-term student research project [17] to model exemplarily related ontologies for three main reference architectures for ESARC. Based on this work we are currently extending our ontology modeling and research to support multidimensional architecture representations and inference processes for diagnostics and optimizations of software architectures in enterprise services computing and extending our work for cloud computing, as in [22] and [23].

Service-oriented architecture SOA [11] is the computing paradigm that utilizes services as fundamental flexible and interoperable building blocks for both structuring the business and for developing applications. SOA promotes a business oriented architecture style, based on best of breed technology of context agnostic business services that are delivered by applications in a business-focused granularity. Early definitions of SOA were technology focused and the differences between SOA and web services were often blurred. SOA technologies emerged due to the expansion of the Web technology during the last years and produced abundance specifications and standards as in [13], [14], [15], and [6], [12], [18], which are developed by open standard organizations like W3C, OMG, OASIS, and The Open Group. The perspective of a service development process is offered by [16] and [10].

Our architecture reference model ESARC relates closely to SOAMMI, which is our previous designed maturity framework for evaluation of enterprise and service-oriented product architectures. Unfortunately most of existing SOA and EA maturity models lack a clear metamodel base. Therefore we have extended CMMI [24] in our previous research, which is a framework for assessments of software processes, and transformed it into a specific framework for the assessment of the maturity of service-oriented enterprise and software architectures [1] and [2].

The main scope of the intuitively specified Architecture Capability Maturity Model (ACMM) [25] framework from TOGAF is the evaluation of enterprise architectures in internal enterprise architecture assessments.

The SOA Maturity Model in [26] considers intuitively multidimensional aspects of a SOA.

The SOA Maturity Model from Sonic [27] distinguishes five maturity levels of a SOA, and associates them - in analogy to a simplified metamodel of CMMI - with key goals and key practice. Key goals and key practices are reference points in SOA maturity assessments.

The SOA Maturity Model of ORACLE in [28] characterizes in a loose correlation with CMMI five different maturity levels and associates them with strategic goals and tactical plans for implementing SOA. Additional capabilities of a SOA are referenced with each maturity level: Infrastructure, Architecture, Information & Analytics, Operations, Project Execution, Finance & Portfolios, People & Organization, and Governance.

V. Conclusion and Future Work

Our approach for architecture evaluation and optimization of service-oriented enterprise software architectures is based on ESARC - a special architecture reference model, an associated architecture metamodel and on architecture patterns. In our research we have motivated the necessity to extend both existing architecture reference models and service-oriented maturity models to accord to a clear metamodel approach due to the well understood and verified CMMI model. Our approach provides a sound basis from theory for practical evaluations of service oriented standard platforms in heterogeneous environments with four major global acting technology vendors. Future work has to consider conceptual work on both static and dynamic architecture complexity, and in connecting architecture quality procedures with prognostic processes on architecture maturity with simulations of enterprise and software architectures. Additional improvement idea deals with patterns for visualization of architecture artifacts and architecture control information to be operable on an architecture management cockpit. To improve semantic-based navigation within the complex space of EAM-visualization and service-oriented enterprise software architecture management we are working on ontology models for the ESARC – The Enterprise Software Architecture Reference Cube.

ACKNOWLEDGEMENT

This work was initiated and partially supported by the SOA Innovation Lab, a major research and innovation network on Enterprise Architecture Management for Services and Cloud Computing in Germany and Europe.

REFERENCES

[1] H. Buckow, H.-J. Groß, G. Piller, K. Prott, J. Willkomm, and A. Zimmermann, "*Analyzing the SOA-ability of Standard Software Packages with a dedicated Architecture Maturity Framework*", EMISA 2010: October 7– 8, 2010 - Karlsruhe, Germany, GI-Edition - Lecture Notes in Informatics (LNI), P-172-2010, pp. 131-143, 2010.

[2] A. Zimmermann, "*Method for Maturity Diagnostics of Enterprise and Software Architectures*", in A. Erkollar (Ed.) ENTERPRISE & BUSINESS MANAGEMENT, A Handbook for Educators, Consulters and Practitioners, Volume 2, Tectum 2010, ISBN 978-3-8288-2306-8, pp. 129-172, 2010.

[3] A. Zimmermann, H. Buckow, H.-J. Groß, F. O. Nandico, G. Piller, K. Prott, "*Capability Diagnostics of Enterprise Service Architectures using a dedicated Software Architecture Reference Model*", in SCC 2011 - IEEE International Conference on Services Computing, July 4-9, 2011, Washington DC, USA, IEEE Proceedings of the SCC 2011, ISBN 978-0-7695-4462-5/11, IEEE Computer Society USA, pp. 592-599, 2011.

[4] F. Buschmann, R. Meunier, H. Rohnert, P. Sommerlad, and M. Stal, „*Pattern-oriented Software Architecture“,* Wiley. 1996.

[5] A. Zimmermann, F. Laux, R. Reiners, "*A Pattern Language for Architecture Assessments of Service-oriented Enterprise Systems*", in PATTERNS 2011 - The Third International Conferences on Pervasive Patterns and Applications, September 25-30, 2011 Rome, Italy, ISBN 978-1-61208-158-8, IARIA Proceedings of PATTERNS 2011, pp. 7-12 (2011).

[6] The Open Group, *"Service-Oriented Architecture Ontology"*, Technical Standard, 2010, https://www2.opengroup.org/ogsys/jsp/publications/PublicationDetails.jsp?catalogno=c104, last access: March 20th, 2012.

[7] L. Bass, P. Clements, and R. Kazman, "*Software Architecture in Practice"*, Second Edition, Addison Wesley, 2003.

[8] TOGAF "*The Open Group Architecture Framework*" Version-9, The Open Group, 2009.

[9] *Essential Architecture Project*, http://www.enterprise-architecture.org, last access: June, 19th, 2011.

[10] T. Erl, "*SOA Design Patterns*", Prentice Hall, 2009.

[11] T. Erl, "*Service Oriented Architecture"* Prentice Hall, 2005.

[12] The Open Group "*SOA Governance Framework*", August 2009.

[13] C. M. MacKenzie, K. Laskey, F. McCabe, P. F. Brown, and R. Metz, OASIS "*Reference Model for Service Oriented Architecture*" 1.0, OASIS Standard, 12 October 2006.

[14] J. A. Estefan, K. Laskey, F. G. McCabe, and D. Thornton, OASIS "*Reference Architecture for Service Oriented Architecture*" Version 1.0, OASIS Public Review Draft 1, 23 April, 2008.

[15] The Open Group, *"SOA Reference Architecture"*, Technical Standard, 2011, https://www2.opengroup.org/ogsys/jsp/publications/PublicationDetails.jsp?catalogno=c119, last access: March 20th, 2012.

[16] G. Engels, A. Hess, B. Humm, O. Juwig, M. Lohmann, J.P. Richter, M. Voß, and J. Willkomm, „*Quasar Enterprise*“ dpunkt.verlag, 2008.

[17] S. Bourscheidt, T. Breuer, T. Brunner, B. Fetler, G. Fogel, "*ESARC-Referenzmodell und Ontologie für Enterprise Architecture Management (EAM)*", Research Report, Reutlingen University, Enterprise Services Architecture Reference Lab, 2012.

[18] OMG, "*Meta Object Facility (MOF) Core Specification*", Version 2.0, Object Management Group, 2006.

[19] N. F. Noy, D. L. McGuineness, "*Ontology Development 101: A Guide to Creating Your First Ontology*", Stanford University, 2001

[20] D. Gasevic, D. Djuric, V. Devedzic, "Model Driven Engineering and Ontology Development", 2nd Edition, Springer Verlag, 2009.

[21] M. Horridge, and H. Knoblauch, A. Rector, R. Stevens, C. Wroe, S. Jupp, G, Moulton, N. Drummond, *"A Practical Guide To Building OWL Ontologies Using Protégé 4 and CO-ODE Tools"*, Edition 1.2, The University of Manchester, 2009.

[22] L. Youseff, M. Butrico, D. Da Silva, "*Towords a Unified Ontology of Cloud Computing*", in Grid Computing Environments Workshop (GCE# 08), Nov. 2008, pp. 1-10, 2008.

[23] H. Zhou, H, Yang, A. Hugill, "*An Ontology-Based Approach to Reengineering Enterprise Software for Cloud Computing*", in 34th Annual IEEE Computer Software and Application Conference, 2010, pp. 383-388, 2010.

[24] CMMI-DEV-1.3 2010 "*CMMI for Development, Version 1.3*" Carnegie Mellon University, Software Engineering Institute, CMU/SEI-2010-TR-033, 2010.

[25] ACMM, "*Architecture Capability Maturity Model*", in TOGAF Version 9, The Open Group Architecture Framework, The Open Group, 2009, pp. 685-688.

[26] S. Inaganti and S. Aravamudan, "*SOA Maturity Model*" BP Trends, April 2007, pp. 1-23, 2007.

[27] Sonic "*A new Service-oriented Architecture (SOA) Maturity Model*" http://soa.omg.org/Uploaded%20Docs/SOA/SOA_Maturity.pdf, last access: March 20th, 2012.

[28] Oracle "*SOA Maturity Model*", http://www.scribd.com/doc/2890015/oraclesoamaturitymodelcheatsheet, last access: March 20th, 2012.

Determine the Market Position for VTS Service Systems Based on Service Value Position Model Using Novel MCDM Techniques

Chia-Li Lin
Department of Resort and Leisure Management, Taiwan Hospitality and Tourism College
No. 268, Chung-Hsing St. Feng-Shan Village, Shou-Feng Township, Hualien County, 974, Taiwan, ROC
e-mail: linchiali0704@yahoo.com.tw

***Abstract*—With the development of information and communication technology, People are gradually replaced map with the Vehicles telematics system (VTS). In addition, navigation is not the only function of VTS nowadays, new generation VTS even provides various functions and system services. This study tries to discuss the development trend of VTS products/services and customers' needs of navigation and location services, audio-video and entertainment services, communication and information services and safety and security services. However, automobile manufacturers and VTS products/service providers need to determine the new product development strategy based on their customers' needs. Because customers' needs will influence the automobile sales and give an impact on the development of VTS service in the future. This study also proposes NRM to determine the service improvement path and the improvement strategy based on the NRM. The aspect of the Safety and Security Services (SS) is the primary dominating aspect, the aspect of Communications and System Services (CS) is the most important influencing aspect in the NRM of the VSI. The aspect of the Payment Channel (PC) is the primary dominating aspect and the aspect of the Payment Method (PM) is the most important influencing aspect of the PSI.**

Keywords-Vehicles telematics system (VTS); Service system; Market position; DEMATEL; VIKOR.

I. INTRODUCTION

The original purpose of the GPS technology was for the military. It was gradually released to private enterprises. The related map information systems were thoroughly built. The application service of GPS technology had begun to move toward diversification [1-3]. New generation automobile not only emphasizes the mechanical efficiency, but also the VTS system functions. In order to satisfy the service needs of mobile environment for different customer groups, the VTS service system needs to be integrated and offers diverse service functions. Therefore some integrated functions have been appeared, such as the information and communication service, vehicles monitor service. According to the U.S.A., navigation devices can be classified as vehicle, aviation and marine navigation devices. Vehicle navigation devices are widely applied to END (Embedded Navigation Device), PND (Portable Navigation Device), PDA (Personal Digital Assistants) smart phones and a diversity of mobile devices [4, 5]. The car has become an open mobile service system from a closed transportation tool. The automobile has increased the added value of navigation, safety, security, information, communication, and entertainment functions. The VTS service system can aid drivers and passengers to contact with call center and access the navigation information. The VTS service system can not only improve driving convenience but also ensure the safety of vehicles under the appropriate monitoring. The innovation of ICT technologies has attributed to the diverse system service, and devices have become more and more thinner and lighter. Besides the VTS service system continually increase service functions, and the VTS service devices generally become similar to customers. This kind of trend attributes the VTS service system/device to become more and more popular in the current year, and to become the necessary goods from luxury goods. The navigation technology and electric service function of VTS service system has rapid progress continually for achieving the satisfaction level of customer needs. In the highly competition market of vehicle telematics service, the customers can't only be pleased by the improvement of the device hardware and electronic map software services. It is important for VTS service operators to identify what is consumers' needs develop new generation VTS service system to satisfy customers' needs. Because this will not only influence sales volumes for automobile operators but also impact the service market of end user for automobile market.

The remainder of this paper is organized as follows. In Section 2, product development and market position for VTS service system are reviewed and discussed. In Section 3, the evaluation model of market position for VTS service system is built. In Section 4, a novel MCDM technique is used to solve the market position decision problem (i.e., customers' preferences). The performance of the VTS service system is then discussed and an empirical study is demonstrated for the novel MCDM model. Finally, in Section 5, the conclusions and remarks are presented.

II. THE DISCUSSION OF PRODUCT DEVELOPMENT AND MARKET POSITION FOR VTS SERVICE SYSTEM

The GPS applications are quite extensive. According to the use environment, the navigation devices can be classified into many navigation devices of vehicle, aviation

ISBN: 978-1-61208-215-8

and marine and so on. The vehicle navigation devices are widely applied to END (Embedded Navigation Device), PND (Portable Navigation Device), PDA (Personal Digital Assistants) Smart phone and diverse mobile devices [4, 5]. This study discusses the service needs of customers, defines the functions/utilities of VTS service system, and analyzes the difference of customer needs. Besides, this study also arranges the service functions based on different customers' attributes, and generalizes five main value evaluation aspects (Location and navigation and services, LN; Safety and security services, SS; Communications and system Services, CS; Multimedia and entertainment services, ME; Image and customer relationship, IR) and four price evaluation aspect (Service Fee Rate, SF; Package Pricing, PP; Payment Method, PM; Payment Channel, PC) for VTS service systems/devices. This study determines the customer preference of VTS service systems using user questionnaire survey, and builds the network relation map by DEMATEL (Decision Making Trial and Evaluation) technique. Then this study uses the ANP (Analytic Network Process) technique to determine the aspect weights and builds the group component among the aspect by PCA (Principal Component Analysis) technique. Finally, this study uses the VIKOR (VlseKriterijumska Optimizacija I Kompromisno Resenje, VIKOR) [6] to analyze the VSI (Value satisfaction index) and PSI (Price satisfaction index) for VTS service systems/devices [7]. This study proposes an integrated evaluation model to analyze the current service gap of VTS service systems/devices, and illustrates four real commercial types VTS services/devices to test the proposed model. This proposed model can aid VTS system service operators to determine the market position of VTS service systems/devices by MPM (market position map) approach, determine the service improvement paths using network relation map (NRM) approach and can provide product/service development strategy of VTS systems/devices for the VTS service providers and automobile operators in the future. In the VTS service systems/devices, the system service providers need to play the integrated role to provide the user various vehicle navigation and mobile services applications (Table I).

TABLE I. DESCRIPTION OF CRITERIA AND ITS CODEWORD FOR EVALUATING VTS'S FUNCTIONS

Aspects / Criteria	Descriptions
Value Satisfaction Index (VSI)	
Location and Navigation and Services (LN)	
Voice-Guided Navigation Services (LN1)	The more precise voice-guided navigation services improves the efficiency of driving and reduces driving time.
Traffic Information (LN2)	More correct traffic situation information and more driving time to save, helps users realize the immediate road conditions, and comply with traffic signals.
Electronic Map Information (LN3)	More accurate map information allows drivers to handle and estimate the distance to the destination.
Safety and Security Services (SS)	
Safety and Emergency Services (SS1)	To prevent accidents and provide rescue assistance. Also clarifies the responsibilities for investigations after an accident.
Remote Central Control Services (SS2)	Remote door lock or unlock services to assure the safety of passengers and car security.
Vehicle Location Services (SS3)	To search and locate a stolen vehicle or a towed car.
Car Security Services (SS4)	To prevent the vehicle from being stolen and provide prior warning.
Vehicle Diagnosis and Maintenance Services (SS5)	To handle the operating conditions of vehicle devices, and provide maintenance suggestions.
Communications and System Services (CS)	
Mobile Information Services (CS1)	Enable consumers to manage e-commerce, get real-time information and access the Internet while moving.
User Interface (CS2)	Friendlier and more numerous choices will increase the convenience of use.
Platform Integration Services (CS3)	Integrating different platforms will increase compatibility of systems, and save replacement costs.
Information Security Protection (CS4)	Stricter security protection, more safeguards for the privacy of personal data, to prevent the criminal use of personal data.
Information Update Frequency (CS5)	More immediate and quick information updates to ensure more accuracy and precision.
Multimedia and Entertainment Services (ME)	
Real-Time Multimedia Services (ME1)	More choices for real-time multimedia services, enabling more current access to fashion and entertainment.
Vehicle Multimedia Playing System (ME2)	Larger screen size, support for various multimedia formats, and larger storage capacity enables consumers to enjoy more comfortable audio-video services.
Game Services (ME3)	Various choices of game services allow for more fun.
Personal Platform Services (ME4)	Personalized set-up functions of multimedia. Consumers can enjoy personalized services.
Image and Customer Relationship (IR)	
Product Design (IR1)	More popular product design that is more selective and easier to carry can stimulate the desire to buy the product.
Brand Image (IR2)	Better brand image, more confidence in the quality of the services provided.
After-Sales Services (IR3)	More after-sales service locations and wider channels, consumers will feel confident about maintenance and warranty services.
Privacy Policy (IR4)	More stringent privacy protection policies can avoid the leakage of credit and personal information for criminal use.
Price Satisfaction Index (PSI)	
Fee Rate and Payment Method	
Service Fee Rate (SF)	The service fee rate and promotion terms.
Package Pricing (PP)	The different pricing items and pricing methods which users prefer.
Payment Method (PM)	Flexible payment methods can satisfy consumers with different spending habits.
Payment Channel (PC)	More payment channels can enhance consumers' convenience.

III. THE EVALUATION MODEL OF MARKET POSITION FOR VTS SERVICE SYSTEMS

The analysis process of the service development strategy model for VTS service system is based on a novel MCDM technique as the following five steps: (1) introduces the research idea and market position map, (2) applies FCM to construct the network relations map (NRM); (3) constructs the motivation of needs by using PCA; (4) uses ANP to analyze the group weights; and (5) evaluates the performance gaps of digital music service systems using the FIM , discussing the development trend and related studies suggestion on the VTS services system.

A. *The concept of market position map*

Some studies have pointed out a trade-off relation between price and value [7, 8]. Some researchers used regression analysis to come out the relation between benefit and price [8]. The advantage of regression analysis is to handle the location analysis and to extract the price data easily. However, the position analysis of a multi-benefit/function is hard to be handled by a regression analysis, and we need to consider the value and price respectively. Some study adopted the value satisfaction index and price satisfaction index to solve the multi-benefit/function problems, and used the MCDM technique to evaluate varying tangible and invisible benefit/functions in the mobile phone market [7]. Therefore, this research adopts the MCDM techniques to analyze our research problem and extend the value-price map into the market position map (Figure 1).

The axes of the market position map include the value satisfaction index and price satisfaction index. The value satisfaction index consists of the aspects increasing the customer value satisfaction, while the price satisfaction index consists of aspects improving customer price satisfaction. As shown in Figure 1, the X axis is the value satisfaction index (VSI) and the Y axis is the price satisfaction index (PSI). In this study, the value satisfaction index includes four aspects: Location and navigation and services (LN), Safety and security services (SS), Communications and system Services (CS), Multimedia and entertainment services (ME) and Image and customer relationship (IR), and the price satisfaction index includes four aspects: Service Fee Rate (SF), Package Pricing (PP), Payment Method (PM) and Payment Channel (PC). The market position map is divided into four sections or market segmentations [Common and luxurious, (H, H); high price and gorgeous, (H, L); Low price to penetrating market (L, H); No or limited choice (L, L)].

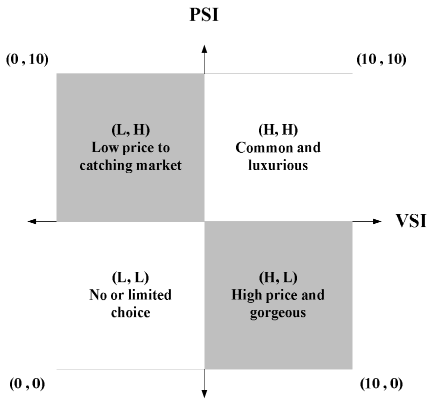


Figure 1. The concept of the market position map model.

B. Builds the DEMATEL model

The basic concept of DEMATEL technique was initiated for the Science and Human Affairs Program by Battelle Memorial Institute of Geneva between 1972 and 1976 to solve complex problems. This study uses the concept DEMATEL method to build the evaluation structure of network relation map (NRM) for creating an e-era VTS system. When making decisions, the decision-maker has to consider the criteria in detail and all the interrelations between them. What the decision-maker has to do is to find out the key criteria, modify them and then the whole performance of satisfaction will be enhanced. Therefore, when the decision-maker copes with lots of criteria being changed, the best solution is to determine the key criteria that affect the other criteria mostly and modify them. Eventually, the results of the evaluation will become more and more precise. Therefore, some recent studies considered the DEMATEL techniques for solving complex studies, such as the expectation model of service quality [9], and value-created system of science (technology) park [10]. The steps of the DEMATEL method are described as follows: (1) Calculates the original influence matrix by average of expert-respondents, (2) Calculates the degree of direct influence matrix, (3) Calculates the total degree of indirect influence matrix, (4) Calculates the total degree of direct and indirect influence matrix, and (5) Determines the Network Relation Map (NRM).

(1) Calculates the initial average matrix

Respondents were asked to indicate the influence that they believe each aspect exerts on each of the others; according to scoring scales ranging from 0 to 4, where “0” means no influence and “4” means “extremely strong influence.” For the question between aspect/criterion; “1”, “2”, and “3” mean “low influence”, “medium influence” and “high influence,” respectively. As the data shows in Table II, the influence of Multimedia and entertainment services (ME) on “Image and customer relationship (IR)” is 2.300, which means “medium influence”. On the other hand, the influence of Image and customer relationship (IR) on Multimedia and entertainment services (ME) is 1.950, which means “low influence”.

TABLE II. ORIGINAL INFLUENCE MATRIX.

Aspects	LN	SS	CS	ME	IR	Total
Location and navigation and services (LN)	0.000	1.625	2.825	1.575	2.300	8.325
Safety and security services (SS)	1.725	0.000	2.525	1.350	2.300	7.900
Communications and system Services (CS)	2.975	2.350	0.000	2.075	2.350	9.750
Multimedia and entertainment services (ME)	1.575	1.350	2.200	0.000	2.300	7.425
Image and customer relationship (IR)	2.250	2.025	1.950	1.950	0.000	8.175
Total	8.525	7.350	9.500	6.950	9.250	-

(2) Calculates direct influence matrix

The initial direct influence matrix (*D*) can be calculated by Eq. (1) and Eq. (2). (*A*) is the initial average influence matrix, and can produce the initial direct influence matrix (*D*) through the process of Eq. (1) and Eq. (2). Matrix *D* represents each direct influence, and in the Matrix, the numbers on the diagonal are 0 and the sum of each column and row is 1 in maximum (only one equals 1). Adding the sums of each row and column in the Matrix results in the direct influence value:

$$\boldsymbol{D} = s\boldsymbol{A}, \quad s > 0 \tag{1}$$

where

$$s = \min_{i,j}\,[1/\max_{1\le i\le n}\sum_{j=1}^{n} a_{ij}, 1/\max_{1\le j\le n}\sum_{i=1}^{n} a_{ij}],\ \ i, j = 1, 2, ..., n \tag{2}$$

and $\lim_{m\to\infty} \boldsymbol{D}^m = [0]_{n\times n}$, where $\boldsymbol{D} = [x_{ij}]_{n\times n}$, when $0 < \sum_{j=1}^{n} x_{ij}, \sum_{i=1}^{n} x_{ij} \le 1$ at least one $\sum_{j=1}^{n} x_{ij}$ or $\sum_{i=1}^{n} x_{ij}$ equal one, and only one row sum or column sum equal one .So we can guarantee $\lim_{m\to\infty} \boldsymbol{D}^{m-1} = [0]_{n\times n}$.

From Table II, we processed the "original influence matrix (*A*)" by using Eq. (1) and Eq. (2) and obtained the "direct influence matrix (*D*)". As shown in Table III, the diagonal items of *D* are all 0, and the sum of a row is 1, at most. Then we calculated Table IV by adding up the rows and columns. In Table IV, the sum of the rows and columns for "Communications and system Services (CS)" is 1.974, which is the most important influence aspect. On the other hand, the sum of the rows and columns for Multimedia and entertainment services (ME) is 1.474, which is the least important influence aspect.

TABLE III. DIRECT INFLUENCE MATRIX ($\boldsymbol{D}$)

Aspects	LN	SS	CS	ME	IR	Total
Location and navigation and services (LN)	0.000	0.167	0.290	0.162	0.236	0.854
Safety and security services (SS)	0.177	0.000	0.259	0.138	0.236	0.810
Communications and system Services (CS)	0.305	0.241	0.000	0.213	0.241	1.000
Multimedia and entertainment services (ME)	0.162	0.138	0.226	0.000	0.236	0.762
Image and customer relationship (IR)	0.231	0.208	0.200	0.200	0.000	0.838
Total	0.874	0.754	0.974	0.713	0.949	

TABLE IV. COMPARISON TABLE OF DIRECT INFLUENCE MATRIX.

Aspects	Sum of row	Sum of column	Sum of row and column	Importance of Influence
Location and navigation and services (LN)	0.854	0.874	1.728	3
Safety and security services (SS)	0.810	0.754	1.564	4
Communications and system Services (CS)	1.000	0.974	1.974	1
Multimedia and entertainment services (ME)	0.762	0.713	1.474	5
Image and customer relationship (IR)	0.838	0.949	1.787	2

(3) Calculates Indirect Influence Matrix

The indirect influence matrix can be derived from Eq. (3), as shown in Table V.

$$\boldsymbol{IT} = \sum_{i=2}^{\infty} \boldsymbol{X}^i = \boldsymbol{X}^2(\boldsymbol{I} - \boldsymbol{X})^{-1} \tag{3}$$

TABLE V. INDIRECT INFLUENCE MATRIX ($\boldsymbol{ID}$)

Aspects	LN	SS	CS	ME	IR	Total
Location and navigation and services (LN)	1.126	0.965	1.137	0.916	1.131	5.275
Safety and security services (SS)	1.047	0.943	1.092	0.879	1.077	5.037
Communications and system Services (CS)	1.191	1.060	1.351	1.013	1.275	5.890
Multimedia and entertainment services (ME)	0.990	0.874	1.038	0.846	1.015	4.762
Image and customer relationship (IR)	1.044	0.920	1.133	0.875	1.142	5.115
Total	5.399	4.762	5.750	4.528	5.640	-

(4) Calculates full influence matrix

Full influence matrix *T* can be derived from Eqs. (4) or (5). Table VI is the calculated full influence matrix. As shown in Table VI, the full influence matrix *T*, consists of multiple elements, indicated as Eq. (6). The sum vector of the row value is { d_i }, and the sum vector of the column value { r_i }; then, let $i = j$, the sum vector of row value plus column value is { $d_i + r_i$ }, which means the full influence of the matrix *T*. As the sum of the row value plus the column value { $d_i + r_i$ } is higher, the relationship of the dimension or criterion is stronger. The sum of the row value minus the column value is { $d_i - r_i$ }, which means the net influence relationship. If $d_i - r_i$ > 0, it means the degree of influencing others is stronger than the degree to be influenced; otherwise, $d_i - r_i$ <0. Formulations show as follows:

$$\boldsymbol{T} = \boldsymbol{X} + \boldsymbol{IT} = \sum_{i=1}^{\infty} \boldsymbol{D}^i \tag{4}$$

$$\boldsymbol{T} = \sum_{i=1}^{\infty} \boldsymbol{D}^i = \boldsymbol{D}(\boldsymbol{I} - \boldsymbol{D})^{-1} \tag{5}$$

$$\boldsymbol{T} = [t_{ij}], \quad i, j \in \{1, 2, ..., n\} \tag{6}$$

$$\boldsymbol{d} = \boldsymbol{d}_{n\times1} = [\sum_{j=1}^{n} t_{ij}]_{n\times1} = (d_1, ..., d_i, ..., d_n) \tag{7}$$

$$\boldsymbol{r} = \boldsymbol{r}_{n\times1} = [\sum_{i=1}^{n} t_{ij}]'_{1\times n} = (r_1, ..., r_j, ..., r_n) \tag{8}$$

As shown in Table VII, the aspect of Communications and system Services (CS) has the highest degree of full influence ($d_3 + r_3$ =13.62) and the aspect of Multimedia and entertainment services (ME) has the lowest degree of full influence ($d_4 + r_4$ =10.77). The aspect of Safety and Security Services (SS) has the highest degree of net influence [($d_3 - r_3$)=0.331]. The order of other net influences is listed as follows: the Multimedia and entertainment services (ME) aspect ($d_1 - r_1$ =0.283), Communications and system Services (CS) aspect ($d_1 - r_1$ = 0.165), Location and navigation and services (LN) aspect ($d_3 - r_3$ = -0.144), and the last one, the Product Image (PI) aspect ($d_5 - r_5 =$ -0.635).

TABLE VI. FULL INFLUENCE MATRIX

Aspects	LN	SS	CS	ME	IR	Total
Location and navigation and services (LN)	1.126	1.132	1.427	1.078	1.367	6.130
Safety and security services (SS)	1.224	0.943	1.351	1.017	1.313	5.847
Communications and system Services (CS)	1.496	1.301	1.351	1.226	1.516	6.890
Multimedia and entertainment services (ME)	1.152	1.012	1.264	0.846	1.251	5.524
Image and customer relationship (IR)	1.275	1.128	1.333	1.075	1.142	5.954
Total	6.274	5.516	6.725	5.241	6.589	-

TABLE VII. DEGREE OF FULL INFLUENCE

Aspects	{ d_i }	{ r_i }	{ $d_i + r_i$ }	{ $d_i - r_i$ }
Location and navigation and services (LN)	6.130	6.274	12.40	-0.144
Safety and security services (SS)	5.847	5.516	11.36	0.331
Communications and system Services (CS)	6.890	6.725	13.62	0.165
Multimedia and entertainment services (ME)	5.524	5.241	10.77	0.283
Image and customer relationship (IR)	5.954	6.589	12.54	-0.635

(5) Determines the network relationship map (NRM)

According to the aspects/criteria defined in Table I, some experts were invited to discuss the relation and influence levels of criteria under the same aspects/ criteria and to score the relation and influence among criteria based on the DEMATEL technique. Aspects/criteria are divided into different types, so the experts could answer the questionnaire in areas/fields with which they were familiar. The net full influence matrix, T_{net} , is determined by the Eq. (9).

ISBN: 978-1-61208-215-8

$$\boldsymbol{T}_{net} = [t_{ij} - t_{ji}], \qquad i, j \in \{1, 2, ..., n\} \tag{9}$$

The diagonal items of the matrix are all 0. In other words, the matrix contains a strictly upper triangular matrix and a strictly lower triangular matrix. Moreover, while values of strictly upper triangular matrix and strictly lower triangular matrix are same, their symbols are opposite. This property helps us that we only have to choose one of strictly triangular matrix.

TABLE VIII. THE NET INFLUENCE MATRIX FOR VTS SERVICE SYSTEM

Aspects	LN	SS	CS	ME	IR
Location and navigation and services (LN)	-				
Safety and security services (SS)	0.092	-			
Communications and system Services (CS)	0.069	-0.050	-		
Multimedia and entertainment services (ME)	0.075	-0.005	0.038	-	
Image and customer relationship (IR)	-0.092	-0.185	-0.183	-0.176	-

We can understand the related influence structure of NRM for the VTS service system from Figure 2, the figure shows that the aspects of Safety and Security Services (SS), Communications and system Services (CS) and Multimedia and entertainment services (ME) are mainly influencing aspects, and the aspect of Location and navigation and services (LN) and Image and customer relationship (IR) are mainly influenced aspect. Therefore, this study wants to assist decision-makers to build an improvement process, and conducts calculus on the net (be received) influence matrix using the full influence matrix (TABLE VIII). The evaluated method can integrate the degree of influence of the aspect, sand gain the net influence relation of the five aspects. From TABLE VIII and Figure 2, the IR aspect has net influence on the aspect of SR, PM and PF, the aspect of SR had net influence on the aspect of PM, PF and PP. The aspect of PM had net influences on the aspect PF and PP, and the aspect of PF influences the aspect of PP. The aspect of IR should be improved firstly, then the aspect of SR, PM and PF should be improved secondly. The aspect of PP is the least improvement item among all aspects. It’s indirect for improving the service performance of the system.

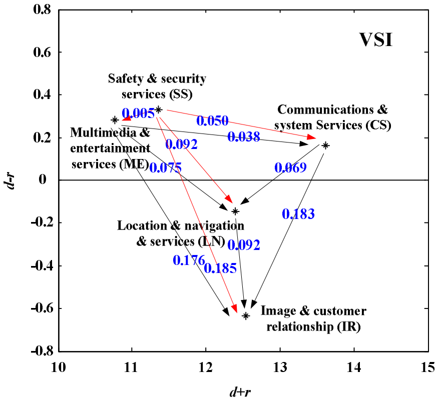


Figure 2. The improvement strategy map for VTS service system based on VSI

C. Principal Component Analysis (PCA)

This study uses PCA to analyze the original data of importance degree. It can be used to simplify the large number of criteria and it also can satisfy the hypothesis of AHP/ANP on the independence/dependence of criteria included in system aspect. However, the founder of ANP, Professor Saaty, didn’t explicitly define it [11]. From the paper analysis of AHP/ANP, it can be figured out that the hypothesis is that criteria in aspects are independent /dependent. That’s why we use this technique in this study. There are two components that can be extracted: PM (Platform design and maintenance) and the square sum (88.092%), and named the major elements, as shown in TABLE IX. Support of device system (PM3), frequencies of content update (PM2), and system protection (PM1) can be integrated into the first major component PMP1 (Platform design and maintenance). System stability (PM5) and system protection (PM4) can be integrated into the second component PMP2 (System stability and security).

