eKNOW 2014

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

ISBN: 978-1-61208-329-2

March 23 - 27, 2014

Barcelona, Spain

eKNOW 2014 Editors

Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway
Dirk Malzahn, OrgaTech GmbH, Germany

eKNOW 2014

Foreword

The Sixth International Conference on Information, Process, and Knowledge Management (eKNOW 2014), held between March 23-27, 2014 in Barcelona, Spain, continued a series of events related to information, process and knowledge management.

The variety of the systems and applications and the heterogeneous nature of information and knowledge representation require special technologies to capture, manage, store, preserve, interpret and deliver the content and documents related to a particular target.

Progress in cognitive science, knowledge acquisition, representation, and processing helped to deal with imprecise, uncertain or incomplete information. Management of geographical and temporal information becomes a challenge, in terms of volume, speed, semantic, decision, and delivery.

Information technologies allow optimization in searching and interpreting data, yet special constraints imposed by the digital society require on-demand, ethics, and legal aspects, as well as user privacy and safety.

Nowadays, there is notable progress in designing and deploying information and organizational management systems, expert systems, tutoring systems, decision support systems, and in general, industrial systems.

Capturing, representing, and manipulating knowledge was and still is a fascinating and extremely useful challenge from both theoretical and practical perspective. Using validated knowledge for information and process management and for decision support mechanisms raises a series of questions this conference addressed.

We take here the opportunity to warmly thank all the members of the eKNOW 2014 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 eKNOW 2014. 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 eKNOW 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that eKNOW 2014 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 information, processes and knowledge management.

We are convinced that the participants found the event useful and communications very open. We hope that Barcelona, Spain, provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

eKNOW 2014 Chairs:

Dirk Malzahn, OrgaTech GmbH, Germany
Roy Oberhauser, Aalen University, Germany
Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway

eKNOW Special Area Chairs

Technological foresight and socio-economic evolution modeling
Andrzej M.J. Skulimowski, AGH University of Science and Technology - Krakow, Poland

eKNOW 2014

COMMITTEE

eKNOW Advisory Chairs

Dirk Malzahn, OrgaTech GmbH, Germany
Roy Oberhauser, Aalen University, Germany
Conceição Granja, Norwegian Centre for Integrated Care and Telemedicine, University Hospital of North Norway, Norway

eKNOW Special Area Chairs

Technological foresight and socio-economic evolution modelling
Andrzej M.J. Skulimowski, AGH University of Science and Technology - Krakow, Poland

eKNOW 2014 Technical Program Committee

Gil Ad Ariely, Interdisciplinary Center Herzliya (IDC), Israel
Werner Aigner, Institute for Application Oriented Knowledge Processing – FAW / University of Linz, Austria
Panos Alexopoulos, iSOCO, Spain
Jesus Manuel Almendros Jimenez, Universidad de Almería, Spain
Amin Anjomshoaa, Vienna University of Technology, Austria
Zbigniew Banaszak, Warsaw University of Technology, Poland
Ladjel Bellatreche, LISI- ENSMA/ Poitiers University, France
Alejandro Bellogin Kouki, Centrum Wiskunde & Informatica (CWI), Netherlands
Peter Bellström, Karlstad University, Sweden
Jorge Bernardino, Polytechnic Institute of Coimbra, Portugal
Yaxin Bi, University of Ulster - Jordanstown, UK
Marco Bianchi, Fondazione Ugo Bordoni, Italy
Grzegorz Bocewicz, Koszalin University of Technology, Poland
Sabine Bruaux, Picardie Jules Verne University, France
Elżbieta Bukowska, Poznan University of Economics, Poland
Martine Cadot, University of Nancy1, France
Massimiliano Caramia, University of Rome "Tor Vergata", Italy
Shu-Ching Chen, Florida International University, USA
Yu Cheng, IBM TJ Watston Research Center, USA
Chi-Hung Chi, CSIRO, Australia
Dickson K.W. Chiu, Dickson Computer Systems, Hong Kong
Marco Cococcioni, University of Pisa, Italy
Ting Deng, Beihang University, China
Ioan Despi, University of New England, Australia
Ali Eydgahi, Eastern Michigan University, USA
Francesca Fallucchi, Guglielmo Marconi University, Italy
Abed Alhakim Freihat, University of Trento, Italy
Susan Gauch, University of Arkansas, USA

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Analyzing the Effectiveness of Investment Strategies through Agent-based Modelling: Overconfident Investment Decision Making and Passive Investment Strategies

Hiroshi Takahashi
Graduate School of Business Administration
Keio University
Email: htaka@kbs.keio.ac.jp

***Abstract*—This article analyzes the effectiveness of investment strategies through agent-based modeling. In this analysis, we will focus on the performance of a passive investment strategy (which is one of the most popular investment strategies in the asset management business) under conditions where overconfident investors trade. As a result of intensive experimentation, it was concluded that overconfident investors could achieve a positive excess return in the market where there are no passive investors. However, our agent-based simulation shows that overconfident investor could not survive in a market where passive investors exist. These results suggest the effectiveness of a passive investment strategy. The results are of both academic interest and practical use.**

Keywords-Finance; Agent-based Modelling; Behavioral Economics; Overconfidence; Asset Management.

I. INTRODUCTION

The financial system plays a significant role in society and the economy. The role of investors providing capital to companies has become more important than ever. Financial markets contribute to efficient capital allocation, and a great amount of research regarding financial markets has been carried out.

In the last years, there has been rising interest in a field called behavioral finance, which incorporates psychological methods in analyzing investor behavior. There are numerous arguments in behavioral finance that investors' decision making bias can explain phenomenon in the financial market which until now had gone unexplained. Such arguments often point out the limit of arbitrage and the existence of systematic biases in decision-making [12][23][24][32]. Behavioral finance has examined a wide range of phenomena in the market and among investors, drawing a number of provocative conclusions. There are, for example, studies which suggest that overconfident investors could survive in the market.

Market efficiency is a central hypothesis of traditional financial theory. Indeed, the efficiency of the market lies at the heart of traditional financial theory. For example, in the Capital Asset Pricing Model (CAPM), one of the most popular asset pricing theories, equilibrium asset prices are derived on the assumption that markets are efficient and investors rational [21]. CAPM indicates that the optimal investment strategy is to hold a market portfolio. Since it is very difficult for investors to get an excess return in an efficient market, it is assumed to be difficult to beat a market portfolio even if the investment strategy is firmly based on public information. A passive investment strategy, which tries to maintain an average return using benchmarks based on market indices, is consistent with traditional asset pricing theories and is considered to be an effective method in efficient markets.

With this background in mind, the purpose of this research is to analyze the performance of overconfident investors in a market where passive investors exist. To address this problem, we have employed agent-based modeling in this analysis [2][3][7][8]. Agent based modeling is an effective method of analyzing the relationship between Micro-rules and Macro-behavior. Agent-based modeling is an attempt to explain the macro-behavior of systems by local rules. As a result of applying this model to Social Science, it has been found that a variety of different macro-behaviors emerge bottom-up from local micro-rules. Agent-based modeling has many applications, and none more suitable than for the creation of an artificial market. For example, Arthur et al. [1] analyse the market under conditions where heterogeneous investors trade and concluded that complex conditions emerge. Using agent-based modeling, Takahashi et al. [25] found that irrational traders could survive in the market. Takahashi [29] suggests that the combination of behavioral biases and financial constraints causes a significant deviation from fundamental values. Analyses which attempt to replicate realistic market conditions and dynamics present a greater challenge to the researcher than traditional forms of research. Due to the efficacy of this type of advanced analysis, there is a greater demand for research conducted employing these kinds of models. There is, therefore, a need for analyses using this more current approach, in addition to original methods. Agent-based modeling is making an increasingly valuable contribution to financial research.

The next section describes the model used in this analysis. Section III shows the results of the analysis. Section IV summarizes this paper.

II. MODEL

A computer simulation of the financial market involving 1000 investors was used as the model for this research. Shares and risk-free assets were the two types of assets used, along with the possible transaction methods. Several types of investors exist in the market, each undertaking transactions based on their own stock evaluations [1][13][28][30][31]. (see Fig. 1). This market was composed in three major stages; (1) generation of corporate earnings, (2) formation of investor forecasts, and (3) setting transaction prices. The market advances through repetition of these stages (see Fig. 2).

 ISBN: 978-1-61208-329-2

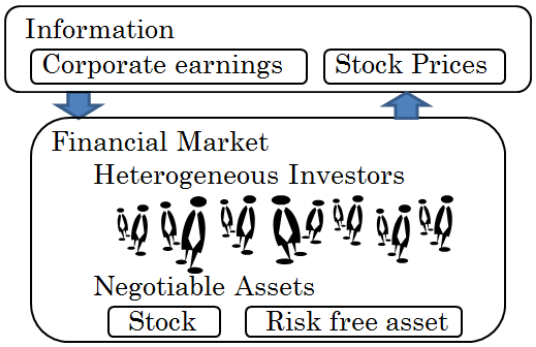


Figure 1. Basic architecture of financial market simulator

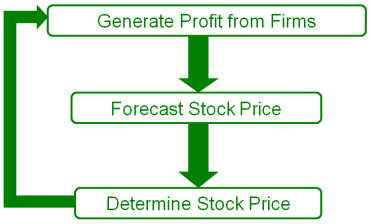


Figure 2. Simulation steps of market transactions

The following sections describe negotiable transaction assets, modeling of investor behavior, transaction price setting, and rules of natural selection in the market.

A. Negotiable assets in the market

This market has both risk-free and risk-associated assets. There are risk-associated assets in which all profits gained during each term are distributed to shareholders. Corporate earnings (y_t) are expressed as $y_t = y_{t-1} \cdot (1 + \varepsilon_t)$. However, they are generated according to the process of $\varepsilon_t \sim N(0, \sigma_y^2)$ with share trading being undertaken after the public announcement of profits for the term [5][19]. Each investor is given common asset holdings at the start of the term with no limit placed on debit and credit transactions (1000 in risk-free assets and 1000 in stocks). Investors adopt the buy-and-hold method for the relevant portfolio as a benchmark to conduct decision-making by using a one-term model. The buy-and-hold method [20] is an investment method to hold shares for medium to long term.

B. Modeling investor behavior

Each type of investor handled in this analysis is organized in Table I. This analysis covers most major types of investor [6][24]. The investors can be classified into two categories: active investors (Type 1-4) and a single passive investor type (Type 5). Active investors in this market evaluate transaction prices based on their own forecasts of market movements, taking into consideration both risk and return rates when making decisions. Passive investors employ a buy-and-hold strategy [14]. A passive investment strategy is one of the most popular investment strategies in the asset management business. Each active investor determines the investment ratio (w_t^i) based on the maximum objective function$(f(w^i t))$, as shown below [10][15].

$$f(w_t^i) = r_{t+1}^{int,i} \cdot w_t^i + r_f \cdot (1 - w_t^i) - \lambda(\sigma_{t-1}^i)^2 \cdot (w_t^i)^2. \quad (1)$$

Here, $r_{t+1}^{int,i}$ and σ_{t-1}^i in the eq. (1) express the expected rate of return and risk for stocks as estimated by each investor i. r_f indicates the risk-free rate. w_t^i represents the stock investment ratio of the investor i for term t. λ shows degree of investor risk aversion. The value of the objective function $f(w_t^i)$ depends on the investment ratio(w_t^i). The investor decision-making model here is based on the Black/Litterman model that is used in the practice of securities investment [4][16][17][18].

The integrated expected rate of return for shares is calculated as follows [4]:

$$r_{t+1}^{int,i} = \frac{c^{-1}(\sigma_{t-1}^i)^{-2} r_{t+1}^{f,i} + (\sigma_{t-1}^i)^{-2} r_t^{im}}{c^{-1}(\sigma_{t-1}^i)^{-2} + (\sigma_{t-1}^i)^{-2}}. \quad (2)$$

Here, $r_{t+1}^{f,i}$, r_t^{im} in the eq. (2) express the expected rate of return, calculated from short-term expected rate of return, and risk and gross current price ratio of stocks respectively. c is a coefficient that adjusts the dispersion level of the expected rate of return calculated from risk and gross current price ratio of stocks [4].

The short-term expected rate of return $(r_t^{f,i})$ is obtained where $(P_{t+1}^{f,i}, y_{t+1}^{f,i})$ is the equity price and profit forecast for term $t+1$ is estimated by the investor, as follows: $r_{t+1}^{f,i} = ((P_{t+1}^{f,i} + y_{t+1}^{f,i})/P_t - 1)$.

The price and profit forecast$(P_{t+1}^{f,i}, y_{t+1}^{f,i})$ includes the error term $(P_{t+1}^{f,i} = P_{t+1}^{f,typej} \cdot (1 + \eta_t^i)$, $y_{t+1}^{f,i} = y_{t+1}^{f,typej} \cdot (1 + \eta_t^i)$, where $\eta_t^i \sim N(0, \sigma_n^2))$ reflecting that even investors using the same forecast model vary slightly in their detailed outlook. The stock price $(P_{t+1}^{f,i})$, profit forecast $(y_{t+1}^{f,i})$, and risk estimation methods are described in the following paragraph.

The expected rate of return obtained from stock risk and so forth is calculated from stock risk (σ_{t-1}^i), benchmark equity stake (W_{t-1}), degree of investor risk aversion (λ), and risk-free rate (r_f), as follows [22]: $r_t^{im} = 2\lambda(\sigma_{t-1}^i)^2 W_{t-1} + r_f$.

1) Stock price forecasting method: The fundamental value is estimated by using the discounted cash flow model (DCF), which is a well known model in the field of finance. Fundamentalists estimate the forecasted stock price and forecasted profit from profit for the term (y_t) and the discount rate (δ) as $P_{t+1}^{f,typej} = y_t/\delta, y_{t+1}^{f,typej} = y_t$.

Forecasting based on trends involves forecasting the next term 's stock prices and profit through extrapolation of the most recent stock value fluctuation trends. Stock price and profit of the next term are estimated from the most recent

TABLE I. LIST OF INVESTOR TYPES

No.	Investor types
1	Fundamentalist
2	Forecasting by past average (most recent 10 days)
3	Forecasting by trend (most recent 10 day)
4	Latest Price
5	Passive investor

 ISBN: 978-1-61208-329-2

trends of stock price fluctuation (a_{t-1}) from time point $t-1$ as $P_{t+1}^{f,typej} = P_{t-1} \cdot (1+a_{t-1})^2, y_{t+1}^{f,typej} = y_t \cdot (1+a_{t-1})$.

Forecasting based on past averages involves estimating the next term stock prices and profit based on the most recent average stock value.

C. Risk Estimation Method

Stock risk is measured as $\sigma_{t-1}^{s,i} = s_i \cdot \sigma_{t-1}^h$. In this case, σ_{t-1}^h is an index that represents stock volatility calculated from price fluctuation of the most recent 100 steps, and s_i is the degree of overconfidence. The presence of a strong degree of overconfidence can be concluded when the value of s_i is less than 1, as estimated forecast error is shown as lower than its actual value. The investors whose value of s_i is less than 1 tend to invest more actively. For example, when such investors predict that stock prices will increase, they invest more in stock than ones whose value of s_i is 1.