TABLE IX. THE PCA ANALYSIS OF NAVIGATION AND LOCATION SERVICES (NL) ASPECT

			Components	
Aspects	Components	Criteria	1	Community
Location and navigation and services (LN)	Traffic Information and Navigation Services (NLP1)	Voice-Guided Navigation Services	0.908	0.825
		Traffic Information	0.861	0.742
		Electronic Map Information	0.824	0.678
Engen-value λ			2.245	
% of Variance (contribution)			74.843	
Cumulative contribution (%)			74.843	
Cronbach's α			0.831	

D. Analytic network procedure model (ANP)

Saaty (1996) proposed the concepts of ANP in 1996, to solve the issue that the AHP method is too ideal to evaluate the problems correctly. The ANP method [12-14] can cope with the dependence and feedback relation in the problems. The evaluation is closer to the actual adoption. The following three steps are undertaken to evaluate the decision problems with the ANP method: (1) builds the network hierarchical structure, (2) calculates the weighing of factors in each hierarchy, and (3) calculates the weighting of the whole hierarchy structure. In this study, the ANP steps are introduced as follows: (1) clarifies the problems and build the structure based on NRM, (2) designs the questionnaire and survey, (3) builds the weightings of pair-wise comparisons, calculates the weightings of factors, and test the consistency, (4) calculates the super-matrix.

E. Vlse Kriterijumska Optimizacija I Kompromisno Resenje (VIKOR)

After establishing the evaluation model, including criteria and given weights in each criterion, the next step is to evaluate and improve the performance of benchmarked alternatives. The more utilities/functions of the service system of VTS, the more expensive it is. Thus, among the evaluation model of service system for VTS, the functional criteria are mutually conflicted with the cost criteria. The VIKOR method is used to evaluate, improve and rank the performance of benchmarked alternatives. The VIKOR method is a multi-criteria decision making (MCDM)

ISBN: 978-1-61208-215-8

method, and is applied to solve a discrete decision problem with non-commensurable and conflicting criteria. This method focuses on ranking, improving and selecting the best alternative from a set of alternatives, and determines the compromise solution for a problem with conflicting criteria, which can help the decision-makers to reach a final best decision. Here, the compromise solution is a feasible solution closest to the ideal point (or closest to the aspired/desired levels in each criterion), and a compromise means an agreement established by mutual concessions. Thus, the VIKOR method would be applied to rank, evaluate and improve the performance of the best service systems of VTS. The basic concept of VIKOR is to identify the positive-ideal solution (the aspired/desired level) and the negative-ideal solution (the worst level). The positive solution is the best solution that satisfies the most required criteria, and the opposite is the negative-ideal solution. The VIKOR method could rank, improve and determine the difference of negative and positive ideal solutions between services/utilities of the existing service systems of VTS. When calculating the distance between the ideal solution and the proposed service systems of VTS, the scores of each criterion should be summarized. The gaps between the consumers' most satisfied one and most unsatisfied one is also analyzed, with respect to services/utilities of the existing service systems of VTS. The VIKOR method was started with the form of the $L_p - metric$, which was used as an aggregating function in a compromise programming method and it developed the multi-criteria measure for compromise ranking [15, 16]. VIKOR provided a maximum group utility of the "majority" and a minimum individual regret of the "opponent". The compromise solutions could be the base for negotiation, involving the decision makers' preferences by criteria weights (Figure 3).

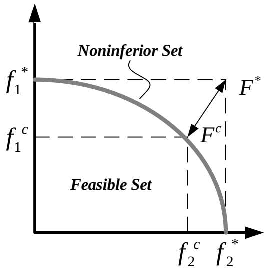


Figure 3. Ideal and compromise solutions

where: F^* is the ideal solution. f_1^* represents the ideal value (or called the aspired/desired level) of criterion 1. f_2^* represents the ideal value (the aspired/desired level) of criterion 2. The compromise solution, F^c, is a feasible solution that is "closest" to the ideal F^*. A compromise means an agreement established by mutual concessions. The VIKOR method is presented with the following steps:

(1) Determines the best f_k^* value and the worst f_k^- value in criterion i.

$$f_k^* = \left\{\left(\max_k f_{ik} \mid k \in I_1\right), \left(\min_k f_{ik} \mid k \in I_2\right); \text{or setting the aspired level for } i \text{ criterion}\right\}, \ \forall k \quad (10)$$

$$f_k^- = \left\{\left(\min_k f_{ik} \mid k \in I_1\right), \left(\max_k f_{ik} \mid k \in I_2\right); \text{or setting the worst level for } i \text{ criterion}\right\}, \ \forall k \quad (11)$$

where: k is the kth alternative; i is the criterion; f_{ik} is the performance value of the ith criterion of kth alternative; I_1 is the cluster of utility-oriented criteria; I_2 is the cluster of cost-oriented criteria; f_i^* is the positive-ideal solution (or setting the aspired level); and f_i^- is the positive-ideal solution (or setting the worst level).

(2) Computes the values S_k and Q_k, $k = 1, 2, \cdots, m$, using the relations.

Let r_{ik} be $r_{ik} = (| f_i^* - f_{ik} |)/(| f_i^* - f_i^- |)$. Before we formally introduce the basic concept of the solutions, let us define a class of distance functions.

$$d_k^p = \left\{\sum_{i=1}^{n} [w_i(| f_i^* - f_{ik} |)/(| f_i^* - f_i^- |)]^p\right\}^{1/p} = \left\{\sum_{i=1}^{n} [w_i r_{ik}]^p\right\}^{1/p}, p \ge 1 \quad (12)$$

$$S_k = d_k^{p=1} = \sum_{i=1}^{n} w_i r_{ik} \quad (13)$$

$$Q_k = d_k^{p=\infty} = \max_k \{r_{ik} \mid i = 1, 2, ..., n\} \quad (14)$$

where S_k shows the average gap for achieving the aspired/desired level; Q_k shows the maximal degree of regret for prior improvement of gap criterion. w_i is the weight of the criterion i and $i = 1, 2, ..., n$, expressing the relative importance value of the criteria gained via the application of the ANP method, based on NRM.

(3) Computes the index values R_k, $k = 1, 2, \cdots, m$, using the relation.

$$R_k = v(S_k - S^*)/(S^- - S^*) + (1-v)(Q_k - Q^*)/(Q^- - Q^*) \quad (15)$$

$$S^* = \min_k S_k, \ S^- = \max_k S_k$$

$$Q^* = \min_k Q_k, \ Q^- = \max_k Q_k$$

where $S^* = \min_k S_k$ (showing the minimal average gap is the best, we also can set $S^* = 0$), $S^- = \max_k S_k$ (we can set $S^- = 1$), $Q^* = \min_k Q_k$ (showing the minimal degree of regret is the best, we also can set $Q^* = 0$), $Q^- = \max_k Q_k$ (we can set $Q^- = 1$). We also can re-write Eq. (15), $R_k = vS_k + (1-v)Q_k$.

(4) Ranks the alternatives.

In addition, $0 \le v \le 1$ when $v > 0.5$, this indicates S is emphasized more than Q in Eq. (15), whereas when $v < 0.5$ this indicates Q is emphasized more than S in Eq. (15). More specifically, when $v = 1$, it represents a decision-making process that could use the strategy of maximum group utility; whereas when $v = 0$, it represents a decision-making process that could use the strategy of minimum individual regret, which is obtained among maximum individual regrets/gaps of lower level dimensions of each project (or aspects/objectives). The weight (v) would affect the ranking order of the dimensions/aspects/criteria and it is usually determined by the experts or decision making. In this paper, R_k (here, v =0.5) is applied to determine the customer satisfaction index (CSI). R_k could also consider the index of the maximum group utility and the minimum individual regret of the "opponent", where R_k smaller is better and $0 \leqq R_k \leqq 1$.

IV. THE EMPIRICAL ANALYSIS OF MARKET POSITION FOR VTS SERVICE SYSTEMS

As shown in Table X and Figure 4, the Type-G (0.524, 0.549) is the highest value in VSI and PSI, and Type-G is located in the common and luxurious (H, H). The Type-A (0.426, 0.492) and Type-O (0.423, 0.492) are the lowest value of PSI, the Type-O (0.423, 0.759) is the lowest value of VSI. Besides, the Type-T is located in the low price to catching market (L, H) and the Type O and Type-A are located in the no or limited choice (L, L). Accounting research result, the development trend of VTS service systems/device that generally move from the no or limited choice (L, L) to the low price to catching market by improving the PSI (Price satisfaction index), then move to the common and luxurious (H, H) through improving the VSI (Value satisfaction index). The development strategy of VTS service systems is to reduce the price of VTS service system, and to increase the value of VTS service system. Therefore, Continuing conducting product design innovation and service process improvement can increase the customers' satisfaction degree and create new value position helping VTS service operators to leave red sea market of high price competition into the blue sea market of low price competition.

TABLE X. THE VSI AND PSI OF VTS PRODUCTS UNDER v =0.5

	Type-T	Type-O	Type-A	Type-G
v=0.5	Taiwan	America	Europe	Japan
R_{vk}	0.556	0.577	0.574	0.476
VSI	0.444	0.423	0.426	0.524
VSI Rank	(2)	(4)	(3)	(1)
R_{pk}	0.475	0.508	0.508	0.451
PSI	0.525	0.492	0.492	0.549
PSI Rank	(2)	(3)	(3)	(1)

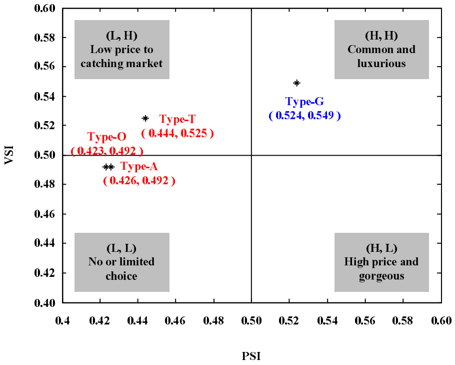


Figure 4. The map of service value position based on VSI and PSI

V. CONCLUSIONS

Vehicles telematics system (VTS) market includes five service operators (hardware suppliers, software operators, vehicles telematics system service operators, digital content service providers and telecommunications service operators), and the VTS service operators pay a key role in systems integration for diverse VTS system services. In some successful experience of service operators, there is a huge challenge about how to understand customers' needs. In the service system of VTS, the European and American VTS service operators laid particular stress on the service function regarding safety and security services, while the Japanese VTS service operators focus on navigation and map information services. VTS system service operators gradually strengthen the service function about communication and information services, and multimedia and entertainment services in the recent years. Taiwan's VTS service operators adopt open hybrid system based on Japanese navigation technologies and on Taiwanese users' needs such as high rate of car theft, frequent stowaway execution. Therefore, Taiwan's VTS service operators provide various service functions with location and navigation services and safety and security utilities such as guarding against burglary, assuring the safety, detection of vehicles towing. Considering the regional differences of market characteristics, the users' needs will be differed based on their environment. This study deliberates about the VTS service system development of some developed regions such as America, Europe, Japan, and analyzes the local condition and attributes of Taiwan. This study tries to find a suitable direction of VTS service market based on the VTS service system survey, and compares four regions commercial VTS service systems including America, Europe, Japan and Taiwan.

REFERENCE

[1] W. Lechner and S. Baumann, "Global navigation satellite systems," *Computers and Electronics in Agriculture,* vol. 25, no. 1-2, pp. 67-85, 2000.

[2] B. Sadoun and O. Al-Bayari, "Location based services using geographical information systems," *Computer Communications*, vol. 30, no. 16, pp. 3154-3160, 2007.

ISBN: 978-1-61208-215-8

[3] A. Theiss, D. C. Yen and C.-Y. Ku, “Global Positioning Systems: an analysis of applications, current development and future implementations,” *Computer Standards &amp; Interfaces,* vol. 27, no. 2, pp. 89-100, 2005.

[4] S. Pace, “The global positioning system: policy issues for an information technology,” *Space Policy,* vol. 12, no. 4, pp. 265-275, 1996.

[5] P. Zheng and L. Ni, "Introduction to Smart Phone and Mobile Computing," *Smart Phone and Next Generation Mobile Computing*, pp. 1-21, Burlington: Morgan Kaufmann, 2006.

[6] S. Opricovic and G. H. Tzeng, “Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS,” *European Journal of Operational Research,* vol. 156, no. 2, pp. 445-455, 2004.

[7] C. L. Lin C. W. Chen and G. H. Tzeng, “Planning the development strategy for the mobile communication package based on consumers' choice preferences,” *Expert Systems with Applications,* vol. 37, no. 7, pp. 4749-4760, 2010.

[8] R. A. Daveni, “Mapping your competitive position,” *Harvard Business Review,* vol. 85, no. 11, pp. 110-120, Nov, 2007.

[9] M. L. Tseng, “A causal and effect decision making model of service quality expectation using grey-fuzzy DEMATEL approach,” *Expert Systems with Applications,* vol. 36, no. 4, pp. 7738-7748, 2009.

[10] C. L. Lin and G. H. Tzeng, “A value-created system of science (technology) park by using DEMATEL,” *Expert Systems with Applications,* vol. 36, no. 6, pp. 9683-9697, 2009.

[11] M. P. Niemira and T. L. Saaty, “An Analytic Network Process model for financial-crisis forecasting,” *International Journal of Forecasting,* vol. 20, no. 4, pp. 573-587, 2004.

[12] C. W. Chang, C. R. Wu, C. T. Lin and H. L. Lin, “Evaluating digital video recorder systems using analytic hierarchy and analytic network processes,” *Information Sciences,* vol. 177, no. 16, pp. 3383-3396, 2007.

[13] J. L. Yang, H. N. Chiu, G. H. Tzeng and R. H. Yeh., “Vendor selection by integrated fuzzy MCDM techniques with independent and interdependent relationships,” *Information Sciences,* vol. 178, no. 21, pp. 4166-4183, 2008.

[14] I. Yüksel and M. Dagdeviren, “Using the analytic network process (ANP) in a SWOT analysis - A case study for a textile firm,” *Information Sciences,* vol. 177, no. 16, pp. 3364-3382, 2007.

[15] P. Yu, L., “A class of solutions for group decision problems,” *Management Science (pre-1986),* vol. 19, no. 8, pp. 936, 1973.

[16] M. Zeleny, "Multiple Criteria Decision Making," McGraw-Hill, ed., New York, 1982.

ISBN: 978-1-61208-215-8

Objective and Perceptual Impact of TCP Dropping Policies on VoIP Flows

Erika P. Alvarez-Flores
Centro de Estudios Superiores del Estado de Sonora
Hermosillo, México
erika.alvarez@cesues.edu.mx

Juan J. Ramos-Munoz, Jose M. Lopez-Vega and
Juan M. Lopez-Soler
Research Center on Information and Communications
Technologies (CITIC), University of Granada
Granada, Spain
jjramos@ugr.es, jmlvega@ugr.es and juanma@ugr.es

Abstract— **The integration of voice and data has allowed the development of new services and applications on the network. Despite of the growth of voice over IP applications, main active queue management (AQM) algorithms primarily focus on TCP traffic. An AQM scheme implicitly adopts a policy for selecting the packet to be marked or dropped. However, the traditional packet selection policies do not take into account the effect on the Quality of Service for different traffic types. In this case, is not always satisfied the quality of service of applications such as Voice over IP (VoIP), which require limited packet delay and loss rate. In this work we study the effect of TCP dropping policy over VoIP for scenarios with mixed traffic conditions. We evaluate their objective and perceptual impact on VoIP flows. As main result, we show that the adoption of one of the proposed AQM dropping procedures improves the user's perceptual score for VoIP, with no penalty on the TCP throughput and loss rate.**

Keywords - Active Queue Management; QoS; VoIP

I. INTRODUCTION

The adoption of Voice over IP (VoIP) technology has allowed the use of the same infrastructures for voice and data, saving installing and managing costs, thus enabling new services. Customer relationship systems (CRM), private branch exchange (PBX), or simply affordable international calls services profit with VoIP. However, IP networks were not designed for carrying multimedia data with real-time requirements. Therefore, the best-effort service that IP networks provides does not fit the VoIP needs.

In particular, one of the most impacting problems of IP for this type of traffic are the network congestion episodes: if routers become congested, the router may start dropping packets. While for TCP traffic it means that packets have to be resent and the transmission rate decreased, for real-time, UDP-based traffic the loss of packets may degrade the quality of the quality perceived by the final users. In addition, congestion may increase the end-to-end packet delay and the average packet jitter, impairments which also affects the VoIP flows quality.

To overcome this limitation and preventing congestion, Active Queue Management (AQM) schemes such as the well known Random Early Detection (RED) [1] have been proposed. AQM schemes monitor the router queue, triggering countermeasures to alleviate the congestion by marking or dropping packets. A TCP packet drop results in the decreasing of the TCP flow throughput, since the TCP congestion algorithm will be initiated. However, UDP flows such as VoIP do not react to these packet losses.

Generally, AQM schemes do not differentiate TCP and UDP traffic. In addition, in shared AQM queues, an inherent problem is that all packets will be exposed to the same drop probability regardless its source pattern. We consider that identifying the responsible source which is causing congestion may alleviate this issue. Unfortunately, up to the author's knowledge, no scalable solution for doing this has been found.

After a number of simulations, we have checked that a good approach for selecting the packet is to drop among both reactive (TCP) and non-reactive (UDP) packets. Otherwise, TCP traffic would be unfairly punished, and consequently non-reactive (UDP) traffic would be incorrectly favoured.

Our approach intends to reach a trade-off that avoids that some flows monopolize the available bandwidth and consequently penalize other flows in active queue management (AQM).

For this goal, in this work we study different dropping strategies for shared AQM queues and evaluate their impact on the VoIP quality of service (QoS). We will experimentally show that if we appropriately select the packet to be dropped, the network level QoS and the VoIP subjective end-user quality will be enhanced. In addition, we show that our approach is TCP friendly. That is, we prove that the VoIP traffic improvements have little impact on the TCP traffic performance.

The remainder of the paper has been organized as follows. In Section II, we briefly describe the AQM RED scheme. Different AQM packet dropping procedures are explained in Section III. The VoIP traffic evaluation framework is detailed in Section IV. Next, in Section V, we report the performance of the studied dropping strategies after some simulations. Finally, the paper is concluded in Section VI.

II. RANDOM EARLY DETECTION AND RELATED WORKS

The Random Early Detection queue management was first described in the seminar paper [1] by Floyd and Jacobson. RED gateways drop or mark each arriving packet with a certain probability, where the exact probability is a function of the average queue size. Its effectiveness is heavily dependent on the setting of its parameters.

Let Avg, Min_{th}, Max_{th} and Max_p be defined respectively as the RED average queue size, minimum threshold, maximum threshold and maximum packet drop probability. As detailed in [1], the packet drop probability is given by

$$P = Max_p \frac{Avg - Min_{th}}{Max_{th} - Min_{th}} \qquad (1)$$

RED estimates Avg as the exponentially weighted moving average, expressed as

$$Avg(t+1) = (1 - w_q)Avg(t) + w_q Q \qquad (2)$$

in which, $w_q \in [0,1]$ weights the current queue size (denoted by Q), and $(1-w_q)$ weights the previous long-term average value ($Avg\ (t)$).

RED scheme has motivated a significant number of interesting works. Different aspects have been widely studied in the literature since it was proposed. For instance, [2] and [3] have addressed the stability issue, giving as results the Stabilized RED algorithm (SRED) and improvements in Additive Increase and Multiplicative Decrease/RED systems, respectively. Some AQM schemes such as Balanced RED (BRED) [4] and Fairness-Improvement for RED (FI-RED) [5] have been also proposed to deal with fairness. The settings of RED's parameters have been also addressed, giving as results adaptive schemes such as the Self-Configuring RED [6] or frameworks to find the optimal value of Max_p in RED gateways [7]. Moreover, in [8] alternative packet dropping strategies, as the Drop Front Strategy, have been investigated. Finally, the provision of a better control over the burstiness traffic level has been considered in [9], providing as result the Modified RED (MRED) scheme.

However, the majority of the aforementioned schemes do not take into account the nature of the processed traffic when discarding packets, degrading thus the performance for certain types of traffic.

III. DROPPING STRATEGIES

For victim selection, we use an AQM variant referred to as the Drop-Sel algorithm [10]. Interestingly, Drop-Sel can be integrated into any AQM scheme. Drop-Sel defines three classes of traffic: real-time flows (VoIP), other UDP (O-UDP) and elastic flows (TCP). Drop-Sel records the queue occupancy of each class. When a new packet arrives to the router, and the AQM scheme decides to drop a packet, the Drop-Sel algorithm chooses a packet from the traffic with the highest consumption of memory space in the queue.

The basic idea behind Drop-Sel is to discard the packet belonging to the class of traffic which most contributes to congestion.

Additionally, to improve the QoS perceived by the end user, especially over connections with long propagation delays, it is important to adopt the packet dropping policy with improves the user´s perception.

In this occasion, we evaluate five procedures for TCP packet victim selection. Three of them consider the overall traffic pattern, and the other two are flow based. The policy for the UDP class is always to discard the UDP packet nearest to the front of the queue, whereas for TCP class, the following dropping schemes were considered:

- First-TCP: selects the TCP packet nearest to the front of the queue.
- Last-TCP: selects the TCP packet nearest to the tail of the queue.
- Random-TCP: randomly selects a TCP packet.
- Flow-TCP: selects the packet nearest to the front of the queue belonging to the most populated TCP flow.
- First-Flow-TCP: selects 2 packets. The TCP packet nearest to the front of the queue, and the packet nearest to the front of the queue belonging to the most populated TCP flow. In other words, we combined First-TCP and Flow-TCP. If the algorithm confirms that the selected packets are the same, it considers only one drop.

Fig. 1 shows a packet discarding example for each victim selection procedure.

IV. VOIP QOS ASSESSMENT

For a given voice communication system, the evaluation of the perceptual quality is a costly process, which even could be hardly reproducible. However, for the multimedia communication general case, and for VoIP applications in particular, QoS should somehow definitively include the final perceived user quality.

For quantifying the effect of the aforementioned dropping strategies on the transmission quality, we adopt the E-model and ITU-T Recommendation G.107 [11]. The E-model was initially conceived for network planning design purposes; it predicts the subjective effect of combinations of impairments using stored information on the effects of individual impairments. However, it has also been adopted to estimate the subjective QoS perceived by the user in many voice transmission systems.

For this purpose, the model is usually simplified for the sake of practicality. Henceforth, we adopt the E-model setup

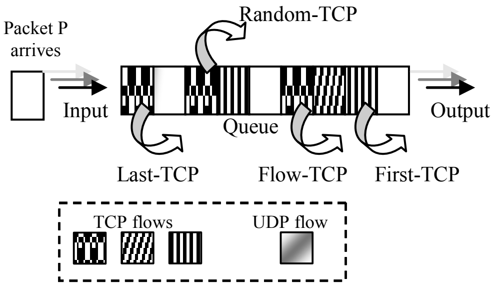


Figure 1. Dropping Strategies Example.

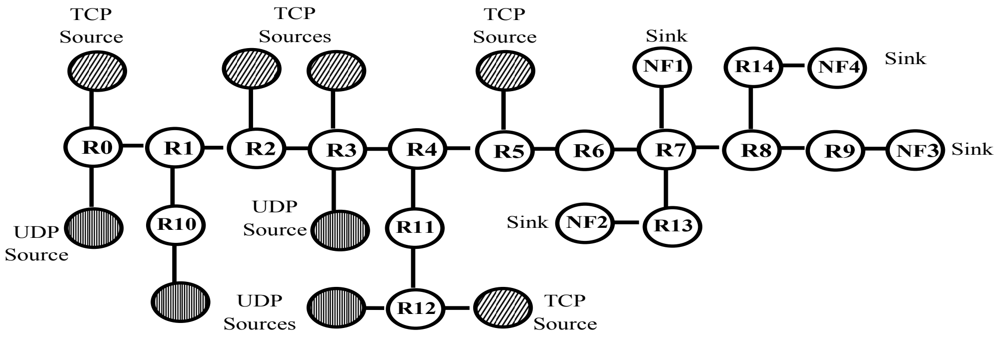


Figure 2. Simulated Topology.

proposed in [12], and obtain the R factor using (3), defined as:

$$R=94.2-0.11(d-177.3)\,H - 0.024d - 30\log(1+15p) \quad (3)$$

where d -expressed in milliseconds- is the end-to-end average delay of the VoIP packets, p is the loss packet probability, and H shapes the delay contribution according to the following equations,

$$\begin{aligned} H &= 0\, if\,(d-177.3) < 0 \\ H &= 1\, if\,(d-177.3) \geq 0 \end{aligned} \quad (4)$$

To provide more readable subjective evaluations, the R factor can be mapped to MOS score [11]. Like [13], we show the impact of the dropping strategies on VoIP traffic in terms of the MOS scale.

V. Experimental Evaluation

A. Experimental Setup

In order to properly assess both the network-level and the user-level impact of the proposed dropping strategies, we conducted a number of simulations with the ns-2 simulator [14].

Althought Drop-Sel algorithm can be used in conjunction with any AQM scheme, we adopted the RED scheme in our simulations. More precisely, we evaluated the selected AQM RED scheme implementing the five alternative victim selection algorithms described in Section III.

We simulated the reference topology shown in Fig. 2. Links bandwidths are all set to 10 Mbps, except for link R6-R7. To cause network congestion at the AQM node, the R6-R7 link bandwidth is restricted to a narrower bandwidth of 4.5 Mbps. Therefore, R6-R7 is the "bottleneck" link.

The simulated topology represents a complex reference scenario in which a number of elastic TCP and voice over UDP flows compete for the resources of the AQM router.

Since sources are located at different distances from the AQM node, they will undergo a range of different round-trip time (RTT) delays for the TCP flows, and end-to-end delays for the different VoIP flows. The resulting end-to-end delays are close to the maximum allowed delay, 300 ms, so the delay introduced at the router may cause useless expired VoIP packets [15].

TABLE I UDP FLOWS SPECIFICATION

Flow	Source connected	Sink	Interpacket period	Packet size	End-to-end delay
A	R0	NF1	30ms	292 bytes	234 ms
B	R10	NF2	10ms	132 bytes	225 ms
C	R10	NF2	10ms	132 bytes	225 ms
D	R3	NF3	60ms	532 bytes	218 ms
E	R12	NF4	30ms	292 bytes	231 ms

The audio sources generate RTP packets encapsulated into UDP datagrams during random periods, in accordance with the configuration detailed below. The FTP sources also generate the TCP traffic at random variable periods.

The VoIP and FTP applications are modeled as ON/OFF traffic sources. The ON period for the VoIP application lasts 180 seconds, and the OFF period lasts 100 seconds. The FTP traffic follows a Pareto distribution with a shape parameter of k=1.4, an average ON period equal to 2 seconds, and an OFF period that follows an exponential distribution with an average duration of 1 second.

During their ON period, the FTP sources generate TCP segments with a length of 1500 bytes at 1 Mbps. Since different TCP flavours may lead to different results [16], we will evaluate two different implementations. They differ in the ACK algorithms employed by the TCP receiver; the classical ACK (noted hereafter as SACK(1)) and delayed ACK algorithm (called hereafter as SACK(2)) will be tested.

In our setup VoIP sources generate constant bit rate flows that represent voice streams encoded with a G.711 vocodec [17]. To provide a scenario with heterogeneous VoIP applications, the VoIP flows may generate several G.711 frames per packet. This configuration is described in more detail in Table I.

The generated flows go from the node connected to a VoIP or a FTP traffic generator to one of the sink nodes. Specifically, from node R0 to the NF1 sink node, from R3 to NF3, from R12 to NF4, and from nodes R2, R5 and R10 to sink node NF2. Each simulation lasts 500 seconds.

TABLE II LOSS RATE OF AUDIO PACKET BY FLOW

		SACK(1)					SACK(2)				
Flow	Metric	Last TCP	Random TCP	First TCP	Flow TCP	First Flow TCP	Last TCP	Random TCP	First TCP	Flow TCP	First Flow TCP
A	Drop rate at AQM (%)	1.10	1.08	0.98	1.04	0.66	0.43	0.52	0.56	0.49	0.32
	Useless packets (%)	24.24	20.21	17.37	16.15	10.43	19.88	16.56	15.33	15.51	11.02
	Total	25.34	21.29	18.35	17.19	11.09	20.31	17.08	15.89	16.00	11.34
B	Drop rate at AQM (%)	0.93	1.04	0.92	0.93	0.66	0.50	0.48	0.48	0.55	0.31
	Useless packets (%)	12.51	7.83	4.72	4.87	2.25	9.65	6.26	4.79	5.29	2.81
	Total	13.44	8.87	5.64	4.80	2.91	10.15	6.74	5.27	5.84	3.12
C	Drop rate at AQM (%)	1.54	1.43	1.46	1.38	0.99	0.86	0.68	0.70	0.79	0.47
	Useless packets (%)	12.28	7.66	4.60	4.77	2.19	9.44	6.23	4.64	5.24	2.77
	Total	13.82	9.09	6.06	6.15	3.18	10.30	6.91	5.34	6.03	3.24
D	Drop rate at AQM (%)	1.01	0.88	0.70	0.63	0.50	0.42	0.48	0.35	0.40	0.22
	Useless packets (%)	4.71	1.20	0.52	0.80	0.17	3.35	1.18	0.72	0.88	0.23
	Total	5.72	2.08	1.22	1.43	0.67	3.77	1.66	1.07	1.28	0.45
E	Drop rate at AQM (%)	0.90	0.68	0.67	0.84	0.70	0.42	0.38	0.38	0.38	0.28
	Useless packets (%)	19.05	14.39	11.39	10.43	6.01	15.52	11.98	10.13	10.35	6.78
	Total	19.95	15.07	12.06	11.27	6.71	15.94	12.36	10.51	10.73	7.06
Average results all flows	Drop rate at AQM (%)	1.17	1.13	1.08	1.07	0.78	0.60	0.54	0.54	0.59	0.36
	Useless packets (%)	14.12	9.61	6.70	6.58	3.51	11.10	7.84	6.37	6.81	4.08
	Total	15.29	10.74	7.78	7.65	4.29	11.70	8.38	6.91	7.40	4.44

B. *Experimental Results*

Given that VoIP quality depends on the loss rate, we evaluated in particular this factor. In our scenario there are two sources of packet losses: firstly, packets are dropped at the AQM router for notifying or preventing congestion; and secondly, useless late packets – those which accumulate an end-to-end delay exceeding 300 ms- are also dropped at the final user node.

To show the effect of the packet victim selection procedures on the packet loss probability, Table II gives the results obtained for the different VoIP flows of the simulated scenario.

It is experimentally shown that in both the drop rate at the AQM router and the drop rate at the final user due to the useless packets, Last-TCP causes the highest packet loss probability at the RED queue. For instance with SACK(1), First-TCP produces an average total loss rate of 7.78% audio packets, while Last-TCP generates 15.29%. Although this degraded performance also occurs with SACK(2), this is less severe. In that case, First-TCP produces an average total loss rate of 6.91% while Last-TCP generates 11.70%

On the other hand, First-TCP, Flow-TCP, and First-Flow-TCP algorithms (with both SACK TCP variants) achieve a significant decrease in the number of useless packets at final user's site, compared to Random-TCP. For example, it is shown for flows A and E, the ones with largest end-to-end delay, that loss rate fluctuates from 11.98% to 20.21% with Random-TCP while it varies from 6.01% to 17.37% with the other three procedures. This result can be explained because of dropping packets from the front generates an empty slot at the head of the queue, and it reduces the delay of all queued packets behind the dropped packet. Therefore, it provides a reduction in the overall end-to-end VoIP delay, making it possible to diminish the number of useless packet.

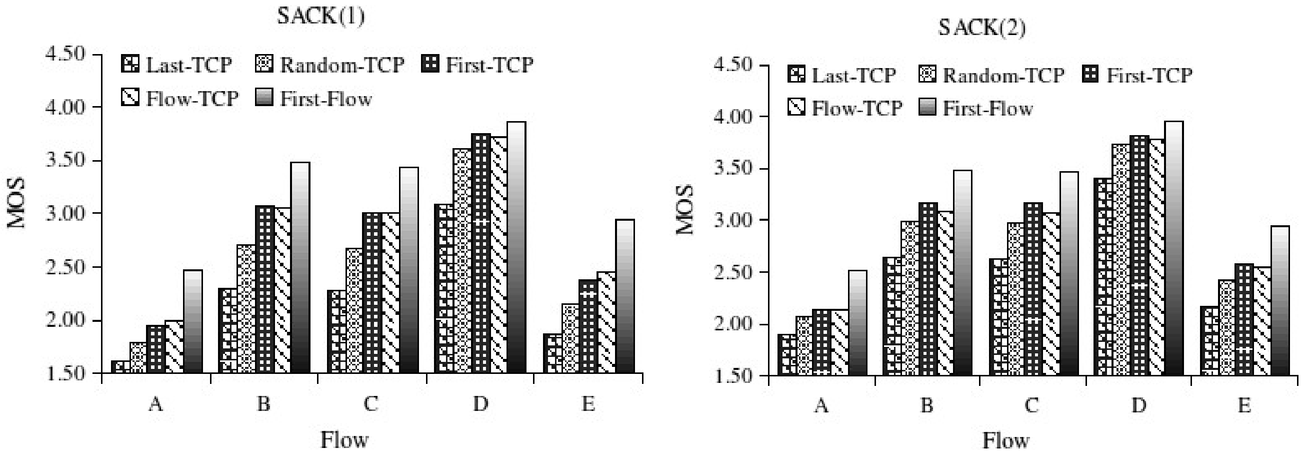
However, the most significant reduction is obtained with First-Flow-TCP. This algorithm applies a more aggressive dropping policy that causes double packet drops, providing an earlier congestion detection and notification simultaneously to several flows. An interesting point is that First-Flow-TCP controls more effectively the sending rate of the TCP connections with the shortest RTT, generating additional empty slots in the queue at the same time. We see in Table II that this dropping strategy outperforms the other procedures, regardless of how far the sources are from the AQM router. Note that it improves both the drop rate at the router and the number of useless packets percentage. Thus, it causes the lowest loss probability between the evaluated dropping strategies with both SACK TCP variants.

Note that even though sometimes two different TCP flow simultaneously reduce their sending rate, the global TCP performance is not severely degraded. Table III and IV show the TCP loss rate at the AQM router and the TCP arrival rate at the final user, respectively. Note that there is no significant difference between the results obtained by the discarding procedures for the same TCP flavour.

If we reduce the packet loss at the router and at the final user side, we could expect a significant improvement in the subjective end-user perceived quality of VoIP. Fig. 3(a) and 3(b) show the MOS values obtained for each flow with different packet victim selection procedures.

As it can be observed, First-Flow-TCP achieves the best results for VoIP traffic. For all flows, the highest MOS values are obtained using the First-Flow-TCP procedure. This means that using a more aggressive dropping policy with TCP traffic, better QoS is provided for VoIP traffics.

On the other hand, without imposing an aggressive dropping rate, the experiments carried out show that the discarding of packets nearest to the front of the queue achieves significantly improvement in MOS values.



(a)

(b)

Figure 3. E-MODEL based MOS evaluation of dropping policies.

TABLE III LOSS RATE OF TCP PACKET (%)

	Last TCP	Random TCP	First TCP	Flow TCP	First Flow
SACK(1)	4.00	3.82	3.84	3.36	3.64
SACK(2)	2.06	1.95	1.94	1.84	1.94

TABLE IV ARRIVAL RATE OF TCP PACKET (MBPS)

	Last TCP	Random TCP	First TCP	Flow TCP	First Flow
SACK(1)	4.098	4.110	4.117	4.132	4.099
SACK(2)	3.879	3.891	3.924	3.985	3.861

As expected, MOS results obtained for the different procedures have a rational correlation with network-level results.

VI. CONCLUSION AND FUTURE WORK

In this work, we have studied different TCP packets dropping policies for the Drop-Sel victim selection scheme. We have assessed their impact on the quality of service perceived by end users. We have evaluated their performance by means of simulation, obtaining network parameters and perceptual scores which experimentally demonstrate the benefits of using specific policies for TCP packets. Such benefit results in an enhanced perceived quality of the VoIP flows, without degrading the TCP flows performance.

The experiments also show that despite of using an aggressive discard policy for TCP, as the First-Flow-TCP procedure does, discarding up to two packets at once, the TCP performance is not significantly penalized. Those results have been validated for two SACK TCP variants.

As future work, we plan to evaluate the impact of AQM schemes on VoIP and other real-time media flows such as video streaming. For instance, the use of TCP and HTTP as transport protocols for streaming media with time constraints [18] will be also considered. Additionally, we will study Drop-Sel based algorithms to cope with that class of TCP traffic.

ACKNOWLEDGMENT

This work was supported by the "Ministerio de Ciencia e Innovación" of Spain under research project TIN2010-20323.

REFERENCES

[1] S. Floyd and V. Jacobson, “Random early detection gateways for congestion avoindance,” IEEE/ACM Transaction on Networking, August, 1993. doi: 10.1109/90.251892.

[2] T. J. Ott, T. V. Lakshman, and L. Wong, “SRED: Stabilized RED,” IEEE, 1999, pp. 1346-1355. doi: 10.1109/INFCOM.1999.752153.

[3] L. Wang , L. Cai , X. Liu, and X. Shen, “Stability and TCP-friendliness of AIMD/RED system with feeback delays,“ Computer Networks 51, 2007, pp. 4475-4491. doi: 10.1016/j.comnet.2007.06.022.

[4] F. M. Anjum and L. Tassiulas, “Fair bandwidth sharing among adaptive and non-adaptive flows in the internet,” IEEE, 1999, pp. 1412-1420. doi: 10.1109/INFCOM.1999.752161.

[5] H. Ohsaki, T. Eguchi, and M. Murata, “FI-RED: AQM Mechanism for improving fairness among TCP connections in tandem networks,” Proc. IEEE SAINT ’05, 2005. doi: 10.1109/SAINT.2006.33.

[6] W. Feng, D. Kandlur, D. Saha, and G. Kang, “Self-configuring RED gateway,” Proc. IEEE INFOCOM, 1999. pp. 1320-1328. doi: 10.1109/INFCOM.1999.752150.

[7] B. Zheng and M. Atiquzzaman, “A framework to determine bounds of maximum loss rate parameter of RED queue for next generation routers,” J. Network and Computer Applications, 2008. doi: 10.1016/j.jnca.2008.02.003.

[8] T. V. Lakshman, A. Neidhardt, and T. J. Ott, “The drop front strategy in TCP and in TCP over ATM,” IEEE, 1996. doi: 10.1109/INFCOM.1996.493070.

[9] G. Feng, A. Agarwal, A. Jayaraman, and C. Siew, “ Modified RED gateways under bursty traffic,” IEEE Commun Lett, 2004, pp. 323-325. doi: 10.1109/LCOMM.2004.827427.

[10] E. P. Alvarez-Flores, J. J. Ramos-Munoz, P. Ameigeiras, and J. M. López-Soler, “Selective packet dropping for VoIP and TCP flows”, Telecommunication System Journal, Vol. 46 No.1, 2011, pp. 1-16. doi: 10.1007/s11235-009-9252-z

[11] ITU-T, Recommendation G-107, “The E-model, a computational model for use in transmission planning,” March 2005.

[12] R. G. Cole and J. H. Rosenbluth, “Voice over IP performance monitoring,” Proc. SIGCOMM Comput. Commun. Rev. 31, 2001, pp. 269-275. doi: 10.1145/505666.505669.

[13] V. A. Reguera, F. F. Alvarez Paliza, W. Godoy Jr., and E. M. García Fernández, “On the impact of active queue management on VoIP quality of service,” Computer Communications 31, 2008 pp. 73-87. doi: 10.1016/j.comcom.2007.10.016.

[14] Network Simulator ns2. [Online]. Available from: http://www.isi.edu/nsnam/ns/.

[15] C. Perkins, O. Hodson, and V. Hardman, “A survey of packet-loss recovery techniques for streaming audio,” Proc. IEEE Network, 1998, pp. 40-48. doi: 10.1109/65.730750 .

[16] M. Allman, V. Paxson, and W. Stevens, “TCP Congestion Control”, IETF RFC 2581, April, 1999.

[17] ITU-T, Recommendation G.711, “Pulse code modulation (PCM) of voice frequencies”.

[18] R. Pantos, "HTTP Live Streaming", IETF Internet-Draft, September 30, 2011.

Workload Characterization for Stability-As-A-Service

Vaishali Sadaphal and Maitreya Natu
Systems Research Lab, Tata Research Development and Design Centre (TRDDC), Pune.

***Abstract*—*Stability As A Service* is critical for data centers and workload characterization is one of the essential services within *Stability As A Service*. Different solutions have been proposed in the past for characterizing various system properties, however most of these approaches often require high levels of instrumentation and intrusiveness. This makes their use difficult in real-world production systems. In this paper, we argue that many interesting insights can be derived to characterize system workload simply by analyzing the transaction logs captured at different layers in the system such as access logs at application servers, SQL logs at database servers, among others. We present two approaches that analyze various dimensions and measures of the transaction logs in order to characterize workload. We demonstrate the effectiveness of the proposed approach through real-world case-studies and show how the findings of the two approaches complement each-other.**

Keywords: Stability As A Services, Performance and Capacity Management, Workload Characterization

I. Introduction

Today's computing requirements make data centers large and complex. After the initial infrastructure and application planning, the data centers keep evolving and growing to accommodate newer requirements. The continuously evolving nature and the ad-hoc growth of the data centers increases the complexity manifold. In one of the leading telecommunication companies, we observed the number of users multiplying every 15 days resulting in addition of infrastructure components, and installation of more application instances. Due to continuous evolution and ad-hoc growth, the data centers become brittle. Even a small change in an application or infrastructure carries a risk of breakdown and instability. Compromising stability of the system and breach of SLAs (Service Level Agreements) result into incurring large amount of financial losses. All these factors have made *Stability As A Service* critical for data centers. Goal of *Stability As A Service* is to ensure zero down-time, provide consistent performance, and quickly detect and resolve any instability.

A quintessential requirement for providing *Stability As A Service* is the understanding of the as-is state of system operations. The data center operators need to understand the workload patterns, performance areas, system hotspots, bottlenecks, among others. In this paper, we address one part of this puzzle by focusing on the problem of workload characterization. We present algorithms to demonstrate how the state of the art logs can be analyzed in a systematic, efficient, and automated manner to extract various properties of system workload. Workload characterization can be exposed as a service to provide the analysis of system workload properties at various layers. This service can extract properties such as workload patterns, heavy hitters, anomalous behavior, etc. The service can be used in the overall offering of *Stability As A Service* to obtain insights to better understand system behavior, do base-lining of as-is behavior, identify heavily-used and poor-performing areas, and plan for growth and optimization.

Problems related to workload characterization have been addressed in the past [11], [6]. Solutions have been proposed based on techniques such as analysis of paths followed by requests, generation of resource signatures for the request types, observing temporal communication patterns using frequent subgraph discovery, etc. Effectiveness of most of these solutions depends on the availability of wide instrumentation and use of intrusive techniques. However, the data center administrators are reluctant to introduce intrusive solutions in the production environment because of the risk of modification of system behaviour. This makes many prior solutions difficult to deploy in real-world operations. A practical solution for workload characterization is to perform analysis using logs that are commonly captured in most operational systems with minimal instrumentation and intrusiveness. The workload is commonly captured at various layers in a system hierarchy in the form of access logs. For instance, the application servers store the transaction logs, database servers store SQL logs, and disks capture the read and write operations. We argue that analysis of these logs, obtained by simple, non-intrusive, at-a-point monitoring solutions, can also provide a practical and easy-to-use solution while providing many useful insights.

In this paper, we target analysis of semi structured or structured application monitoring logs. These logs are parsed to construct a multi-column dataset where each column maps to a metric being monitored. Each row contains the value of the observed metric at a time instance. The metrics are classified into two types viz. measures and dimensions. The measures are numeric representatives of various system properties such as latency, arrival rate, etc. The dimensions are categorical attributes that provide structured labeling information to the measures. The dimensions thus provide a mechanism to filter, group, and label the measures. For instance, consider a web transaction log that contains an entry of each request received by the web server containing the request time stamp, HTTP method (GET/POST), URL, client IP address, and the time to respond. In this log, the time to respond metric is a measure while HTTP method, URL, and client IP address are the dimensions. In this paper, we present algorithms to analyze hidden relationships between these measures and dimensions. Extracting such relationships helps answering questions such as (1) which set of requests should be optimized to result in maximum improvement, (2) requests with which dimension value require largest amount of resource, etc.

The workload or application monitoring log typically contains a dump of all types of activities observed in the system. Various details of the activities such as workload patterns, heavy hitters, anomalous behavior, etc., tend to get lost in the monitoring logs. For instance, in the example of web transactions log, all types of transactions are reported in a single log. Various properties of these individual groups of transactions are thus lost in this collective log. This paper addresses the problem of understanding the workload characteristics captured in the monitoring logs.

Existing literature in the area of data-mining can be leverages to analyze the dimensions and measures of these logs. Techniques for clustering, feature selection, subgroup discovery, detection of components in Gaussian mixture models, can be leveraged for developing workload characterization services. In the following, we present two approaches for workload characterization:

Consider a database P, with N records, where each record has a measure L and k dimensions $D = \{D_1, D_2, \ldots, D_k\}$. Each dimension D_i consists of its domain $Domain(D_i) = \{d_1^i, \ldots, d_n^i\}$ that represent all possible values of the dimension D_i. A dimension-value pair is referred as $(D_i = d_j)$. A descriptor θ denotes a given record type which consists of one or more dimension-value pairs.

Multi-modal analysis (MMA): Using the observation that the real world processes follow Gaussian distributions, we discover Gaussian distributions in the measure L to characterize the workload with respect to a measure. Given these modes of measure values, we next use the dimension information to tag each mode with a descriptor θ that best represents the records belonging to a mode. *Interesting subset discovery (ISD):* In contrast to the multi-modal analysis, in this approach we first explore the space of dimensions. We first identify possible combinations of dimension descriptors, identify the set of records that are explained by the dimension-descriptors, and then use the measure L to compute interestingness of the subset.

Both approaches complement each other in their findings as follows:

- While multi-modal analysis analyzes the entire spectrum of measure values, interesting subset discovery only focuses on the extreme ends of interestingly high and low measure values.
- Multi-modal analysis tries to find the dominant property and tries to describe that property with a representative dimension-value pair. The set of descriptors, thus identified, most often represent large sets. Interesting subset discovery, on the other hand, identifies the sets that are interestingly different in their statistical properties. The set of descriptors, thus identified, represent anomalies an most often represent small and medium sets.

We discuss the approach for multi-modal analysis in Section II and interesting subset discovery in Section III. We demonstrate the utility of the proposed approaches through real-world case-studies in Section V. We summarize our contributions in Section VI.

II. Multi-modal analysis (MMA)

In order to develop this approach, two sub-problems need to be solved - (1) how to construct modes for a workload measure?, and (2) how to identify descriptors that best describe the requests that belong to one mode? We address these two problems below:

A. Construction of modes

Given a measure L with n values and mean and standard deviation as $< \mu, \sigma >$, we identify the set M consisting of m modes $M = \{M_1, M_2, \ldots, M_m\}$ that best classify the values of L. A mode M_i is a set of values in measure L that belong to one mode or Gaussian distribution. A 3-tuple $< \mu_i, \sigma_i, w_i >$ describes each mode M_i where μ_i, σ_i is the mean and standard deviation of values in mode M_i and w_i is the weight defined as the fraction of points belonging to the mode M_i given by $\frac{|M_i|}{n}$ where $n = \sum_{i=1}^{i=m} |M_i|$.

We use Expectation maximization algorithm [5] to estimate the parameters of the modes such as mean and standard deviation. Host of algorithms exist to identify number of modes in a given set of values [12]. However, the modes produced by these algorithms need to be refined for various reasons, such as (1) Many modes may be very similar in their mean values and are required to be merged. (2) Modes may have very large standard deviation and they need to be split to form multiple modes with smaller standard deviation. The algorithms proposed in the past require manual tuning of a parameters that merge or split the modes.

We address this problem by defining a self-tunable threshold T_{merge} on the basis of which we merge or split the modes. We use coefficient of kurtosis to tune the threshold T_{merge}. The coefficient of kurtosis is a statistical measure of values that signifies the peakiness of a distribution. For Gaussian distributions coefficient of kurtosis is known to be near-zero. We compute coefficient of kurtosis of the entire set of values and infer if the distribution is uni-modal or multi-modal. We merge the modes aggressively in case the computed values of kurtosis is near zero assuming that the distribution in uni-modal else we assume that the distribution is multi-modal and adopt a conservative approach while merging the modes. Another way is to compute coefficient of kurtosis for each identified mode and split the mode in case the distribution is inferred to be multi-modal on the basis of value of coefficient of kurtosis.

B. Identifying mode-descriptors

Given a set M of modes, we find the descriptor $\theta = (D_i = d_j)$ that has largest probability of explaining the values of measure L observed in a mode. We present two approaches to identify mode-descriptors. We first present a score-based approach that assigns scores to each dimension descriptor to find the best descriptor. We then present a less compute-intensive feature-selection-based approach that uses classification and regressions trees to identify descriptors. The basic idea is to evaluate each dimension on how well can its descriptors describe all modes.

1) Score-based approach: We first compute the probability that the descriptors $(D_i = d_j)$ explains a mode and then present an approach to compute score of dimension using score of descriptors.

Computing probability that a descriptor explains a mode: The probability p^x_{ij} that a descriptor $(D_i = d_j)$ explains a mode $M_x \in M$is computed by identifying the percentage of values in M_x that hold property $D_i = d_j$.

Formally, if r_k is the k^{th} record, $Value(r_k, D_i)$ is the value in the dimension field D_i of the record r_k, and $Value(r_k, L)$ is the value in the measure field L of record r_k. then,

$$p^x_{ij} = \frac{|M_x = \{d_j | D_i = d_j)|}{N} \tag{1}$$

where M_x is the set of records in which $Value(r_k, D_i) = d_j$.

Computing score of a dimension: The score of a dimension D_i indicates how well is each of the mode described by the descriptor of dimension D_i. In order to calculate the score of a dimension D_i, first the probability of best descriptor of D_i is calculated for each mode. For a mode M_x the best D_i descriptor $(D_i = d_j)$ is defined as the descriptor with maximum probability p^x_{ij}. The score of a dimension D_i is then calculated by taking an average of the probability of the best descriptor for each mode.

The score-based approach computes the score for each dimension using the above approach. The dimension with maximum score is chosen as the best dimension. For each mode, the descriptor of the best dimension that has maximum probability is selected as the best descriptor.

2) Feature selection based approach: The score-based method is computationally intensive since it requires inspecting the records for every value of every dimension. We next propose an efficient algorithm that uses Classification and Regression Trees (CARTs) [3] to identify the best dimension to describe all the modes. The basic idea is to assign a class label to each record based on the mode to which the record belongs. We then use CART to identify dimensions that best classify these class labels. CART is constructed such that the class labels form the leaves and the dimension descriptors form the intermediate nodes.

CARTs can be used in two ways: One approach is to construct the entire CART and select the dimension that dominates across the non-leaf nodes. A simpler and more efficient approach is to select the dimension used as the root node. While building a CART, the root node is chosen as the node that provides the largest improvement in the classification accuracy of the leaf nodes.

Incorporating multiple dimensions: In certain cases, no single dimension best describes all the modes. To address such scenarios, we extend the proposed approach to analyze multiple dimensions to identify the best descriptor for every mode. We do this by identifying multiple candidate dimensions that can act as classifiers. Instead of using only one dimension with largest improvement in classification accuracy in CARTs, we propose to also use next few candidate dimensions with respect to the improvement in classification accuracy. We then identify the best descriptors across these selected multiple dimensions.

III. Interesting Subset Discovery (ISD)

The basic idea behind the proposed approach is to exploit the workload descriptors such as URLs, client IPs, and other dimensions of the dataset to construct subsets and then use the workload measures such as workload, performance, throughput, and other measures to define interestingness of the subsets. In order to develop this approach, two sub-problems need to be solved - (1) how to compute interestingness of a subset? (2) how to efficiently navigate through the large search space of all possible descriptors and their combinations? We address these two problems below:

This problem is similar to the subgroup discovery problem, which is well-known in data mining [2], except that the *class label* column (e.g. response time) in continuous in our case, rather than a finite discrete set. In the past, we have used the problem of interesting subset discovery in the domain of ticket analysis for IT infrastructure support [10]. In this paper, we present an application of interesting subset discovery for workload characterization and demonstrate how its findings complement the findings of multi-modal analysis.

A. Interestingness of a subset

Let descriptor θ denote a given record type and let $D[\theta]$ be the set of record of type θ. We build multisets $L[\theta]$ and $\overline{L}[]$ consisting of only the values of measure L for the records in the set $D[\theta]$ and its complement $\overline{D}[\theta]$) respectively. We use the 2-sample 2-tailed Student's t-test to compare the multisets $L[\theta]$ and $\overline{L}[\theta]$. Student's t-test makes a null hypothesis that both these sets of L values are drawn from the same probability distribution. It computes a t-statistic for two sets X and Y ($L[\theta]$ and $\overline{L}[\theta]$ in our case) as follows:

$$t = (X_{mean} - Y_{mean})/\sqrt{(S_x^2/|X| + S_y^2/|Y|)}$$

S_X, S_Y denote the unbiased estimators of the standard deviations of the values in X and Y The denominator is a measure of the variability of the data and is called the *standard error of difference*. Another quantity called the p-value is also calculated. The p-value is the probability of obtaining the t-statistic more extreme than the observed test statistic under null hypothesis. If the calculated p-value is less than a threshold chosen for statistical significance (usually 0.05), then the null hypothesis is rejected; otherwise the null hypothesis is accepted. Rejection of null hypothesis means that the means of two sets do differ significantly. A positive t-value indicates that the set X has higher values than the set Y and negative t-value indicates smaller values of X as compared to Y.

B. Construction of subsets

Thus given a specific descriptor, we can use the t-test to decide whether or not the descriptor defines an interesting set. The main question now is how to systematically and efficiently search the space of all possible sets to identify interesting sets.

We build subsets of records in an incremental manner starting with level 1 subsets and increase the descriptor size in each iteration. The subsets built in first iteration are level 1 subsets. These subsets correspond to the descriptors $(D_i = u)$ for each dimension $D_i \in D$ and each value $u \in DOM(D_i)$. The subsets built at level 2 correspond to the descriptors $\{(D_i = u), (D_j = v)\}$ for each pair of distinct dimensions $D_i, D_j \in D$, for each value $u \in DOM(D_i)$ and $v \in DOM(D_j)$.

The brute-force approach is to systematically generate all possible level-1 descriptors, level-2 descriptors, ..., level-k descriptors. For each descriptor θ construct subset D_θ of D and use the t-test to check whether or not the subsets L_θ and $\overline{L_\theta}$ of their measure values are statistically different. If yes, report D_θ as interesting. Clearly, this approach is not scalable for large datasets, since a subset of N elements has 2^N subsets. We next propose various heuristics to limit the exploration of the subset space.

1) The size heuristic: The *t-test* results on the subsets with very small size can be noisy leading to incorrect inference of interesting subsets. Small subset sizes are not able to capture the properties of the record dimensions represented by the subset. Thus by the size heuristic we apply a threshold M_s and do not explore the subsets with size less than M_s.

2) The goodness heuristic: While identifying interesting subsets of records that have performance values greater than the rest of the records the subsets with the performance values lesser than the rest of the records can be pruned. In the web transaction records case, as we are using the case of identifying the requests that perform significantly worse than the rest of the requests in terms of the *response time*, we refer to this heuristic as the goodness heuristic. By the goodness heuristic, if a subset of records show significantly better performance than the rest of the records then we prune the subset. We define a threshold M_g for the goodness measure. In the case of the web transaction records database with *response time* as the performance measure, a subset is pruned if the *t-test* result of the subset has a *t-value* < 0 and a *p-value* $< M_g$.

3) The p-prediction heuristic: A level k subset is built from two subsets of level $k - 1$ that share a common $k - 2$ level subset and the same domain values for each of the $k - 2$ dimensions. The p-prediction heuristic prevents combination of two subsets that are statistically very different, where the statistical difference is measured by the *p-value* of the *t-test*. We observed that if the two level $k - 1$ subsets are statistically different mutually, then the corresponding level k subset built from the two sets is likely to be less different from the rest of the data.

Consider two level $k - 1$ subsets D_{θ_1} and D_{θ_2} of the database D. Let the p-values of the *t-test* ran on performance data of these subsets and that of the rest of data are p_1 and p_2 respectively. Let p_{12} be the mutual p-value of the *t-test* ran on the performance data L_{θ_1} and L_{theta_2}. Let D_{θ_3} be the level k subset built over the subsets D_{θ_1} and D_{θ_2} and p_3 be the p-value of the *t-test* ran on the performance data L_{θ_3} and $\overline{L_{\theta_3}}$. Then the p-prediction heuristic states that $if(p_{12} < M_p)$ *then* $p_3 >$ *min*(p_1, p_2), where M_p is the threshold defined for the p-prediction heuristic. We hence do not explore the set D_3 if $p_{12} < M_p$.

4) Beam search strategy: We also use the well known beam search strategy [4], in that after each level, only top b candidate descriptors are retained for extension in the next level, where the beam size b is user-specified.

5) Sampling: The above heuristics reduce the search space as compared to the brute force based algorithm. But for very large data set (in the order of millions of records) the search space can still be large leading to unacceptable execution time. We hence propose to identify interesting subsets by performing sampling of the data set and using the above mentioned heuristics on the samples. The algorithm then retains only the most frequently occurring subsets in results obtained from several samples.

C. Algorithm for interesting subset discovery

Based on the above explained heuristics, we present Algorithm ISD for discovery of interesting subsets in an efficient manner. The algorithm builds a level k subset from the subsets at level *k-1*. A level *k-1* descriptor can be combined to another level *k-1* descriptor that has exactly one different dimension-value pair.

Before combining two subsets, the algorithm applies the p-prediction heuristic and skips the combination of the subsets if the mutual p-value of the two subsets is less than the threshold M_p. The subsets that pass the p-prediction heuristic test are tested for their size. Subsets with very small size are pruned. The remaining sets are processed further to identify records with the dimension-value pairs represented by the subset-descriptor. The interestingness of this subset of records is computed by applying the *t-test*. The interesting subset-descriptors are identified in the result subset L.

The algorithm then applies the goodness heuristic on each of the level k subset-descriptors to decide if the subset descriptor should be used for building subset-descriptors in subsequent levels.

IV. Related work

Workload characterization has been addressed in various different ways in the past. Graph mining techniques have been used to characterize requests based on similarity of request paths [6]. Requests have been characterized by identifying signature of resource demands of requests using machine learning techniques such as blind source separation [11]. We present a practical approach to provide a first-cut understanding of the system using the basic and most commonly available transaction logs. The insights captured complement the previous approaches. Also, the results can provide guidelines for capturing more information in order to use other sophisticated techniques.

The problem of automatically discovering interesting subsets is well-known in the data mining community as *subgroup discovery*. Much work in subgroup discovery [1], [7], [9], [8], is focused on the case when the domain of possible values for the performance measure column is finite and discrete. In contrast, we focus on the case when the domain of the performance measure column is continuous. Also, many quality measures are used to evaluate the interestingness of subgroups and to prune the search space. A new feature of our approach is the use of Student's t-test as a measure for subgroup quality.

Unlike subgroup discovery and other techniques such as identifying anomalies, we propose to analyze the complete spectrum of the measure values. Furthermore, most of existing approaches construct groups from the perspective of dimensions. In contrast, we propose to form groups based on the measure values and then identify attributes describing these groups.

V. Application on real-world case-studies

We applied the proposed algorithms in various real-world case-studies. In this section, we present how the proposed algorithms successfully derived many useful insights across a wide-variety of case-studies. We have masked or not disclosed some part of the datasets due to privacy reasons.

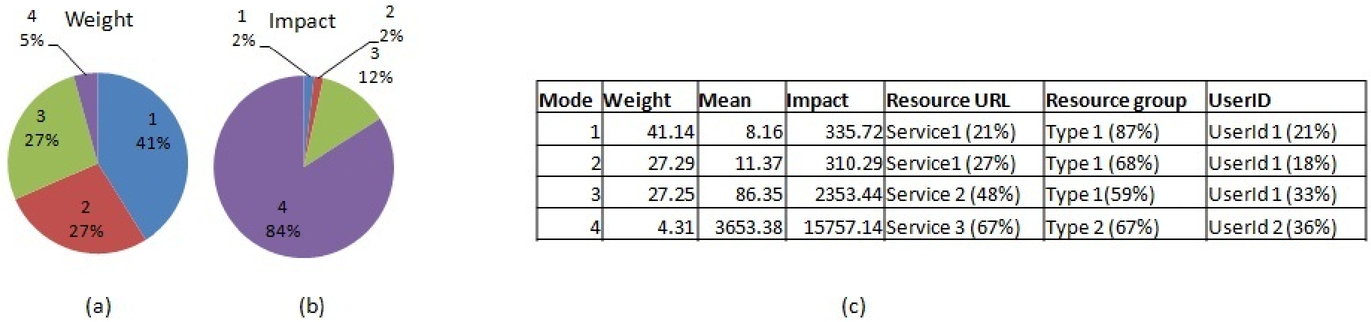
A. Analysis of web-transactions

We present a case-study of transactional system, where a data center hosts the IT system of an on-line retail system. The monitoring log of a transactional system contains information about various workload and performance properties. We show how workload characterization service can be used to mine these logs and derive meaningful insights and actionable recommendations for performance management of the system. These insights can thus contribute to the higher-level objective of *Stability As A Service*. During on-line shopping clients perform various operations such as browsing, comparison of items, shopping, redeeming of vouchers, etc. Each request received by the data center is associated with various attributes such as client IP address, Host name, date and time of request, URL name, etc. The requested URL can be further split to obtain derived attributes. For instance a URL *http://abc.com/retail/AddToCart.jsp* can be split to extract *http://abc.com*, *retail*, *AddToCart* and *jsp*. Similarly date and time of the request can be split to derive more attributes such as Day of the week, Date of the Month, Month of the year, etc. Each request is associated with a performance measure of response time.

Figure 1 presents the multimodal analysis of the transactions log. The response time values of transactions belong to four different modes. The weight and impact of these modes are shown in Figure 1(a) and Figure 1(b) respectively. As can be seen, there exists a small set (5%) of requests that have very high response time (3653ms), and this set forms the highest impacting set. The next high impact set consists of 27% requests and mean response time of 86.36ms. Figure 1(c) further characterizes these modes by identifying dominant properties of the requests falling in each mode. The mode with highest response time (3635ms) contains 67% requests with $ResourceURL = Service3$. The mode with second highest response time (86.36ms) contains 48% requests with $ResourceURL = Service2$. Figure 1(c) presents more such insights to further characterize modes based on Resource URL, Resource group, and User ID. In this case study, improving Service 3 would have largest impact on decreasing the average latency of the system. Furthermore, requests with User ID=2 and Resource group = Type 2 are indicative of performance bottlenecks and demand further investigations.

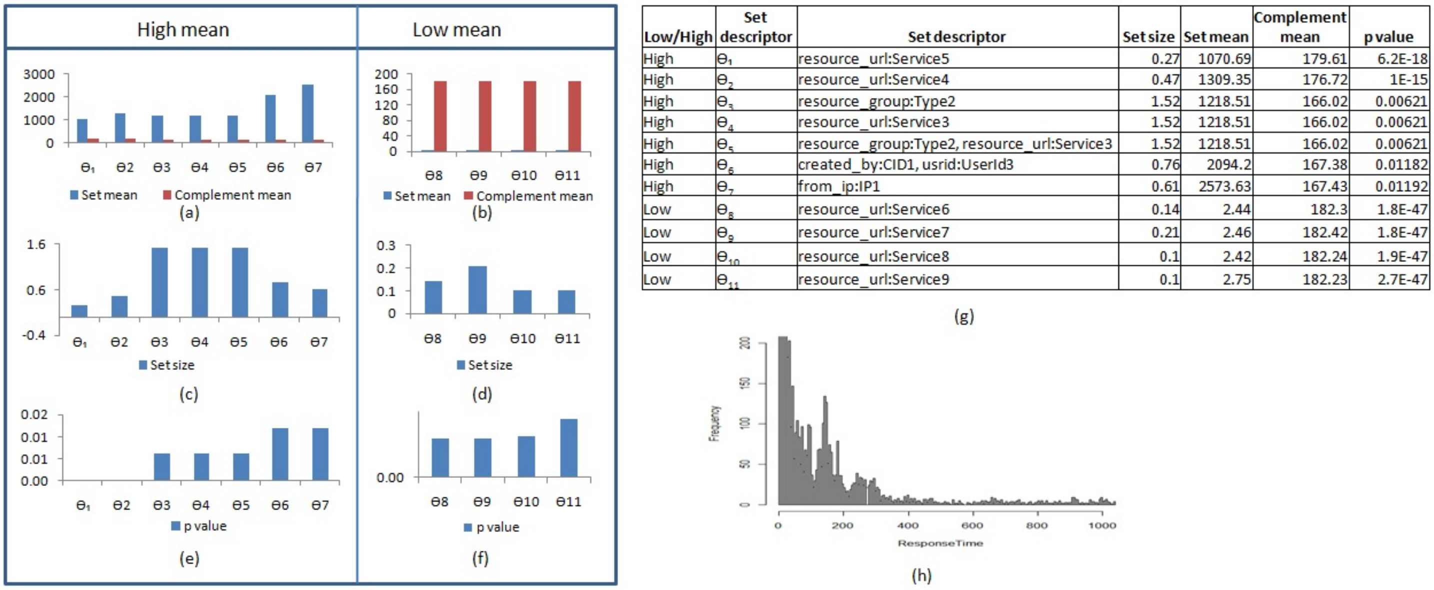
Figure 2 (g) presents the discovered interesting subsets in the transactions log with respect to response time. Interesting subsets are discovered with respect to high and low response time as shown in Figure 2(g). Results show the descriptor and the statistics for each set. These are also shown in the form of plots in Figure 2 (a,b,c,d,e,f). The results show (1) percentage of requests belonging to the set, (2) average response time of the requests in the set. (3) average response time of the requests in the complement of the set, and (4) the probability of statistical similarity, p value, between the set and its complement. The algorithm identifies that the requests (0.27% of total requests) with $ResourceURL = Service5$ have significantly high response time (1070.69ms) than the rest of the requests (179.61ms). Other interesting sets with high response time are $ResourceURL = Service4$, $ResourceURL = Service3$. The algorithm also identifies sets described by multiple dimensions such as requests with $Createdby = CID1\ and\ UserId = UserId3$. Similarly, requests that perform significantly better than the rest of requests are also identified in Figure 2. For instance, interesting subsets of requests with low response time are $ResourceURL = Service6$, and $Resourcegroup = Type1$.