D. Determination of transaction prices

Transaction prices are determined as the price where stock supply and demand converge $(\sum_{i=1}^{M}(F_t^i w_t^i)/P_t = N)$. In this case, the total asset (F_t^i) of investor i is calculated from transaction price (P_t) for term t, profit (y_t) and total assets from the term $t-1$, stock investment ratio (w_{t-1}^i), and risk-free rate (r_f), as $F_t^i = F_{t-1}^i(w_{t-1}^i(P_t+y_t)/P_{t-1} + (1-w_{t-1}^i)(1+r_f))$.

E. Natural Selection in the Market

Investors who are able to adapt to and, hence, profit from the market as it fluctuates will remain in the market and their position will grow stronger. Conversely, investors who are unable to do this will drop out of the market. Such a pattern is very suggestive of what might be termed Natural Selection in the market. The driving force behind this Natural Selection is the desire for cumulative excess profit [9]. Two aspects of this pattern are of particular interest: (1) the identification of investors who alter their investment strategy, and (2) the actual alteration of investment strategy [25].

Each investor must decide whether he should change investment strategies based on the most recent performance of each 5 term period (after 25 terms have passed since the beginning of market transactions). The higher the profit rate obtained most recently is, the lesser the possibility of strategy alteration becomes. The lower the profit, the higher the possibility becomes. (In the actual market, evaluation tends to be conducted according to baseline profit and loss.) Specifically, when an investor could not obtain a positive excess profit for the benchmark portfolio profitability, they are likely to alter their investment strategy with the probability below:

$$p_i = \min(1, \max(-100 \cdot r_i^{cum}, 0)). \quad (3)$$

Here, however, r_i^{cum} in the eq. (3) is the cumulative excess profit for the most recent benchmark of investor i. Measurements were conducted for 5 terms, and the cumulative excess profit was calculated as a one-term conversion. For example, if excess profit over a 5 term period is 5 %, a one term conversion would show this as a 1 % excess for each term period.

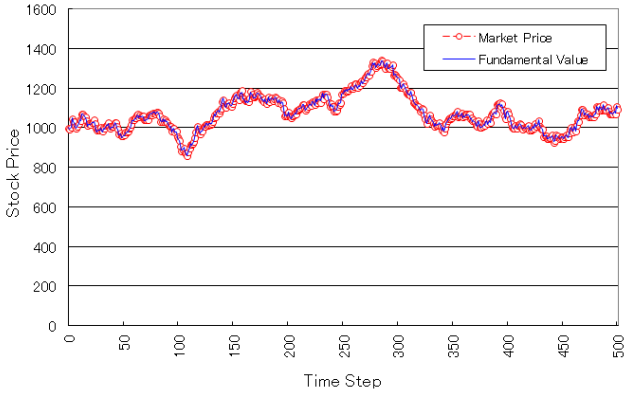


Figure 3. Price Transitions (Fundamentals)

When it comes to deciding on a new investment strategy, an investment strategy that has a high cumulative excess profit for the most recent five terms (forecasting type) is 'naturally' more likely to be selected. Where the strategy of the investor i is z_i and the cumulative excess profit for the most recent five terms is r_i^{cum}, the probability p_i that z_i is selected as a new investment strategy is given as $p_i = e^{(a \cdot r_i^{cum})} / \sum_{j=1}^{M} e^{(a \cdot r_j^{cum})}$. Selection pressures on an investment strategy become higher as the coefficients' value increases. Those investors who altered their strategies make investments based on the new strategies after the next step.

III. Results

The first set of results is from a model in which no passive investors are analyzed. The second set presents a situation in which passive investors are present.

A. Case 1: No Passive Investors

At first, this section analyzes a situation where all investors make investment decisions based on fundamental values (Table 1, Type 1). Fig. 3 shows the transitions of transaction prices. The horizontal axis in the graph shows time steps and the vertical axis shows stock prices. Two transitions are shown: Fundamental values and transaction prices, and it can be seen that transaction prices are consistent with fundamental values throughout the entire transaction period. These results are consistent with traditional financial theory. Looking at transitions in the degree of overconfidence, a strengthening degree of overconfidence can be seen in the behavior of the remaining investors as market transactions move forward (Fig. 4). These results suggest that there is something going on in the market which allows overconfident investors - with their biases in investment decision-making - to survive. This would be in clear contradiction of traditional financial theory.

Similar results are seen when there are equal numbers of four types of investors (Table 1, Type 1-4). Figs. 5-7 show these results. Fig. 5 shows the transitions of transaction prices and Fig. 6 shows the transitions of the numbers of each type of investor. Fig. 7 shows the transition in the degree of overconfidence and shows that investors with a strong degree of overconfidence tend to survive in the market even under these conditions. These results further support the suggestion

 ISBN: 978-1-61208-329-2

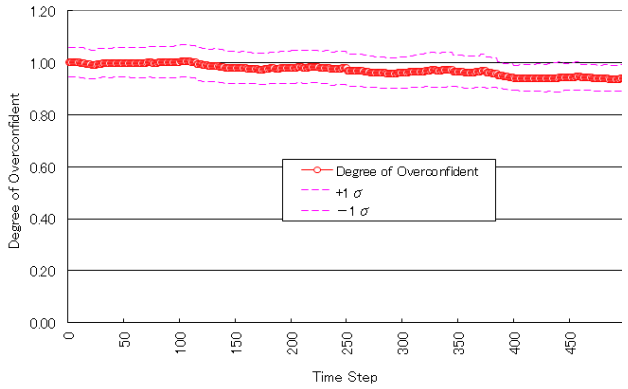


Figure 4. Transition of degree of overconfidence (Fundamentals)

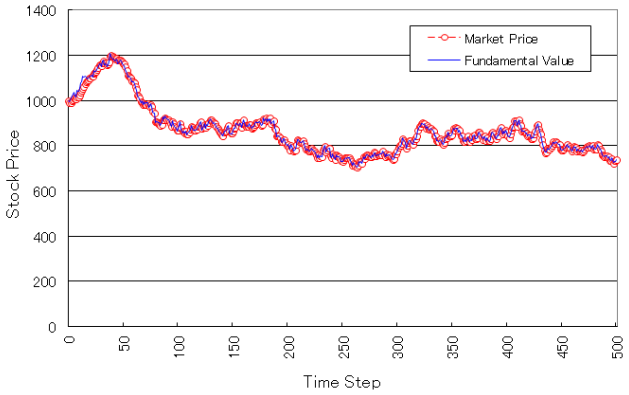


Figure 5. Price Transitions (Fundamentals, Latest, Trend, Average)

that overconfident investors can survive in the market. For details of the influence of overconfident investors, refer to [27].

B. Case 2: Passive Investors

1) Fundamentalists and Passive Investors: This section analyzes a situation where passive investors invest in the market. At first, this section analyzes the market where the same number of fundamentalists and passive investors trade (Table 1, Type 1 and 5). Figs. 8, 9, and 10 show the transition of share prices, the number of investors, and the degree of overconfidence, respectively. Fig. 9 shows that, as time steps go, the number of passive investors increases. These results speak to the effectiveness of passive investment strategies. However, it is also the case that market prices reach a point where they begin to deviate from fundamental values (see Fig. 8). These latter result indicates possible drawbacks of passive investment strategies. As for the details of the analysis focusing on passive investment strategies, please refer to [26]. Yet from the data (see Fig. 10), a clear conclusion may be drawn: in this model overconfident investors do not survive, and a passive investment strategy is superior in its effectiveness.

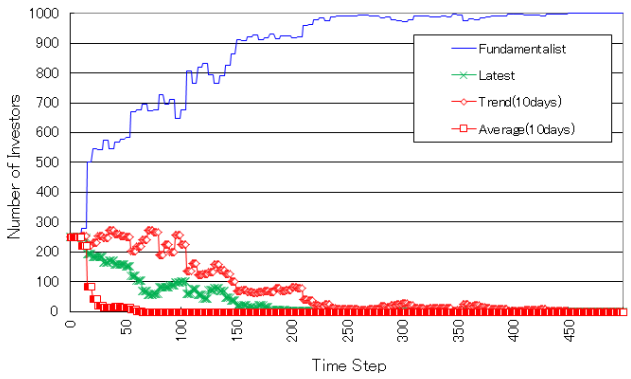


Figure 6. Transitions of Number of Investors (Fundamentals, Latest, Trend, Average)

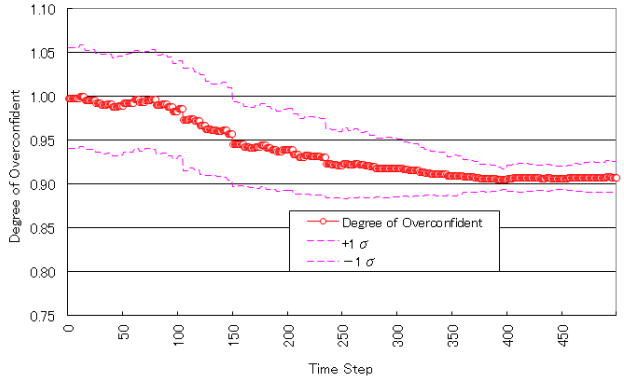


Figure 7. Transition of degree of overconfidence (Fundamentals, Latest, Trend, Average)

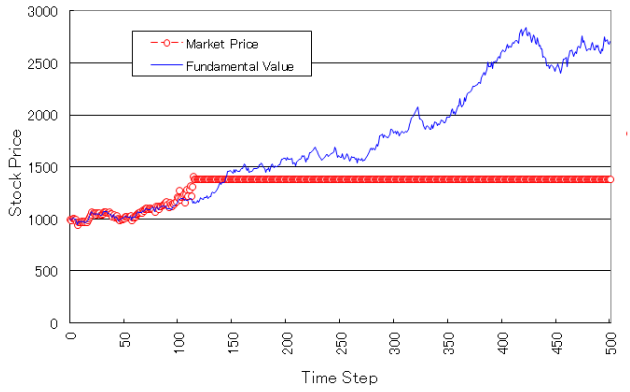


Figure 8. Price Transitions (Fundamentals, Passive)

2) Introducing extra investor types: This section analyzes the case where the same number of five types of investors including passive investors trade in the market (Table 1, Type 1-5). Figs. 11-13 show the results (Transitions of stock prices, number of investors and the degree of overconfident). Fig. 12 shows that passive investors survive in the market, and Fig. 11 shows that market prices reach a point where they begin to deviate from fundamental values, as is in the previous section. Fig. 13 shows that investors who survive in the market do not have a tendency towards overconfidence. These results also suggest the effectiveness of passive investment strategies. Although very simple, passive investment strategies

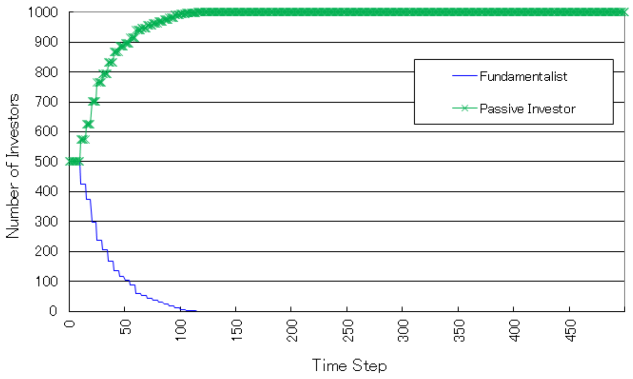


Figure 9. Transitions of Number of Investors (Fundamentals, Passive)

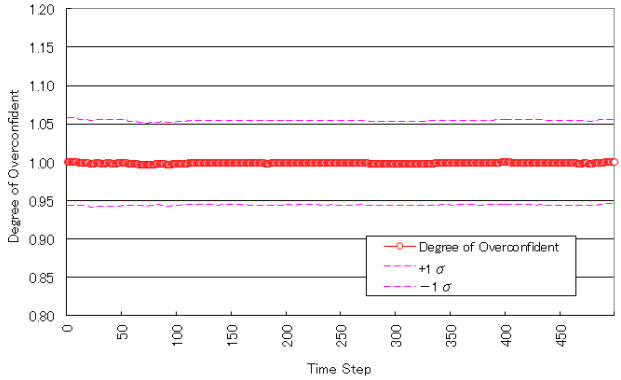


Figure 10. Transition of degree of overconfidence (Fundamentals, Passive)

show impressive flexibility, adaptability and resilience.

IV. SUMMARY

This article examines the effectiveness or otherwise of passive investment strategies, utilizing agent-based modeling. As a result of intensive experimentation, this paper confirms that a passive investment strategy is effective under conditions where overconfident investors invest. This conclusion is of interest in itself and merits further study. Future analyses will focus on examining the effect on markets of practical changes.

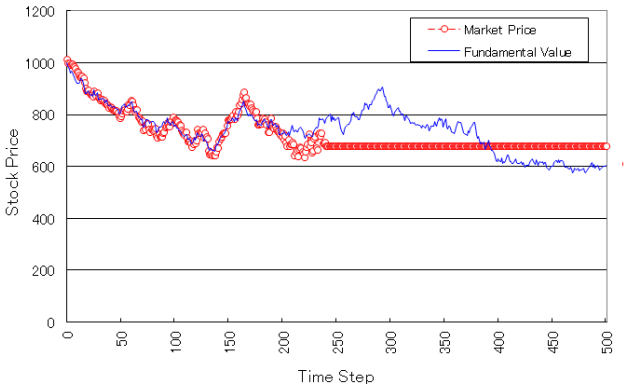


Figure 11. Price Transitions (Fundamentals, Latest, Trend, Average, Passive)

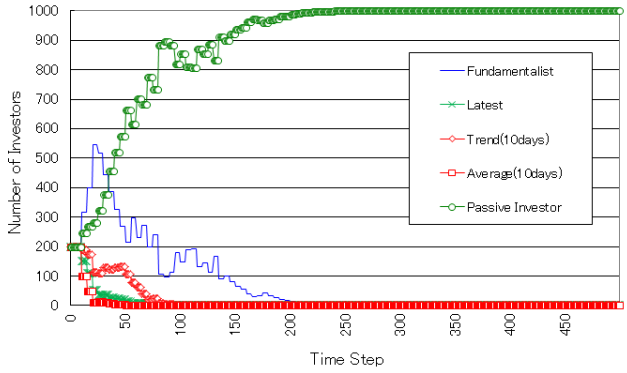


Figure 12. Transitions of Number of Investors (Fundamentals, Latest, Trend, Average, Passive)

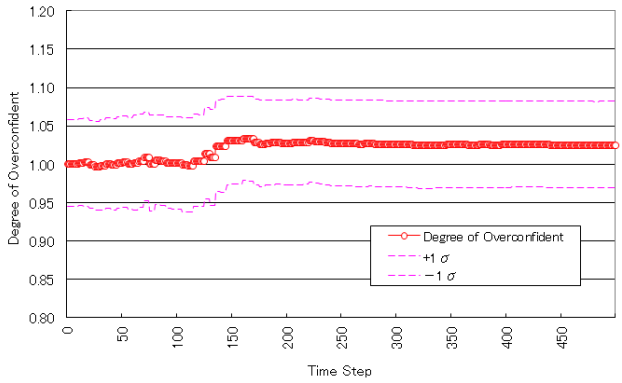


Figure 13. Transition of degree of overconfidence (Fundamentals, Latest, Trend, Average, Passive)

This research focuses on passive investors and overconfident investors. A more detailed analysis would consider both types of investment behavior under more realistic market conditions. This is a matter for further research.