The results of multi modal analysis characterize the entire spectrum of the values of response time by grouping the response time values into different modes. ISD, on the other hand, focuses only on the two ends of the spectrum by identifying subsets with significantly high and low response times. The two approaches thus have some common results such as $ResourceURL = Service3$ and $Resourcegroup = Type2$ are identified as high response time requests. Similarly, requests with $Resourcegroup = Type1$ are identified as requests with low response time. However, the two results also complement each other. For instance, MMA shows that the entire range of response time values when grouped in four modes, can be explained by $ResourceURL = Service1$ (low values), $ResourceURL = Service2$ (intermediate values), and $ResourceURL = Service3$ (high values). ISD complements this results by further exploring the two ends. It identifies additional descriptors that explain very high and very low response times. For instance, $ResourceURL = Service5$, $Createdby = CID1 and UserId = UserId3$ show significantly high response time, and $ResourceURL = Service6$, $ResourceURL = Service7$ show significantly low response



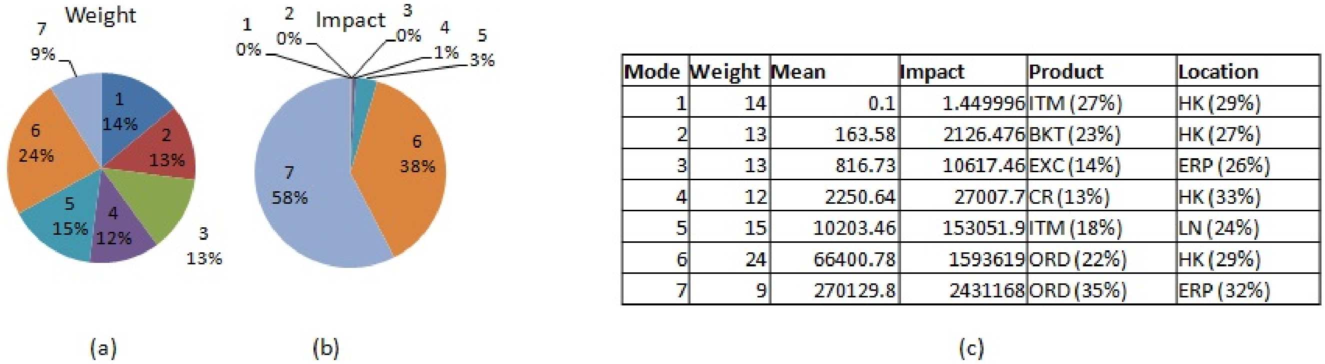
Mode	Weight	Mean	Impact	Resource URL	Resource group	UserID
1	41.14	8.16	335.72	Service1 (21%)	Type 1 (87%)	UserId 1 (21%)
2	27.29	11.37	310.29	Service1 (27%)	Type 1 (68%)	UserId 1 (18%)
3	27.25	86.35	2353.44	Service 2 (48%)	Type 1(59%)	UserId 1 (33%)
4	4.31	3653.38	15757.14	Service 3 (67%)	Type 2 (67%)	UserId 2 (36%)

Fig. 1. Multi-modal analysis of transactions log. (a) Modes by weight, (b) Modes by impact, (c) Mode characteristics.



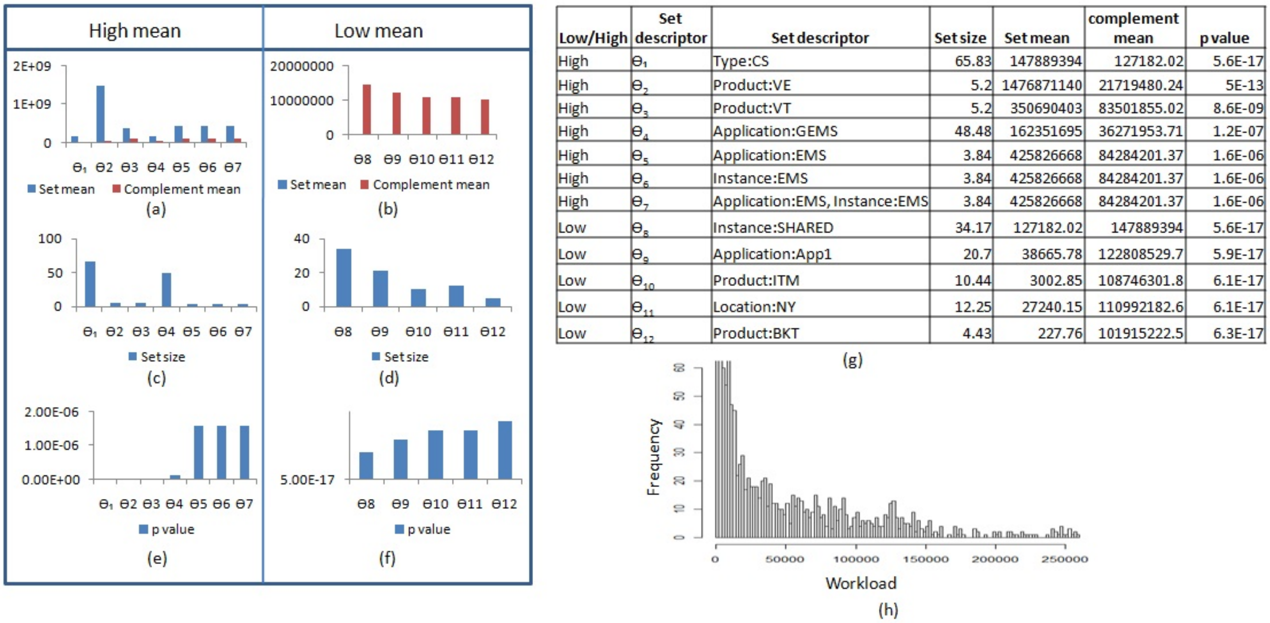
Low/High	Set descriptor	Set descriptor	Set size	Set mean	Complement mean	p value
High	Θ_1	resource_url:Service5	0.27	1070.69	179.61	6.2E-18
High	Θ_2	resource_url:Service4	0.47	1309.35	176.72	1E-15
High	Θ_3	resource_group:Type2	1.52	1218.51	166.02	0.00621
High	Θ_4	resource_url:Service3	1.52	1218.51	166.02	0.00621
High	Θ_5	resource_group:Type2, resource_url:Service3	1.52	1218.51	166.02	0.00621
High	Θ_6	created_by:CID1, usrid:UserId3	0.76	2094.2	167.38	0.01182
High	Θ_7	from_ip:IP1	0.61	2573.63	167.43	0.01192
Low	Θ_8	resource_url:Service6	0.14	2.44	182.3	1.8E-47
Low	Θ_9	resource_url:Service7	0.21	2.46	182.42	1.8E-47
Low	Θ_{10}	resource_url:Service8	0.1	2.42	182.24	1.9E-47
Low	Θ_{11}	resource_url:Service9	0.1	2.75	182.23	2.7E-47

Fig. 2. ISD analysis of transactions log. (a,c,e) Sets with high mean: (a) Set mean and Universe mean, (c) Set size, (e) p value (b,d,f) Sets with low mean: (a) Set mean and Universe mean, (c) Set size, (e) p value (g) Tabular form of ISD analysis result, (h) Histogram of response time of requests.



Mode	Weight	Mean	Impact	Product	Location
1	14	0.1	1.449996	ITM (27%)	HK (29%)
2	13	163.58	2126.476	BKT (23%)	HK (27%)
3	13	816.73	10617.46	EXC (14%)	ERP (26%)
4	12	2250.64	27007.7	CR (13%)	HK (33%)
5	15	10203.46	153051.9	ITM (18%)	LN (24%)
6	24	66400.78	1593619	ORD (22%)	HK (29%)
7	9	270129.8	2431168	ORD (35%)	ERP (32%)

Fig. 3. Multi-modal analysis of workload log. (a) Modes by weight, (b) Modes by impact, (c) Mode characteristics.



Low/High	Set descriptor	Set descriptor	Set size	Set mean	complement mean	p value
High	θ_1	Type:CS	65.83	147889394	127182.02	5.6E-17
High	θ_2	Product:VE	5.2	1476871140	21719480.24	5E-13
High	θ_3	Product:VT	5.2	350690403	83501855.02	8.6E-09
High	θ_4	Application:GEMS	48.48	162351695	36271953.71	1.2E-07
High	θ_5	Application:EMS	3.84	425826668	84284201.37	1.6E-06
High	θ_6	Instance:EMS	3.84	425826668	84284201.37	1.6E-06
High	θ_7	Application:EMS, Instance:EMS	3.84	425826668	84284201.37	1.6E-06
Low	θ_8	Instance:SHARED	34.17	127182.02	147889394	5.6E-17
Low	θ_9	Application:App1	20.7	38665.78	122808529.7	5.9E-17
Low	θ_{10}	Product:ITM	10.44	3002.85	108746301.8	6.1E-17
Low	θ_{11}	Location:NY	12.25	27240.15	110992182.6	6.1E-17
Low	θ_{12}	Product:BKT	4.43	227.76	101915222.5	6.3E-17

Fig. 4. ISD analysis of transactions log. (a,c,e) Sets with high mean: (a) Set mean and Universe mean, (c) Set size, (e) p value (b,d,f) Sets with low mean: (a) Set mean and Universe mean, (c) Set size, (e) p value, (g) Tabular form of ISD analysis result, (h) Histogram of workload.

time.

These results can provide powerful levers to contribute to *Stability As A Service* as follows:

1) The analysis can provide opportunities for optimization and improvement by identifying requests that can have highest impact on the overall performance. For instance, improvement of Service 3 would result in highest impact on the overall improvement.
2) It identifies performance bottlenecks. For instance, requests with User ID=2 and Resource group = Type 2 are identified as performance bottlenecks.
3) It allows estimation of the contribution of individual request types and the impact of their improvement on the overall system performance stability.
4) Analysis on different log dimensions enables localization of the performance problems to specific URLs, Client IPs, location, time, among others.

B. *Analysis of application workload*

We next present the analysis of workload observed at the application tier of a loan-processing application over a period of one month. The application monitoring logs consist of per-day workload observed at the application tier along with various details of the requests such as application type, location, the requested service, etc. The dimensions in this log can be used to analyze the workload properties of the system.

Figure 3 presents the multi-modality analysis of the workload log on the basis of amount of workload. The observed workload values belong to 7 different modes. The smallest amount of workload is observed in mode 1 with average workload of 0.1 requests per day. This workload type is dominated by Product = ITM (27%), Location = HK (29%). Mode 7 represents heavy workload of 270,000 requests per day. This workload type is dominated by Product = ORD (35%), Location = ERP (32%).

Figure 4(g) presents the interesting subsets are discovered with respect to high and low workloads. Results show the descriptors and the statistics for each set. Workload with $Type = CS$ (65% of total records) is significantly higher (147889393.97) than the rest of the workload (127182.02). Other sets with high workload are described by $Type = CS$, $Product = VE$, $Application = GMS$, among others. It also identifies sets described by combinations of dimensions such as $Instance = NYD, andLocation = NYD$. Similarly, sets with significantly low workload are also identified in Figure 4. For instance, $Instance = SHARED$, $Product = ITM$, $Location = HK$.

Comparing the ISD and MMA results, it can be seen that both results share some common findings. For instance, Both identified the sets $Product = ITM$, and $Location = HK$ to have significantly low workload. Some findings, on the other hand, complement each-other. For instance, MMA identifies $Product = ORD$ to describe the high workload sets. MMA thus identifies the most dominating high workload component. ISD, on the other hand, complements this result by providing other not-so-dominating high workload components such as $Product = VE$, $Product = VT$. These sets are not very large to describe an entire mode, but nevertheless have significantly higher workload.

These insights can be used for efficient resource allocation such that the resources to serve request types with low workload can be rationed while the request types with heavy workload can be supplied adequate resources. Better load-balancing and workload distribution policies can also be derived using such analysis. In the given case-study, resources at location HK for serving requests with Product = ITM or BKT can be rationed while sufficient resources need to be provided at the location ERP to serve requests with Product = ORD. Also, in order to avoid location bottlenecks recommendations can be given to investigate the possibility of migrating the workload from Location ERP to other locations.

These insights can be used for capacity analysis such as efficient resource allocation, load balancing etc. For instance,

1) The resources to serve request types with low workload can be rationed while the request types with heavy workload can be supplied adequate resources. In the given case-study, resources at location HK for serving requests with Product = ITM or BKT can be rationed while sufficient resources need to be provided at the location ERP to serve requests with Product = ORD.
2) Better load-balancing and workload distribution policies can also be derived using such analysis. A workload distribution policy can be quickly worked out on the basis of modes such that resource requirements of low and high workload are met. In the given case-study workload of product VE can be moved to resources that service workload of product BKT.
3) In order to avoid location bottlenecks, recommendations can be given to investigate the possibility of migrating the workload from Location ERP to other locations.

VI. Conclusion

Stability As A Service is critical for data centers and workload characterization is one of the essential services within *Stability As A Service*. Different solutions have been proposed in the past for characterizing various system properties, however most of these approaches often require high levels of instrumentation and intrusiveness. This makes their use difficult in real-world production systems. In this paper, we argued that many interesting insights can be derived to characterize system workload simply by analyzing the transaction logs captured at different layers in the system such as access logs at application servers, SQL logs at database servers, among others. We presented two approaches that analyze various dimensions and measures of the transaction logs in order to characterize workload. We demonstrated the effectiveness of the proposed approach through real-world case-studies and showed how the findings of the two approaches complement each-other.

References

[1] M. Atzmueller and F. Puppe. Sd-map: a fast algorithm for exhaustive subgroup discovery. In *Proc. PKDD 2006*, volume 4213 of *LNAI*, pages 6 – 17. Springer-Verlag, 2006.

[2] M. Atzmuller. *Knowledge intensive subgroup mining: Techniques for automatic and interactive discovery*. Aka Akademische Verlagsgsellschaft, 2007.

[3] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth Intl., Belmont, CA, 1984.

[4] P. Clark and T. Niblett. The CN2 induction algorithm. *Machine Learning*, 3(4):261–283, 1989.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[6] Y. Jin, E. Sharafuddin, and Z.-L. Zhang. Unveiling core network-wide communication patterns through application traffic activity graph decomposition. In *SIGMETRICS*, 2009.

[7] B. Kavšek, N. Lavrač, and V. Jovanoski. Apriori-sd: adapting association rule learning to subgroup discovery.

[8] N. Lavrač, B. Cestnik, D. Gemberger, and P. Flach. Subgroup discovery with cn2-sd. *Machine Learning*, 57:115 – 143, 2004.

[9] N. Lavrač, B. Kavšek, P. Flach, and L. Todorovski. Subgroup discovery with cn2-sd. *Journal of Machine Learning Research*, 5:153 – 188, 2004.

[10] M. Natu and G. Palshikar. Discovering interesting subsets using statistical analysis. In G. Das, N. Sarda, and P. K. Reddy, editors, *Proc. 14th Int. Conf. on Management of Data (COMAD2008)*, pages 60–70. Allied Publishers, 2008.

[11] A. Sharma, R. Bhagwan, M. Choudhury, L. Golubchik, R. Govindan, and G. M. Voelker. Automatic request characterization in internet services. In *1st HotMetrics Workshop, Association for Computing Machinery*, 2008.

[12] M.-H. Zhang and Q.-S. Cheng. Determine the number of components in a mixture model by the extended ks test. In *Journal Pattern Recognition Letters, Volume 25 Issue 2, 19 January 2004 , Elsevier Science Inc. New York, NY, USA*.

4R of Service Innovation: Research, Requirements, Reliability and Responsibility

Anastasiya Yurchyshyna
CUI, University of Geneva
Battelle - Batiment A
7, Route de Drize
CH - 1227 Carouge
+41 22 379 02 59
Anastasiya.Yurchyshyna@unige.ch

Abdelaziz Khadraoui
CUI, University of Geneva
Battelle - Batiment A
7, Route de Drize
CH - 1227 Carouge
+41 22 379 02 31
Abdelaziz.Khadraoui@unige.ch

Michel Léonard
CUI, University of Geneva
Battelle - Batiment A
7, Route de Drize
CH - 1227 Carouge
+41 22 379 77 77
Michel.Leonard@unige.ch

Abstract—**This research aims at resolving the challenges arising from the processes of service innovation and evolution. It continues and extends our works concerning different aspects of services innovation and analyzes the "4R" phenomena of service engineering: service research, service requirements, reliability and responsibility in services. By studying the interdependence of these aspects for services development, we investigate the phenomenon of corporate social responsibility and discuss its development through the information kernel of services. In order to integrate the 4R-analysis for service innovation, we propose a framework for innovation in services and illustrate the role of responsibility in the development of its information kernel. Finally, we discuss a usage scenario implementing this approach, which is related to the treatment of authorization requests for building at the canton of Geneva.**

Keywords-service; service innovation; requirements; reliability; compliance; corporate social responsibility

I. INTRODUCTION

Today, services have become the key element for all aspects of business and corporate activities, whilst the importance of their innovation process can hardly be overestimated. Indeed, dynamic collaborative processes of services innovation and evolution form the kernel for development and communication in the organizational context of an enterprise, and define the main principles of the corporate management within the services society [1].

Based on the current state of the art on modeling initiatives and services and studying the phenomena of innovation and evolution for supporting services, this research is within a series of works aiming to define the notion of corporate social responsibility, identify the main challenges for services innovation, study the interdependence of different layers of services and develop a methodology for services innovation while taking into account the challenges in both conceptual modeling and practical implementation.

By underlining the increasing importance of collaborative creation in services and the role that responsibility plays in it, we show how such collaborative creation becomes one of the main factors in innovation in services in services society.

This paper is structured as follows. In Section 2, we present the general state of the art concerning the problem of service innovation and analyze its complex aspects. Section 3 describes our approach for service innovation and focuses on its main axis: research in services, requirements for services development and engineering, reliability in services, the notion of responsibility in services innovation, and analyses their interdependence through the phenomenon of corporate social responsibility. A conceptual framework representing different (R-)layers of services and their interdependences is introduced in Section 4. Section 5 describes a simplified usage scenario illustrating the proposed approach. Finally, section 6 concludes the paper and discusses the envisaged future works and perspectives of this research.

II. GENESIS AND STATE OF THE ART

We start by introducing the existing definitions of the main concepts related to the problem of corporate social responsibility and services through innovation and corporate management, by analyzing and arguing these notions and thus, by constructively describing their semantics in the context of our research.

A. Service

In our previous research [15], we described a service as the result of a process of acquiring knowledge in the context of the IS (information systems) engineering, defined at the junction of the organizational domain, the ontological domain, the technological domain and the informational domain.

Consequently, it is based on four dimensions: (i) *ontological* dimension; (ii) *informational* dimension; (iii) *technological* dimension; and (iv) *organizational* dimension. The ontological dimension of a service describes not only all the invariants of the information system domain, in particular knowledge and concepts, but also some business rules and roles of actors, which are independent of the information system development. The informational dimension of a service defines the information semantics necessary for defining services. This dimension of a service describes the static aspects, the dynamic aspects and the integrity constraints aspects. The organizational dimension of a service relates to the business rules, the organizational roles, the responsibility zones and business processes inside an enterprise/organization. It allows one to clarify the decisions and responsibilities inside the enterprise/organization. The technological dimension of a service allows one to study the implementation of the specified entities. It is a challenge of choosing the appropriate technology, the informatics architecture and the corresponding environment, in order to implement this service.

B. *Innovation and Evolution of Services*

There is a certain ambiguity in general understanding of the phenomenon of innovation. Traditionally, it is seen as introduction of something new: a new material, way of doing, a new concept, etc. This definition is however different from the widely used meaning of the notion of innovation – the process that aims at bringing new features into an existing entity (concept, good), renewing something that already exists, i.e., evolution of an existing thing. On analyzing these aspects [15], we identify different but interdependent phenomena: innovation and evolution. Innovation represents the process that allows the change of state of the component of a system, so precipitating the emergence of a system, whose characteristics or behaviors are different from the previous time [3]. It can be thus viewed as the source of evolution. In other words, we envisage innovation as a dynamic and participative process that leads to co-creation and value creation of a product (artifact, method, etc.) thanks to its evolution.

Defined by both *dynamic* and *collaborative* characteristics, each of these processes (i.e., innovation and evolution) generally leads to *sustainability* of a product (good, process, service, etc.). In their interdependence, they also contribute to enriching the semantics and usability of related services and knowledge bases. Indeed, lets us take an example of the process of evolution of e-government services. While they are developed and modified, the corresponding regulatory ontologies are also modified and enriched, and new organizational contexts are identified and adopted. In other words, both innovation and evolution processes result with added value to a product, service, related knowledge bases, information systems and services in their dynamic environment.

C. *Requirements for Service Engineering*

By meeting the challenge of designing sustainable services, service engineering addresses the specification, the compliance and the management of interoperability of services across public institutions or governmental organizations.

The service compliance aims to enhance the quality of services to offer to the stakeholders. We consider that the verification of compliance is based on three criteria: (i) how to build a stable service from the organizational context?; (ii) how to support evolution of services?; (iii) how the interoperability of services is managed? [5].

The phenomenon of service compliance with legal aspects is nowadays gaining increasing importance.

Indeed, the institutional activities are governed by legal sources represented by a set of laws, which regulates their execution. The compliance of services with legal aspects is a crucial issue for each public administration. This issue becomes more difficult with the fast-evolving dynamics of laws [7].

Another challenge is to describe how to get service engineering and legal compliance in closer interaction. The main issue is to consider the requirement for service compliance analysis and for service overlap management when we engineer services.

In order to face these new challenges, we proposed a methodological approach for service specification [2]. It aims to describe the interactions among the organizational layer, the informational layer of a service and the responsibility dimension in order to verify the compliance of services.

D. *Reliability*

Reliability is normally seen as the ability of a service to perform a required function, under stated conditions, for a stated period of time. In service design, reliability reflects a measure of how long an IT-service can perform its function without interruption or how likely required outputs will be delivered within a stated period of time.

According to the definition of the Technical Committee on Communications Quality & Reliability [14], the service reliability is a complex notion, which combines the 3 following characteristics: (i) accessibility – a service is available when it is required; (ii) continuity – a service has an uninterrupted duration when required by a customer; (iii) performance – a service is designed and able to meet the customer's expectations.

It is important to note that our vision of reliability concerns both its interdisciplinary and temporary aspects: a service and an information system are known to be reliable if they are coherent for different information systems and/or services integrating them, as well as for different time frames of the lifecycle of the same information system/service.

E. *Responsibility*

To define the phenomenon of responsibility, we rely on the research [5]. The authors propose to model it as a state assigned to an actor to signify him its accountabilities (accountability defined at [10] as a process of being called to account to some authority for one's actions) concerning a business activity and the capabilities and the right necessary to perform it (the right represents the resources provided by the company to an employee required to perform the accountability. Moreover, the authors point to its interdependence with the phenomena of capability, right and accountability, and put these concepts into the core of the methodological approach for services specification, by particularly specifying the links between the organizational and informational layers of services and by enriching the model with the responsibility dimension.

In this context, the concept of responsibility is composed of the accountabilities to perform an obligation in a business activity and it specifies, at the same time, the rights and the capabilities that are required therein.

In the further research [2], these conceptual findings allowed to develop a meta-model that show the interdependencies between the four layers of services (i.e., ontological, organizational, technical and informational) and the concept of responsibility, by integrating capabilities, rights and accountability.

Such a complex model enlightens the fact that the responsibility dimension contributes to the added-value of a service since it facilitates the alignment between the different layers. Indeed, on the organizational layer, responsibility is assigned to a role that performs business activities, the informational layer identifies the responsibilities-required knowledge, the responsibility on the technical layer can partially be incorporated in security and or accessibility characteristics, and the ontological layer allows one to correctly specify the business rules, the ontological roles and the fundamental concepts dedicated to specify this service, whilst taking into account the responsibility requirements.

F. Research in service innovation: towards collaborative decision constructing

It is difficult to overestimate the importance of innovation in the development of new services that are aimed at answering the challenges our services-based society brings today.

By relying on the interactive exchange and functioning of interoperable services [1], our services-oriented society, is, to some extent, dependant from the progress and innovations in services. Services are becoming "mirrors" of specific competences (e.g., knowledge, skills, technologies) of one economic entity for the benefit of another economic entity. Hence, whilst value creation occurs when a resource is turned into a specific benefit, the classical supply chain is thus re-conceptualized as a network of service systems, the service value creation network [9].

To answer come of these challenges, in [8], we proposed our approach for innovation in services thanks to collaborative decision constructing, and demonstrated how decision constructing can be supported by the processes of knowledge actionalising during the process of service innovation.

Introduced in this work, we consecutively developed the methodology for collaborative decision constructing, which is practically implemented in the CPS, a collaborative platform for services innovation.

III. TOWARDS CONCEPTUALIZING THE PROCESS OF SERVICE INNOVATION: 4R-ANALYSIS

This research aims at answering the challenges of the process of service innovation and evolution. Based on our previous works [2], [8], [15], its motivation is to investigate the interdependence of "4R" aspects of service innovation: service research, service requirements, reliability and responsibility in services.

A. Service Research

Recently, the complex problem of research in the services domain has attracted significant interest in both academic and business worlds.

In the context of this work, we focus on selected aspects of this problem, which highlight the importance of co-creation in services development and identifying the information kernel of a service, by taking into account its different layers (i.e., ontological, organizational, informational and technical).

Based on our approach for innovation in services, we claim that the information kernel is developed thanks to the collaborative decision constructing process [8] and is defined by the interdependence of its layers [6].

This leads to an (initially) intuitive idea to study if the responsibility dimension provides an additional added value to decision-constructing processes. Consequently, it is debated that this makes the process of innovation in services more efficient.

B. Service Requirements

By analyzing service requirements, we focus on the 3 main processes: identification of services, service compliance and service specification. The interoperability characterizing these processes and their interdependence ensure the quality of service integration into existing and developed information systems, as well as allows creating added value in the process of service engineering.

The problem of service identification in the processes of services innovation and evolution was discussed in [15] where we analyzed the role of initiatives for developing the services information kernel, which describes the knowledge required for creation of a service.

The process of service identification benefits from another important aspect of service engineering: the compliance of services with legal aspects.

In [7], we introduced a novel approach that allows establishing and clarifying the links between laws and services, in particular the alignment between the amendment of laws and the evolution of services. In other words, we use laws as a source of knowledge to analyze and construct the ontological level of an institutional domain. The exploitation of these sources of knowledge permits one to find stable concepts and invariant concepts. The analysis of the legal framework permits one to identify the main characteristic concepts of the domain, the ontological roles and the ontological business rules (cf. Figure 1). We use the ontology model elaborated from the legal framework to design the IS Kernel.

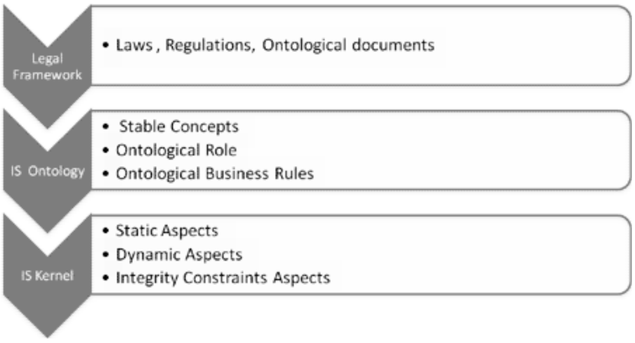


Figure 1. IS kernel and the compliance of services with legal aspects

One of the main advantages of such an approach is to explicitly match the legal framework, which provides the basis of the activities of a public administration to the developed services.

Another challenge in the context of service engineering concerns the specification of a service upon multiple already existing IS. Linking ISs with regulations / laws they have inherited from is a primary importance for people in charge of managing legacy ISs.

In [6], an approach for the specification of domain services upon multiple existing information systems is proposed. This approach is based on the construction of a referential around the services and the analysis of its required data. More precisely, it helps to preserve the legacy information system by creating services upon them via a common base capturing the overlap between all related information systems.

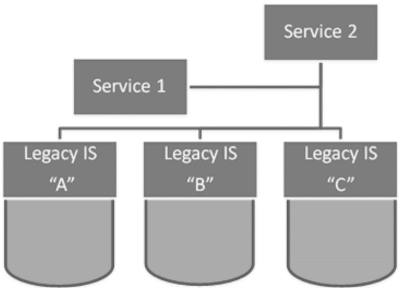


Figure 2. Specification of domain services upon multiple existing information systems.

In fact, this approach is adapted to transform legacy information system, since organizations are forced to continue to operate taking into account these existing applications and legacy information systems (cf. Figure 2).

C. Service Reliability

In the context of services society, where the problem of sustainable development needs to be considered and addressed, it is also important to study how to use services to identify the sources of added value, to elaborate an approach aimed at facilitating effective diffusion of scientific knowledge and technology transfer, and to develop knowledge infrastructure and networks.

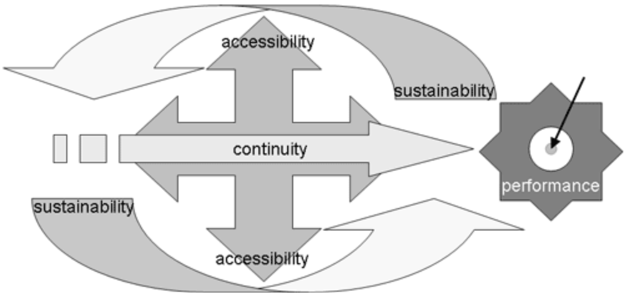


Figure 3. Reliability and its components.

We thus propose to enrich the notion of service reliability by integrating the characteristics of sustainability (cf. Figure 3). Indeed, today we are interested in how services might contribute to sustainable development, how they should be developed to increase the added value of related processes, which is the approach for making them most adaptable for business environments.

The accent is put on sustainable services, which we envisage as services that are capable of adapting to their environment, to dynamically integrate the ever-changing conditions of the environment, and as such to be sustainably coherent with its evolving challenges.

Analogically, it seems promising to enrich the notion of service reliability by integrating the sustainability criterion. Service reliability can be envisaged as a complex notion combining the 4 interdependent characteristics: (i) accessibility (availability when required by an actor/service); (ii) continuity (an uninterrupted duration when required by different actors/services); (iii) performance (being designed and implemented in the way to meet the customer's requirements); and (iv) sustainability (possibility to dynamically and coherently integrate the ever-changing conditions of the environment).

D. Responsibility in Service Innovation

Analogically to Corporate Social Responsibility (CRS), there is no unique definition of the Responsible Research and Innovation. It is often seen as a transparent, interactive process by which societal actors and innovators become mutually responsive to each other with a view to the (ethical) acceptability, sustainability and societal desirability of the innovation process and its marketable products [12].

In order to explore the phenomenon of responsibility in service innovation, we first need to concretize the notion of a service in the context of service engineering. We propose to envisage a service as the result of a process of acquiring knowledge in the context of the information system engineering. It corresponds to an action or series of actions to characterize the relationships between the stakeholders.

While adapting the approach for responsible research and innovation, the services are thus developed by respecting the following aspects [13]: (i) the deliberate focus of research and the products of innovation to achieve a social or environmental benefit; (ii) the consistent, ongoing involvement of society (incl. public & non-governmental groups) during the innovation process; (iii) assessing and effectively prioritizing social, ethical and environmental impacts, risks and opportunities, alongside the technical and commercial; (iv) where oversight mechanisms are better able to anticipate and manage problems and opportunities and which are also able to adapt and respond quickly to changing knowledge and circumstances; (v) where openness and transparency are an integral component of the research and innovation process.

E. Corporate Social Responsibility: interdependence 4R's through services layers

Initially, the notion of Corporate Social Responsibility was defined in [4] as "the continuing commitment by business to behave ethically and contribute to economic development while improving the quality of life of the workforce and their families as well as of the local community and society at large".

At the same time, despite the main motives – to improve qualitatively (the management of people and processes) and quantitatively (the impact on society) – are the same, and an international standard to provide guidelines for adopting and disseminating social responsibility (ISO 26000 – Social Responsibility) is being developed by the ISO, it is up to companies to define their own CSR objectives.

In the context of this research, we argue that corporate social responsibility for the problem of innovation in services is defined as the interdependence of layers of services: ontological, organizational, informational and technical. Moreover, the responsibility dimension provides a promising added-value since it facilitates the alignment between the different layers : the organizational layer (a responsibility is assigned to a role that performs business activities), the informational layer (the responsibilities required information), and the technical layer (the responsibility has an existence at the technical layer: e.g., by assigning the role of a facilitator during the decision-constructing process or by defining the criteria for accepting or refusing the initiative in the process of service innovation [15]).

IV. CONCEPTUAL 4R-FRAMEWORK FOR INNOVATION IN SERVICES

In this section, we introduce a conceptual framework for innovation in services by focusing on the 4 layers and the responsibility dimension.

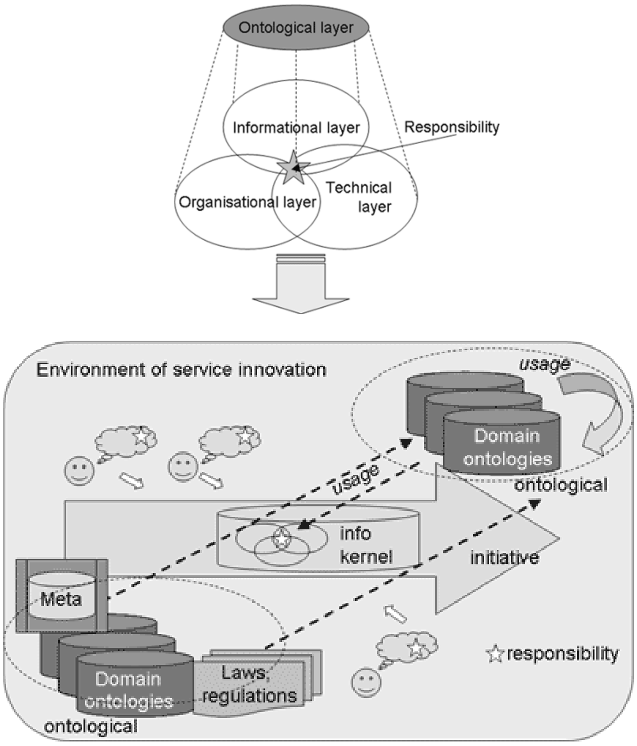


Figure 4. Responsibility in the framework for innovation in services.

Based on the analysis of the main service layers and the responsibility dimension [2], we argue that responsibility is the phenomenon that is concretized thanks to the interdependence of the ontological, organizational, informational and technical layers of a service and defines the essential conceptual information related to this service. In other words, the responsibility dimension allows one to identify the information kernel of the service.

From a different viewpoint, our model for constructing the collaborative environment for services innovation [8] is developed thanks to the process of knowledge actionalising, which helps to identify the information kernel of a corresponding initiative leading to service development. It also means that the information kernel of an initiative serves as a starting point for the process of service innovation, which, thanks to multiple interactions with the decision-making actors, the existing ontologies and the usage-based feedback, leads to concretizing the information kernel of a developed service.

V. USAGE SCENARIO

In order to illustrate the 4R-analysis for service innovation, let us consider the following case related to the treatment of authorization requests for building at the canton of Geneva. This case study was done in the context of our project collaboration with the Center of Information Technology at the Canton of Geneva (Switzerland).

At the Canton of Geneva, requests for building permits are submitted to the department responsible for construction. As soon as the legal conditions are met, the department shall issue the building authorization.

Authorization requests are subject to an advisory notice to municipalities, departments and agencies concerned. The Department's decision is based primarily on the notice of architectural committee or on that of the Committee on monuments, nature and sites. It takes into account those issued by the municipality or the competent department.

There are several stakeholders concerned by this services: (i) Directorate of building permits; (ii) Directorate General of Water; (iii) the architectural committee; (iv) Committee on monuments, nature and sites; (v) Department of Geology, soils and wastes; (vi) Land Registry; (vii) Energy Department; and (viii) Protection Administration of the population

In order to facilitate the access to the information, a national commission has been charged to define the appropriate service.

Research. We use the legal framework, which describes the conditions to manage the access to confidential and public information related to authorization requests. The fundamental concepts and relationships (e.g., who decides if a building is compliance to existing norms; which are the accessibility requirements to take into account, etc.) are defined on the ontological level, the roles and business activities are manages at the organizational level (e.g., which is the procedure for certification, timescale for handling the authorization request, etc.). All related knowledge is specified on the informational level and processed on the technical level. As the result, the decision-constructing environment handling authorization requests is developed as a specification of the global framework of the cross-pollination space, and it is enabled by services interoperable with the existing legal framework.

Requirements. The answers to the main challenges for requirements modeling are in fact acquired thanks to the dynamic development of the cross-pollination space for handling authorization requests. Indeed, (i) "how to build a stable service from the organizational context?" is ensured by the fact that the information kernel of the CPS is based on the organization context itself; (ii) "how to support evolution of services?" is possible thanks to the dynamic evolution of the information kernel according to the feedback from the usage; and (iii) "how the interoperability of services is managed?" is ensured at the ontological layer of the CPS for its semantics, at the informational layer for its statics, dynamics and integrity, as well as at the technical layer for services realization.

Reliability and responsibility. These characteristics are closely connected to the ability of a service to handle authorization requests according to the updated legal base (ensured at the ontological layer) and for a stated period of time (handled at the technical layer). Moreover, thanks to the dynamic enrichment of the CPS by the knowledge of decision-makers, one can witness a certain shift in defining the quality of a service: from "technical" reliability to "human" responsibility of individual and collective decision-makers.

In other words, the interdependence of services layers of a service for handling authorization requests permits the capitalization of contributions from each of them (i.e., ontological, organizational, informational, and technical) and as such – ensures the corporate social responsibility for development of such a service.

VI. Conclusion

This paper was developed in the scope of research concerning different aspects of innovation in services and particularly focuses on the analysis of the 4Rs of service innovation: service research, requirements, reliability and responsibility in services. By analyzing these phenomena, we showed that it is thanks to their interdependence that the responsibility dimension in the process of service innovation can be identified. From a different viewpoint, we argued that the responsibility dimension within the decision-constructing process is dynamically constructed as the information kernel of a service initiative. This research thus demonstrated how the responsibility dimension defines and guides the process of developing the information kernel of an initiative leading to the creation of a service, and as such, the process of service innovation itself.

Among the main perspectives envisaged for this research, we are to focus on: (i) the formalization of the responsibility dimension and the study of its interoperability in the both contexts of services layers and the decision-constructing model; (ii) the analysis of the measurability of the responsibility dimension and its integration into a services lifecycle; and (iii) more complex and more heterogeneous case studies with further conceptualization and generalization of the acquired results.

References

[1] H. Demirkan, R. Kauffman, J. Vayghan, H. Fill, D. Karagiannis, and P. Maglio. "Service-oriented technology and management: Perspectives on research and practice for the coming decade". In Electronic Commerce Research and Applications 7(4): 356-376, 2008

[2] C. Feltus, A Khadraoui, A. Yurchyshyna, M. Léonard, E. Dubois, "Responsibility aspects in service engineering for eGovernment", in Proceedings of the Interoperability for Enterprise Systems and Applications conference (I-ESA'12) Workshop Service Science and the Next Wave in Enterprise Interoperability, Valencia, Spain, 2012.

[3] K. Frenken, "Innovation, Evolution and Complexity Theory" In: Cheltenham UK and Northampton MA: Edward Elgar. 2006

[4] L. Holme, and R. Watts, "Corporate Social responsibility: making good busness sense". WBCSD (World Business Council for Sustainable Development), available at www.wbcsd.ch, last accessed 9/02/2012

[5] A. Khadraoui, C. Feltus, "Service Specification and services compliance: How to consider the Responsibility Dimension?" Submitted for publication, Journal of Service Science Research (JoSS), 2012.

[6] A. Khadraoui, W. Opprecht, M. Léonard, and C. Aïdonidis, "Service Specification Upon Multiple Existing Information Systems", 2011 Fifth International Conference on Research Challenges in Information Science RCIS 2011, May 19-21, Gosier, Guadeloupe.

[7] A. Khadraoui, W. Opprecht, C. Aïdonidis, and M. Léonard, "Laws-Based Ontology for e-Government Services Construction Case Study: The Specification of Services in Relationship with the Venture Creation in Switzerland", PoEM 2008: 197-209.

[8] M. Leonard and A. Yurchyshyna, "Decision constructing as conceptualisation of service innovation" Proc. of IJCSS2011, the International Joint Conference of Service Sciences, 25-27 May 2011, Taipei, Taiwan

[9] R.F. Lusch, S.L. Vargo, and G. Wessels. "Towards a conceptual foundation for service science: Contributions from service-dominant logic". IBM Systems Journal, Vol. 47, No. 1, (2008)

[10] R. Mulgan, "Accountability: An Ever-Expanding Concept?" Public Administration 78 (3), 2000: 555–573.

[11] A. Murray, K.Haynes, L. Hudson, "Collaborating to Achieve Corporate Social Responsibility and Sustainability? Possibilities and problems". Sustainability Accounting, Management and Policy Journal 2010, 1(1), 221-231.

[12] von Schomberg (2011) "Prospects for Technology Assessment in a framework of responsible research and innovation" in: M. Dusseldorp and R. Beecroft (eds). Technikfolgen abschätzen lehren: Bildungspotenziale transdisziplinärer Methoden,Wiesbaden: Vs Verlag

[13] H. Sutcliffe. What is Responsible Research & Innovation? Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/rri-report-hilary-sutcliffe_en.pdf, last accessed 9/02/2012

[14] The Technical Committee on Communications Quality & Reliability, available at http://committees.comsoc.org/cqr/, lasr accessed 22/03/2012

[15] A. Yurchyshyna, A. Khadraoui, W. Opprecht, and M. Léonard, "Innovation and Evolution of Services: Role of Initiatives" Proc. ePart 2011: 262-273, 2011

Visualization Method Based on Cloud Computing for Real Estate Information

Mingyuan Yu
College of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
E-mail: yu_mingyuan@163.com

Donghui Yu, Lei Ye, Xiwei Liu
College of Computer Science and Technology
Zhejiang University of Technology
Hangzhou, China
E-mail: donghui_hz@qq.com,
yelei@zjut.edu.cn,
lxwcc_2011@163.com

***Abstract*—In order to accelerate the computing speed and display efficiency of large-scale information visualization, this paper proposes a visualization method based on cloud computing for real estate information. The core idea is combining cloud computing and information visualization, making the data processing stage of information visualization on the cloud computing platform, so as to enhance the ability of parallel processing of big data and accelerate the computing speed and display efficiency. To verify the validity of this idea, we design and implement a visualization prototype system based on cloud computing. This system uses *Hive* to get the query results, uses *MapReduce* to process data in the background, and returns the result data to the server, and then the server generates relevant results by using information visualization technologies and back to the Web.**

Keywords-Cloud Computing; Information Visualization; Real Estate Information; MapReduce; Hive.

I. INTRODUCTION

Cloud computing has already become a hotspot of research and application on information services. Since Google's cloud computing model emerging in 2006, more and more companies participate in cloud computing, and cloud computing has become the mainstream model of new network service applications. Amazon, IBM, and other large companies have proposed their own cloud platforms, which can be regarded as a service. Such as Amazon's Elastic Compute Cloud (EC2) [1], users can build their own private cloud [8] platforms on it. The open-source and free framework--*Hadoop*--has been the most widely used, and some small companies use *Hadoop* to build their own private cloud entirely. However, the private cloud also faces challenges. Although it is easy to set up, ensures the data security, and users can utilize all the resources in the private cloud, it requires too many physical resources. The resources which the cloud obtains are their own cloud equipment resources. It costs too much. The public cloud [8] is the future mainstream framework of cloud computing development.

Compared with grid computing and distributed computing, cloud computing has many advantages. First, it cost less, which is the most prominent advantage. Second, it is supported by virtual machines, so that it can handle some things easily which is difficult in grid. Third, it can execute mirror deployment, which makes us to deal with heterogeneous process expediently. In addition, cloud computing emphasizes services, which is more suitable for commercial operation.

Information visualization [9] is also a popular emerging field of visualization. Information visualization combines several theories and methods such as scientific visualization, human-computer interaction, data mining, image, graphics, cognitive science, and gradually develops [10]. The traditional scientific visualization generally pays attention to three-dimensional visualization, including areas such as medicine, biology and architecture. While Information Visualization attaches importance to display information to users by any diagrams, so that users can get the information conveniently. A good interactive technology can show the information and data batter, and facilitate the user's operation. Therefore, human-computer interaction [11] in information visualization is particularly important. With the increasing of the information and data, the traditional Web applications cannot store all the information data in cache, and data processing efficiency is obviously insufficient to address these issues. We use parallel processing to solve this problem in the cloud. Users submit the demands to the system on the Web. Then process data in the cloud and return the results to the Web. Although this method increases the additional overhead of interaction time, it is faster than the traditional method. Besides, it helps solve the problem of information and data processing of big data.

In the field of information visualization, it generally has the following six data types: multi-dimension (including 1D, 2D), time, space, hierarchical (tree) structure, the network (figure) structure, text and so on. For different data types, you need to use different visualization techniques. In this paper, we mainly use the hierarchical structure--*TreeMap*.

TreeMap is a very space-saving way to display hierarchical data [15]. The display space is divided into a series of rectangles, and each rectangle is a data item. If one data item contains other data in the hierarchy, the rectangle is subdivided into smaller rectangles. In the past decade, *TreeMap* is widely used in the field of information visualization.

In this paper, we use *Hadoop* to build our private cloud platform, and design a visualization prototype system for real estate information. When users login our prototype system in

ISBN: 978-1-61208-215-8

the Web, they can apply for the released service, and use the released data to apply for a new service. They can also upload the relevant data and then apply for a new service. After users submitting the demands to the system, the system will analyze data in the background, and return the results to the Web. Only have a browser, can users access the system and get results.

Section II describes the related work of cloud computing and information visualization, including using cloud computing to solve current information services problems, information visualization progress at home and abroad, and the application progress of combining cloud computing and engineering visualization. In Section III， we describe the basic idea and system framework of the real estate information visualization model based on cloud. And we implement this model and do some experiments in Section IV. The last section draws our conclusions and future research directions.

II. Related Work

Cloud computing is a hotspot in the current field of information services. Zhou et al. analyzed the cloud services [12], and exampled the three mainstream services, Data as a Service (DaaS), Network as a Service (NaaS) and Identity Policy Management as a Service (IPMaaS). Chen et al. applied cloud service to computer-aided design [13], and designed intelligent house. Liu et al. applied cloud computing to the elasticity public VPN service model [14], which was different from the traditional VPN. This model would pre-assess the resource consumption and dynamically adjust.

Hoang et al. proposed an elastic cloud storage system—ecStore [17], which supported automatic data partitioning and replication, load balancing, efficient range query and transaction access. Kossmann et al. proposed a modular cloud storage system—Cloudy [18], which provided a highly flexible architecture for distributed data storage, and manipulated various workloads. Based on a common data model, Cloudy could be customized to meet different application needs. Because of more and more applications and their data placed on mobile devices, independent storage of mobile systems had become a key issue. Yuan et al. proposed a wireless Network File System—RFS [19], which realized a device-aware cache management, data security of customers and privacy protection.

Nowadays, the space-time data used in the information visualization are becoming larger and reaches to the scale of TB or PB level, and the combination of cloud computing and information visualization is the general trend. In 2010, Ma et al. published Multi-GPU volume rendering using *MapReduce* [3]. Then he transplanted this system to the Web, and proposed the interface design for future cloud-based visualization services [4]. Many scholars applied cloud computing to geographic information services [5][6][16], which were appropriate to solve the problem that large GIS data processing at one computer was slowly and could not meet user demands. As we know, the data of Google Earth is large and updated in real time, which requires us to explore other ways to meet this demand. Google Earth is built upon the Google distributed systems based on cloud computing.

Combination of the design concept of above-mentioned areas, we have designed and implemented a house information visualization system for the computational speed and display efficiency of large-scale information visualization, which combines cloud computing and information visualization. We use *MapReduce* to process data in the background, generate results through *TreeMap*, and return the final results to the Web.

III. Visualization Model Based on Cloud Computing for Real Estate Information

A. Basic Idea

The basic idea of this model is using *MapReduce* to analyze and process real estate data in the background, and using *TreeMap* to display results on the Web.

After login the system, users can see the released services and released data in the menu--*Released Services*. If the released services meeting users' needs, user can apply for the released services in the menu--*Applying Services*. If the released services cannot meet users' needs, the user can choose to apply for a new service. Users submit the detailed description of the service and the data service uses, and wait for the administrator to open this new service.

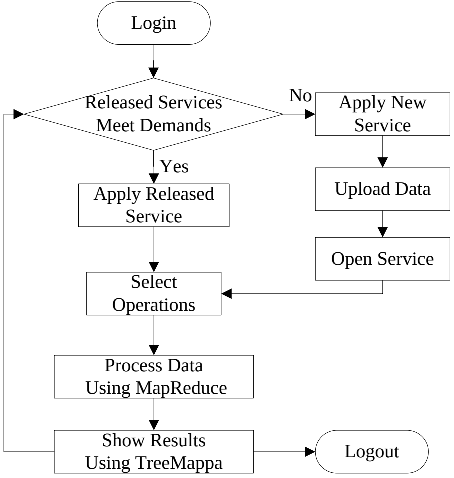


Figure 1. System flowchart

After applying for the service, users will see their own services in the Applying Services menu, such as Average House Prices. Click on this option enter Selecting Operation interface which provides the selections as city, year, showing effect etc. After submitting the selection results, users can see the relevant results on the Web. The system flowchart is shown in Figure 1.

This system is divided into three steps. Firstly, upload the data to HDFS. Secondly, use *MapReduce* to process data. Finally, display the results on the Web.

1) Data uploading

Uploading data also can be divided into two steps--Uploading data to the Tomcat server and then uploading to HDFS.

Released data format is shown in TABLE I.

TABLE I. DATA FORMAT

Column Names	Data Types
House	String
City	String
District	String
Type	String
Year	Int
Month	Int
HousePriceperQuaremeter	Float
BuildingArea	Float
HousePrice	Float
RecordCount	Int

If users want to apply for released services, the data users upload must meet the above format. If users apply for a new service, the data format can be any form of text files. The administrator will open the related services in accordance with the uploaded data.

2) Data processing

The uploaded data are divided into several independent blocks, and parallel processed by *Map* function. According to the different demands of users, specify different fields as *Key* and *Value* of *Map* functions. For example, if users want to see the average house prices of every city, specify *House* and *City* as *Key*, *HousePriceperQuaremeter* as *Value*. If users want to see the average prices of every city and every year, specify *House*, *City* and *Year* as *Key*, *HousePriceperQuaremeter* as *Value*. *MapReduce* framework would sort the *Map* output, and transmit to *Reduce*. *Reduce* function returns different results according to different demands.

If users want to see the average prices of every city, *Map* and *Reduce* function is as follows.

```
Map function
Input: LongWritable key, Text value, Context context
Output: Null
change Value to String;
change String to StringTokenizer;
for(i=0;hasMoreTokenizer;i++)
  specify nextToken as Key;
  specify nextToken as Value;
write context (Key, Value);

Reduce function
Input: Text key, Iterable<IntWritable> values,
         Context context
Output: Null
sum = 0, num = 0;
for (IntWritable value : values)
num++;
sum += get value;
set result (sum/num);
write context (key, result);
```

3) Results displaying

Users do not concern about the background how to deal with the data. What they care about is whether the results meeting user demands or not, intuitive and easy to understand, so that users can see the desired data directly from the results. In our system, we display the results by *TreeMap*. At this stage, rewrite *TreeMap* configuration file and data file according to the different demands, and then display the results on the Web, as shown in Figure 4.

B. Model Architecture

In Figure 2, the cloud cluster structures above several servers. First, we build a Hadoop distributed file system (HDFS), consists of a namenode and several datanodes. Above HDFS, we build a distributed database (Hbase) for data management. Users can login the system through various devices and apply for services.

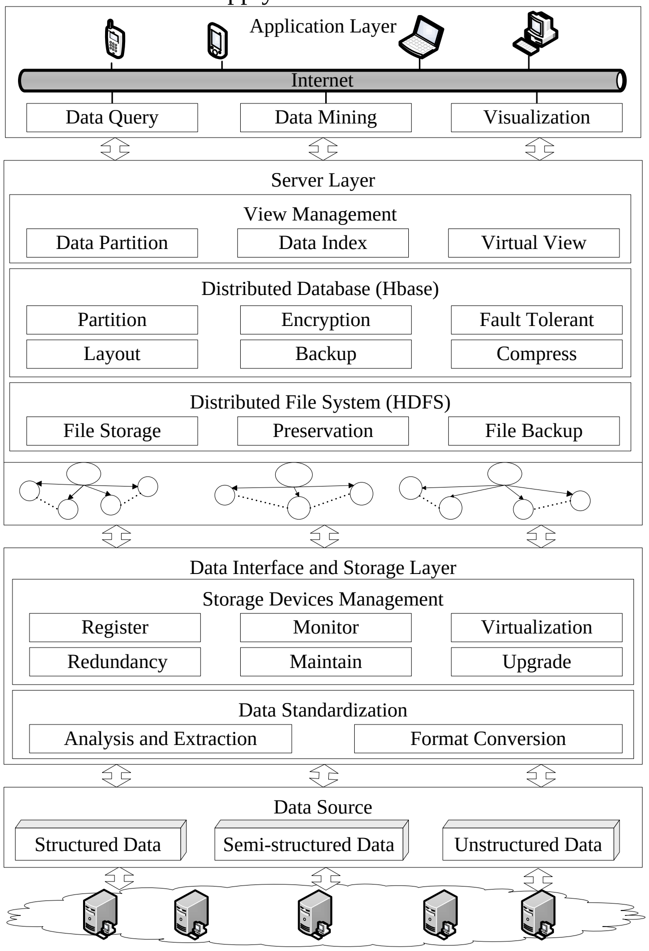


Figure 2. Cloud platform architecture

1) Data Source (DS)

The data source is the data center that provides a wide range of structured, semi-structured or unstructured data.

Commonly used data source includes, observational data—the practical observational and survey data, including field survey data and record data of observation stations; analyze data—using chemistry, physics and other scientific methods analysis of the survey data; graphic data—graphic data of medical, aerospace and other industry demand; survey data—various types of survey reports, social survey data, and so on. At present, Chinese huge number of data source produce the amount of data which achieve TB or PB level, and have different structures, the dynamic changes and updated in real time.

2) Data Interface and Storage Layer (DISL)

DISL completes the standardization of data acquisition provided by the data source and management of storage devices.

Due to the various structures of the data provided by the data source, the data have to be standardized analyzed, extracted and format converted in DISL and data, and convert to a standard format for storage devices storage. The huge number of storage devices in the cloud storage distribute in different regions, connecting with each other through the WAN, Internet or Fibre Channel (FC) network. There is a unified management system of storage devices in DISL, which manages logic virtualization of storage devices, registers storage devices, virtualizes storage devices, manages multi-link redundancy, and monitors, maintains, upgrades storage devices.

3) Data Server layer (DSL)

In the DSL, we build a distributed file system using *Hadoop*, which uses master/slave architecture, consisting of a namenode and several datanodes. The namenode is a central server responsible for managing the file system and client access to files. The datanode is distributed in the cluster responsible for managing the storage of its own node. Internally, a file is divided into one or more blocks, and each block is stored in one datanode. The namenode is responsible for file system operations, such as open, close, rename files and directories, and decide the block mapping to the specific datanode. Under the command of the namenode, datanode creates, deletes, and copies blocks.

Above *HDFS*, we build a distributed database (Hbase) for data management. Through data partition and data layout technology, data can be stored in *HDFS*. And then through data encryption, fault tolerance, compression, backup technologies and measures, we ensure that data will not lose in the cloud, and the cloud is safe and stable.

In addition, this layer can do data partition, create data index and virtual view in order to efficient query.

4) Data Application layer (DAL)

In the DAL, we can design efficient parallel data query optimization algorithms, data mining and analysis algorithms to reduce the query response time. Through mobile phones, PDAs, PCs and others, users login our system and achieve data access, data analysis services.

IV. PROTOTYPE SYSTEM AND EFFICIENCY ANALYSIS

A. System Implementation

The prototype system shown in Figure 3 provides several services, such as *Released Services*, *Applying Services*, *Being Nodes* and others. *Released Services* displays the released services and released data. If users want to apply for the released services, they can apply for the services directly in *Applying Services*. If the released services cannot meet user needs, users can choose to apply for a new service and upload related data, and the administrator will open the service in time.

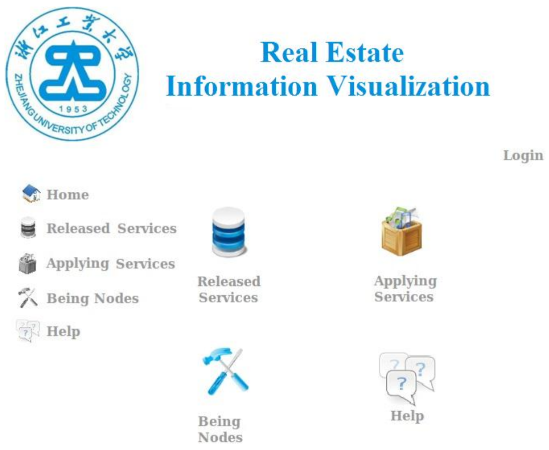
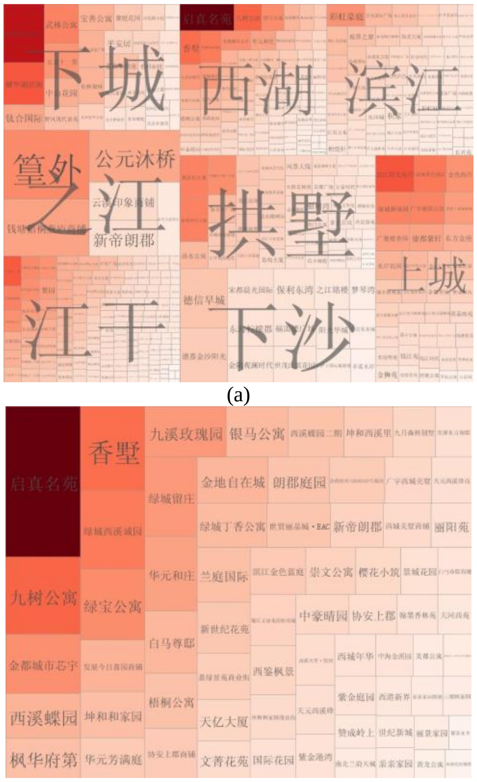


Figure 3. Prototype system

After the services opened which users apply for, users can see the specific services in *Released Services*, such as *Average House Prices*. If users select *City='Hangzhou'*, they can get average house prices in Hangzhou, as shown in Figure 4.



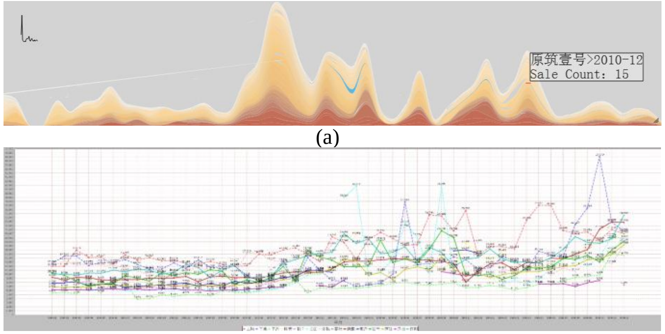
(a)

(b)

Figure 4. (a)Average house prices in Hangzhou; (b)Average house prices of Xihu district in Hangzhou

In the *Treemap* diagram shown in Figure 4, both the size of the rectangle and the color depth stand for the house average prices. The deeper the color, larger rectangle, means higher house prices.

In addition, users can also select other visualization methods like *StringGraph* and *Line Chart* to see results, shown in Figure 5.



(a)

(b)

Figure 5. (a) StringGraph for sale counts; (b) Line chart for average house prices of every district in Hangzhou

B. Efficiency Analysis

Hardware environment: one Dell PowerEdge R410 server (Intel Xeon quad-core 5600 processors, 4G memory, 160G hard disk), seven Dell Optiplex 780 PCs (Intel Core 2 Duo E7500 processors, 4G memory, 160G hard drive) .

Software environment: Ubuntu10.10, JDK1.6.0_24, Hadoop-0.20.2, Hbase-0.20.6, Zookeeper-3.3.3.

TABLE II. MAPREDUCE COSTS

Data Size	Map (s)	Reduce (s)	MapReduce (s)
64M	38	36	74
1G	40	38	78
8G	41	42	83
32G	43	45	88
128G	49	56	105

Firstly, use the server as the namenode and virtual machines created on other PCs as datanodes to build HDFS. Secondly, build Hbase above HDFS. TABLE II shows the average time that this cluster process data using *MapReduce*.

Seen from TABLE II, the cost using *MapReduce* computational framework to deal with 64MB of data and 128GB of data is little different. Therefore, using *MapReduce* to deal with big data has obvious advantages.

We also do the experiment on traditional single-server--PowerEdge R410 server (Intel Xeon quad-core 5600 processors, 4G memory, 160G hard disk). Figure 6 shows the contrast about the costs of above two experiments.

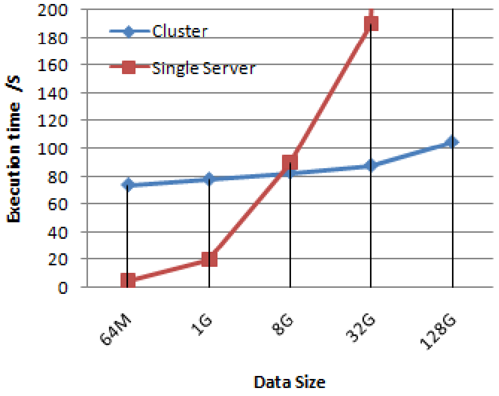


Figure 6. The execution time of cluster and single server

Seen from Figure 6, when the data size is about 8GB, the execution time of cluster and single server is nearly. For small-size data (less than 8GB), the traditional single-server service model has advantages. When the data is larger than 8GB, the advantages using *MapReduce* computational framework to parallel process are very obvious, while single server costs too much time. Therefore, the cloud computing model is more suitable for big data processing.

In addition, in accordance with the format of TABLE I, we use Hive to create partition table of real estate information. The HiveQL is as follows.

CREATE TABLE HouseInfo_Partition (House String, District String, Type String,Year Int, Month Int, HousePriceperQuarem-eter Float, BuildArea Float, HousePrice Float, RecordCount Int) PARTITIONED BY (City String);

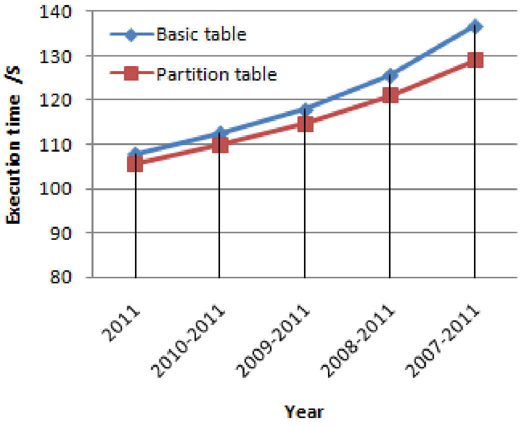


Figure 7. The query execution time of partition table and basic table

User's operation can be directly transformed to Hive query. In Hive, a partition of the table corresponds to a directory. Which partition table we create partitioned by *City*. Some queries need not all data in the table. For example, when we query the prices of Hangzhou real estates, we do not have to search for redundant information of other cities. Seen from Figure 7, the query execution time of partition table is slightly less than basic table.

V. Conclusion and Future Work

In this paper, we propose a visualization method based on cloud computing for real estate information. The core idea is the combination of cloud computing and information visualization, making the data processing stage of information visualization on cloud computing platform, to speed up the calculation efficiency. Use *MapReduce* to process data and use *TreeMap* to show the results.

In addition, we design and implement a prototype system for real estate information visualization. Users can apply for the relevant real estate information visualization services on our system. The experimental analysis shows that the advantages using *MapReduce* computational framework to parallel process big data are very obvious, which provides a new approach for massive information visualization.

This system also has some unsatisfactory areas. For example, human-computer interaction is not convenient, users' waiting time is too long and the system also has some limitations. Therefore, in future, it is necessary to consider the improvement of human-computer interaction, but also consider the results effect and user-friendly operation.

VI. ACKNOWLEDGMENT

This work was partly supported by the National Natural Foundation of China (NSFC) project under grant No. 60703002, Major Program of Zhejiang Provincial Natural Foundation under grant No.Z1090630, the Natural Science Foundation of Zhejiang Province (No.Y1110768, Y110950), and Dr start-up fund research of Zhejiang University of technology (No: 119001229).

REFERENCES

[1] G. Turcu, I. Foster, and S. Nestorov. “Reshaping text data for efficient processing on Amazon EC2.” Scientific Programming, 2011, 19(2-3), pp. 133-145.

[2] J. Dean, and S. Ghemawat. “Map/Reduce: Simplied Data Processing on Large Clusters.” Communications of the ACM, 2004, 50(1), pp. 107-113.

[3] Jeff A. Stuart, C.K. Chen, K.L. Ma, and John D. Owens. “Multi-gpu volume rendering using mapreduce.” 1st International Workshop on MapReduce and its Applications, 2010.

[4] T. Yuzuru, C.K. Chen, M. Stephane, and K.L. Ma. “An Interface Design for Future Cloud-based Visualization Services.” 2nd IEEE International Conference on Cloud Computing Technology and Science, 2010, pp. 609-613.

[5] X.Q. Yang, and Y.J. Deng. “Exploration of Cloud Computing Technologies for Geographic Information Services.” 2010 18th International Conference on Geoinformatics, 2010, pp. 1-5.

[6] C.W. Yang, M. Goodchild, Q.Y Huang, D. Nebert, R. Raskin, Y. Xu, M. Bambacus, and D. Fay. “Spatial cloud computing: how can the geospatial sciences use and help shape cloud computing.” International Journal of Digital Earth, 2011, 4(4), pp. 305-329.

[7] W.G. Teng, W.H. Wen, and Y.C. Liu. “From experience to expertise: digesting cumulative information for informational web search.” Journal of information science and engineering, 2012, 28(1), pp. 161-176.

[8] H. Brian. “Cloud computing.” Communications of the ACM, 2008, 51 (7), pp. 9-11.

[9] Nahum D. Gershon, and Stephen G. Eick. “Information Visualization.” ACM Interaction, 1998, 5(2), pp. 9-15.

[10] M. Kanae, Y. Masato, and S. Hideki. “A Proposal of Framework for Information Visualization in Developing of Web Application.” 2011 IEEE/IPSJ 11th Infornational Symposium on Applications and the Internet(SAINT), 2011, pp. 457-462.