APPENDIX

This section lists the major parameters of the financial market designed for this paper. The explanation and value for each parameter is described.

M: Number of investors (1000)
N: Number of shares (1000)
F_t^i: Total asset value of investor i for term t ($F_0^i = 2000$: common)
W_t: Ratio of stock in benchmark for term t ($W_0 = 0.5$)
w_t^i: Stock investment rate of investor i for term t ($w_0^i = 0.5$: common)
y_t: Profits generated during term t ($y_0 = 0.5$)
σ_y: Standard deviation of profit fluctuation ($0.2/\sqrt{200}$)
δ: Discount rate for stock($0.1/200$)
λ: Degree of investor risk aversion (1.25)
σ_n: Standard deviation of dispersion from short-term expected rate of return on shares (0.05)
c: Adjustment coefficient (0.01)

REFERENCES

[1] W. B. Arthur, J. H. Holland, B. LeBaron, R. G. Palmer, and P. Taylor, " Asset Pricing under Endogenous Expectations in an Artificial Stock Market, " in The Economy as an Evolving Complex System II, W. B. Arthur, S. N. Durlauf and D. A. Lane, Eds. Addison-Wesley, 1997, pp. 15-44.

[2] R. Axelrod, The Complexity of Cooperation -Agent-Based Model of Competition and Collaboration, Princeton University Press. 1997.

[3] R. Axtell, " Why Agents? On the Varied Motivation For Agent Computing In the Social Sciences, " The Brookings Institution Center on Social and Economic Dynamics Working Paper, November, 17, 2000.

[4] F. Black and R. Litterman, " Global Portfolio Optimization, " Financial Analysts Journal, September-October, 1992, pp. 28-43.

[5] R. Brealey, S. Myers, and F. Allen, Principles of corporate finance 8th edition, The McGraw-Hill Companies Inc., 2006.

[6] M. Brunnermeier, Asset Pricing under Asymmetric Information, Bubbles, Crashes, Technical Analysis and Herding, Oxford University Press, 2001.

[7] J. M. Epstein and R. Axtell, Growing Artificial Societies Social Science From The Bottom Up, MIT Press, 1996.

[8] N. Gilbert, Agent-Based Models, Sage Publications: London, 2007.

[9] D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley, 1989.

[10] J. E. Ingersoll, Theory of Financial Decision Making, Rowman & Littlefield, 1987.

[11] R. Jarrow, " Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices, " Journal of Finance, 35 (1980) pp. 1105-1113.

[12] D. Kahneman and A. Tversky, " Prospect Theory of Decisions under Risk, " Econometrica, 47, 1979, pp. 263-291.

[13] M. Levy, H. Levy, and S. Solomon, Microscopic Simulation of Financial Markets, Academic Press, 2000.

[14] B. G. Malkiel, " Passive investment strategies and Efficient Markets, " European Financial Management, 9, 2003, pp. 1-10.

[15] H. Markowitz, " Portfolio Selection, " Journal of Finance, 7, 1952, pp. 77-91.

[16] L. Martellini and V. Ziemann, " Extending Black-Litterman Analysis Beyond the Mean-Variance Framework: An Application to Hedge Fund Style Active Allocation Decisions, " Journal of Portfolio Management, 33, 2007, pp. 33-45.

[17] A. Meucci, " Beyond Black-Litterman in Practice, " Risk, 19, 2006, pp. 114-119.

[18] A. Meucci, " Enhancing the Black-Litterman and Related Approaches: Views and Stress-test on Risk Factors, " Journal of Asset Management, 10, 2009, pp. 89-96.

[19] P. O'Brien, " Analysts' Forecasts as Earnings Expectations, " Journal of Accounting and Economics, January, 1988, pp. 53-83.

[20] R. G. Clarke, M. T. FitzGerald, P. Berent, and M. Statman, " Market Timing with Imperfect Information, " Financial Analysts Journal, November-December, 1989, pp. 27-36.

[21] W. F. Sharpe, " Capital Asset Prices: A Theory of Market Equilibrium under condition of Risk, " The Journal of Finance, 19, 1964, pp. 425-442.

[22] W. F. Sharpe," Integrated Asset Allocation, "Financial Analysts Journal, September-October, 1987, pp. 25-32.

[23] R. J. Shiller, Irrational Exuberance, Princeton University Press, 2000.

[24] A. Shleifer, Inefficient Markets, Oxford University Press, 2000.

[25] H. Takahashi and T. Terano, " Agent-Based Approach to Investors' Behavior and Asset Price Fluctuation in Financial Markets, " Journal of Artificial Societies and Social Simulation, 6, 2003.

[26] H. Takahashi, S. Takahashi, and T. Terano, " Agent-Based Modeling to Investigate the Effects of Passive Investment Strategies in Financial Markets, " in Social Simulation Technologies: Advances and New Discoveries(Representing the best of the European Social Simulation Association conferences), Bruce Edmonds, Cesario Hernandes, and Klaus Troitzsch, Eds. Idea Group Inc., 2007, pp. 224-238.

[27] H. Takahashi and T. Terano, " Analyzing the Influence of Overconfident Investors on Financial Markets Through Agent-Based Model, " Lecture Note in Computer Science 4881, Springer-Verlag, 2007, pp. 1042-1052.

[28] H. Takahashi, S. Takahashi, and T. Terano, " Analyzing the Validity of Passive Investment Strategies Employing Fundamental Indices through Agent-Based Simulation, " in Agent and Multi-Agent Systems: Technologies and Applications, Lecture Note in Artificial Intelligence, J. O ' Shea, N. T. Nguyen, K. Crockett, R. J. Howlett, and L. C. Jain Eds. Springer-Verlag 6682, 2011, pp. 180-189.

[29] H. Takahashi, " An Analysis of the Influence of dispersion of valuations on Financial Markets through agent-based modeling, " International Journal of Information Technology & Decision Making, 11, 2012, pp. 143-166.

[30] H. Takahashi, " A Analyzing the influence of dispersion of fundamentalists' valuations on the effectiveness of passive investment strategy under financial constraints, "International Journal of Intelligent Systems Technologies and Applications, 2, 12, 2013, pp. 111-127.

[31] L. Tesfatsion, " Agent-Based Computational Economics, " Iowa Sate University Economics Working Paper, 1, 2002.

[32] A. Tversky and D. Kahneman, " Advances in. prospect Theory : Cumulative representation of Uncertainty, " Journal of Risk and Uncertainty, 5, 1992, pp. 297-323.

 ISBN: 978-1-61208-329-2

Knowledge Representation Support for Substantive Patent Law Precedents

Sigram Schindler
TELES Patent Rights International
TU Berlin
Berlin, Germany
schindler.board@teles.de

***Abstract* — The paper outlines a KR (Knowledge Representation) based IES (Innovation Expert System), for testing a claimed – classical as well as emerging – technology invention under the SPL (Substantive Patent Law) of any NPS (National Patent System), in particular under the 4 §§ 101/102/103/112 of 35 USC (United States Code), as interpreted by the Supreme Courts' *KSR/Bilski/Mayo* decisions. Already the IES prototype is capable of indicating the amazing power of the "Patent Technology" induced by this US Highest Courts' SPL precedents as to such tests for a claimed invention. It works semi-automated when testing in explorative mode and fully automated/real-time when testing in confirmative mode. Developing this powerful Patent Technology has been enabled by performing substantial Mathematical KR research about recent US Highest Courts' patent precedents – published by Mathematical KR research papers and Amicus Briefs submitted to the US Supreme Court and the US CAFC (Court of Appeals for the Federal Circuit) as to KR insights so obtained into the problems of SPL precedents, e.g., when dealing with claimed emerging technology inventions, in particular CIIs (Computer-Implemented Inventions).**

Keywords* — *SPL (Substantive Patent Law); KR Based IES Prototype; Emerging Technologies; Supreme Court's KSR/ Bilski/Mayo Decisions; Inventive Concept; Preemptivity; Abstract Idea; CAFC's Recent Precedents

I. INTRODUCTION

Internationally and nationally, inconsistencies in SPL precedents have increased with the advent of claimed inventions dealing with subject matter in emerging technologies areas. As compared to classical technologies SPL precedents and its allegedly clearly understood pragmatics, applying SPL on emerging technologies inventions encounters new kinds of pragmatics not yet understood. Inconsistencies arise, as these new pragmatics of emerging technology inventions come together with their being intangible and invisible just as their subject matters. This requires replacing both by a purely mental model, the invention and its base of notions, that is, functionality provider. They will be called "model-based" from here on.

Examples of such advanced alias model-based technologies underlying recent patenting are business technology, nano technology, pharmaceutics technology, genetics (DNA (DeoxyriboNucleic Acid) technology), software technology. All these model-based technologies raised fundamental questions of thinking about the creativity embodied by inventions, about SPL stimulating and protecting it, and about the subject matter of inventions dealing with them, which required decisions by the respective national Highest Courts. For this paper, of particular importance is the US Supreme Court's famous, as direction pointing, line of *KSR/Bilski/ Mayo/Myriad* landmark decisions [32]-[34]. They deal with creativity/business respectively software/pharmaceutical/ DNA respectively life science technologies – whereby especially the *Mayo* decision [33] provides clear guidelines.

These fundamental questions of thinking are hardly analyzable without applying Mathematical Knowledge Representation, customized to dealing with it - e.g. by simplifying accordingly the notion of "concept". While this notion is well-known in "Advanced IT"[1)] (e.g., DL/KR/... [3][4]) it is in this form far too complex to enable meticulously mathematically modelling SPL and SPL precedents - and also for ever becoming broadly accepted by the several million patent lawyers/examiners/judges/inventors.

This means that classical SPL precedents is not really applicable to model-based claimed inventions, as its classical claim construction assumes a tangible/visible subject matter, hence allegedly always patent-eligible, i.e., there no need existed to separate patent-eligible from non-eligible inventive concepts – potentially existing with intangible/invisible emerging technologies' subject matters. For dependably achieving this separation and understanding its implications, a refined claim construction is indispensable, as shown by the inconsistencies evolved already (see below). Yet, defining this refined claim construction precisely and completely, as required and clearly outlined by the Supreme Court's *Mayo* decision though with a broad brush only, involves serious intricacies. Removing them dependably is possible by KR Technology, partly by only Mathematical KR, as shown in [5].

To put this quite unmistakably: KR Technology indeed managed to identify the reasons for the notional inconsistencies of recent SPL precedents. It achieved this by removing them by defining, for a claimed invention, be it of classical or of emerging technologies, the refined claim construction precisely and completely. This is probably hitherto the most important contribution of KR Technology to solving a problem otherwise seemingly unsolvable: Of catering innovations in emerging technologies, and hence of technology depending societies. Anyway, the amazingly powerful "Patent Technology" outlined by this paper could not have been developed without this Mathematical KR (Technology) or the US Highest Courts' SPL precedents inducing it.

 ISBN: 978-1-61208-329-2

II. On SPL and SPL precedents

This paper belongs to a series of other papers – as shown by the reference list – of an R&D (Research and Development) project, namely the FSTP (Facts Screening and Transforming Processor) project, dealing with supporting precise SPL interpretation and SPL precedents by Advanced IT[1)] [3][4], e.g., for making patents, in particular on emerging technologies subject matters - and the IES described by this paper - "Highest Court proof". While this paper is self-contained, its terms/ notions yet are hard to understand without their detailed discussions in these other papers. Hence, there are many cross-references to them – though this paper's basic ideas are independent of the precise knowledge of this "context".

The *Mayo* decision [33] showed that describing a model-based claimed invention by its "inventive concepts" facilitates isolating/recognizing its new pragmatics in spite of its new mental problems due to its and its service provider's intangibility/invisibility. The term/notion of "concept" is similarly used since ever in Advanced IT[1)] [3][4]. I.e., the far reaching potentials of the term/notion of "concept" are commonly known and fundamental in probably all branches of Advanced IT since dozens of years. But there this term/notion has been developed to a degree of sophistication completely clouding its potential usefulness for SPL precedents. But, the *Mayo* decision shows that only the next to trivial kernel of this notion is used by the notion of "inventive concept", which makes it apt for SPL. This is confirmed, again, by the Supreme Court's recent invitation of Amicus Briefs as to the question of patent-eligibility of computer-implemented inventions [24][25].

Thus: Here the US SPL is taken exemplarily, i.e., the 4 §§ 101/102/103/112 of 35 USC, but the SPL of any other NPS could have been taken also, e.g., in the EU the §§ 52-57, 69 of the EPC (European Patent Convention). The inconsistencies in the US SPL precedents indispensably imply reconsidering in all NPSes their "claim construction alikes" for emerging technologies' inventions.

Proceeding as the US Supreme Court's *KSR/Bilski/Mayo* decisions [32][33] require is possible in all other national/regional SPLs, too. But this implies getting familiar with the "scientification" coming up in testing a claimed invention this way, especially with the *Mayo* decision's [33] new key terms/notions "inventive concepts" and "preemption"/"abstract idea" – as they facilitate separating in any SPL its concerns (= requirements) from each other [10][18]. Additionally, they fully compensate the impossibility of graphically supporting the presentation of the properties of a model-based claimed invention [19]. They thus enable showing/proving that these properties meet the separated SPL requirements/concerns.

The transition – from the classical claim construction to a refined claim construction by using these additional, new, and more purposeful terms/notions in i) interpreting an SPL, ii) describing the properties of the invention to be tested under this SPL, and iii) showing that these properties meet these requirements/concerns – is a "paradigm refinement", as explained in detail in [18].

Summarizing the message conveyed by this section: This paper is focused on showing i) that the groundbreaking insights coming together with the Supreme Court introduced terms "inventive concept" and "preemption"/"abstract idea" just leverage on Mathematical KR [5] but completely avoid confronting a user with any Mathematics ii) the huge advantages that the so by the US Highest Courts induced "Patent Technology" provides to every patent practitioner's professional life, by outlining the powerful functionalities of the IES.