[11] V. Rantanen, T. Vanhala, and O. Tuisku. “A Wearable, Wireless Gaze Tracker with Integrated Selection Command Source for Human-Computer Interaction.” IEEE Transactions on Information Technology in Biomedicine. 2011, 15(5), pp. 795-801.

[12] M.Q. Zhou, R. Zhang, D.D. Zeng, and W.N. Qian. “Services in the Cloud Computing Era: A Survey.” 2010 4th International Universal Communication Sysposium, 2010, 10, pp. 40-46.

[13] S.Y. Chen and Y.F. Chang. “he cmputer-aded dsign sftware for sart hme dvice based on coud cmputing service.” 010 Second WRI World Congress on Software Engineering, 2010, 7, pp. 273-278.

[14] Q. Liu and W.Q. Gu. “An Elastic Public VPN Service Model Based on Cloud Computing.” 2011 IEEE 2nd International Conference on Software Engineering and Service Science(ICSESS), 2011, pp. 290-294.

[15] A.B. Vigas, W. Martin, V.H. Frank, K. Jesse, and M. Matt. “Many eyes: A site for visualization at internet scale.” In Proceedings of Infovis, 2007, 13(6), pp. 1121-1128.

[16] K. Andrews and M. Lessacher. “Liquid Diagrams: Information Visualisation Gadgets.” Information Visualisation(IV), 2010 14th International Conference, 2010, pp. 104-109.

[17] H.T. Vo, C. Chen, and B.C. Ooi. “Towards Elastic Transactional Cloud Storage with Range Query Support.” Proceedings of the VLDB Endowment, 2010, 3(1-2), pp. 506-514.

[18] D. Kossmann, T. Kraska, S. Loesing, S. Merkli, R. Mittal, and F. Pfaffhauser. “Cloudy: A Modular Cloud Storage System.” Proceedings of the VLDB Endowment, 2010, 3(1-2), pp. 1533-1536.

[19] Y. Dong, H. Zhu, J.Z. Peng, F. Wang, M.P. Mesnier, D.W. Wang, and S.C. Chan. “RFS: A Network File System for Mobile Devices and the Cloud.” ACM SIGOPS Operating Systems Review, 2011,45(1), pp. 101-111.

[20] A. Thusoo, S.J. Sen, N. Jain, Z.Shao, P. Chakka, N. Zhang, S. Antony, H. Liu, and R. Murthy. “Hive – A Petabyte Scale Data Warehouse Using Hadoop”. 26th IEEE International Conference on Data Engineering (ICDE 20-10), 2010, pp. 996-1005.

User Centred Automated Composition in Telco 2.0

Armando Ordonez, Juan C. Corrales
Telematics Engineering Group
University of Cauca
Popayán,
{jaordonez,jcorral}@unicauca.edu.co

Paolo Falcarin
ACE School
University of East London
London, UK
falcarin@uel.ac.uk

Jaime Guzman
Intelligent Web System Group
Universidad Nacional de Medellín
Medellín, Colombia,
jguzman@unalmed.edu.co

***Abstract*— Services composition in Telco 2.0 is known as convergent or unified composition. This composition is a very particular process with special features and complexity associated to technical differences between Web and Telco. As available services grow exponentially and are updated on the fly, it has become impossible for human capacity to analyse all of them and generate manually a composition plan. This paper presents a service architecture for user centred automated composition in Telco 2.0. Our approach is based on artificial intelligence planning considering the context information from the user and its access device using a cost function. Finally, a prototype of the user centred planning module is presented which takes as input the request based in the user in natural language and returns the service to execute.**

Keywords-Convergence; End User Service Composition; Automated Composition.

I. INTRODUCTION

Telco 2.0 can be described as a model that relates concepts, services and technologies of Web 2.0 with traditional telecommunication services. This combination is known as convergent or unified composition. Most common approach for services composition is performed manually at design time. As available services grow exponentially and are updated on the fly, it has become impossible for human capacity, to analyse all of them and generate manually a composition plan [1]. In the other side, using new technologies in services composition field (like semantic annotations and AI planners) allow visualizing an automated composition.

Previous works deal with this problem and present some techniques and architectures; some of these approaches, coming from European projects like SPICE [2], OPUCE [3] and OMELETTE **¡Error! No se encuentra el origen de la referencia.**, do not automate the whole service composition process. Besides, these approaches lack of ways for expressing user request through voice and do not include user context information in plan generation. In the other side, some AI Planning approaches propose customize the planning process extending planning languages like PDDL (Planning domain definition language) like [5]. From Telco 2.0 perspective, this approach makes it very complex to generate planning domains. In this context, the main contributions of this paper are (1) service architecture for automated composition in Telco 2.0 considering issues associated with convergent composition for the whole process, (2) a technique to customize the planning process using cost function using the LSP (Logic Scoring preference) method [6].

This paper is organized as follows: Section 2 the motivating scenario. Section 3 presents the issues related to convergent composition. Section 4 describes the proposed approach. Section 5 presents an evaluation of the planning technique, Section 6 presents the related work and Section 6 draws the conclusion and future work.

II. MOTIVATING SCENARIO

To illustrate our proposal, we present a case study: an environmental management system (see Fig. 1). The Environmental manager is on charge of decision making about environmental alarms and crops. In order to do so, the manager has information from sensor networks; equally, he can use Telco and Web services to process basic data and send information to all the farmers and sensors. Reuse of functionalities is a very important issue for some developing countries where budgets for technologies are limited.

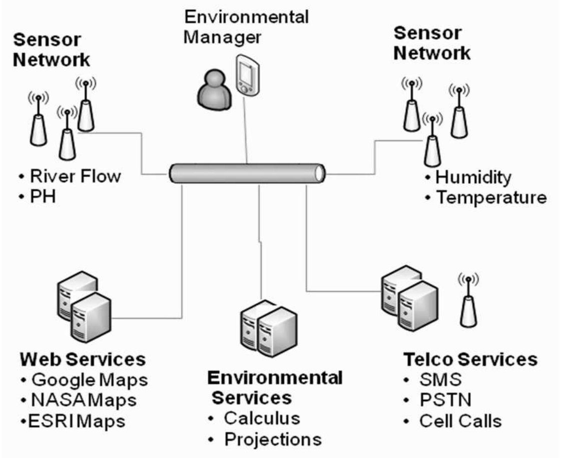


Figure 1. Telco and Web services interaction in environmental management systems.

Usually, environmental manager is an ecologist or biological expert. Therefore, his knowledge about underlying technologies is usually low. Commonly the user expresses his request in an informal way; we will illustrate this with two examples:

- "I need calculate hydrological balance of zone one and receive the resulting map to my mobile".

- “If the river flow of zone two is greater than 15% of average, emit an alarm to every farmer within a radius of 2 miles from the river”

The first request can be entered through voice, by a mobile device. Next, the system gathers information of sensors of zone one (preconfigured in the system or specified by coordinates). The system uses hydrological services from Internet, sending sensor data and maps from Google maps. Finally the resulting image is sent by MMS to the user mobile. In the second request, sensor data is evaluated, if necessary, an emergency map is generated. This map is created drawing a radius of 2 miles from the sensor. To do so, the system uses GIS (Geographical information systems) services and maps from internet. Finally the system informs about the alarm to farmers inside the emergency area; the best way to send the information is selected: SMS, Cell Phone call, fixed telephone call, voice message. In both cases, services from Web and Telco are used. These services work together and in coordination to save lives or help to make decisions about crops. Besides, take the services composition to the end user level, leads to ease of interaction with the system for non-experts users without IT infrastructure or personal required.

III. ISSUES FOR SERVICES COMPOSITION

Services composition in telecommunication networks is fundamentally different from Web services composition [7]. Next, some of the main issues for convergent composition are analysed:

Usability Creation: Convergent composition requires request specification in an easy way. Most of the users are not familiar with technologies such as service creation environments (SCE) or languages like Business Process Execution language (BPEL) [8]. In this scenario, application of natural language processing (NLP) techniques can be useful for deriving a formal specification from a request in natural language [9].

Time constraints: In telecommunications domains, there are real-time requirements in protocols and platforms, e.g., post-dial delay is typically bounded. In contrast, a best-effort response time is typically required from Web Services. For convergent composition, time is a crucial factor [7].

Services representation: Commonly end users should not know implementation details of services like URL, protocols or billing processes. In this scenario, Users should interact with abstract representation of services without implementation details [9].

User centred creation: A high-quality composed service might be one that fits to personal preferences. Besides, users can be connected using different devices with different capabilities. A composed service should include all these preferences [5].

Automatic code generation: In Telco domain, time constraints are mandatory. Heavy XML parsing and deep treatment of semantic inference can be used only for services representation. Using efficient mechanisms such as automated code generation which decrease deploy times are more suitable for execution environments [10].

Reconfiguration: Reconfiguration is a crucial factor in the convergent scenario as reliability of Web services is not as high as Telco services [7].

IV. OUR APPROACH

Next, we describe the architecture of our approach for automatic convergent composition (see Fig. 2.). This paper focuses in the user centred planning module and other modules of the architecture have been detailed by separate [11],[12]. In our approach, description of the services is different depending of the component. At the NLP Analyser, “Abstract services” descriptions are used based on OWL-S. While, at Plan Adapter Component: services descriptions are associated to implementation details and real descriptions of “Executable services” (e.g. WSDL). Abstract Services like “get”, “inform” and “gather” are internally associated with real implementation of Executable services that are associated with SOAP or REST implementations. The relation between Executable Services and Abstract Services is created by domain-experts in environmental management through folksonomies [13]. Folksonomies offer a system of grouping services through a collaboratively method for creating and managing tags to annotate and categorize individual services. The architecture includes a user graphic interface where services can be registered. In this interface, domain experts can add tags to the services in order to be discovered later [14].

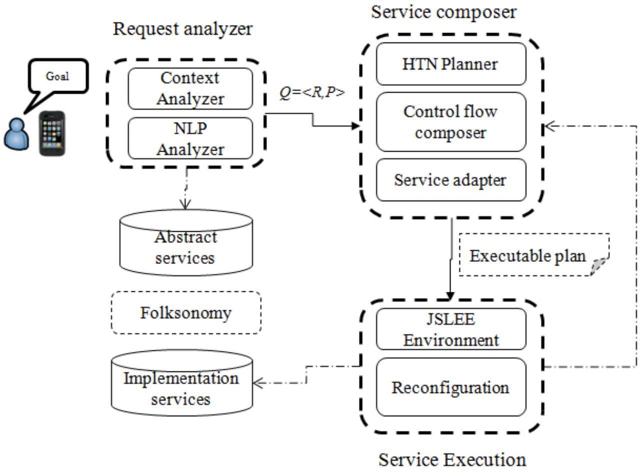


Figure 2. Telco and Web services interaction in environmental management systems.

The NLP Analyser extracts names of Abstract services (e.g., “inform farmers" this service could involve send SMS or establish a voice call) from user request and translates them in a problem for the planner. At the same time, it gathers information about control flow and context. For example, if the request is: “check sensors and inform farmers”. NLP Analyser obtains: (check (Sensors)) AND (inform (farmers)). “Check” and “Inform” are translated to problems for the planner and “AND” describes a concurrent execution that will be considered latter in the Plan Adapter Component.

The context Analyser allows including user context information in the planner; as information from devices. To do so, device reference is checked in the WURFL (Wireless Universal Resource File) and the CC/PP (Composite Capability/Preference Profiles). These repositories hold all the information about capabilities of devices in the market. And provide an effective way for analysing which services can be provided to users.

With this information, a relation factor between services and users is calculated. Next the HTN Planner component obtains a plan decomposing tasks (the higher level task is an Abstract Service from NLP Analyser) in finer atomic tasks (the lowest level task or operator is an Executable Service). This plan is sent to the services discovery component to relate abstract services with implementation services. Next, the plan is sent to plan adapter component, which includes the control flow information from the user request and translate it into an executable code ready to run in Mobicents Communication Platform. Mobicents provides a robust Java Service Logic Execution Environment (JSLEE) [15]. Finally, the reconfiguration component monitors services execution; in case of failure, services selection, or re-planning is performed. Next, a description of each component is presented.

A. Request Analyser Component

This module receives user request and translate it to machine understandable language, this process can be done automatically or include user intervention at each step. This component holds two sub components context and NLP analyser.

1) NLP Analyser: In this component, User Requirements are modelled with a query *Q* specified as a couple *<R;P>*, where *R* represents the request part of the query and *P* represents the user preferences of the query (device, network, position). In turn, *R* is specified as an n-uple *R= [{s1,s2…sn}, F]*. Each *sn* denotes an Abstract Service and F represents control flow information. For its part, each s is composed of 3-uples *<I;O;C>*; where *I* denotes the input data the user provides, *O* denotes required information to be provided as a result of the query and *C* denotes a functionality (associated with Abstract Services). For example:

Q: is a request made by an environmental manager from a cell phone. Where:
P: Cell phone reference and network capabilities.
R: “I need calculate hydrological balance of zone one and receive the resulting map to my mobile”.

From R we can expand:
s1: “calculate hydrological balance of zone one”
s2: “receive the resulting map to my mobile”
F: AND (sequence of actions)

Analysing s1: “calculate hydrological balance of zone one”
I: Zone one. (An internal variable, geo-coded location or set of coordinates of Zone one)
O: Hydrological balance map,
C: Calculate hydrological balance service.

2) Context analyser: This component extracts the context information from the user and its access device. Context is any information that characterizes the situation of an entity that can be person or computational object [16]. Common types of context include the computing context (e.g., network connectivity), the user context (e.g., profile, location), the physical context (e.g., noise levels). These features are known also as Non-functional properties (NFP) and the combination of all these profiles constitutes the User Profile [17]. For our approach we use profile model from Sutterer et.al [18]; this model consider user profiles that allows deal with any parameter that could be necessary according to the domain. In the present architecture, the user profiles define alternatives for notifications delivery based on user location, computing device and network bandwidth, as summarized in Table. 1. The column criteria is extracted from the user and define some services criteria that are assigned to operators.

TABLE I. CRITERIA FOR SERVICES SCORING BASED ON USER CONTEXT

	criteria	**Values**	**Service criteria**	**weight**
User context	Network	GPRS/ WLAN/ GSM	payload size	bytes
	Device	Cell phone, Laptop	payload size	bytes
	Location	Outdoor, indoor	voice , text	integer
User preferences	Data subscription	Yes/No	require data subscription	Boolean
	Only Free services	Yes/No	cost	value
	Voice subscription	Yes/No	voice, text	Boolean
	Delivery quality	low, medium, high	delivery warranty	integer

The weight defines the importance for the user in an specific situation of moment. For example, in case of emergency delivery warranty is most important that cost. And for testing of applications only free services are required for specific services. All these values are included in the cost function with normalized values of the weight and define the cost of a plan as explained further.

B. *Services Composer*

This component receives processed user request from the previous component and translate it into an executable plan for JSLEE environment. This process is performed automatically; however, this module can include user verification of the generated plan.

1) HTN Planner: Previous works have determined benefits of use AI planners in services composition [19]. The HTN planner produce a sequence of actions that perform some activity or task, this sequence is called a plan. Planning proceeds by using functions called methods to decompose tasks recursively into smaller subtasks, until the planner reaches primitive tasks that can be performed directly using the planning operators. For our framework, we use the planning engine SHOP2 [20], improving planning process according to user context through cost functions.

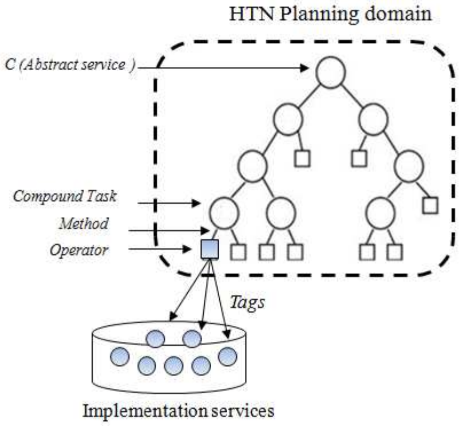


Figure 3. Relation between HTN Operators and implementation services.

A planning domain describes the context of the planning process (see Fig. 3.). This domain consists of a set of methods and operators. A task represents a service or activity to perform. A task may be either primitive or compound. A compound task is one that needs to be decomposed into smaller tasks using a method. Methods indicate how to decompose a compound task into a partially ordered set of subtasks, each of which can be compound or primitive. The highest-level task is extracted from the user request and is associated with variable *C* from *sn* as explained in the previous section. The lowest-level of task is a planning operator, i.e. a primitive service (or task) that cannot be decomposed further and is directly related with implementation Services.

An operator can be associated with many implemented services; the relation between operators and implementation services is optimized using folksonomies. A folksonomy can be described by a tripartite model of tagging activities [21], which consists of tagging entities (i.e., users, tags, and resources). *Tagging =<U, T, R>*; here *U* is the set of users who participate in a tagging activity. In the platform U is represented by a group of domain experts who make this tagging. *T* is the set of available tags; the available tags are the operator names. R is the set of resources being tagged; in this case, the available services. Using folksonomies improves the time for services selection, as the search space is limited to implementation services that best match with related tags. The planning domains for the platform is created at design time based on OWL-S descriptions, so that, a domain is ready at execution time and a HTN planning process can be executed with good time performance.

Each operator also has a cost (the default value is 1). The cost of a plan is the sum of the costs of the operator instances. In our approach, we assign a cost based on the relation between services and user. In this manner cost of plans lets create a ranking of plans; the lowest cost plan is selected. And in case of execution failure the second plan in the ranking can be selected.

The cost of each operator depends on the user context and preferences. For considering relation between services and user preferences, is needed to analyse the importance (weight) of each criteria for each user. For example, a user may establish that they simultaneously need low cost and MMS messages enabled service. For calculating the operator cost, is necessary to calculate a scoring technique. For the present architecture, LSP is selected; LSP is one evaluation method that extends the traditional scoring techniques to consider besides of the weight, the relation between criteria [6]. In the present architecture the cost function, receives the user preferences and returns a value that corresponds to operator cost for each service. This cost is included in the planning process as was explained before.

2) Service Adapter component: This component receives the abstract plan from the previous component, and associate services and control flow (extracted during NLP analysis) creating a composed set of Java components called SBB (services building blocks). SBB are the core components of JSLEE environment and allow create composed services referencing other SBBS. These SBB are precompiled in a repository in order to avoid unnecessary compilation processes. The executable plan includes control points and reconfiguration information included in the code necessary by fault handlers for identifying failures during execution.

C. Services execution component

This component is based on Java Service Logic Execution Environment (JSLEE). JSLEE is an emerging standard specification targeted to host convergent services. In JSLEE: Telco and Web services, as composition of these services can be represented by SBBs. In order to support changes in flow and eventual reconfiguration during execution, a dependency injection method is used to manage changes on the fly without recompilation. Service reconfiguration, is implicitly related to monitoring process, and implies to be aware of services status. When a service has an unwanted behaviour, Mobicents environment activates an alarm. These alarms initialize the reconfiguration process, which would detect the failure and proceed to determine one of three actions: if the error is caused by an atomic service, so the system selects another service and goes on. In the other way, if the problem is caused by the whole process and it can be changed: a new plan from the generated ranking is selected. Finally, if there is a problem in the middle of the executing process, a new planning process is initiated, in order to complete the task beginning from the actual state of the world, i.e., a set of values in the variables that define a system in a given moment.

The goal of the present architecture is to offer a coherent and sound framework for automate the steps in the services composition. This architecture can be used with user intervention at each step in order to verify the automation process. For example, after the Natural language analysis, the user can review the request transformation. Equally, after the composition process, the user may verify the composed plan. Thus the present architecture aims for ease the service composition process.

This architecture is oriented to environmental management domain. Most of the principles and functionalities presented here are appropriated for different domains.

V. EVALUATION OF THE PLANNING APPROACH

In terms of functionality, the planning approaches can respond to a planning problem and returns a plan. However, to provide a suitable mechanism for automated plan generation in convergent Telco 2.0 domain. This mechanism must include user context information and must be fast enough. Therefore, the prototype implementation was used to carry out some simple experiments designed to test the planning performance in a realistic planning scenario. The objective is to determine if the performance time is suitable for Telco environments.

The experiment used a planning domain with variable number of operators that describe Telco Services in the environmental management domain, the cost function relates the criteria presented in Table 1. The basic pattern of the operators is:

(:operator (!send_sms) () () () (call CalculateCost))
(:operator (!generate_map) () () () (call CalculateCost))

Each operator calls the *CalculateCost* function that calculates the related cost between service and user; to do so, the function extracts information of the user and service and apply the scoring technique. The number of plans and the number of operators was modified in order to analyse time response of the prototype. For planning, JSHOP is used, which is a Java version of the SHOP2 [22] planner. Equally the prototype uses MySQL for storing user and services, information and JDK. 1.6.0. All the experiments were performed on a machine equipped with a Pentium Dual Core T4400 2.2 GHz processor, with 4 GB of main memory and running Windows as OS.

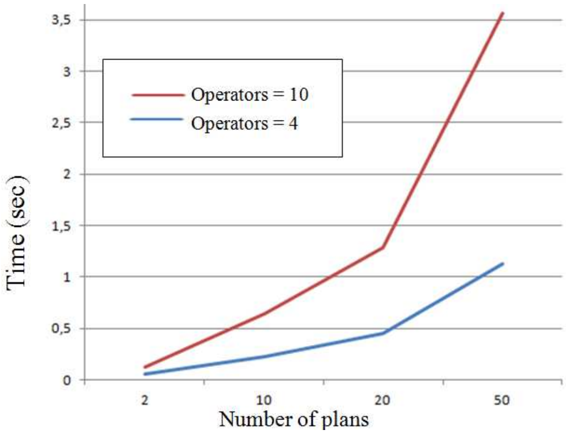


Figure 4. Time performance of the planning module

As expected, Fig. 4 shows that the time increases with the number of plans generated and depending of the operators involved in the planning domain. Convergent services usually have between 4 and 10 operators to avoid extreme complexity, and the resulting ranking usually can have up to 5 plans. In this context, despite the exponential cost of planning module, the figure shows that this one can be used for 5 plans with acceptable execution time (0.5 sec). Future work include decreasing the time for more operators, the last could be done including some heuristics in the planning algorithm.

VI. RELATED WORK

Our approach is focused on automation of user centred service composition, and we intend to apply our framework to the environmental management. Our approach considers all three phases or components for automated service composition: request analysis, services composition and service execution. Previous works have proposed frameworks for automated services composition, like Kim et.al [23] and Rao et.al [24]. They both present phases for automatic composition, focusing only on Web domain without concern on execution. Shia et al. [25] and da Silva et al. [9] present frameworks for automatic composition. These frameworks exploit natural language processing and semantic annotations for services matchmaking based on SPATEL language [25]. They do not address the validation of the non-functional properties and are focused only on the request analysis and plan generation. Our approach deal with all the phases including execution and reconfiguration based on JSLEE environments. Sirin et al. [26], provide an algorithm to translate OWL-S service descriptions to a SHOP2 domain and makes planning based on services. Although, this works lack of details of implementation in real environments and focus only on Web services. Other authors present approaches for Web service composition based on AI planning with preferences [27]; however they propose extensions to standard planning language and adaptation of planners, adding a new level of complexity to the automated composition (other levels include semantic annotation of telecommunication and Web services and basic planning domain definition). Table 2 outlines a comparison of related work, based in the following criteria: first, if the work deals with all the phases in the service composition process including reconfiguration. Second, if the approach for service composition is user-centred. Third, if the approach takes in account convergent considerations.

TABLE II. COMPARISON OF RELATED WORKS

works	Include all phases	User centred	Consider Convergence
[23][24]	No, just the request analysis and the service creation	No	No, they focus on Web domain
[25][5]	No, just the request analysis and the service creation	No	Yes
[26]	No, just the OWL-S based planning	No	No, focused on Web domain
[3]	No, just the planning with preferences	Yes	No, focused on Web domain

VII. CONCLUSION AND FUTURE WORK

Automated convergent composition is a very intensive research area; previous works have worked on some aspects of this process. However, there is not a complete framework to develop this process. None of the above proposals presents details to apply proposed frameworks to Telco environments in a real environment, the works presented by Shia et al. [25] [9] are the most relevant for us in the literature, many elements are similar to our work but our approach has a different direction. They deal with complex treatment of SPATEL language and ontologies in order to reach automated User centred composition. We deal with user profile information to customize AI planning. Equally, our approach is tending to reach low execution times and include mechanisms for automated reconfiguration. Our future works include the development of mechanisms for automation of planning domain creation, and experimentation with other planners in order to consider reconfiguration in different phases of the process and better execution times. Equally we are extending the preferences criteria in order to get a better personalized experience for the user.

ACKNOWLEDGMENT

The authors thank University of Cauca, University of East London and Colciencias for supporting this research.

REFERENCES

[1] S-C. Oh, D. Lee, and S. Kumara, “A Comparative Illustration of AI Planning-based Web Services Composition,” ACM SIGecom Exchanges,vol. 5, Jan. 2006, pp. 1-10, doi: 10.1145/1124566.1124568.

[2] P. Falcarin, “Service Composition Quality Evaluation in SPICE Platform,” High Assurance Services Computing, Springer US, pp. 89-102, 2009.

[3] J.C. Yelmo, J.M. del Alamo, R. Trapero, P. Falcarin, Y. Jian, B. Cairo, and C. Baladron, “A User-Centric service creation approach for Next Generation Networks,” Proc. Innovations in NGN: Future Network and Services (K-INGN), May 2008, pp. 211-218, doi: 10.1109/KINGN.2008.4542268.

[4] O. Chudnovskyy et al., “End-user-oriented telco mashups: the OMELETTE approach, “ Proc. of the 21st international conference companion on World Wide Web (WWW '12 Companion), Abril 2012, pp. 235-238, doi: 10.1145/2187980.2188017

[5] J. Baier and S. McIlraith, “Planning with Preferences,” AI Magazine, vol. 29, 2008, pp. 25-36.

[6] H.Q. Yu and S. Reiff-Marganiec, “A Method for automated Web Service Selection,” Proc. IEEE Congress on Services, July 2008, pp. 513-520, doi: 10.1109/SERVICES-1.2008.8.

[7] G. Bond, E. Cheung, I. Fikouras, and R. Levenshteyn, “Unified telecom and Web Services Composition: problem definition and future directions,” Proc. 3rd International Conference on Principles, Systems and Applications of IP Telecommunications (IPTComm’09), ACM Press, pp. 1-12 ,doi: 10.1145/1595637.1595654.

[8] A. Arkin et al., “Web Services Business Process Execution Language Version 2.0. Committee Draft,” WS-BPEL TC OASIS, December 2005.

[9] E. da Silva¸ L. Ferreira, and M. van Sinderen, “Towards runtime discovery, selection and composition of semantic services,” Computer Communications, vol. 34, Feb. 2011, pp.159-168, doi: 10.1016/j.comcom.2010.04.003.

[10] A. Lehmann et al., “TeamCom: A Service Creation Platform for Next Generation Networks,” Proc. Fourth International Conference on Internet and Web Applications and Services (ICIW’ 09), IEEE Computer Society, May 2009, pp.12-17, doi: 10.1109/ICIW.2009.10.

[11] E. Pedraza, J. Zúñiga, L. Suarez, J.C. Corrales, “Automatic Service Retrieval in Converged Environments Based on Natural Language Request,” Proc. The Third International Conferences on Advanced Service Computing, Sep. 2011, pp. 52-60.

[12] A. Ordonez, J. Corrales, P. Falcarin, “Natural language processing based Services Composition for Environmental management”, Proc. 7th International Conference on Systems of Systems Engineering. July 2012, in press.

[13] I. Peters, “Folksonomies: Indexing and retrieval in Web 2.0,” De Gruyter Saur, Berlin, 2009.

[14] A. Ordonez, J. Corrales, P. Falcarin, “Automated context aware composition for convergent services”, Proc. 7th International Conference on Systems of Systems Engineering. July 2012, in press.

[15] D. Ferry et al. JSR 240, JAIN SLEE (JSLEE) v1.1, July 2008.

[16] A. Dey, “Understanding and Using Context,” Journal of Personal and Ubiquitous Computing, vol. 5, Feb. 2001, pp. 4-7, doi:10.1.1.31.9786.

[17] S. Panagiotakis, M. Koutsopoulou, A. Alonistioti, and S. Thomopoulos, “Context sensitive user profiling for customised service provision in mobile environments,” Proc. 16th International Symposium on Personal, Indoor and Mobile Radio Communications (IEEE PIMRC 2005), IEEE Press, Sept. 2005, pp. 2014-2018, doi: 10.1109/PIMRC.2005.1651793.

[18] M. Sutterer, O. Coutand, O. Droegehorn, K. David, and K. Der Sluijs, “Managing and Delivering Context-Dependent User Preferences in Ubiquitous Computing Environments,” Proc. International Symposium on Applications and the Internet Workshops (SAINT-W '07), IEEE Computer Society, Jan. 2007, pp. 4-4, doi: 10.1109/SAINT-W.2007.60.

[19] J. Xu, K. Chen, and S. Reiff-Marganiec, “Using Markov Decision Process Model with Logic Scoring of Preference Model to Optimize HTN Web Services Composition,” International Journal of Web Services Research, vol.8, 2011, pp. 53-73, doi: 10.4018/jwsr.2011040103.

[20] D. Nau, O. Ilghami, U. Kuter, J. William, D. Wu and F. Yaman, “SHOP2: An HTN Planning System,” Journal of Artificial Intelligence Research, vol.20, Dec. 2003, pp. 379-404, doi: 10.1.1.72.188.

[21] H-L. Kim, J. Breslin, H-G. Kim and J-H. Choi, “Social semantic cloud of tags: semantic model for folksonomies,” Knowledge Management Research & Practice, 2010, pp. 193–202. doi:10.1057/kmrp.2010.10.

[22] D. Nau, O. Ilghami, U. Kuter, J. W. Murdock, D. Wu, and F. Yaman, “SHOP2: An HTN planning system,” Journal of Artificial Intelligence Research, 2003, vol. 20, pp. 379-404, doi: 10.1613/jair.1141.

[23] A. Kim, M. Kang, C. Meadows, and J. Sample, “A Framework for Automatic Web Service Composition,” Technical Report, Naval research Lab USA, 2009.

[24] J. Rao, and X. Su, “A Survey of Automated Web Service Composition Methods,” Proc. First International Workshop on Semantic Web Services and Web Process Composition (SWSWPC 2004), July 2004, pp. 43-54, doi: 10.1.1.122.6700.

[25] M. Shiaa, P. Falcarin, A. Pastor, F. Lécué, E. Silva, and L. Ferreira, “Towards the Automation of the Service Composition Process: Case Study and Prototype Implementations,” Proc. ICT Mobile Summit, June 2008, pp. 1-8.

[26] E. Sirin, B. Parsia, D. Wu, J. Hendler, and D. Nau, “HTN Planning for Web Service Composition using SHOP2,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 1, Oct. 2004, pp. 377-396, doi: 10.1016/j.websem.2004.06.005.

[27] S. Sohrabi, N. Prokoshyna, and S. McIlraith, “Web Service Composition via the Customization of Golog Programs with User Preferences,” Conceptual Modeling: Foundations and Applications, Springer-Verlag Berlin, Heidelberg, pp. 319-334, 2009.

.

An Architecture of Video Object Annotation Platform for Interactive TV

Hyun Namgoong, Kang Yong Lee, Kee Seong Cho, Won Ryu
Smart Screen Convergence Research Department
Electronics and Telecommunications Research Institute
Daejeon, South Korea
{nghyun, kanglee, chokis, wlyu}@etri.re.kr

Hong-Gee Kim
Biomedical Knowledge Engineering Laboratory
Seoul National University
Seoul, South Korea
hgkim@snu.ac.kr

***Abstract*— Current advances of television systems provide interactive user experiences. This paper proposes a Video Object Annotation Platform (VOAP), which enables TV viewers can be provided with additional information during they is watching TV. The platform anchored to semantic data serialization to support a sharing of annotations located to specified object at a moment. With the platform, content providers can publish additional information about objects appeared in a program and the viewer can easily access the information. We expect that the proposed platform will provides a novel business model for advertisement market by providing seamless information transfer among consumer (viewer), advertisers (company) and mediation (broadcasting company). As an ongoing work proposing an information exchanging environment in interactive TV platforms, we implemented a user-side application proving the proposed concept.**

Keywords-IPTV; Interactive TV; Video Object Annotation; Annotation Platform.

I. INTRODUCTION

With the name of IPTV (Internet Protocol Tele-Vision), current advances of TV systems provide people interactive experiences like VOD (Video On Demand) service and time-shifted watching [1]. Broadcasting companies and research groups are trying to put more services showing advantage of such interactive TV system [2], so that, the television viewers can enjoy various experiences in the TV screen. A TV answering user questions is not a new research topic, because, it ensures many new opportunities in commercial Ads. From casual one like “Who is that actor?” to complicated ones “ Where the movie is taken from?” and “ Where can I buy that thing?” the informative questions are easily come up when we watch TV. An environment to provide such additional information to viewers during watching TV is also important to broadcasting companies or content providers. They can not only answer such question with additional information, but also expose them to commercial advertisement at the screen without troubling them.

This paper proposes a video object annotation platform for digitalized and bi-directed TV environment, and it is anchored to current data serialization techniques that let data to be interchangeable and reusable through a standardized data representation format like RDF (Resource Description Framework) [RDF refer]. The proposed platform support a sharing of annotations located to specified object at a moment, so that, content providers publishes additional information related to program. Then, the platform ensures that viewer of the movie or TV series can see and retrieve the information in the interactive TV.

We expect that the proposed platform will provide a novel business model for Ad market by providing seamless information transfer among consumer (viewer), advertisers (company) and mediation (broadcasting company). Contributions of this paper can be summarized as like; 1) a data schema for sharable movie annotation represented, and 2) a technical design of movie annotation platform proposed.