III. On Patents / Inventions

"Patent/SPL Technology" and its refined claim construction – induced by the US Highest Courts' patent precedents – are intellectually only slightly more demanding than the hitherto allegedly sufficient classical claim construction [24][25]. Nevertheless, its "post-Mayo" refined claim construction dramatically reduces by its "purposefulness" [1][10][18] the time for testing a claimed invention under 35 USC §§ 112/102/103/101, i.e., under US SPL, while the classical claim construction is oversimplistic and so creates confusion and invites misuse in many practical cases, in particular if additionally applying the strange BRI (Broadest Reasonable Interpretation) [14][21][24][25].

Patent Technology is an administrative "cross-sectional technology" in that it impacts on decision making in all US institutions below the Supreme Court – but not on the top of this hierarchy, the AIA (America Invents Act) (as erroneously seen, due to its disaggregating the 4 compound legal requirement statements of its 4 §§ 101/102/103/112 into 10 SPL/FSTP tests) [10]. But this administrative view on Patent Technology ignores its impacts on everyday's patent business.

By performing this disaggregation of compound legal concerns/requirements – of the fictional but politically decisive "social contract" underlying SPL – Patent Technology implements the Supreme Court above interpretation of the §§ 101/102/103/112. It maps these §§'s 4 compound requirement statements onto (today) 10 "concerns separating" such statements, checked by 10 simple FSTP/SPL tests (for an invention to be patent-eligible and patentable).

This logically correct mapping – of 4 compound onto 10 elementary legal concern/requirement statements – implies that these 10 simple tests are to be passed by a claimed invention if and only if it is patent-eligible and patentable under the SPL of 35 USC. This mapping onto the 10 simple tests exposes that the Supreme Court's *KSR/Bilski/Mayo* [32][33] and CAFC decisions actually go far beyond their usual decisions impacting on subordinate institutions.

This mapping namely exposes key insights as to basic questions arising in developing a much further reaching "Mathematical Innovation Theory" needed as guide to finding/developing/financing/evaluating/marketing/using, with an efficiency unknown today, useful innovations in all areas of social life – of which the Patent Technology presented here is just a first step. These insights refer to the crucial question, in what way to systematically expand an appropriate KR of a subject matter by inventive concepts such that the resulting knowledge and its KR – both about

 ISBN: 978-1-61208-329-2

the new, so "invented" resulting subject matter – solve a given hitherto unsolvable problem, i.e., these Highest Courts' hints pointed at and inspired starting developing what eventually may become and then would be called "Practical Innovation Technology". Such fundamental technologies – earlier found ones are e.g., building an acre, or a state, or a wheel, or an academy, or an electric conductor, or a computer, … – once recognized are never forgotten.

IV. CONSISTENCY AND PREDICTABILITY OF SPL TESTS

In the international arena of national patent systems (NPSes), the US Highest Courts' patent jurisdiction is just proving its leading role by adjusting the US SPL precedents to the needs of emerging technologies – by accordingly refining the interpretation of 35 USC §§ 112/101/102/103, i.e., the US SPL and hence its precedents' paradigm. This adjustment is important, as the SPL together with its precedents are one of the sources of the wealth of any economically highly developed society, such as the US one.

By KR, this adjustment phenomenon [1][5] is the following. It had started in 2007 with the Supreme Court's interpretation of § 103 in the *KSR* case; but then this seemed to be, for many patent experts, the start of a US internal law administration dispute about the distribution of responsibilities in patent jurisdiction between the Supreme Court and the CAFC. Thereafter this adjustment went on refining the interpretation of 35 USC §§ 101 and 112, according to the Supreme Court's *Bilski/Mayo* decisions [32][33] and the CAFC's *Noah/CLS/Ultramercial/Accenture* decisions [35][36][40]. By today, it is clear that this dispute between the Highest US Courts is much more than a question of distribution of responsibilities in US patent jurisdiction: Namely, that it is an internationally big step forward in getting under control the fundamental problems inevitably arising in classical patent precedents due to purely "model-based" inventions – being totally mental, i.e., of intangible and invisible subject matter, i.e., no longer of "MoT" (Machine-Or-Transformation) type – typical for all emerging technologies. Hence, these problems arise not only in the US but sooner or later in any high tech depending nation, putting the consistency and predictability of its patent precedents into jeopardy, as it happened in the US. With the above decisions the US Highest Courts reacted by starting taking the paradigm underlying US SPL to a higher level of development, which enables consistent and predictable patent precedents also for emerging technology inventions – as the first Highest Courts, worldwide.

This refined US SPL paradigm – underlying the refined SPL precedents and being just a refinement of the classical paradigm – embodies a significant increase of awareness of the intricacies in patenting e.g., business, human genome, pharmaceuticals, nano, and self-replication technology based inventions, and makes it notionally significantly more precise and complete than the classical one. This is recognized easiest by the *Mayo* decision's [33] refinement. Its 3 additional key terms/notions are: "inventive concept", "preemptive", and "abstract idea". But, as to their important meanings, this decision only briefly sketched them[2)]. Yet, the meaning of the term "concept" is known in all branches of Advanced IT[1)] [3][5], of which a simplified version is sufficient here; the precise meanings of the other two key terms follow from elaborating on the outlines provided by the *Mayo* decision [33] in terms of inventive concepts and KR[2)].

In other words, the Supreme Courts' directive to use "inventive concepts" for presenting a claimed invention in increased clarity – its patent-eligible inventive concepts separated from its patent-noneligible inventive concepts, for gaining an increased understanding of its legal aspects, which also enables testing it under 35 USC §§ 101 and 102/103 in a homogenous way – impacts, first of all, the classical interpretation of § 112 to become this section's refined interpretation[2)]. This improved legal understanding of the claimed invention stimulates two important insights into it:

- Its hitherto 4 compound tests under the 4 Sections 101/112/102/103 of 35 USC may be broken down into a set of 10 elementary SPL tests, being logically and legally absolutely equivalent to these 4 intriguingly complex tests.
- Its claim (sloppily just "the claimed invention") is preemptive if and only if it is an abstract idea only, whereby the latter statement is simply testable by the not-an-abstract-idea-only test[10)].

The structurally groundbreaking insights of the preceding paragraph are elaborated on by the 2 following subsections explaining the usability advantages of this term/notion[2)] "inventive concept" and of the new just quoted terms'/notions' "elementary SPL" tests, which comprise this "NAIO (Not-an-Abstract-Idea-Only)" test[10)], testing the claimed invention's (non)preemptivity, as well as the "NANO (Novel-And-Non-Obvious)" test[11)], testing its novelty/(non)obviousness.

A. Inventive Concepts: Basic Advantages

The "misunderstanding" of the Supreme Court's term "inventive concept" among "patent practitioners" got to be removed, first [7, ftn 4.d]. Indeed, the term "concept" as such is ambiguous[3),4)]. I.e., there are

- over the millennia grown broad and sweeping meanings of the term concept, comprising different flavors, being "vagueness tolerating", i.e., colloquially addressing big issues such as "soul, god, love, truth, drama, faith, belief, …, a general principle, a plot of a story, a pattern of events " and
- by IT defined specific meanings of this term concept, also comprising different flavors, but all of them being "details oriented" – as indispensably required for enabling precise statements by them, e.g., "formal specifications", alias "mathematical models" of functional and non-functional properties of any complex system, its modules and their interactions, such as SPL prosecution or litigation cases.

  The first systems, where this notion of the term concept was used for specifying/modelling/configuring them, were large data base systems in the early 70s – then also starting from the above broad notion, but stepwise learning the lesson that it had to

be refined to enable the needed kind of precise descriptions/models of properties of their processes and data structures – and then it migrated from there into other IT research areas, such as AI (Artificial Intelligence), Semantics, KR, DL (Description Logic) [3] [4].

While the use of the above IT notion of concept mostly comes along with the awareness of the pitfalls of human thinking/speaking about complex systems, such as controversial SPL cases – and how these concepts are aggregated therein from other concepts – those who have not undergone the tedious learning process how deficient natural language and thinking often is, e.g., many patent business practitioners, knee-jerkily leap to some historic/vague notions of concept, assuming erroneously it were well-definable and understood by them.

As to these two optional meanings of the term concept the following holds. The *Mayo* decision [33] quite clearly talks of a claimed invention's "details oriented" concepts to be identified as embodied by it, i.e., uses the IT interpretation of this term "inventive concept". By contrast those, worldwide, who disagree with the US Law Maker's and US Highest Courts' broad interpretation of 35 USC § 101 insist that the *Mayo* decision's notion of "inventive concept" uses the above historic/colloquial very vague meaning of this term, although their consequential argument that the Supreme Court had asked for the claimed invention's SINGLE inventive concept evidently contradicts the *Mayo* decision [33] – implying that this insistency strangely claims this were the proof that the US Supreme Court's whole line of *KSR/Bilski/Mayo* decisions [32][33] were untenable, i.e., its breadth of interpreting § 101.

But, if the above "details oriented" notion of the *Mayo* decision's [33] inventive concepts is accepted, a fundamental question remains. Namely, what then are such inventive concepts precisely – defined in terms of the person of pertinent ordinary skill/creativity, comprising some Advanced IT knowledge? This crucial question is answered by the following definition.

Definition: An "inventive concept" of a claimed invention is a notion[2)] disclosed by the claimed invention's specification, the meaning of which meets also the usefulness requirement stated by 35 USC § 101[5),6)].

An inventive concept hence comprises the qualification of its meaning alias its pragmatics to be patent-eligible or not. While it is an indeed trivial mental/fictional construct – after one has understood it – it nevertheless is extremely helpful for clearly presenting and understanding the SPL construct of ideas. This becomes evident after having the following 3 bullet points clarified some basic features of inventive concepts.

- An inventive concept of a claimed invention is not only one of its "technical facts", as disclosed by its patent (application)'s specification, but also the "legal fact" logically underlying it therein. Thus, an inventive concept is a claimed invention's legal fact establishing its respective technical fact, i.e., represents a notional tuple. Inventive concepts hence are artificial notions representing the mental – jointly legal and technical – building blocks of any patent. Every patent business practitioner actually does practically use them every day, when thinking about a patent, also if hitherto not having been aware of them – he/she simply has no alternative but to use these inventive concepts – though normally he would think, at a point in time, about just one of these components of an inventive concept.
- Another evident question seems to pose the relation between terms and inventive concepts in patent precedents, as terms are actually explicitly used in patent specifications' wordings, but inventive concepts hitherto usually not. But the *Mayo* decision's [33] requirement statement for them implies that inventive concepts need not show up explicitly in patent specifications' wordings. *Mayo* [33] implies even stronger: The names of inventive concepts may be freely chosen by the person analyzing the patent at issue to be self-descriptive in natural language (of the person of ordinary skill/creativity). Though, in the future, it would greatly facilitate interpreting a claim claiming an invention, if its specification would explicitly describe the inventive concepts it is made-up of, e.g., in a short section therein of its own.
- Inventive concepts may be compound or elementary. Using a claimed invention's compound inventive concepts when testing it under the SPL is often misleading; then disaggregating them into conjunctions of elementary inventive concepts is indispensable [5]. But, there are several reasons, why for many claimed inventions – especially model-based ones – also not all their technical elementary facts are suitable for its inventive concepts and/or why the sequence of discussing their disclosures matters [5].

Going beyond these clarifications – in testing a claimed invention under SPL – inventive concepts have primarily two advantages over terms, which make this next to trivial refocusing (of the use of the two mental instruments at issue) on "inventive concepts" instead of on "terms" extremely rewarding: This refocusing comes along with intuitively getting/understanding the "SPL construct of ideas" and hence testing a claimed invention therein. Firstly, to an inventive concept usually may be given a self-descriptive name (just as to an atomic concept in DL) unless this is superfluous because the inventive concept's meaning is known under its term's name[2)] to the person of pertinent ordinary skill/creativity. And secondly, an inventive concept's meaning is stated as a useful property of an element – while a term's often identifies a meaning[2)] specified by a negation of a useful property. The first advantage is evident, the second one explained by the next paragraph.

For showing that a claimed invention meets all §§ 101/112 requirements, the

- classical claim construction assumes that the inventivity[7)] of this claimed invention becomes apparent to patent lawyers/examiners/judges by its limitations – ignoring that in their brains, limitations

alone have difficulties to build up respective animate subcortically controlled recognition processes alias "intuition" as to this claimed invention, because limitations totally unnaturally are negations of the properties of this focal object – whereas

- the refined claim construction automatically engages, by its inventive concepts, these patent practitioners' such intuitions while drafting/analyzing/defending a patent's claimed invention – as these inventive concepts expose their contributions to the claimed invention's total usefulness[8)] in a natural way, which makes it for the patent practitioners' brains significantly simpler to build up respective animate subcortically controlled recognition processes of properties of the focal object. This process is stimulated by the brain, as it automatically recognizes that these positive properties are those meanings, with the negations of which it was struggling before.

Such psychological phenomena – psychological preferences, when seeking understanding and/or working with some information, of assuming alleged congruities over concluding analytically, i.e., jumping at a whole over building up this whole – are well known.

This invocation of the patent professional's intuition when testing a claimed invention under SPL does not only counteract any pretence of illegally broadening of terms' meanings by the meanwhile really sophisticated misuse of the BRI guideline [14] and the *Markman/Phillips* decisions [38][39] it is based on [5], but it also animates the sharpness of a patent business professional's ability as to criticism and creativity, thus increasing the comfort and efficiency of his/her work. This makes the refined claim construction based on the claimed invention's inventive concepts by far superior to the classical claim construction based on solely the terms used by the claim's wording.

Thus, in total, there are strong reasons for this superiority of the *Mayo* decision's view of basing the granting of patents, in particular those for emerging technology inventions, on these claimed inventions' inventive concepts – more precisely: for focusing the patent-eligibility and patentability tests of a model-based claimed invention on its inventive concepts, instead on solely its terms. Although inventive concepts as well as terms are subject to interpretations by the person of pertinent ordinary skill/creativity, there is the just outlined and undeniable better appreciation by a human brain of the meanings of inventive concepts than of the meanings of terms[9)].

Evaluating the before said in this subsection: If the notion of inventive concepts at its beginning seemed sophisticated, this only shows how complex the thinking underlying testing a claimed invention under SPL actually is – often not at all recognized by those contemporary discussions clinging to using solely terms to this end, which bars their insights into this complexity. Such consistency and predictability creating insights, as described by the final part of this subsection, are clearly enforced by the Supreme Court by requiring using inventive concepts to this end, i.e., to use them in construing a claimed invention's claim construction as described above.

B. Elementary SPL Tests: Basic Advantages

The *Mayo* decisions [33] inventive concepts also invite breaking down a claimed invention's 4 compound tests under the 4 §§ of 35 USC 101/102/103/112 into 10 elementary "SPL tests" [5][11]. These are scientifically developed, hence their principles are freely available – potentially not their particular applications as "FSTP tests", as they are subject to patent applications.

Advantages these elementary 10 SPL tests offer to patent professionals are outlined below, after first identifying, which "aspects" of a claimed invention's refined claim construction they check – in IT language: what "requirements" alias "concerns" stated by the 4 §§ of 35 USC they may state as being met by the claimed invention – being patent-eligible and patentable iff it passes all 10 SPL tests, i.e., a claimed invention's 4 tests under these 4 §§ is thus refined into 10 tests of

- § 112 for the well-definedness of this claimed invention's inventive concepts, i.e., of their all 1) disaggregation into elementary inventive concepts, and of their 2) lawful disclosures, 3) definitiveness, and 4) enablement;
- §§ 102/103 for the novelty and nonobviousness of this claimed invention, i.e., of its 7) novelty and nonobviousness by its "NANO test", based on its 5) independent and 6) non-equivalent inventive concepts;
- § 101 for the patent-eligibility of this claimed invention, i.e., of its being 8) not a law of nature or natural phenomenon only, 9) not idempotent, and 10) not an abstract idea only by its "NAIO test", i.e., of its claim being nonpreemptive.

The dramatic support of a patent professional working on a patent and its claimed invention – provided by an IES [7] leveraging on these 10 SPL tests – comprises,

- automatically prompting him/her through all steps of exploratively checking, whether they meet these 10 SPL as well as 4 §§ 35 USC respective requirements/concerns by having him/her interactively input or by automatically computing these statements and confirming them (= facts screening), and
- their automatic real-time affirmative execution (= facts transforming) on the user's request. This execution provides to him/her controls for i) access to all information existing in any SPL test of the claimed invention, and ii) crossover from any one item in its patent to its peer in any one document and to any one of their relations, tests respectively their single steps, multiple presentations thereof, …. (and back), and iii) all these services anytime in "dialog real-time".

In so far, the US Highest Courts have taken, by their patent precedents, SPL precedents to a level of development, on which the today notorious problems with emerging technology are overcome, i.e., with model-based inventions. The evolution of classical claim construction to this higher level of evolution – represented by the refined claim construction implied by the Supreme Court's above line of

groundbreaking decisions, which in turn induced the above 10 semi-automatic SPL test – will dramatically increase the productivity of all patent practitioners, be they inventors, research managers, examiners, lawyers, licensers/-sees, investors, or judges [9].

Out of the 10 SPL tests, the NAIO[10] and the NANO[11] test are of particular interest, the latter as to the *KSR* (§102/103) case, the former as to the *Bilski/Mayo/CLS/ Myriad/Accenture* [32]-[35][40] (mostly erroneously understood as being plain § 101) cases. As claim construction up to § 112 is just becoming an issue for the Highest Courts again, the role of the remaining 8 tests will shortly encounter more interest, too, as removing the above loopholes of the *Markman/Phillips* decisions[38][39].

V. THE IES USER INTERFACE

The only prerequisite for applying these 10 FSTP/SPL tests, either exploratively or reconstructively, is appropriately having marked-up all documents involved in a PTR^C's (Problem of TT.0 (Technical Teaching) and RS (Reference Set)) analysis [6]-[9][11]. While this would only rarely happen with the doc.CTs (ConText DOCuments of a PTR), the needs of additional mark-ups in doc.i's are frequently encountered during an explorative FSTP test's iterative executions, in particular if the tested PTR's RS is expanded by a further doc.i or the definition of a cr-C (CReative Concept) is changed [8][10][11]. Such mark-ups will be based on some of the XML (Extensible Markup Language) derivatives currently discussed to this end. Independently thereof, the IES'es UI (User Interface) concisely models the requirements of the NPS'es SPL, of its precedents[10),11)], and potentially also of some application area specificities (such as of communications, software system, lifecycle, DNA, nano, selfreplication, … technologies, including their above quoted pragmatics decisive for their social success).

Figure 1 shows 4 separate windows of the IES'es UI, simultaneously mapped onto one or several screens, in total called "survey window". These 4 windows are identified by their names "o-doc.i", "facts.i", "plcs.i", and "tests" in their top left corners. They serve for the knowledge representations of/about primarily i) the original document.i's in o.doc.i, ii) their "inventive concepts" on their o/BAD/BID-KR-levels (o=Original, BAD=Binary, Aggregated and Disclosed, BID=Binary, Independent and Disclosed) in facts.i, iii) their "patent logic carrying semantics" items on these levels in plcs.i, and iv) the 10 FSTP/SPL tests. They may be arbitrarily zoomed, positioned, and overlapped within the survey window. The graphical items within these 4 windows basically represent inventive concepts and/or their components in these KRs. The lines between these items represent their peering in any KR and indicate interrelations between them. Their arrowheads are exemplary for browsing between them – i.e., all lines may have two arrowheads.

This UI presents in its survey window – functionality top-down in telegram style – the following:

- The middle "tests" window provides access to the use of the claimed invention's inventive concepts by any FSTP test – skipped here but shown to the user on its request by the ANC (anticipates/not-anticipates-and-not-contradicts/contradicts) matrix columns, represented by test specific matrix lines describing in short hand this use.
- On the left lower side, in the "o-doc.i" window, two stacks are shown: Of 3 peer doc.i's (their mark-ups comprising all potential cr-Cs' disclosures) and of doc.CTs (their mark-ups comprising all le-Cs (LEgal-Cs), e.g., law/precedents items to be applied where appropriate, respectively additional information potentially belonging to it, such as explanations/confirmations/warnings/…, all of them independent of the doc.i≠doc.CT, i.e., any pragmatics independent of the TT.i's).
- On the right lower side the "facts.i" window shows a stack of 3 doc.i's/TT.i's – for simplicity assuming doc.i comprised just a single claim, otherwise any claim would be one sub-plane. Per TT.i its elements' (= rectangles) properties (= ovals) are arranged on its plane in concentric "KR rings", delimited by dashed lines. The large/small ovals represent BAD/BED-in-Cs, o-in-Cs are parts of their elements' rectangles. A BED-in-C shows some of its relations to other in-Cs and what all their KR details are, e.g., where in a claim in "o-doc.i" or "test" it is involved in and where in the problem to be solved by TT.i in these windows. The encoding of all KR details and the tests is shown in "plcs.i".
- The "plcs.i" window on top is the IES "brain". It stores all in-Cs' peerings of all subject matter items (cr-Cs) with all legal items (le-C) and all their interrelations. It indeed shows everything the user's brain knows about the PTR: all its objects, as well as all potential and/or actual associations between them, and all the sophisticated structures potentially appended to them (not shown here for brevity).

The quick and total overview about all the documents and their mark-ups of all subject-matter items respectively legal/pragmatic items respectively all to these mark-ups related in-Cs (in o-/BAD-/BID-KR) in doc.i is provided to the user – be it an inventor or patent lawyer or examiner or judge – by the two bottom windows, whereby these stacks' items may be presented non-overlapping and then show the interrelations between their peer items.

The PTR independent counterparts to the cr-Cs, the le-Cs, potentially making cr-Cs to in-Cs are the items on the right of the top window. The respective doc.CT's, their mark-ups, and their items in the plcs.i-window are absolutely the same for all PTRs (in particular for their TT.0s' claim constructions). For a given PTR, all such peerings and the explanations why they happened are the items on the left of the top window.

As usual, the user would access any item of interest in any window by clicking on it and zooming into one or several of its interrelations. Thereby simultaneously several of such interrelations as well as concatenations of them may stay displayed and zoomed as momentarily of interest for the user. What actually is – or ought to be – of interest to him may be determined by him or an additional application not elaborated on, here.

 ISBN: 978-1-61208-329-2

The "test" window, providing access to all FSTP tests (in all their various configurations), is highly configurable for the various needs of the user in particular in real-time confirmation mode for being able to appropriately guiding the user through a test.

In total: The survey window provides e.g.,

- immediate access to ALL information/knowledge existing in any one FSTP/SPL test of the claimed invention.
- immediate and instant crossovers between ALL KRs of ANY ONE subject matter and/or legal item.
- immediate crossover from ANY ONE subject matter item to ANY ONE of its relation – and back.
- immediate crossover from ANY ONE relation to its peer in any TT.i – and back.
- immediate crossover from ANY ONE test using an item or relation to any test and its use thereof.
- immediate information about the impact of a change performed in one of the 4 windows on the other ones.

and all these services instantly, i.e., in "dialog real-time", i.e., necessarily automatically.

An important other powerful feature of the UI of the IES had to be completely skipped in this paper [24][25], due to space limitation. It is its capability to translate all ASTs (Arguable SubTests) of a claimed invention into LACs (Legal Argument Chains) in a variety of multimedia presentations – including natural voice presentations, e.g., using the user's voice – and to enable the user to easily select and control any LAC's presentation in realtime as needed by the user, potentially as to the logics of testing even suggested by the IES.

VI. CONCLUSIONS

No system like the IES exists today – or could only have been thought of without the insights of Mathematical KR presented in [5] and the informal KR ex- or implicitly used in our publications addressing the community of patent law professionals. The kind of KR induced primarily by the US Highest Courts SPL precedents enabled transforming it into this Advanced IT system. While the current IES is only a prototype, even its final version would not yet be capable of acting as an autonomous innovations tracing system, but will be able only of supporting such tracing activities. It is designed as just as a versatile evaluation system of innovations completely identified and specified already – though an amazingly powerful one.

This is made evident in particular by its capability to semi-automatically generate in real-time all argument chains legally correct and technically as correct confirmed – in user controllable verbosity and user controllable multimedia presentations – that may be of actual interest in an invention's test whether it satisfies SPL[24][25].

REFERENCES

[1] S. Schindler, "US Highest Courts' Patent Precedents in Mayo/Myriad/CLS/Ultramercial/LBC: 'Inventive Concepts' Accepted – 'Abstract Idea' Next? Emerging Technology Patents: Intricacies Overcome." 2013, www.FSTP-expert-system.com [accessed March 2014].

[3] R. Brachman, H. Levesque, "Knowledge Representation and Reasoning", M K, 2004.

[4] F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. Patel-Schneider, "The Description Logic Handbook", CUP, 2010.

[5] S. Schindler, "Mathematically Modelling Substantive Patent Law (SPL) Top-Down vs. Bottom-Up", Yokohama, JURISIN-2013, , www.FSTP-expert-system.com [accessed March 2014].

[6] S. Schindler, "The FSTP Expert System", 2012, Patent Application, www.FSTP-expert-system.com [accessed March 2014].

[7] S. Schindler, "The Innovation Expert System, IES, and its PTR Data Structure", 2013, www.FSTP-expert-system.com [accessed March 2014].

[8] J. Schulze, "Tech. Rep. #1.V1 on the '882 PTR and UI of the IES prototype", www.FSTP-expert-system.com [coming soon].

[9] S. Schindler, "Patent Business – Before Shake-Up", 2013, www.FSTP-expert-system.com [coming soon].

[10] S. Schindler, "Amicus Brief in LBC v. Philips", to CAFC, 2013, www.FSTP-expert-system.com [accessed March 2014].

[11] S. Schindler, "Inventive Concepts Enabled Semi-Automatic Tests", 2013, Pat. Appl., www.FSTP-expert-system.com [accessed March 2014].

[14] USPTO/MPEP, "2111 Claim Interpretation; Broadest Reasonable Interpretation [R-9]" , www.FSTP-expert-system.com [accessed March 2014].

[18] S. Schindler, "Amicus Brief in Alice v. CLS.", to Supreme Court, 2013, www.FSTP-expert-system.com [accessed March 2014].

[19] S. Schindler, "Amicus Brief in WildTang. v. Ultramercial", to Supreme Court, 2013, www.FSTP-expert-system.com [accessed March 2014].

[21] K. O'Malley, CAFC, IPO, 2013, www.FSTP-expert-system.com [accessed March 2014].

[23] S. Schindler, "Amicus Brief as to computer-implemented inventions, CIIs", to Supreme Court, 28.01.2014, www.FSTP-expert-system.com [accessed March 2014].

[24] S. Schindler, "Semi-Automatic Generation Customization of All Correct Legal Argument Chains (LACs) in an Invention's Substantive Patent Law Test", 2014, Pat. Appl., www.FSTP-expert-system.com [coming soon].

[25] S. Schindler, "Automatic Derivation of Legal Argument Chains (LACs) from Arguable Subtests (ASTs) of a Claimed Invention's Test for Satisfying Substantive Patent Law (SPL)", 2014, www.FSTP-expert-system.com [accessed March 2014].

[32] "Bilski v. Kappos" - Supreme Court, 2010, www.supremecourt.gov/opinions/09pdf/08-964.pdf [accessed March 2014].

[33] "Mayo v. Prometh." - Supreme Court, 2012, www.supremecourt.gov/opinions/11pdf/10-1150.pdf [accessed March 2014].

[34] "AMP v. Myriad" – Supreme Court, 2013, www.supremecourt.gov/opinions/12pdf/12-398_1b7d.pdf [accessed March 2014].

[35] "CLS Bank v. Alice" - CAFC, 2013, www.cafc.uscourts.gov/images/stories/opinions-orders/11-1301.Opinion.5-8-2013.1.PDF [accessed March 2014].

[36] "Ultramercial v. WildTang." - CAFC, 2013, www.cafc.uscourts.gov/images/stories/opinions-orders/10-1544.Opinion.6-19-2013.1.PDF [accessed March 2014].

 ISBN: 978-1-61208-329-2

[38] "Markman" - CAFC, Supreme Court, 1996, www.courtlistener.com/cafc/4yqk/herbert-markman-and-positek-inc-v-westview-instrum/?q=&case_name=markman+v.+westview&stat_Precedential=on&order_by=dateFiled+desc [accessed March 2014].

[39] "Phillips v. AWH Corp." - CAFC, 2005, www.cafc.uscourts.gov/images/stories/opinions-orders/03-1269.pdf [accessed 2014].

[40] "Accenture v. Guidewire" - CAFC, 2013, www.cafc.uscourts.gov/images/stories/opinions-orders/11-1486.Opinion.9-3-2013.1.PDF [accessed March 2014].

1) "Advanced IT" is a generic term for IT areas such as AI, Semantics, KR, DL, NL.

2) A term together with its meaning is denoted as "notion", its term being the notion's name. A notion hence comprises a definition of the meaning and its name/term. This meaning may be a property of something, e.g., of an element quoted by a claim. Sometimes the meaning of a notion may also be taken as its name.

A notion is called an "inventive concept", if its meaning represents patent pragmatics, i.e., if its meaning also serves the purpose to define the "patent monopoly granting pragmatics, pmgp" determined by the Parliament or the Supreme Court, i.e. if it to this end puts one or several properties' limitations of the invention (of some broader set of such properties' limitations) such that it specifies an item of the invention's "§ 101(usefulness)" – additionally to its "§ 102(novelty)" and /or "§ 103(nonobviousness)", as explained in more detail later.

But already here is evident that an "inventive concept" as such – while being a mental/fictional construct, just as any notion – in no way may be understood as an "abstract idea", as suggested by some patent business practitioners. Also an "abstract inventive concept" cannot be thought as it is a clause contradictory in itself. A concept, just as an inventive concept, is always a concrete and named representative of something.