The rest of this paper is organized as follows. Our basic idea on movie annotation and the other related work are described in Section 2. Overall architecture of Video Object Annotation Platform (VOAP) will be described in Section 3 with two subsections, which explain details of key elements of the platform. Finally, Section 4 concludes the paper.

II. MOVIE ANNOTATION

Video annotation is tasks of associating graphical objects with moving objects on the screen. In existing interactive applications, only still images can be annotated, as in the “telestrator” system [3] used in American football broadcasting. Jeroen [4] developed an adaptive movie annotation with speech recognition techniques. The system helps people easily attach annotation by mapping scripts and a specific moment which the scripted is spoken. In the system the additional information like “who?” and “where” can be obtained from script of the program. A system proposed in [5, 6] helps easy annotating on moving objects in the scene by human. The system supports descriptive labels, illustrative sketches, thought and word bubbles communicating speech or intent, path arrows indicating motion, and hyperlinked regions.

The previous works mostly are focused on how we can easily attach annotations on the movies and how a movie is explained by annotations on divided scenes. Our work is focused on the annotations attached to appeared object at each moment in a movie and how the annotations can be shared and spent by viewers. Of course, a technology for easy and correct creation of annotations is an essential issue to be studied. We rather assume that the issues of creation of annotations could be solved through human labors if we have useful business model or motivation as we can see the Wikipedia [7] case. Therefore, our work is more focused on how the annotations can be used to enhance information viewers and content providers or to provide useful application.

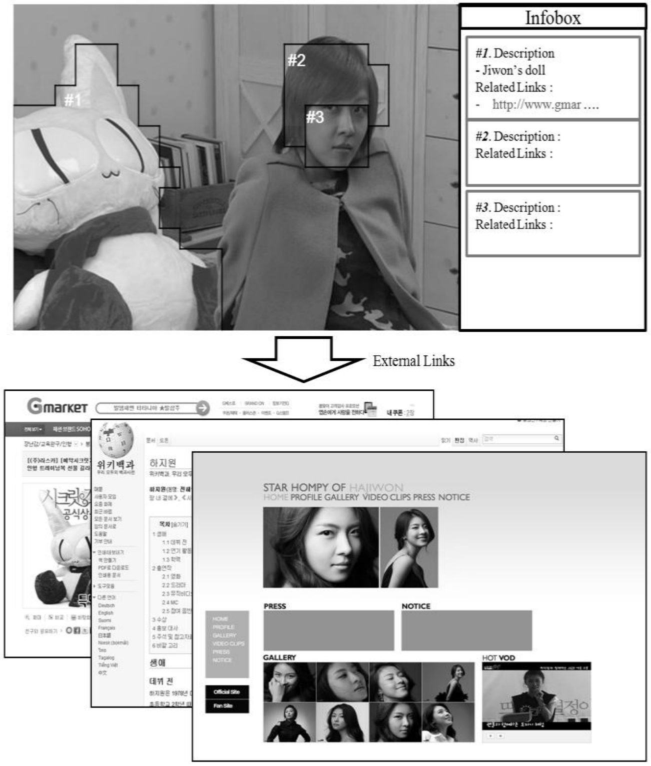


Figure 1. Examplied VOAP application running on IPTV

Figure 1 shows an example use of the attached annotation on the TV program. The annotation includes simple description about the objects we can currently see on the screen. Also, they provide external links containing useful information about the objects. By watching TV, a viewer can make a pause for a moment to have a look on additional information through the mixed annotations provided by broadcasting companies, sponsors of the program, and the other viewers. In the example figure, some questions the viewer may be interested will be seen. For example, an online market which sells the doll in the current movie scene, a name of actress appeared now, and a name of her hairstyle can be provided with useful external links.

The system proposed in this paper named VOAP (Video Object Annotation Platform) enables such exampled application. It consists of a systematic structure and a RDF schema which support representation and sharing of the video annotation data.

III. Video Object Annotation Platform

This section introduces a systematic structure of VOAP supporting share and the use of the video object annotation with semantically exchangeable data format.

A. Architecturel of VOAP

The architecture of VOAP is composed of three layers which represent three actors, namely, provider, server, and client, as shown in Figure 2.

In the provider-side, there are annotation data providers which consist of content (program) provider, economically related actors like sponsor, and ordinary viewers. In our scenario, the content provider take charge of deciding authorities for attaching annotation to a movie. Content provider distributes authorities, divided by each time seconds, to the other actors to reduce garbage data. The other actors can attach annotations to only moments of program they have an authority. Of course, there could be other way to distributes authorities, for example, Wiki [8]. However, we believe that this approach will be effective to minimize garbage annotation data and to secure rich annotation data.

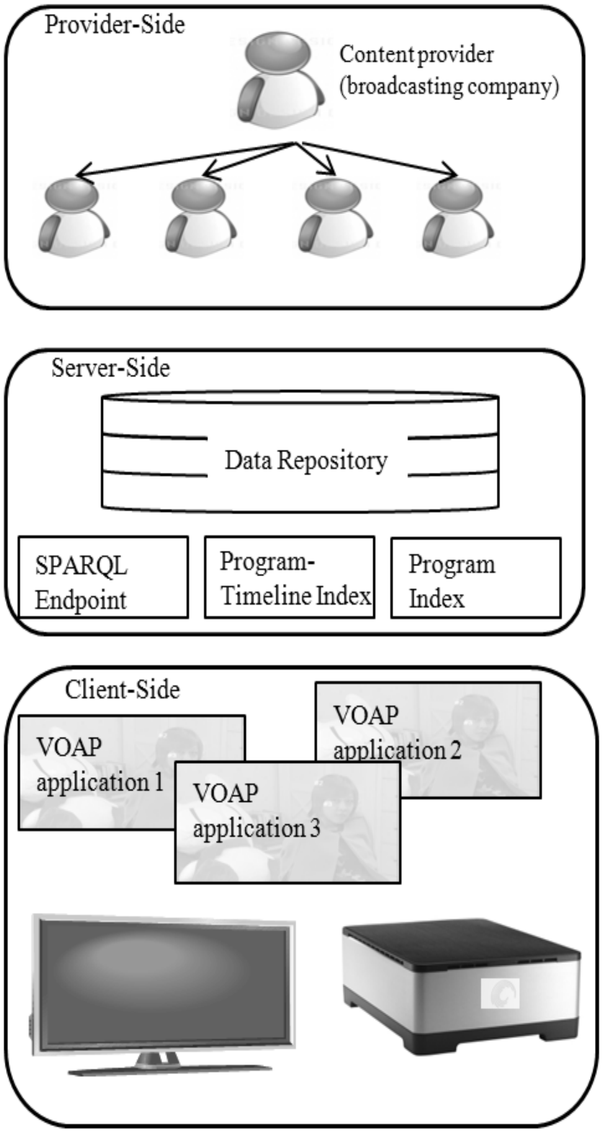


Figure 2. Artecture of VOAP

The server of the platform stores the created annotation data to be accessible to clients like applications run in the digital TVs or set-top boxes. As the annotation data are represented as RDF data formation to enhance interoperability, the server-side is equipped with a semantic data store like Jena [9]. By keeping annotation data persistent, the server has three types of access points; SPARQL, program index, program-timeline index. SPARQL provides SPARQL, as a standardized query language for semantic web data, allows a user to query on data. As proved in previous researches on benchmarks on RDF data stores, the store lacks of stable response time for large scale of data. Some of recently approaches [10] invented with massive computation power to create index data are applicable to build indexes.

The applications in clines-side provide user-oriented services with the shared annotation data as illustrated in figure1. The applications are running on the digital TV or set-top box that both are connected to the internet and are equipped with operating system for various applications.

B. *Schema for Annotation Data*

The video object annotation data are represented as RDF in VOAP. As a serialize-able data format, it promises interoperability and reuse of data among different applications. Figure 3 depicts a RDF data schema used for representing the annotation data. As the figure shows, the schema is related to the BBC Programs ontology [11] and FOAF (The Friend of a Friend) [12]. The first one is usable to universally publish program-related information, whereas, the last one is exploited to represent information about agent or person.

Instances of po:program from the BBC program ontology are linked as a program which have a annotation data, and foaf:person instances are linked to denote who created the annotation. Then, the annotation data which is an instance of voap:annotation has several type of attributes as followings.

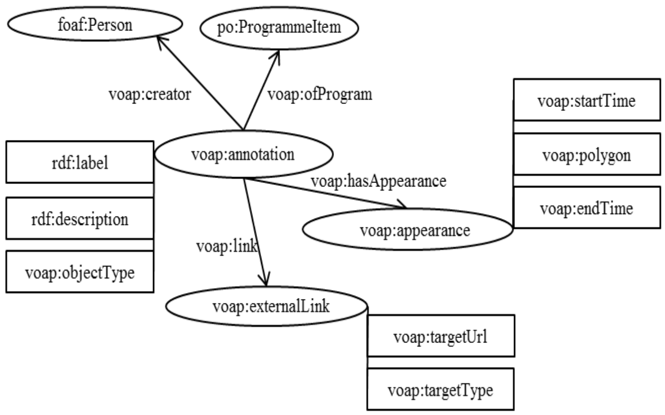


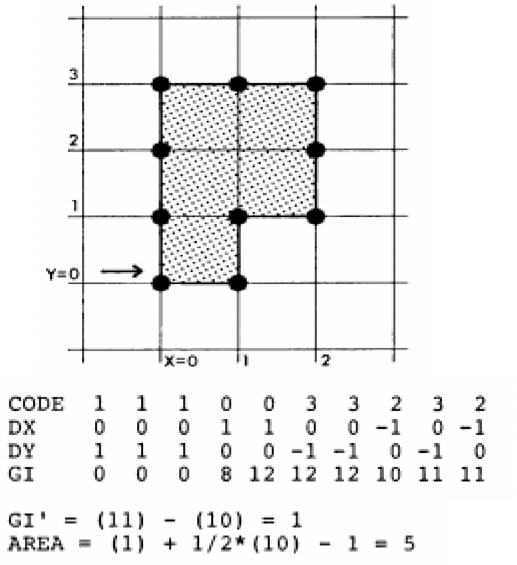
Figure 3. Ontology schema for Video-Object Annotation

TABLE I. A LIST OF ATTRIBUTES APPEARED IN FIGURE3 WHEN A CIRCLE DENOTES AN INSTANCE AND A SQUARE MEANS A ATTRIBUTE.

Attributes	Description
rdf:label	Label of the annotation
rdf:description	Description for the annotation
voap:objectType	A type of objet appeared and selected in the program. e.g., human, doll
Properties	**Description**
voap:creator	Indicates foaf:Person who created the annotation
voap:ofProgramm	Indicates po:ProgrammeItem the annotation is attached
voap:hasAppearance	Indicates an appearance of the annotation with start-end time and rendered polygon
voap:externalLink	Indicate an external link including explanation of the annotated object.

C. *Polygon rendering*

Polygon rendering is employed to designate an exact spot to be located an annotation in a scene. The computation of polygonal areas is a common operation in cartographic systems. The method of area calculation employed is dependent to some extent on the data format. A formula which was proven long before the days of computer-assisted cartography, Pick's Theorem, calculates areas of polygons whose vertices are points in a regular grid. The basic equation is AREA = GI + 1/2*GB - 1 where GI is the count of grid points inside of the polygon and GB is the count of grid points in the polygon's perimeter [13].



```
CODE  1  1  1  0  0  3  3  2  3  2
DX    0  0  0  1  1  0  0 -1  0 -1
DY    1  1  1  0  0 -1 -1  0 -1  0
GI    0  0  0  8 12 12 12 10 11 11

GI' = (11) - (10) = 1
AREA = (1) + 1/2*(10) - 1 = 5
```

Figure 4. Polygon rendering for annotating objects appeared in video

Using the polygon rendering algorithm in Figure 4, the area that includes an object annotated can be encoded when the annotation data is provided. Then, the area can be informed to the clients after decoding the rendered data.

IV. IMPLEMENTATION

To prove a concept of VOAP, we set up an annotation data repository at service-side and implemented a client-side application, so called, Interactive InforBox. The service-side data repository is equipped with Jena TDB [9] for serving queries on RDF data through SPARQL, RDF data query language. When he/she pause the VOD playing, the application invokes a data request query to the data repository with SPARQL constrained with a eclipsed time of current VOD and an URI (Uniform Resource Identifier) assigned to each VOD.

TABLE II. AN EXAMPLE SPARQL QUERY AND ANSWER FOR REQUESTING ANNOTATION DATA

```
SELECT ?annotations, ?apperance ?externalLink
WHERE
{
       ?annotations voap:ofProgram "http://www.etri.re.kr/voap/
programmItem/item1" .
       ?annotations voap:hasApperance ?appearance .
      ?annotations voap: externalLink ?externalLink .
       ?appearance voap:startTime ?startTime.
       ?appearance voap:endTime ?endTime.

 FILTER (?startTime >= "2012-05-28T07:18:11Z"^^xsd:dateTime
&&  ?endTime <= "2012-05-28T07:18:12Z"^^xsd:dateTime)
}
```

```
// Annotation
<rdf:Description
rdf:about="http://www.etri.re.kr/voap/annotation/annotation1">
        <rdf:label> Friends Season4 Episod 1 </rdf:label>
        <rdf:description> Monica is stung by a jellyfish and asks Joey
and Chandler to help her. The three...</rdf:description>
        <rdf:objectType>Information</rdf:objectType>
        <voap:externalLink
rdf:resource="http://www.etri.re.kr/voap/exLink/link1"/>
        <voap:hasApperance
rdf:resource="http://www.etri.re.kr/voap/apperance/apperance1" />
      </rdf:Description>

//Apperance
<rdf:Description
rdf:resource="http://www.etri.re.kr/voap/apperance/appearance1">
        <voap:stratTime>2012-05-28T07:18:09Z </voap:stratTime>
        <voap:endTime>2012-05-28T07:18:13Z </voap:stratEnd>
        <voap:polygon>CODE[1,1,1,0,0,3,3,2,1,7]DX[0,0,0,0,0,1,-1,-
1,1,7,]DY[0,2,0,4,0,1,-2,-1,1,7]GI[1,2,0,12,0,10,24,11,13,17]
</voap:polygon>
</rdf:Description>

//External Link
<rdf:Description
rdf:resource="http://www.etri.re.kr/voap/exLink/link1">
        <voap:targetUrl rdf:resource
="http://en.wikipedia.org/wiki/Jennifer_Aniston"/>
        <voap:targetType>WebLink</voap:targetType>
</rdf:Description>
```

Upon requested, it returns a video object annotation data to be shown over the current scene. The answer for the requested SPARQL includes a set of annotations with appearance of the annotations and external links related to the featured objects as shown in Table 2.

The Interactive InforBox application reads the annotation data and decodes rendered polygons to pop-up InforBox window. As Figure 5 shows, a television view can see the colored buttons with corresponding objects over the current scene. When the button clicked by the viewer, the screen is redirected to a web page containing related information of the object, so that, the viewer can supplied with additional information, such as, details on the object, commercial advertisements, and other related contents .

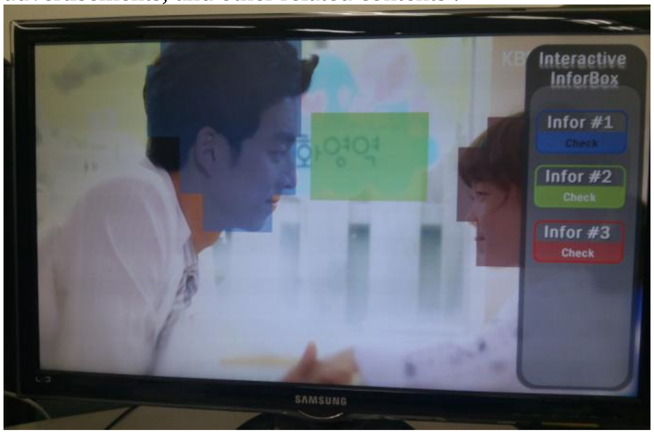


Figure 5. Imange of Intractive InforBox implementation

V. CONCLUSION

This paper introduced the VOAP supporting share and use of the video object annotation. The platform is consisted of a data model for encoding video object annotation data and the data exchange scenario. RDF based data model promises an information exchanging environment universally usable in heterogeneous interactive TV platforms. And, the data exchanging environment can open a business model by adding a new information channel to the television. As a work in progress, we implemented a user-side application proving the proposed concept.

By referring web editorial tools like Wiki, a strategy to resolve multiple annotation attached to the same object at the same moment need to be developed . Further developments on the platform will cover the applications useable in the platform for enhancing the information transfer among consumer, advertisers (company) and mediation. An annotation tool equipped with object tracing technology will be also provided to ensure easier annotating on the moving video object.

ACKNOWLEDGMENT

This work was supported by IT R&D Program of MKE/KEIT/KCC, [Development of Social TV Service Enabler based on Next Generation IPTV Infrastructure].

REFERENCES

[1] Y. Xiao, X. Du, J. Zhang, F. Hu, and S. Guizani, Internet Protocol Television (IPTV): The Killer Application for the Next-Generation Internet, Communications Magazine, IEEE , vol. 45, no. 11, pp. 126-134, November 2007.

[2] B. Schopman, D. Brickley, L. Aroyo, C. V. Aart, V. Buser, R. Siebes, L. Nixon, L. Miller, V. Malaise, M. Minno, M. Mostarda, D. Palmisano, and Y. Raimond, NoTube: making the Web part of personalised TV, In Proceedings of the WebSci10: Extending the Frontiers of Society Online Poster Track. Apr 2010.

[3] H. Jeon, Y. Shin, M. Choi, J. Rho, and M. Kim, User Adoption Model under Service Competitive Market Structure for Next-Generation Media Services, ETRI Journal, vol.33, no. 1, Feb. 2011, pp. 110-120.

[4] Wikipedia, Telestrator, Available at http://en.wikipedia.org /w/index.php?title=Telestrator&oldid=180785495, 2006, [retrieved: May 2012].

[5] J. Vendrig and M. Worring, “Interactive adaptive movie annotation”, IEEE Multimedia, vol 10, issue 3, pp. 33-37, IEEE Computer Society.

[6] D.B. Goldman, C. Gonterman, B. Curless, D. Salesin, and S. M. Seitz, Video object annotation, navigation, and composition, In Proceedings of the 21st annual ACM symposium on User interface software and technology (UIST '08). ACM, New York, NY, USA, pp. 3-12.

[7] Wikipedia The Free Encycropedia, Available at http://wikipedia.org, [retrieved: May 2012].

[8] B. Leuf and W. Cunningham. The Wiki Way: Quick Collaboration on the Web. Addison-Wesley Professional, April 2001.

[9] B. McBride, Jena: a semantic Web toolkit, Internet Computing, IEEE, vol.6, no.6, pp. 55-59, Nov/Dec 2002.

[10] S. Yun, M. Song, H. Namgung, S. Yang, H. Kumar, and H. Kim, Improving the workflow of semantic portals using M/R in cloud platform, In proceedings of KEOD 2010-International Conference on Knowledge Engineering and Ontology Development, 2010, pp. 485-488.

[11] BBC Programmes Ontology, Avalable at http://www. bbc.co.uk/ontologies/programmes/2009-09-07.shtml, [retrieved: May 2012].

[12] M. Graves, A. Constabaris, and D. Brickley, FOAF: Connecting People on the semantic Web. Cataloging & Classification Quarterly, 43(3/4), 2007, pp. 191-202.

[13] H.S.M. Coxeter, Introduction to Geometry, pp. 208-210, John Wiley and Sons, Inc., New York, 1961.

Performance Measurement for CEP Systems

Alexander Wahl and Bernhard Hollunder
Department of Computer Science
Furtwangen University of Applied Science
Robert-Gerwing-Platz 1, D-78120 Furtwangen, Germany
alexander.wahl@hs-furtwangen.de, bernhard.hollunder@hs-furtwangen.de

***Abstract*—Today, Complex Event Processing (CEP) is often used in combination with service-oriented architectures. Several CEP products from different vendors are available, each of them with its own characteristics and behaviors. In this paper we introduce a concept that is able to compare different CEP products in an automated manner. We achieve that by using web service technology. We show how to build a testing environment that includes i) an event emitting component with stable interface, ii) an interchangeable CEP component based on this interface and iii) a measurement and evaluation component. The presented concept is capable to perform different test scenarios fully automated. In this paper three exemplary tests are performed on three selected CEP products.**

***Keywords*-complex event processing; web services; testing architecture; performance testing.**

I. Introduction

Processing of events is a prevalent necessity in todays information systems. Collected events of any kind are detected automatically and systematically analyzed, combined, rated and processed using Complex Event Processing (CEP) [1] systems. The application of CEP in information technology ranges from sensor networks to operational applications, like business activity monitoring (BAM), algorithmic trading systems and service oriented architectures (SOA).

But, what is CEP? According to the Event Processing Glossary, CEP is “Computing that performs operations on complex events, including reading, creating, transforming, abstracting, or discarding them” [2]. In more detail, the two main components of a CEP system are a set of CEP rules and a CEP engine. A CEP rule is a prescribed method for processing events that reflects a certain condition. An event thereby is a kind of indicator to something that happened, like e.g., a financial trade, a key stroke or a method call within an application. By correlations of these events a CEP rule evaluates if a certain condition is satisfied. If a condition is satisfied, the rule fires.

One substantial feature of CEP systems is the realtime processing of events. And in general, the faster the processing of events is performed by the system the better. A result that is detected more rapidly may be a valuable benefit. Today, numerous CEP implementations of different vendors are available. They are widely varying in the supported platforms, the scope of operations, the performance of processing, rule languages, etc.

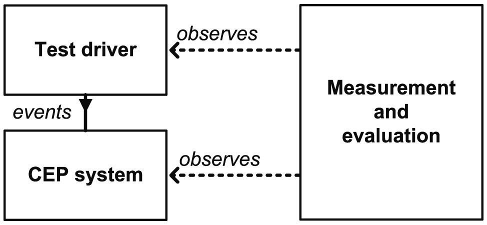


Figure 1. Main components of the concept

In this work, we will introduce a concept to measure the performance of different CEP systems in an automated manner. We thereby focus on i) interchangeability of CEP engines, ii) reproducibility and comparability of the test scenarios, and iii) the automation of measurement without user interaction. Our solution consists of three components, which are shown in Figure 1: i) a test driver component that generates reproducible sets of events, ii) the encapsulated, interchangeable CEP system component, and iii) a measurement evaluation component to calculate measurement results. Based on this solution, three exemplary test kinds are performed: latency test, pollution test and load test.

This paper is structured as follows: After a brief introduction in this section, Section II gives an overview on related work. In Section III, the requirements of the concept will be defined, followed by the description of the concept itself in Section IV. Test candidates and test scenarios are described in Section V. A detailed description of the test setup is provided by Section VI. The test results are shown in Section VII. In the final section, we will conclude the paper.

II. Related work

In the context of event processing systems, there are some frequently stressed benchmarks: Linear Road benchmark [3], NEXMark [4], BiCEP [5] and SPECjms2007 [6]. But, there is no general accepted benchmark for CEP systems.

White et al. [7] described in their work a performance study for the WebLogic event server (now Oracle CEP). In that work, solely latency testing is performed on a single CEP system.

Kounev and Sach [6] provide an overview of techniques for benchmarking and performance modeling of event-based systems. They introduce a benchmark, SPECjms2007, to measure the performance of message-oriented middleware.

In 2007, Bizarro introduced BiCEP [5], a project to benchmark CEP systems. His main goal was to identify core CEP requirements and to develop a set of synthetic benchmarks. In the following years Mendes, Bizarro and Marques built FINCoS, a framework that provides a flexible and neutral approach for testing CEP systems [8]. FINCoS introduces particular adapters to achieve a neutral event representation for various CEP systems. In this work, we wrap the whole CEP system using Web service (WS) technology. In FINCoS, a controller is defined acting as an interface between the framework and the user. Our concept proposes to perform tests automatically and without user interaction.

In a further publication, Mendes et al. [9] use their framework to perform different performance tests on three CEP engines - Esper and two developer versions of not further specified commercial products. They run "micro-benchmarks" while they vary query parameters like window size, windows expiration type, predicate selectivity, and data values. So, they focus on the impact of variations of CEP rules. Beside they perform some kind of load test. The results thereby showed a similar behavior in memory consumption like Esper did in our tests. In summary, Mendes et al. did run performance tests, but their focus is different to ours.

In 2010, Fraunhofer IAO published a comprehensive survey of event processing tools available in the market [10]. It provides extensive information on supported platforms, licensing and features. This study lists several event processing tools, but performance considerations are out of scope.

III. Requirements on the concept

Comparing sets of implementations from different vendors are common tasks. Before a test scenario may be defined it is essential to specify the requirements towards it.

A first requirement is to ensure, that equal sets of events are emitted on every repeted test execution. Ideally, the component that creates events stays the same in all scenarios in order to eliminate any influence on the measurement itself. Next, test scenarios and measurements must be reproducible.

The aim is to create test scenarios that differ solely in the CEP engine used. The solution therefore requires to support the exchange of the CEP engine. However, some CEP engines are restricted to certain platforms. Therefore, it is desirable to keep event generating component and the CEP component separated (following the principle of loose coupling). At best the two components are able to run on different nodes.

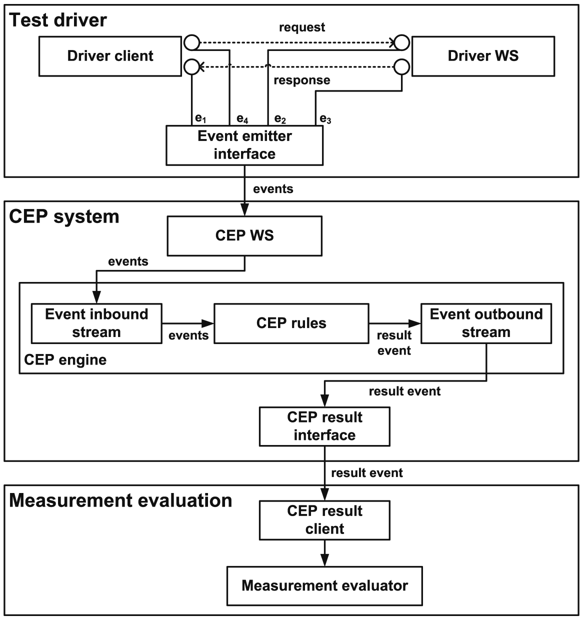


Figure 2. Detailed overview on the components of the concept

The concept requires to perform different test cases. For this work it has to run three exemplary tests, namely latency tests, pollution tests and load tests.

A major requirement is that the test cases need to run automated without user interaction, e.g., by batch script or cron job.

Next, the obtained measurements of the tests need to be analyzed and meaningful statistics have to be generated and presented to the user. In order to ensure comparability of results and statistics the corresponding component is not to be modified at all.

IV. Concept

The requirements suggest the usage of three main components (see Figure 1). One is an interchangeable component wrapping the CEP system. The two others are the test driver component and the measurement evaluation component. Between test driver component and CEP system component, as well as between CEP system component and measurement evaluation component, stable interfaces are introduced. Figure 2 shows the components in more detail.

A. Test driver component

In Figure 2, the uppermost component is the test driver. The responsibility of this component is to produce and emit events. This component reflects the application scenario. We use a WS and a corresponding WS client to reflect a SOA environment as application scenario. In Figure 2, WS and WS client are represented by Driver WS and Driver client. The communication between Driver WS and Driver client follows a request-response pattern based on messages (further named test driver messages) using the Simple Object

Access Protocol (SOAP) [11]. The event emitter, in brief, is linked to the communication path of driver WS and client. It analyzes the interchanged SOAP messages and emits events based on the gathered information. The event emitter itself is implemented as a WS client - as we will see in Section IV-B.

A requirement of the event emitter component is to generate reproducible sets of events. A set of events thereby has two aspects. The first aspect concerns the chronological order, which means that sequence of events and the timespans between these events stay exactly the same at any time a specific set of events is applied. The second aspect is that, using different types of events, the sequence of emitted event types also stays the same.

As we mentioned before, the test driver component generates sets of events. In our application scenario four events, e_1, e_2, e_3 and e_4, are emitted each time test driver messages are exchanged. Suppose that these four events form a set. How does the CEP engine determine which events relate to each other? We, therefore, introduce an ID for each exchange of test driver messages - and thereby also for each set of events. This ID becomes a parameter of the emitted events. With that the CEP engine can determine which events relate to a certain set and afterwards processes these events.

B. CEP engine component

The CEP engine component is formed by four main parts: a CEP engine, CEP rules, CEP WS and CEP result interface. The CEP engine processes incoming events based on CEP rules. The CEP rules and the CEP engine are the core of the CEP system component. CEP rules reflect conditions that are to be satisfied by the system, e.g., the occurence of subsequent events within a certain timespan. Each CEP rule has at least one event inboud stream and fires a result event via an output stream once the condition is satisfied. For the result event a complex event is typically used, which means that it is a more abstract event. A complex event thereby contains a set of incoming events. In our case, a complex event $e_5(e_1, e_2, e_3, e_4)$ is generated.

But, what if more than one driver message exchange is performed? Or in other words: What if more than one set of event is emitted? How can the CEP engine correlate the events of a set? Therefore, the ID parameters of the events are needed. The CEP system is advised to correlates only event with the same ID.

The evaluation of CEP formulas essentially requires events to be passed to the CEP engines input stream. This input stream is exposed to the event emitter component. In more detail, CEP WS defines a unique interface for all CEP systems that accepts events of any structure and routes it to the event inbound stream. The second interface of the CEP system deals with the result events of the CEP system. Next to the event outbound stream is the CEP result interface. A component that is interested in the result events can subscribe itself here. The mechanism behind follows the observer pattern: The CEP result client subscribes at the CEP result interface. Every time a result event occurs, the CEP result client is notified and is provided with the CEP result event. Again, by this mechanism a stable interface, similar to the CEP WS, is created between the CEP system component and the measurement evaluation component. CEP WS and CEP result interface together encapsulate the CEP system and ensure interchangeability.

Interchangeability of the CEP engine component is achieved by defining two stable interfaces: i) Event emitter and CEP WS, and ii) CEP result interface and CEP result client. By the definition of such stable interfaces, and by using WS technology, we achieve the following: i) the CEP component can be designed considering these interfaces, ii) interoperability of the components is achieved by using WS, and iii) no adjustments in code is needed if a CEP component is exchanged. To clarify the latter: Consider the transition from Test driver component to CEP system component. Using different CEP systems each of them can define, for example, its own type of event inbound stream. Therefore adapters between event emitter and event inbound stream are necessary. Once the CEP system is exchanged the event emitter has to use the corresponding adapter, which required to modify the event emitter. By introducing a stable interface the event emitter can use this interface and does not need any changes at all. There is still an adapter needed, but by that means it can be hidden within the CEP system. The adapter then just requires to implement the stable interface. The same applies at the transition between CEP system component and measurement and evaluation component.

C. Measurement evaluation component

The aim of the measurement evaluation component is to perform the calculation necessary to analyze the performance of the individual CEP engine components. Its first component is the CEP result client - a client for the event outbound WS of the CEP engine component. By that client, the component receives the result events that include all the information needed for the performance measurement. Based on these information from the result events, the measurement evaluator determines the performance of a CEP engine component.

D. Automated measurement

As described before, the proposed concept provides three main components. The first component represents the application scenario and acts as the test driver. In our case it consist of an application server including a deployed Driver WS. Automated startup of an application server is a commonly performed task. A typical WS client can be seen as an application, which again can be started easily in an automated manner, e.g., by batch script.

We encapsulated the complete CEP system using WS technology. Thereby, an automated interchange of CEP systems is equivalent to deploying and undeploying the CEP WS of the corresponding CEP system. Again, for an application server this is a commonly performed task and can be automated.

Another option is the usage of several dedicated machines respectively virtual machines (VMs): One for Test driver component and measurement evaluation component, and one per CEP system. Automation of measurement in that case is reduced to start and stop the machines.

V. Test candidates and test scenarios

A decision on the usage of a CEP Engine of a specific vendor rises some interesting questions. For example: If events are emitted to the CEP engine, how long does it take to process the events and to generate a result event? What happens if events that are irrelevant for the CEP rules are emitted? What is the impact on the result calculation time? How does the CEP engine behave on certain event loads? How complex are the rule sets that are necessary to describe a condition?

A. Test candidates

There are several CEP engines available from different vendors. In this work we choose three CEP engines for testing: i) Microsoft SQL Server StreamInsight 1.1 [12], ii) EsperTech Esper 4.3 [13], and iii) JBoss Drools Fusion 5.3.0 [14]. A first difference arises while comparing the rule sets, as we will show in the following paragraphs. We will use an example that fits to the application scenario described before (WS and client). The task is to collect four related events of the driver message exchange. Once all four events are detected a complex event, the CEP result event, is created.