3) These new terms' meanings are here defined only in natural language of the person of pertinent ordinary skill and creativity, i.e., their Mathematical KR definitions – as far as today possible – are here left away [5], though its thinking is occasionally used.

4) In patent law language the meaning of a term[1)] is often called "limitation". One meaning of a term/name may be split into two parts and/or subject to different pragmatics. An inventive concept of a claimed invention is, by the *Mayo* decision, in the US patent precedents always a notion of this "meaning-tuple" kind, whereby any meaning-tuple component may be subject to different pragmatics alias meaning-qualifications. I.e., an inventive concept always comprises a legal concept and a creative concept [5].

5) Notwithstanding that any inventive concept of a claimed invention represents an item of usefulness and creativity of this claimed invention up to § 101, by the Supreme Courts' interpretation of the Constitution it is additionally to determine, whether an inventive concept – as being of an exceptional kind/quality/pragmatics, namely of "law of nature", of "natural phenomenon", or of "abstract idea" – is by § 101 non-patent-eligible, to be described by this inventive concept's legal concept.

Put in Mathematical KR clarity [5]: An inventive concept's "patent monopoly granting pragmatics, pmgp" qualifies this inventive concept as contributing to the claimed invention's pmgp-height (Q^{pmgp}) over prior art and pertinent skill (i.e., as contributing by 1 not only to its Q^{plcs} (plcs = patent law carrying semantics), determined by the claimed inventions NANO test, but also to its Q^{pmgp}). I.e., an inventive concept may contribute by 1 solely to the claimed invention's plcs-height alias semantic height Q^{plcs} over prior art and pertinent skill, but not also to this claimed invention's pmgp-height [6]. In general hence holds $Q^{pmgp} \leq Q^{plcs}$. It evidently is the determination of any inventive concept's pmgp that performs the separation of concerns discussed above [5].

If the claimed invention is not well defined, i.e., does not pass one of the other SPL (e.g., the NAIO) tests, running its NANO test, i.e., determining its Q^{plcs}, is meaningless, anyway.

6) For representing its pragmatics, an inventive concept identifies one or several properties' limitations (of the claimed invention's total set of such properties' limitations) putting it/them such as to specify its "§ 101(usefulness)" – in addition to its "§ 101/102(novelty)" and "§ 103(nonobviousness)". I.e., a well defined claimed invention embodies no inventive concept that does not meet the usefulness requirement stated by § 101, otherwise this claimed invention were not well defined [5] and the question as to its inventivity[6)] were obsolete.

7) The legal meaning of the notion "inventivity" of a claimed invention – i.e., embodied by it – is represented by this claimed invention's total set of limitations of all its elements, i.e., of all elements of the claim claiming the invention described by its specification. The psychological meaning of the notion of inventivity as such, counted in the number of inventive creative ideas it embodies, is not elaborated on, here. It has been clarified in [7], based on a pertinent German Highest Court decision, by the BGH (BundesGerichtsHof).

8) The legal meaning of the notion "usefulness" of a claimed invention – i.e., embodied by it – is, just as its inventivity[6)], represented by this claimed invention's total set of limitations of all its elements. Consequently, from the definition of the inventive concepts making-up this claimed invention follows [5] that any one of them contributes – by its contribution to the total set of limitations of the claimed invention – equally to the claimed invention's usefulness, too, as required by § 101.

The *Mayo* decision invokes, by its inventive concepts, for its refined claim construction for a claimed invention this additional "contribution to its usefulness" minded view at its claimed invention's inventive concepts. This "contribution to the claimed invention's usefulness" minded view at inventive concepts changes nothing with these inventive concepts' and/or their terms' hitherto only "contribution to this claimed invention's total limitations" minded pragmatics – i.e., nothing is changed for the more basic classical claim construction for this claimed invention. It evidently is this additional "contribution to this claimed invention's usefulness" minded pragmatics of the inventive concepts, by which the Supreme Court achieves an increased purposefulness of its refined claim construction.

9) Whether the earlier exclusively used set of "terms" and their error prone interpretations/limitations of a claimed invention ought to be, in its refined claim construction, eventually completely replaced by a set of inventive concepts legally equivalent to them – and their more explicit names and more target-oriented pragmatics, i.e., their better as context sensitive guided interpretations – and hence making the former set redundant, needs no discussion, yet. Such redundancy is often avoiding committing errors of any kind and then to be preserved.

10) The "NAIO test" of a claimed invention was originally suggested incompletely, as ignoring its potentially being pathological, which is fixed here. It is not clear at all, whether a pathological TT.0 exists – which also applies to the NANO test. For a more complete and detailed explanation see [5].

The complete NAIO test – just as the NANO test – would start with disaggregating the compound inventive concepts of the claimed invention into the respective sets of BED-in-Cs, then reduced to maximal sets of BID-in-Cs thereof.

For brevity only considering the BID-cr-Cs of the BID-in-Cs, the NAIO test comprises 4 steps:

1) verifying the TT.0's specification of the patent (application) discloses a problem, P.0 (Problem), described to be solved by the claimed invention/TT.0, the latter described by its refined claim construction's inventive concepts, {BID-cr-C};

2) verifying, using these BID-cr-Cs of **1)**, that the so described TT.0 actually solves this problem of **1)**;

3) verifying for any one KR^ of TT.0 that this problem of **1)**, P^, is in KR^ not solved by any TT^*, derived from TT^ by ignoring therein one of these BID^-in-Cs completely or relaxing its limitation by reducing its TS(d(BID^-in-C)) (TS=Truth Set, d=Domain, in-C=INventive Concept), i.e., not solved by any TT^* $<^{TT}$ TT^ ("$<^{TT}$" being the "less in-C-limited than" relation between TTs resp. Ps in KR*) – all these verifications to be confirmed by the person of posc (person of Pertinent Ordinary Skill and Creativity);

4) stating, if all steps in **1)-3)** are executed successfully, that the so described claimed invention/TT.0 is "not an abstract idea only"

of this problem's solution, otherwise that it is only an "abstract idea" of this problem's solution.

11) The preamble of this NANO test and its respective first steps are the same as the ones described in the above NAIO test – for the given PTR, for its "anticipation combinations, ACs" and their "1 concept modifications, 1-cMs" for anticipating TT.0 each, as well as for an ind{BID-cr-C} (ind=INDependent) describing PTR's TT.0 (For a more complete and detailed explanation see [5]) – and thereafter comprises the steps:

1) the user generating the ANC matrix for all TT.i ϵ RS, i>0, its columns representing the BID-cr-Cs;

2) the user generating, for any entry in the ANC matrix the technical and/or legal justification;

3) automatically deriving from the predicates $\underline{X}$.i.n, 1≤n≤N, 1≤i≤I, and the ANC matrix an AC anticipating TT.0 with a minimal Q^{plcs} of 1-cMs;

4) automatically delivering Q^{plcs} as TT.0's semantic height over RS and <Q^{plcs}, {all justifications for AC's 1-cMs}>.

 ISBN: 978-1-61208-329-2

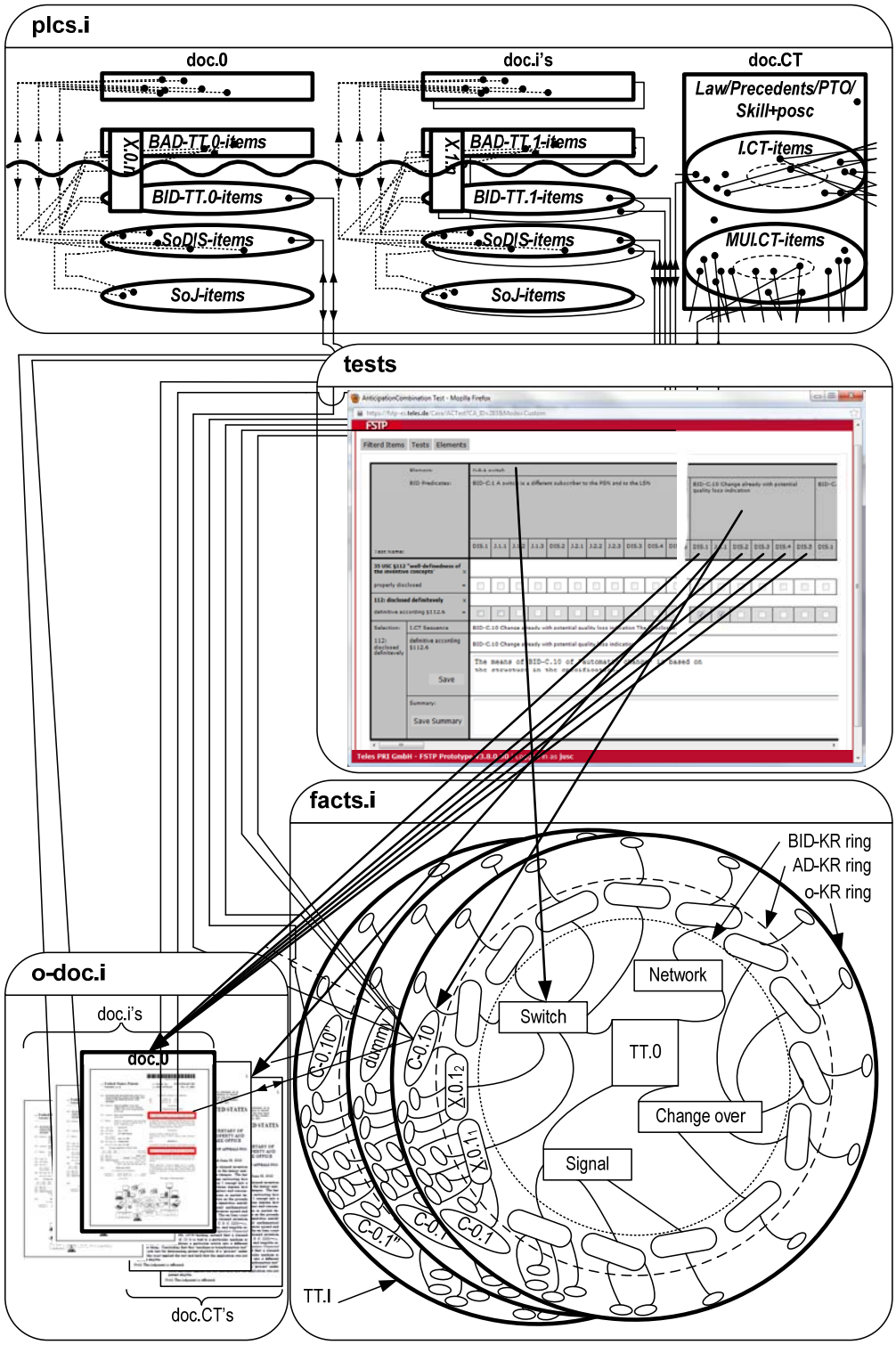


Figure 1: 4 separate UI windows of the IES

An Intelligent Robotic Engine Using Digital Repository of the DSpace Platform

Rafael Luiz de Macedo, Elvis Fusco
Computing and Information Systems Research Lab
Center University Euripides of Marilia
Marilia, Brazil
e-mails: {rafaelldemacedo@gmail.com, fusco@univem.edu.br}

***Abstract* - This article presents an Artificial Intelligence Markup Language as a set of eXtensible Markup Language, which is able to represent and relate expressions in natural language. This will allow the creation of intelligent robotic engines capable of maintaining a simple dialogue; however, robotic engines are limited in the amount of questions they can answer, thus, failing to maintain a simple dialogue for a long time. This paper solves the problem of the limitation of robotic engines with the development of a software that will perform a search in the repository to find information on files for generating new questions and answers for robotic engines.**

Keywords-Artificial Intelligence; Engines Robotic; Semantic Web; Repository.

I. INTRODUCTION

Current technologies have increasingly allowed the perception that it is possible to make a machine be intelligent enough to answer questions asked by the user through robotic engines built with AI (Artificial Intelligence).

The ways in which humans and computers normally communicate are very different. Humans tend to spend a lot of time in chit-chat and informal dialogue with little or no effect. Computers are known to give accurate answers, true and logical. The rate of exchange of information of most human dialogues is very low, no more than 1kbit per second, but the computer communication is faster. Alicebot/AIML is an attempt to bridge this divide [2][10][18][19][20].

One of the languages that enables the creation and development of these intelligent robotic engines capable of maintaining a dialogue is the **A**rtificial **I**ntelligence **M**arkup **L**anguage (AIML) [10][12][19][20].

Robotic engines developed in AIML language have a great limitation on the amount of questions they can answer; the main reason is that they were developed based on one or more contexts and are not able to dynamically extend the amount of questions and answers in an information bank.

With the emergence of the Internet, information is increasingly accessible, so that, only with a machine that has access to the Internet it is possible to search for new information.

The amount of information that is available daily on the Web has facilitated an increasing access for people. Consequently, the number of online stores that have articles, monographs, newspapers, magazines and other types of digital files has grown. A platform that offers that kind of repository is the DSpace [14][15][17], which is rather used as an open source platform and has several tools available to be used in the repository, such as search engine, to access files through search requests to Web servers.

The main objective of this paper is to develop a metadata parser software of digital repositories of DSpace platform, which performs searches of digital files in repository metadata to find the files related to the topics the users are seeking [13][14][15][17]. The Analyzer software will generate new knowledge database of questions and answers for the engines on top of these robotic metadata from the repository; this new information might be useful for debugging purposes, or resolving the limitation of robotic engines in only answering questions that are contained in the context that were developed.

For the development of the metadata Analyzer software, the digital repository of Center University Euripides of Marilia – UNIVEM, called Univem Aberto [1] will be used. This is because the digital repository of Univem Aberto is based on the DSpace platform.

This article is structured as follows. Section 2 presents a preview of AIML. Section 3 shows the importance of institutional repository (university). In Sections 4 and 5, the types of metadata utilized in this work and in the repository of platform DSpace are shown. In Section 6, the results obtained in this work are presented. Section 7 presents a conclusion of the article.

II. METHODOLOGY

In this paper, the methodology used for the development is divided into a few steps, as follows: (i) the bibliographical study of languages AIML and XML, (ii) a survey of the types of institutional repositories platforms (iii) the development of the application, and (iv) the related works, testing and validating the results.

A. The Bibliographical study of languages AIML and XML

The study of markup languages are required for the knowledge on how intelligent robotic engine work and for the development of the application proposed in this paper.

B. The survey of the types of institutional repositories platforms

Among the types of repository platforms, we will choose only one to be used as a basis of collecting scientific files, and we specifically use an institutional repository to develop the application on this work platform.

 ISBN: 978-1-61208-329-2

C. The development of the application

In this work, an application was developed that searches files in the server of an institutional repository; for the development of the application, the programming language Java was used.

D. Related works

The related research is an important step. The articles and documents found helped in understanding markup languages AIML and XML. The study also helps to evaluate the results obtained in this work.

E. Testing and validation of the results

The application developed in this work was tested in a real scenario to search files from the institutional repository at Univem Open of Center University Euripides of Marilia.

III. ARTIFICIAL INTELLIGENCE MARKUP LANGUAGE

The AIML language is a set of XML elements capable of representing and linking natural language expressions, allowing the creation of engines capable of maintaining a simple dialogue. Each set of AIML elements has one or more elements referred to as category. The categories are developed on top of a context; a category is formed by all the elements and template pattern, which are the elements responsible for interpreting the message sent by the user and send a reply message back to the user.

The Chatterbots (Chat = Chat and bot = Bot), as currently known, are robotic intelligent engines that interact with users by means of questions and answers. Currently, the amount of intelligent robotic engines used in online consultations and through voice communication devices is growing.

Companies have used these robotic engines to perform an auto-customer service for the purpose of improving the attendances and saving time. The use of these robotic engines is because they recognize the issue the customer wants to deal with in the company and redirect the customer towards the competent sector. Robotic engines are used for online sessions and for telephone calls.

A very well known robotic engine used to educate people in caring for the environment is the Ed robot, belonging to the company Conpet, along with the support of PETROBRAS [21].

The emergence of the AIML language was the result of the development of the software of PLN known as The Artificial Linguistic Internet Computer Entity (ALICE) [2]; the license of AIML is under the GNU GPL.

The robotic engines developed in AIML have two modules, namely, language and engine. The language is all the knowledge that these engines have, natural language and information developed in AIML language. The engine performs the communication between the two languages, natural and AIML, the recognition of the information contained in the two languages, so that the robotic engines can recognize written questions made by humans and answer these questions.

Robotic engines are limited in the amount of questions they can answer because the robotic engines can not reshape new questions dynamically, as a human. With this limitation, the robotic engines do not maintain a long dialogue with a human.

Alan Turing (1912-1954), in his famous essay Computing Machinery and Intelligence [12], suggested that, instead of asking whether machines can think, we must ask if machines can undergo a behavioral intelligence test, which came to be called the Turing test [11].

The Turing Test is performed by means of questions and answers made to a machine with robotic and intelligent engine for a human; the machine and the human are located in separate rooms. An evaluator will hold the questions for these two rooms, but the evaluator does not know in which room is the machine with the robotic engine.

According to the responses that the evaluator will receive, one can find out which one is the machine room and which one is the room with the human. Intelligent robotic engines that recognize natural language can't create new questions and answers autonomously, causing a disadvantage by not being able to recognize a question that is outside of the knowledge or repeat an answer already used in different questions.

The machine that can pass the Turing test is considered smart. However, currently, there are no machines capable to go through this test.

The AIML language is open source [2][19][20], thus enabling the use in research of improvement in creations of robotic intelligent engines able to recognize information written in natural language.

To create a robotic motor using the artificial intelligence markup language, you need the use of a development platform. There are several open sourceplatforms, paid and developed in various programming languages such as Java, Python, C and C + +.

The platform ProgramD is the most used platform when it comes to developing robotic engines in AIML language and the most complete resource on the language, because the platform being developed in Java language and be open source [4].

A. Funcionality of engine with AIML

In this subsection, we present some commands (elements) of AIML used in the creation of a new base of questions and answers for robotic motors.

In the creation of a new base of questions and answers, it is necessary to utilize standard elements of the language AIML. The standard elements are: category, pattern, template, star, aiml and xml.

Figure 1 shows the base of questions and answers organized with elements of AIML language, utilized for robotic engines.

A new database of questions and answers is started with elements XML and AIML. These elements define which versions of XML and of AIML language is being used at the base [20].

The category element is a set of questions and answers; each category has elements pattern and template.

The pattern element is where a possible question is declared that the human can ask to chatterbot.

 ISBN: 978-1-61208-329-2

The template element is localization of an answer of a question declared on a pattern element.

```
<?xml version='1.0' encoding='ISO-8859-1'?>
    <aiml version='1.0.1'
xmlns='http://alicebot.org/2001/AIML-1.0.1'
xmlns:html='http://www.w3.org/1999/xhtml'
xmlns:xsi='http://www.w3.org/2001/XMLSche
ma-instance'
xsi:schemaLocation='http://alicebot.org/2001/AI
ML-1.0.1
http://aitools.org/aiml/schema/AIML.xsd'>
    <category>
        <pattern>Hi</pattern>
        <template>Hi, okay</template>
    </category>
    <category>
        <pattern>Hi, okay and you?</pattern>
        <template>I'm fine, what is the
name?</template>
    </category>
    <category>
        <pattern>My name is <star></pattern>
        <template>My name is
Jose.</template>
    </category>
    </aiml>
```

Figure 1. Based of questions and answers.

IV. DIGITAL REPOSITORY

With the need to disseminate the works produced by institutions, without relying on a Publisher to publish the work, institutions began to spread this information on the Internet on their own by creating several tools called digital repositories [13][14][15][17].

The first digital repositories by institutions began to be developed in 2002; the digital institutional repositories began to play the role of the Publisher.

Digital repositories have the main functions of storage, dissemination and durability of digital files, making it easy to access the files that are submitted to these tools.

The institutions have used the repositories of different platforms, disseminating their scientific journals, monographs, Ph.D. theses and other work carried out in the institutions over 11 years; after the development of the first digital repository, repositories used is the platform DSpace.

DSpace is an Open Source software platform, responsible for the storage, dissemination and durability of digital files. It was created by MIT (Massachusetts Institute of Technology) and Hewlett-Packard [14].

The DSpace platform is geared to the academic area, the purpose of which is serving as a basis for the future development to address the long-term preservation of files and access problems [14].

The registration site of providers of OPEN ARCHIVES repositories is a table with information where digital repositories are registered; they are currently registered and the record repository #2140 is the Center University Euripides of Marilia, Univem Aberto Repository.

The Center University Euripides of Marilia-DATA has deployed in 2012 the digital repository DSpace platform for the purpose of disseminating the scientific papers, monographs, theses of master's and doctoral degrees from the faculty and other work carried out by its students, teachers and researchers.

The Institutional Repository of UnB is a set of services offered by the Central Library for the management and dissemination of scientific and academic production of the University of Brasilia. The content is publicly available; it is widely accessible [8].

To include their scientific production in the repository, teachers, researchers and students graduating from UnB must complete and sign a term of authorization and return it to the Management of Digital information (GID). This document is signed, scanned and sent along with the file by email.

With the huge amount of files that are made available in institutional digital repositories, it is necessary to standardize rules to manage and identify each file that is submitted in the repositories.

When a file is submitted for digital repository DSpace platform, some information about the file is needed.

All the information entered in the fields of metadata is used to catalog a file in the repository, using the standard Dublin Core metadata (DC) [3][13].

To integrate intelligent robotic engines developed in AIML, questions with the information about the files that the repository user is seeking should be defined. With this information collected, robotic engines integrated with metadata Analyzer software developed in this work carry out requests for search of Univem Aberto files in the repository. Metadata Analyzer software called XML2AIML treatment will apply to the information that is contained in the response from the digital repository Univem Aberto in order to generate new knowledge bases to intelligent robotic engines, in the form of questions and answers.

V. THE STANDARD METADATA DUBLIN CORE

With the great increase of digital files posted on the Internet, there was the need to develop standards that identify the exact description of each piece of information from the files, i.e., to develop metadata standards.

Metadata means data about data. The metadata are forms of cataloging all information from a file, the same way they are made in real libraries: each book has a tumble of identification, are arranged by area and alphabetically by title. The metadata are intended to document and organize digital files in a structured way, making it possible to identify files through standardized data such as author, title, and summary.

Metadata is defined here as given that describe attributes of a resource, characterized their relations, and enables its recovery and effective use and its existence in electronic

 ISBN: 978-1-61208-329-2

environment. Metadata usually consists of a set of data elements, where each element describes an attribute of the resource, its administration or use [5][9][16].

All the reasons why the indexing and cataloging are required for printed sources apply even more strongly to the metadata for electronic documents [6][7].

Digital repositories have metadata for cataloguing all files that are submitted. The standard used in digital repositories is the standard Dublin Core which are collections of metadata [6][7].

DC is not for replacing some richest models with AACR2/MARC Code, but only provides basic sets of elements of description, which can be used by catalogers or non-catalogers for a simple description of information resources [3][20].

With the files catalogued on the use of metadata, the same files end up being most used by users than the files that have no metadata cataloging, because of the ease in which the metadata provide the tools of Internet searches to find these files.

The concept of metadata is not something new, but the use of this term in digital environments and a variety of patterns and shapes is new. The bibliographic records that have been created by information workers in a long time must be regarded essentially as metadata. They provide analytical and descriptive information about an object [5].

Some different metadata standards to identify certain files are presented below:

- Government Information Locator Service (GILS) – Government information;
- Federal Geographic Data Committee (FDDC) – description of geospatial data;
- Machine Readable Card (MARC) – bibliographical cataloging;
- Dublin Core (DC) – data on Web pages, and
- Consortium for the Interchange of Museum Information (CIMI) – information about Museums.

New standards of metadata can be created according to the informational needs of an organization and contribute to documenting the data of the digital file.

These patterns can be viewed as metadata content standards, standards for exchanging data by electronic means and, in lastly, standards for data models [9].

Metadata is structured data blocks; each block contains information, such as author, title, where it was published, etc.; also, some information is a field that can be set for field name, type of field information, the format that is accepted by the field and other descriptions that identifies the information passed in this field.

Creating a metadata schema must establish a standard framework and terminology. Declarations of labels such as creator, author, sculptor or composer have little representatively if these fields, who all have the same function, cannot be mapped to the same unique concept. A form should be established, either through a list of authority or Affairs of a controlled vocabulary standard and so relationships will map out alternative ways to the established form [5].

In digital repositories, it is possible to develop software able to get files through requests made to repositories. These requests make use of DC metadata sets that are passed by the request.

VI. DUBLIN CORE

The DSpace platform enables the digital repository web server to receive requests from external media, off the server. With requests, we are able to have access to all digital files that are submitted to the repository.

When we perform a request, sets of metadata elements are used, i.e., Dublin Core is used to define what are the parameters that will be passed by the address, the values passed by all parameters will be used to identify which file, or group of files is being sought in the repository.

The request made to the web server is parsed by the repository and all files that fit the information passed by DCs will generate an XML response that is sent to the agent (software) that made the request. The software developed in this work is called XML2AIML; it will be responsible for making requests to the web servers of the repositories, which in this case is the open repository web server.

An XML response from the repository contains information from the catalog file, such as author (s), title, and publication date, date on which it was submitted to the repository, summary, keywords and other information used to catalogue. This information is recorded in fields defined the DC.

The DCs are used for adding in the DSpace platform keywords (metadata) that reference the main catalog information from a file, i.e., it is possible to locate a particular file via their own metadata.

The variables used in the request address of search of the web servers of the repositories are standardized by the DSpace platform, but you can create new metadata (DC) directly from a function that offers digital repositories. This way, each institution may define multiple metadata according to each need, but for which the XML2AIML software that is developed in this paper can do search request to the servers of the repositories, you will use the metadata already created standards for the DSpace platform.

VII. RESULTS

The main objective of this work is to cover the knowledge of robotic engines developed in AIML language, with new bases of knowledge created autonomously by software responsible for seeking information in digital repositories, via metadata, and treated to generate the new bases.

Digital repositories have a limitation in the material search tool, for not being very precise in finding the files when a user uses the repository. So, the software developed in this paper solves two problems: making use of Chatterbots to help repository users in searches of files and , with the deployment of Chatterbots in repositories, the information contained in the Dublin Core (metadata) of the repositories will be used by the software to cover with new AIML knowledge bases.

 ISBN: 978-1-61208-329-2

The creation of new questions and answers to the robotic AIML language engines are required to use new bases of information in order to analyze and treat, to generate new questions and answers on top of new information.

Figure 2 presents the operation of the chatterbot with metadata software analyzer XML2AIML.

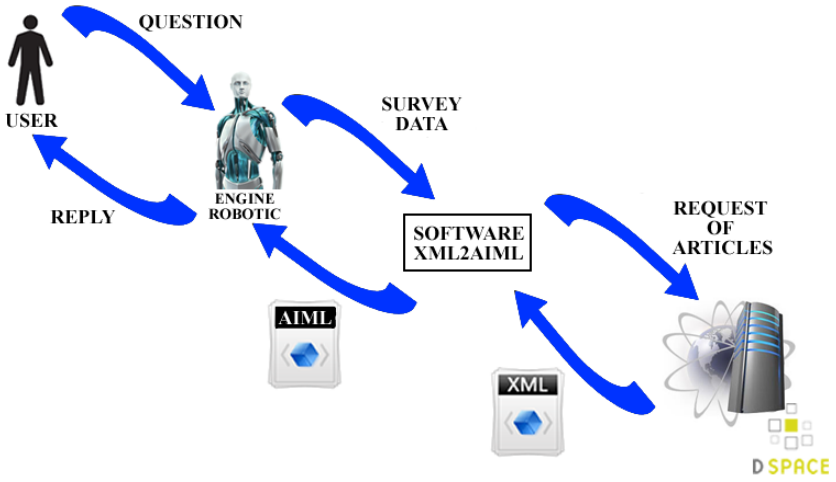


Figure 2. Model search functionality in digital repository file through a chatterbot AIML.

The functionalities are divided into four steps, namely (i) collecting user data, (ii) requesting files search, (iii) presenting questions and answers, and (iv) creating new knowledge base. The scenario creating (to test) a file search conversation in the repository is shown below.

A. *Collecting user data*

The first step is to gather as much information from the user about the material he/she is looking for. One could get this information as the chatterbot already contains questions that were created to make this collection. As the search will be driven by DC variables, the chatterbot has to collect information that can be used in these variables, that is, information such as author, title, publication date, area or a period of publication.

In the test scenario, chatterbot explains to the user whether he/she wants to look for a material, and what rules to follow. These rules were set for the chatterbot as it can understand what kind of information the user is passing/willing to obtain.

When a user is looking for materials development in AIML, the chatterbot saves that information in a variable and then to asks the user if he has any more information about this area. As the user does not have any more information, the chatterbot triggers a script element, which is responsible for calling the method colect_information (author, title, start_date, end_date, area); this method passes all the information of the chatterbot to XML2AIML software.

B. *Request files search*

The user information was passed to the XML2AIML software; with this information, the software analyzes what types of information was received, according to the type of information, and the correct method is called.

In this scenario, the user passes information to the chatterbot on the material sought, then the software will invoke the method responsible for creating an ADDRESS that contains the variables parameter DC author and title area; even though the chatterbot has not passed information about the author and title. The XML2AIML is already scheduled to put DC in the ADDRESS; only the variables that are assigned some value in time to call the method are created in the scope of the parameter passing method.

The software with the address ready will call the functions InputStreamReader and BufferedReader of the Java library, which are responsible for making the request to the repository server and receive the response from the server. [22] is a link generated by software XML2AIML for search about the articles on institutional repository.