1) StreamInsight: Microsoft StreamInsight is based on the Microsoft SQL Server and has a high-throughput stream processing architecture. It is equipped with several adapters for input and output event streams. LINQ is used as query language to specify the CEP rules. The corresponding CEP rule is as follows:

```
var filtered = from e in inputstream
  group e by e.ProcessId into con
  from item in con.HoppingWindow(
    TimeSpan.FromSeconds(180),
    TimeSpan.FromSeconds(3),
    WindowInputPolicy.ClipToWindow,
    HoppingWindowOutputPolicy.ClipToWindowEnd)
    select new TimeAggregation()
    {
      Id = con.Key,
      Value = (int)item.Count(),
      TimeStart = item.Min(t=>t.Timestamp),
      TimeEnd = item.Max(t=>t.Timestamp),
    };
var filteredFromValue = from e in filtered
                        where e.value>=4
                        select e;
```

2) Esper: EsperTech Esper is an open-source CEP engine available under GNU General Public License (GPL) license. It is available for Java as Esper and for .NET as NEsper. It offers the Event Processing Language (EPL) to specify CEP formulas. The CEP rule is defined by

```
String expression =
  ``insert into CEP.LatencyEvent'' +
  ``select one.timestamp,'' +
         ``two.timestamp,''+
         ``three.timestamp,''+
         ``four.timestamp''+
  ``from CEP.inputStream1.win::time(180) as one,'' +
       ``CEP.inputStream2.win::time(180) as two,'' +
       ``CEP.inputStream3.win::time(180) as three,'' +
       ``CEP.inputStream4.win::time(180) as four'' +
  ``where four.id = two.id and `` +
        ``three.if = one.id'';
statement =
  epService.getEPAdministrator().createEPL(expression);
```

3) Drools: Fusion is a module for JBoss Drools to enable CEP. Like Esper, it is open-source software. It can run Java and .NET and supports a variety of input adapters. The Drools Rule Language (DRL) is used to specify CEP formulas. The CEP rule is

```
rule''cep-formula''
when
  $pay:InternalPayload($ide:identifier,
                       $step:step,
                       step>=4)
       from entry-point StoreOne
  $minTime:Number()
  fromaccumulate(InternalPayload(identifier==$ide,
                                 step==1,
                                 $id:identifier,
                                 $time:timestamp)
                 from entry-point StoreOne,
                                  init(long tm=0;),
                                  action(tm=$time;),
                                  result(tm))
then
  System.out.println(``step is: ``+$ide+
                     '' min ``+$minTime'');
End
```

B. Test scenarios

1) Latency test: The following setup applies to our latency test setup: driver WS and its client both emit two events - one for inbound SOAP message and one for outbound SOAP message. All four messages include the timestamp of their creation, and are passed to the CEP component. Within the CEP engine, the four events are correlated and combined to a complex event, the result event. Once the result event occurs at the CEP result client, the measurement evaluator calculates the overall latency and estimates the CEP component latency.

2) Pollution test: In a pollution test setup, the CEP engine is confronted with events that are relevant for a CEP rule and events that are irrelevant. In the pollution test scenario a constant rate of events is emitted to the CEP component. During the test the rate of irrelevant events is increased step by step. The aim of the CEP component is to filter out the relevant events. The observed parameters were changes in latency, CPU load and memory usage.

3) Load test: A load test in brief applies different event rates to the CEP component. The test starts with one driver WS and client pair. During the test the amount of driver WS and client pairs is continuously increased. The observed parameters were latency, CPU load and memory usage.

VI. Test setup

The testing was performed using different virtual machines (VM). As virtualization product Oracle VirtualBox [15] was used. As an alternative several dedicated machines can be used. The VMs were interconnected using a virtual network. The operating systems of VM1 and VM2 was Microsoft Windows 7 with Service Pack 1. VM1 includes the event emitter component and the measurement component. For the implementation of the event emitter component WS technology was used. Communication between event emitter and CEP WS was based on SOAP messages.

The event emitter were implemented as message inspectors. Each time a SOAP message passes the message inspector its content is analyzed and an event is emitted. Each event thereby includes a timestamp and an unique ID, as described before. In our case emitting such an event means that an event is embedded in a SOAP message and is sent to the CEP WS of the CEP system component.

The measurement component is part of VM1, which reduces the number of VMs needed. In consistence with the driver WS and its client the measurement component also uses WS technology. The measurement and evaluation results are provided in database tables.

Within the second VM (VM2) the CEP system component is implemented. Several instances of VM2 exist, one for each tested CEP system. The individual instances of VM2 are completely interchangeable and can potentially even run in parallel - with slight modifications at the event emitter and separate IP addresses for each instance.

The technology used to implement the event inbound WS depends on the CEP engine used. In our work either WCF technology or JAX-WS is used. For interoperability reasons the binding type 'basicHttpBinding' (following the WS-I Basic Profile [16]) was used for all WS.

VII. Test results

As described before, we performed three exemplary tests using three different CEP systems. For completeness the results are displayed in this section.

A. Latency test

Comparing the latency all of the three CEP systems showed similar result. As the main source for fluctuation the driver message exchange was identified. The CEP system latency showed no significant differences.

CPU usage of B and C was similar. A, in contrast, has a significantly higher CPU consumption. In terms of memory usage A and B showed a modest behavior compared to C.

Table I
Pollution test results

CEP system	Pollution	Latency	CPU	Memory
A	10events	2.5ms	86%	27M
	25events	1.5ms	81%	28.5M
	50events	1.7ms	83%	28.4M
	100events	2.3ms	89%	29.6M
	200events	1.9ms	92%	33.9M
B	10events	0.1ms	24%	85.1M
	25events	1ms	25%	88.5M
	50events	1ms	23%	89.3M
	100events	1ms	24%	97.2M
	200events	3ms	24%	93.4M
C	10events	0.1ms	41%	122M
	25events	1ms	38%	160M
	50events	1ms	50%	164M
	100events	1ms	73%	165M
	200events	1ms	79%	165M

Table II
Load test results

CEP system	Threads	Events/sec	CPU	Memory
A	1	8	2%	18.1M
	5	40	12%	18.2M
	10	80	16%	18.1M
	20	128	30%	18.4M
	30	128	20%	18.4M
B	1	8	1%	24M
	5	40	4%	30M
	10	80	7%	33.4M
	20	128	10%	42M
	30	128	10%	43.2M
C	1	8	1%	65M
	5	40	4%	67M
	10	80	5%	113M
	20	128	9%	120M
	30	128	10%	126M

B. Pollution test

The pollution test showed interesting differences between the CEP systems (see Table I). Regarding the latency change A showed an overall pretty constant latency. However, in comparison with B and C the overall latency was significantly higher. At the beginning B and C were almost equal - with slight advantages for C. But at higher pollution rate the latency of B highly increased, in contrast to C with a constant latency.

Concerning the average CPU usage, A showed a slight increase with increasing pollution rate. Its overall CPU usage was higher that with B and C. For B the average CPU usage did not change at any pollution rate. With C an increase in pollution rate resulted in a significant increase of average CPU usage. A had a low memory usage at any pollution rate. With B memory usage, again, increased with pollution rate. C has a constant memory usage for all pollution rates, but at high level.

C. Load test

Table II shows the results of the load test. We could not identify significant differences to the results of the latency test applying load to the test candidates. We will therefore reinvestigate on load tests in more detail in future work.

For all candidates, the applied load and the CPU usage correlated. The same applies to the memory usage with B and C. For A the memory consumption was almost constant.

VIII. Conclusion

Today, CEP is commonly used and several products from different vendors are available in the market. This paper described a platform to compare different products under controlled and equal conditions. The test thereby were performed in an automated manner, which means that no user interaction was necessary.

We described that a complete CEP system can be encapsulated using WS technology. Thereby interchangeability of CEP systems can be achieved, which further simplifies for automation of performance tests. The paper also reproduced characteristics concerning performance and resource consuption already described elsewehere.

In summary, CEP systems can be encapsulated as a whole by means of WS technology. Thereby, interchangeability of complete CEP systems can be achieved. With interchangeability of CEP systems fully automated performance testing is available.

Acknowledgment

We would like to thank the anonymous reviewers for giving us helpful comments.

This work has been partly supported by the German Ministry of Education and Research (BMBF) under research contract 17N0709.

References

[1] D. Luckham and B. Frasca, “Complex event processing in distributed systems,” Stanford University, USA, Tech. Rep. CSL-TR-98-754, Mar. 1998. [Online]. Available: http://citeseer.nj.nec.com/luckham98complex.html

[2] D. C. Luckham and R. Schulte, “Event Processing Glossary – Version 2.0,” 2011. [Online]. Available: http://www.complexevents.com/2011/08/23/event-processing-glossary-version-2-0/, last accessed: May 13, 2012.

[3] A. Arasu, M. Cherniack, E. Galvez, D. Maier, A. S. Maskey, E. Ryvkina, M. Stonebraker, and R. Tibbetts, “Linear road: a stream data management benchmark,” in *Proceedings of the Thirtieth international conference on Very large data bases - Volume 30*, ser. VLDB ’04. VLDB Endowment, 2004, pp. 480–491. [Online]. Available: http://dl.acm.org/citation.cfm?id=1316689.1316732, last accessed: May 13, 2012.

[4] J. Li, J. Maier, V. Papadimos, P. Tucker, and K. Tufte, “NEXMark Benchmark.” [Online]. Available: http://datalab.cs.pdx.edu/niagara/NEXMark/, last accessed: May 13, 2012.

[5] P. Bizarro, “BiCEP - Benchmarking Complex Event Processing Systems,” in *Event Processing*, ser. Dagstuhl Seminar Proceedings, Mani Chandy, Opher Etzion, and Rainer von Ammon, Eds. Dagstuhl and Germany: Internationales Begegnungs- und Forschungszentrum fuer Informatik (IBFI), Schloss Dagstuhl, Germany, 2007. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2007/1143, last accessed: May 13, 2012.

[6] S. Kounev and K. Sachs, “Benchmarking and Performance Modeling of Event-Based Systems Modellierung und Bewertung von Ereignis-basierten Systemen: It - Information Technology,” *it - Information Technology*, vol. 51, no. 5, pp. 262–269, 2009.

[7] S. White, A. Alves, and D. Rorke, “WebLogic event server: a lightweight, modular application server for event processing,” in *Proceedings of the second international conference on Distributed event-based systems*, ser. DEBS ’08. New York and NY and USA: ACM, 2008, pp. 193–200.

[8] M. R. N. Mendes, P. Bizarro, and P. Marques, “A framework for performance evaluation of complex event processing systems,” in *Proceedings of the second international conference on Distributed event-based systems*, ser. DEBS ’08. New York and NY and USA: ACM, 2008, pp. 313–316.

[9] M. Mendes, P. Bizarro, and P. Marques, “A Performance Study of Event Processing Systems,” in *Performance Evaluation and Benchmarking*, ser. Lecture Notes in Computer Science, R. Nambiar and M. Poess, Eds. Springer Berlin / Heidelberg, 2009, vol. 5895, pp. 221–236.

[10] K. Vidackovic, T. Renner, and S. Rex, *Marktuebersicht Real-Time-Monitoring-Software Event-Processing-Tools im Ueberblick*. Stuttgart: Fraunhofer-Verlag, 2010.

[11] W3C, “SOAP Version 1.2,” 2007. [Online]. Available: http://www.w3.org/TR/soap/, last accessed: May 13, 2012.

[12] Microsoft, “Microsoft StreamInsight.” [Online]. Available: http://www.microsoft.com/sqlserver/en/us/solutions-technologies/business-intelligence/complex-event-processing.aspx, last accessed: May 13, 2012.

[13] EsperTech, “Esper - Complex Event Processing.” [Online]. Available: http://esper.codehaus.org/, last accessed: May 13, 2012.

[14] JBoss, “Drools Fusion.” [Online]. Available: http://www.jboss.org/drools/drools-fusion.html, last accessed: May 13, 2012.

[15] Oracle, “VirtualBox.” [Online]. Available: https://www.virtualbox.org/, last accessed: May 13, 2012.

[16] Web Service Interoperability Organization, “Basic Profile Version 1.2.” [Online]. Available: http://www.ws-i.org/Profiles/BasicProfile-1.2.html, last accessed: May 13, 2012.

Using Machine Learning in Web Service Composition

Valeriu Todica, Mircea-Florin Vaida, Marcel Cremene
Technical University of Cluj-Napoca, Cluj-Napoca, Romania
valeriu.todica@gmail.com, Mircea.Vaida@com.utcluj.ro, cremene@com.utcluj.ro

***Abstract*—Web service composition is the process of aggregating a set of existing web services in order to create new functionality. Current approaches to automatic web service composition are based, in general, on various Artificial Intelligence (AI) techniques, in particular search algorithms. Automated planning is one of the most frequently used. A limitation of these approaches is that they do not involve learning from previous attempts in order to improve the planning process. A new approach for automatic web service composition, based on Reinforcement Learning, is proposed. The method is suited for problems that do not define a particular goal to be reached but a reward that has to be maximized.**

Keywords - Web Services; Reinforcement Learning.

I. Introduction

The process of composing web services according to a workflow enables business-to-business and enterprise application integration [1]. Automatic web service composition and execution is a multidisciplinary issue, involving three important domains: Service Oriented Architecture (SOA), Semantic Web (SW), and Artificial Intelligence (AI).

Service Oriented Architecture represents a style of building distributed systems that deliver functionality as services, with the additional emphasis on loose coupling between interacting services. Web services technology is a popular choice for implementing SOA [1].

Semantic Web [2] is a vision for a web that may be used by humans and also by computers. SW has at its core the concept of ontology, defined as a “formal representation of a set of concepts within a domain and the relationships between those concepts”. Semantic Web services are combining the versatility of web services with the power of semantic based technologies. A semantic web service processes data that is described using terms formally defined in common ontology. Semantic web services offer new possibilities such as: automatic discovery, automatic invocation, automatic composition and automatic monitoring.

The field of Artificial Intelligence has a long history. At the beginning, the expectations from AI were very high. Even if these expectations were not fulfilled, AI offers valuable techniques for knowledge representation and problem solving. The semantic web vision promotes the concept of software agent provided with “intelligence”. A software agent is usually considered to be a complex software entity capable of a certain degree of autonomy in order to accomplish tasks on behalf of human or other software agents. A semantic web agent is capable of accessing, processing and managing web resources and web services. A classification of software agents based on their abilities to cope with the environment is proposed in [3]:

- **Simple reflex agents**. A simple reflex agent is acting based on what it perceives from the environment and ignores the past.
- **Model-based reflex agents**. The agent maintains a time-based environmental model and acts based on the current situation taking into account the old model of the environment.
- **Goal-based agents.** In this category are included agents driven by a particular goal. The goal-based agent uses an environment model. The actions to be taken in order to achieve the goal are automatically discovered.
- **Utility-based agents.** A utility function is a more general performance measurement than a goal. The utility-based agent tries to maximize one or more utility functions that are considered known.
- **Learning agents.** A learning agent is superior to the agents presented before because it can operate also in an unknown environment. Such an agent is able to learn from its actions and becomes more competent by learning.

Learning is one of the most important feature of a human being. Machine learning tries to employ this mechanism by offering various techniques. Such techniques can be used to develop software agents capable of learning, in order to better interact with their environment.

There is a wide range of applications that can be addressed using web services composition. In many cases, the problem is specified by a particular goal that needs to be reached. For instance, a semantic web-enabled software agent may receive a request for buying a particular product. In this case, the agent will try to produce a workflow containing services for searching products and also services for secure payment. A solution based on classical search algorithms or AI Planning is suited for this scenario. Our previous work in this area can be found in [4].

However, in many other scenarios a concrete goal cannot be explicitly specified. Instead, it is possible to compute a reward/penalty (from the environment) depending on the actions performed by the agent. In these cases, the objective of the agent will be to maximize the total reward. Such scenarios may be addressed by reinforcement learning techniques.

The rest of this paper is organized as follows: Section 2 presents a scenario addressed using machine learning techniques, Section 3 provides an introduction to reinforcement learning and related methods. Section 4 presents a composition technique based on reinforcement learning and semantic web technologies. The proposed composition system architecture and the model for the semantic web services are described also in this section. Section 5 presents some related work and Section 6 concludes the paper.

II. A MOTIVATING EXAMPLE

Consider a software system responsible for generating advertisement information for various users accessing an e-commerce Web site. The system is based on web services. Each web service is associated with a particular category of advertisement. The system is responsible for composing a set of web services into a business process. A problem is how to select particular web services, in order to offer relevant advertisement information for the user. Such a problem is not suited to be addresses using AI Planning techniques, since the goal of the problem cannot be easily specified in a formal manner.

When a user enters the web site for the first time, the system is not aware about his interests. In this case, the system will create a random list of web services and will show the generated advertisement to the user. If the user access some parts of the advertisement information, the related web services will be associated with a positive reward. The next time the user enters the web site, the system will select with a higher probability the services associated with the positive reward. Each web service will be associated with a particular probability for each user. In the initial case all web services have the same probability. These probabilities will be updated during the service usage depending on the user actions. Thus, the system will learn to provide the user with more relevant advertisement information.

This example may be seen as a particular case of the *n-armed bandit problem* [5]. In the Probability Theory, the n-armed bandit problem is the problem in which one is faced repeatedly with a choice among *n* different options. Each option is associated with a reward. The reward for a particular choice is known only after that choice is selected. A probabilistic reward is also possible. The objective of the problem is to maximize the expected total reward over a period of time. The solution to this problem should respect a tradeoff between exploration and exploitation. Exploration means that new options are selected during the service usage, while exploitation is a greedy method of selecting the best options already tried.

III. REINFORCEMENT LEARNING

Reinforcement learning is informally defined as a learning method that tries to maximize a kind of reward. The reward is usually a numerical value and it is application specific. A software agent based on reinforcement learning has no rules to tell him what to do. Instead, it must discover what actions yield to a maximum reward. Reinforcement learning do not characterizes some learning methods but a learning problem [6].

A reinforcement learning system is characterized by three elements [6]: a) a policy, b) a reward function and c) a value function. The policy defines the behavior of the agent. The policy can be seen as a relation between the environment state and the actions that can be taken. For simple problems the policy is just a lookup table. The reward function defines the goal of the learning agent. The reward function maps the state of the environment to a reward, indicating the desirability of that state. The objective of the learning agent is to maximize the total reward. The value function specifies what is “good” for the agent considering a long term, while the reward indicates the short term desirability. The rewards are directly given by the environment while the value functions must be in general estimated by the agent.

A reinforcement learning model may be formally defined as a Markov Decision Process (MDP) [6]:

- A finite set of environment states S;
- A finite set of actions A;
- The probability that action a in state s at time t will lead to state s'at time $t+1$:

$$P_a(s,s') = Pr(s_{t+1} = s' | s_t = s, a_t = a) \qquad (1)$$

- The expected reward associated with the transition from the state s to the state s' with the transition probability $P_a(s,s')$:

$$R_a(s,s') = E(r_{t+1} | s_t = s, a_t = a, s_{t+1} = s') \qquad (2)$$

In MDP, the agent, also called the decision maker, interacts with the environment at each of a sequence of discrete time steps, $t = 0,1,2,\ldots,n$. At each time t the agent perceives the state of the environment $s_t \in S$. The agent also receives a numerical reward, $r_{t+1} \in R$ when it enters a new state s_{t+1}. MDP allows to model uncertainty in a sense that the actions can be stochastic.

The agent implements a policy, a mapping from states to actions, called π_t. For instance, $\pi_t(s,a)$ is the probability that $a_t = a$ if $s_t = s$.

The reinforcement learning problem is to find an optimal policy that will maximize the expected return. The expected return is defined as some specific function of the reward sequence. Typically, the expected return is defined as a discounted sum over the individual rewards:

$$\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1}) \qquad (3)$$

ISBN: 978-1-61208-215-8

The discount factor γ satisfies $0 \leq \gamma \leq 1$ and the action a_t is given by the policy π.

Reinforcement learning was successfully employed in various applications such as: elevator scheduling, inverted cart-pole, robot control, game playing, and others. Reinforcement learning was particularly successful in the game of Backgammon [7] for instance.

IV. PROPOSED COMPOSITION METHOD

A. *System architecture*

The proposed system architecture is depicted in Figure 1. The system is based on a closed-loop mechanism. The closed-loop, also known as feedback loop, allows the composer component to use the feedback generated by the executor component.

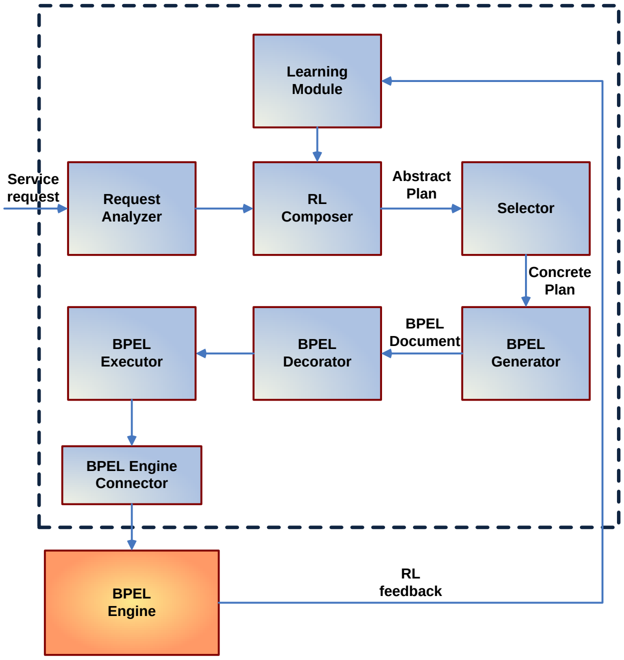


Figure 1. Proposed system architecture.

The proposed architecture is based on the architectural pattern called "*pipes and filters*". The composition request will usually indicate to the system what web services have to be used and what reward should be considered. The architecture modules are described further.

Request Analyzer component verifies if the composition request is valid. *RL Composer* component receives the composition request from *Request Analyzer* and tries to aggregate a workflow containing web services based on the functional requirements, defined by the request. This workflow is abstract (it contains abstract web services).

Selector component is responsible with the concrete services selection.

RL Composer uses the *Learning Module* component to take into account the feedback from previous executions of the generated workflow. *BPEL Executor* component is responsible for executing the workflow and for generating the feedback. The feedback is used by the *Learning Module* to acquire new information about the problem.

The composition system uses *BPEL* (*Business Process Execution Language*) [8] for representing workflows based on Web services. *BPEL Generator* component is responsible for generating the workflow BPEL description. *BPEL Executor* component is responsible for executing the BPEL workflow and it uses the *BPEL Engine Connector* component representing the interface to a *BPEL Engine*, which is a software container designed to run BPEL processes.

BPEL Decorator component is responsible for modifying the BPEL workflow by inserting calls to a web service exposed by *Learning Module* component. This web service, called *RegisterRewardService*, is invoked after the invocation of a web service associated with a reward. This mechanism for collecting the reward is presented in Figure 2. Considering a workflow containing three web services, the *BPEL Decorator* will insert two calls to *RegisterReward-Service*, one for each Web service associated with a reward.

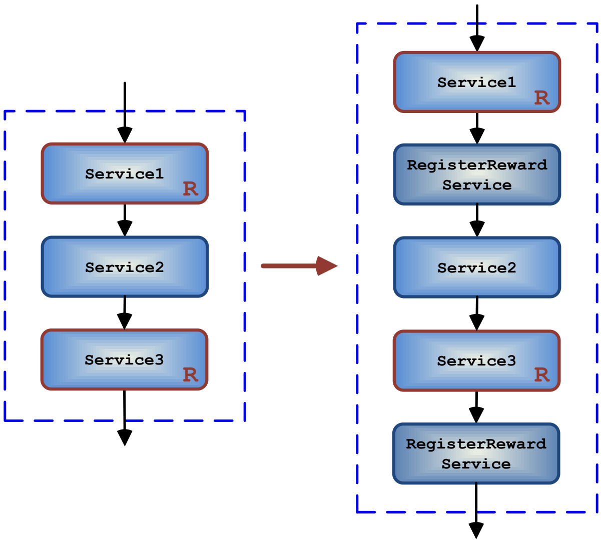


Figure 2. The mechanism for collecting the reward.

The call to *RegisterRewardService* will contain the ID of the workflow, the name of the web service invoked, the name of the web method called, the name of the parameter associated with the reward and the value of the reward (the value of the parameter associated with the reward). This mechanism for collecting the reward is independent of the used BPEL engine.

B. *Web service model*

One important aspect of the composition process is how the *Executor* component determines the rewards during the execution of web services. A call to a Web service corresponds to the execution of an action in the reinforcement learning model. The semantic web service is modeled by its input parameters, output parameters, preconditions and effects (Figure 3).

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-215-8

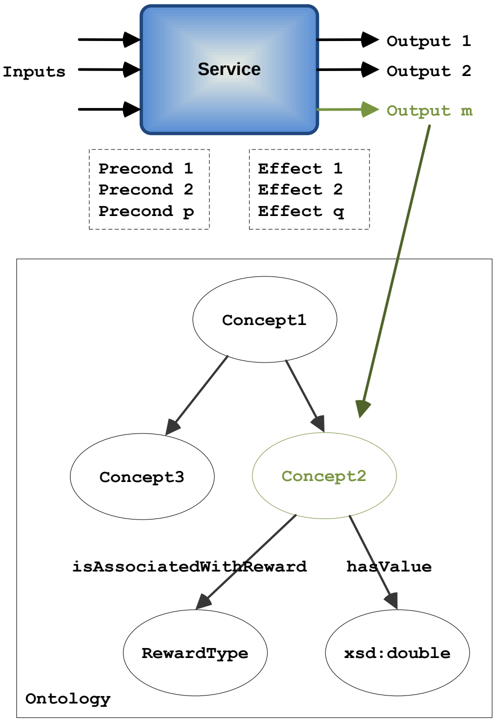


Figure 3. The model of a semantic Web service.

The preconditions define what conditions related to the environment has to hold in order to call the service. The effects define what conditions are true after the service is executed. Preconditions and effects are addressing design time requirements and they are not changing during the runtime. On the other hand, inputs and outputs are known only at runtime. One reason for using preconditions and effects for the Web service model is to combine reinforcement learning with AI planning techniques.

Considering this model of a Web service, the reward corresponding to such a Web service will be associated with one of the output parameters (Figure 3). The output parameter is associated with a formal concept defined by an ontology. The composition request has to contain the reward parameter. While executing a Web service workflow, the *Executor* component will determine the reward associated with each Web Service call. At the end of the execution, the list of rewards will be sent to the *Learning Module*. It is not necessary that all web services contained in the executable workflow to be associated with a reward. There can be web services that have no reward parameter. The reward itself is computed by dedicated web services.

C. *Composition algorithm*

The algorithm used to guide a reinforcement learning agent is trying to find out which actions are “good” by considering the past experience. Many reinforcement learning algorithms are based on estimating the *value function*. A value function is a function of state-action pair that estimates the future rewards that can be expected for the agent that selects a particular action, in a given state. A specific value function is associated with a particular policy. Formally, a value function is defined as:

$$V^{\pi}(s) = E_{\pi}\{R_t|s_t = s\} = E_{\pi}\left\{\sum_{k=0}^{\infty}\gamma^t r_{t+k+1}|s_t = s\right\} \quad (4)$$

This equation defines the value of a state s under the policy π, denoted $V^{\pi}(s)$. E_{π} specifies the expected value considering that the agent uses the policy π.

The algorithm used for composition is based on value iteration algorithm. Value iteration algorithm recursively calculates the value function so that in the end the optimal value function is computed:

$$V_{k+1}(s) = max_a \sum_{s'} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V_k(s')\right] \qquad (5)$$

The idea is to select the best action, denoted by max_a, according to the sum of the expected rewards over the possible states available from that action. The sum is discounted, $0 \leq \gamma \leq 1$. $P^{a}_{ss'}$ represents the probability that action a in state s will lead to state s' while $R^{a}_{ss'}$ represents the expected immediate reward. After the optimal or near optimal value function is computed the optimal policy is generated with the following method:

$$\pi(s) = arg\ max_a \sum_{s'} P^{a}_{ss'}\left[R^{a}_{ss'} + \gamma V_k(s')\right] \qquad (6)$$

The proposed algorithm is the following:

```
1  V ← vector
2  threshold ← small positive value
3  delta ← 0
4  optimalPolicy ← NULL
5
6  foreach state s in S
7      V(s) = rand()
8  end foreach
9
10 do
11     delta ← 0
12     v ← V(s)
13     V(s) ← ComputeValueFunction()
14     delta ← Max (delta, |v - V(s)|)
15 while threshold > delta
16
17 optimalPolicy ← ComputeOptimalPolicy()
```

Firstly, the value function vector **V** is randomly generated (lines 6-8). After that, the value functions for each state are computed until the difference between two consecutive iterations is smaller than a predefined threshold (lines 10-15). In the end, the optimal or near optimal value function will be available. Finally (line 17), the optimal policy is computed based on the optimal value function.

In this phase, we are considering only a limited set of candidate problems. These limitations are related to the fact that the value iteration method has to know the expected rewards and the transition function between states. Currently we are investigating also some other methods for solving the reinforcement learning problem. For instance, temporal difference methods seem to be good candidates for addressing the limitations of the proposed approach.

V. Related work

The majority of existing solutions for automatic web service composition is based on classical search or AI planning combined with logic formalisms. Hendler et al. [9] and Sirin et al. [10] present a method for automatic web service composition using Hierarchical Task Network (HTN) planning. HTN planning is an approach to automated planning in which the dependency among actions is given in the form of hierarchical networks.

Yuhong et al. [11] and Yang et al. [12] present two methods for automatic web service composition using the “Graphplan” algorithm. The Graphplan algorithm uses a data structure called planning graph to search for a solution to a given planning problem. This method of planning is very fast and many AI planners are based on this approach.

McDermott [13] extends the PDDL (Planning Domain Definition Language) planning language in order to associate web services with planning operators. The extended planning operator is able to represent the messages exchanged by a web service. This approach transforms the composition problem into an AI planning problem.

Hongbing et al. [14] propose a reinforcement learning algorithm for web services composition based on logic of preference. However, this solution seems to have limited applicability.

Web service composition using Markov Decision Processes was approached by Gao et al. [15]. In this case, the composition is based on QoS description for Web services. Workflow patterns like sequential, conditional or parallel constructs are modeled using Markov Decision Processes.

VI. Conclusion

A new approach for web service composition, based on reinforcement learning is proposed. There are cases when a concrete goal is not possible to specify. In these cases, a software agent tries to learn from the environment in order to maximize the total reward.

The proposed solution uses a value iteration algorithm in order to find an optimal or near optimal policy. In order to employ this method we consider a semantic web service model. The difference compared with other approaches is that our composition system uses a web service model based on Semantic Web technologies. The composition system uses an architecture based on a closed-loop mechanism. An original mechanism for collecting the reward, independent of the BPEL execution engine, is also proposed.

Composition based on reinforcement learning can be used in scenarios where a concrete goal cannot be explicitly specified. Instead, it is possible to compute a reward/penalty depending on the actions performed by the agent. Markov Decision Process is used for modeling uncertainty. This is another advantage over composition methods based on classical planning.

Since this research is still a work in progress, the next objective is to implement a prototype for demonstrating the proposed idea. Another future development will be to combine the reinforcement learning method presented here with AI planning techniques.

Acknowledgment

This project was supported by the national project code TE 252, contract no. 33/2010, financed by the Romanian Ministry of Education and Research CNCS-UEFISCDI.

References

[1] Alonso, G., Casati, F., Kuno, H., and Machiraju, V., “Web Services: Concepts, Architecture and Applications”, Springer Verlag, 2004.

[2] Berners-Lee, T., Hendler, J., and Lassila, O., “The semantic web”, Scientific American, 284(5): pp. 34–43, 2001.

[3] Russel, S. and Norvig, P., “Artificial Intelligence: A Modern Approach”, Prentice-Hall Inc., 3rd Edition, 2009.

[4] Todica, V., Cremene, M., and Vaida, M., “A Framework for Developing Complex Systems of Services”, Coping with Complexity COPCOM, pp. 77-88, 2011.

[5] Berry, D. and Fristedt, B., “Bandit problems: Sequential allocation of experiments, Monographs on Statistics and Applied Probability”, London: Chapman & Hall, 1985.

[6] Sutton, R. S. and Barto, A. G., “Reinforcement Learning: An Introduction”, MIT Press, 1998.

[7] Tesauro, G. J., “TD-gammon, a self-teaching backgammon program, achieves master-level play”, Neural Computation, 6(2): pp. 215-219, 1994.

[8] Weerawarana, S., Curbera, F., Leymann, F., Storey, T., and Ferguson, D., “Web Services Platform Architecture: Soap, Wsdl, Ws-Policy, Ws-Addressing, Ws-Bpel, Ws-Reliable Messaging and More”, Prentice Hall PTR, Upper Saddle River, NJ, USA, 2005.

[9] Hendler, J., Wu, D., Sirin, E., Nau, D., and Parsia, B., “Automatic web services composition using SHOP2”. In: Proceedings of The Second International Semantic Web Conference (ISWC), 2003.

[10] Sirin, E., Parsia, B., Dan, W., Hendler, J., and Nau, D., “HTN Planning for Web Service Composition Using SHOP2”, International Semantic Web Conference, Sanibel Island, Florida,USA, pp. 20-23, 2003.

[11] Yuhong, Y., Poizat, P., and Ludeng, Z., “Self-Adaptive Service Composition Through Graphplan Repair”. In Proceedings of the 2010 IEEE International Conference on Web Services (ICWS '10). IEEE Computer Society, Washington, DC, USA, 624-627, 2010.

[12] Yang, B. and Qin, Z., “Semantic Web service composition using Graphplan”, Industrial Electronics and Applications, ICIEA 2009, 4th IEEE Conference, pp.459-463, 2009.

[13] McDermott, D., “Estimated-regression planning for interactions with Web services”, In the 6th International Conference on AI Planning and Scheduling, (Toulouse) France, AAAI Press, 2002.

[14] Hongbing, W., Pingping, T., and Hung, P., “RLPLA: A Reinforcement Learning Algorithm of Web Service Composition with Preference Consideration”, IEEE Congress on Services Part II, pp. 163 – 170, 2008.

[15] Gao, A., Yang, D., Tang, S., and Zhang, M., “Web Service Composition Using Markov Decision Processes”, Advances in WebAge Information Management, Vol. 3739, pp. 308-319, 2005.