The server receives the values of variables from parameters passed in the address and searches the repository database for any file related to that information; after the search, it generates an XML to send as reply for the machine that made the request.

Figure 3 presents a response of a request from a server for software XML2AIML. The response received is in XML format.

 ISBN: 978-1-61208-329-2

```
<metadata>
▼<oai_dc:dc xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
  http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
  ▼<dc:title>
     Interpretador AIML Alimentado com TAGS HTML5 - Manual de Comandos do AIML
   </dc:title>
   <dc:creator>Macedo, Rafael Luiz de</dc:creator>
  ▶<dc:description>...</dc:description>
   <dc:date>2012-11-21T11:56:02Z</dc:date>
   <dc:date>2012-11-21T11:56:02Z</dc:date>
   <dc:date>2012-11-21</dc:date>
   <dc:type>Dissertação</dc:type>
   <dc:identifier>http://hdl.handle.net/11077/804</dc:identifier>
   <dc:language>pt_BR</dc:language>
  </oai_dc:dc>
</metadata>
```

Figure 3. Response received of server of institutional repository [22].

The software XML2AIML, with the request response assigned to a variable of type BufferedReader, first parses the response that contains information on any file. In the case when it does not contain the response, the software will pass to the next step, which is to create the new information base; only after containing this new basis of questions and answers related to the point that no file was found in the repository within the information the user passed to the chatterbot.

C. *Preset questions and answers*

When the XML2AIML software parses and generates the new questions and answers on the information contained in the metadata of the XML response, consultations are held to a MySQL database; only information for pre-defined questions and answers is used in this work.

The database contains parts of questions and answers, which will be concatenated with certain kinds of information from the metadata that is contained in the XML to generate the questions and answers that will be used in the new AIML knowledge bases.

The goal of the use of a database in the software XML2AIML is to focus the new predefined questions and answers in a single location, using the database, to make it faster and to improve and increase the amount of these questions and answers that will be generated for each file found in the search of the repository.

This database also uses a table to focus the questions and answers, which are responsible for collecting information about the material that the repository user is seeking. These questions and answers have suffered many changes during the development of the software, to make the user clearly understand the questions and answers that the chatterbot will use during the dialogue.

D. *Creating new knowledge base*

The XML2AIML software parses the XML for the metadata for each file that is contained in the response. As seen in Sections 2 and 3, metadata is used to catalog every file, and, in the digital repository DSpace platform, the default DC metadata is used.

XML digital repositories are described according to the type of information about the file found, such as dc: title, dc: creator, dc: date, dc: type and other types of DC. DCs are passed by parameter to the method responsible for generating the questions and answers.

The parameter responsible for receiving the information about the file found in the digital repository performs a query to MySQL database software to pick up words that will be part of each set of questions and answers.

The questions and answers already concatenated with the information on the file are inserted automatically in the new AIML knowledge base; but, for each set of question and answer, XML2AIML inserts the elements that are part of the structure of a category to which the chatterbot can recognize and use these new questions and answers.

Each set of metadata represents a file found in the search in the repository. For each set, the software generates the same questions and answers that were used in the first set of metadata.

When the software has finished doing all the analyzing and applying the treatment of questions and answers, in all the sets of metadata, the software completes the new AIML knowledge base, with the element </aiml>.

Figure 4 presents a new base of questions and answers generated by XML2AIML software after analyzing the response received from the server.

 ISBN: 978-1-61208-329-2

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<aiml>
<category>
    <pattern>Author The AIML interpreter
    Powered with TAGs HTML5 - Commands
    Manual AIML.</pattern>
    <template>Author Macedo, Rafael Luiz
de</template>
</category>
<category><pattern>LINK The AIML
interpreter Powered with TAGs HTML5 -
Commands Manual AIML.</pattern>
      <template>Follow the link of article:
http://hdl.handle.net/11077/804</template>
</category>
</aiml>
```

Figure 4. New base of questions and answers generated by software XML2AIML.

The software generated XML2AIML file is passed to the chatterbot, which submits the search result to the user from the repository. The chatterbot will continue with the conversation telling the user who found the material.

In the test scenario, the user askes questions about the material found by chatterbot, and the user terminates with the question that he wants to access the complete material. The chatterbot gives an answer the access link to the page of the material.

VIII. CONCLUSION AND FUTURE WORK

The AIML markup language has a limitation in creating robotic conversational engines, limiting the number of questions these robotic engines can be asked. Due to this limitation, the robotic engines cannot keep a simple dialogue with a human for a long time because the questions are held outside the context of knowledge in which these robotic engines were developed, and cannot be answered.

Another problem is users accessing digital repositories of DSpace platform; the users encounter difficulties in being able to obtain the materials. Then, with the help of a chatterbot to interact with users and gather information about the desired materials, users can learn in a short time if the repository has a particular material or not.

The objective of this work was to encompass the knowledge of robotic AIML language engines with new bases of knowledge generated by the XML2AIML Analyzer software, using the data contained in the metadata of digital repositories of DSpace platform. In addition, we also improved the search of files the user can perform in these digital repositories.

As future work, XML2AIML will be implemented in repository server at the Center University of Euripides Marilia and it will be an available software to users of the repository of the university.

Another future work is to enable the XML2AIML to find files in others institutional repositories.

REFERENCES

[1] http://aberto.univem.edu.br/ [retrieved: March 13, 2014]

[2] http://www.alicebot.org/aiml.html, [retrieved: March 13, 2014]

[3] Rachel C. V. A. Semantic Web: A Focused Analisy using metadata, Dissertation (Master of Science in Information) – Faculty of Sciences – University Estadual Paulista – UNESP, Marilia, 2005.

[4] http://aitools.org/Programd, [retrieved: March 13, 2014].

[5] E. Fusco, Conceptual data models as part of the process of cataloging: use of FRBR perspective in developing digital bibliographic catalogues, Doctor thesis, UNESP – Marilia, pp. 60-74, 2010.

[6] D. Hillmann, Using Dublin Core – The Elements, Making it easier to find information, DCMI, 2005.

[7] J. Milstead and S. Feldman, Metadata Projects and Standards, vol. 23, no. 1, pp. 32-40, 1999. ISSN : 0146-5422.

[8] http://repositorio.unb.br/, [retrieved: March 13, 2014].

[9] G. P. Ribeiro, Digital Geospatial Metadata, Workshop on non-conventional databases, Niterói, Brazil, 1995.

[10] E. Rich and K. Knight, Artificial Intelligence – Second Edition. São Paulo, 1993.

[11] S. Russell and P. Norvig, Artificial Intelligence – Translation of Second Edition. Rio de Janeiro, 2004.

[12] http://apprendre-math.info/portugal/historyDetail.htm?id=Turing, [retrieved: March 13, 2014]

[13] M. I. F. Souza, and L. G. Vendrusculo and Geane C. Melo, Metadata for the description of electronic information resources: using the Dublin Core standard, Ci. Inf., Brasília, vol. 29, no. 1, p. 93-102, jan./abr, 2000.

[14] R. Tansley, et al. The DSpace Institutional Digital Repository System: Current Functionality, Joint Conference on Digital Libraries, pp. 87-97, EUA, 2003.

[15] M. I. Tomael and T. E. of Silva, Institutional Repositories: guidelines for information policy, VIII ENANCIB – National meeting of research in information science, Salvador, Bahia, Brazil, 2007.

[16] S. L. Vellucci, Bibliographic relationships. International Conference On The Principles And Future Development Of AACR. Toronto: American Library Association: Library Association Publishing, pp. 105-147, 1998.

[17] A. M. M. Vieira and G. David, Digital Repository and process management Platform: an integrated architecture, Faculty of Enginneering, University of Porto FEUP, 2011.

[18] R. S. Wallace, The Element of AIML Style, © ALICE A. I. Foundation, Inc., October, 2003.

[19] http://www.alicebot.org/documentation/aiml-reference.html, [retrieved: March 13, 2014]

[20] S. Weibel, The Dublin Core: a simple content description model for electronic resources. Bulletin of the American Society for Information Science, pp. 9-11, Oct/Nov, 1997.

[21] http://www.ed.conpet.gov.br/br/converse.php [retrieved: March 13, 2014]

[22] http://aberto.univem.edu.br/oai/request?verb=ListRecords&from=2012-11-22T00:00:01Z&until=2012-11-23T08:00:01Z&metadataPrefix=oai_dc [retrieved: March 13, 2014]

 ISBN: 978-1-61208-329-2

Designing a Situation-aware Movie Recommender System for Smart Devices

Mhd Irvan, Na Chang, Takao Terano
Dept. of Computational Intelligence and Systems Science
Tokyo Institute of Technology
Yokohama, Japan
irvan@trn.dis.titech.ac.jp, changna@trn.dis.titech.ac.jp, terano@dis.titech.ac.jp

***Abstract*—With the growing number of people using SmartTVs and Smartphones, designing a recommender system for on-demand streaming media, such as movie streaming, has been an attractive, yet challenging work. There are many factors that influence people to enjoy a movie. Smart devices provide many kinds of data from its sensors that can help us deduce, for example, whether it is the time, the day, the location, or the combination of those that makes a great experience in watching a particular movie. However, designing the algorithm to consider all these factors can lead into a very complicated decision tree. To address this issue, we propose a simple evolutionary computational approach that can be used to search through those huge numbers of possible combinations of solutions, and find the relevant factors when recommending a movie to particular type of users.**

Keywords-Recommender System; Classifier System; Genetic Algorithm; SmartTV; Smartphone

I. INTRODUCTION

Smart Devices, such as SmartTVs and Smartphones, provide the integration of various web services into televisions and mobile phones. One of such services is on-demand streaming media. On-demand streaming allows users to choose shows or movies they would like to watch. However, there are countless choices available from the streaming providers. This may lead users to confusion, as they might not know what would be interesting to watch [2].

Recommender systems address this problem. Rather than waiting for users to choose a movie, the system recommends movies that it thinks they will like. The recommendations are usually generated through learning from users' past behavior [3], or from other similar users' interest [5].

Many current recommender algorithms rely on users' feedback, such as rating, or profile. For example, when a user liked a movie, the system will search for other users who liked the same movie, and then, recommend other movies that those users also liked. While this might work very well for shopping websites, it might not be suitable for media streaming.

When a user enjoys a movie, there are many factors that affect his/her enjoyment at that moment. For example, a user who usually enjoys action movies on weekend might prefer watching drama movies on other days at night to relax after getting tired from work. A user might really like science fictions, but s/he only enjoys watching them from a large screen TV at home, and never on smartphones due to the small screen.

In other words, even if a user liked a movie, s/he might not be going to enjoy the same movie, had s/he watched it under different circumstances. Locations, devices, days, times, and other factors contribute to whether she will enjoy a movie or not. SmartTVs and smartphones can provide all these details, and it is only natural to use the information as basis for recommender systems. However, designing a recommender system to consider all these factors using typical recommender algorithms will end with a very complicated decision-making process. This paper offers a simple algorithm to address this issue using Learning Classifier System (LCS) [7], implementing genetic algorithm (GA) [1] and reinforcement learning (RL) [4].

LCS maintains a population of classifiers that predicts the best action given its input. The input we use is information that is available from smart devices, such as sensory data, geographical and device information, as well as date and time. GA is used to search the possible solution space to figure out which part of the inputs, or what kind of input combination affects the viewing experience. Solutions proposed by the GA are evaluated by RL, giving feedback whether they are accurate or not. During training, LCS repeat this process over and over again until it has a good population set with high average accuracy.

This paper starts with the introduction to recommender systems in Section I and reviews some of the literatures related to this field in Section II. We define our proposed method using LCS for recommending items in Section III. Finally, we put our conclusion and the discussion about future work in Section IV.

II. LITERATURE REVIEW

Mukherjee [3] proposed a movie recommender system using voting method. Their system tracks users' preference, such as favorite actors, actress, genres, etc. Each attribute of the preferences is given a weight value, which reflects the relative importance of those attributes. The voting system calculates these weights according Bayesian learning scheme and returns a ranking of alternatives when the user asked for a recommendation.

Salter [2] combined two popular recommender algorithms, Collaborative Filtering (CF) and Content-Based Filtering (CBF), into one system. The CBF was used to address the cold-start problem with CF not being able to make recommendation for new items.

Symeonidis [5] developed a recommender system with explanations. Theirs system gives the ability to a user to

 ISBN: 978-1-61208-329-2

check the reasoning behind a recommendation. This allows users to accurately predict their true opinion of an item.

Those systems managed to make good predictions about movies that users would like. However, those systems were designed before the smart devices and streaming media went mainstream. The way people watch movie has changed, some people like to watch at home, some prefer to watch on mobile devices while commuting. They did not consider these possible factors and other information that can be provided by smart devices. We tackle this issue with our proposed method.

III. PROPOSED METHOD

A. Learning Classifier System

LCS [1] is a machine learning paradigm, in which an intelligent agent is interacting with an environment. LCS keeps a collection of classifiers, referred as population set. Each classifier is essentially a rule of condition-action set. The classifiers have a parameter that predicts the reward that the agent will receive, should it choose the action proposed by the relevant classifiers. LCS agent learns to perform the best action based on the condition. Whenever the agent performs an action, it receives feedback from the environment to inform the quality of the action.

There are many models of LCS available today, such as ZCS [6] and XCS [7]. Different models have different criteria as what is the "best" action. ZCS model trains the system to chase high rewards. Over time, the population set evolves into a set of classifiers that predicts high reward only. The downside of this model is, although it gets huge reward when it does predict correctly, many of the classifiers often predict incorrectly. This led into inconsistent accuracy [6].

XCS model tackles the issues related to ZCS. Each classifier in XCS maintains an additional parameter, referred as accuracy parameter. This parameter job is to keep track of how often its classifier made inaccurate predictions. XCS agents prefer actions proposed by classifiers with high accuracy value, although they may predict low reward. Thus, XCS is more suitable to problems where consistent accuracy is important [7]. For this reason, we choose XCS model as the basis of our proposed method.

Recommendation using LCS means that the systems can, unlike CF and CBF method [2], consider more factors in deciding which item to recommend. CF concerns only about similar users, while CBF concerns about similar items. While they are good for users of web shopping sites, they might not be suitable for users of media streaming services, where users do not simply like an item, but mood factors affect in a sense for example they might have different preferences in morning and night time. LCS can be used to consider these factors when making recommendations. In addition to recommender systems, our proposed method has also been applied to simulate security patrol [8].

B. Generating Initial Classifiers

The condition part of the classifiers is a string of input reflecting the situation that the agent encounters. In our recommender system, the input string consists of information that can be provided by smart devices: Day, time, user's age, gender, type of device, location, movie's release date, movie genres, movie stars, movie ID (Figure 1).

From the training set, when a user "Like"d a movie, the system generates several classifiers that represent the situation. It takes into consideration the information mentioned above.

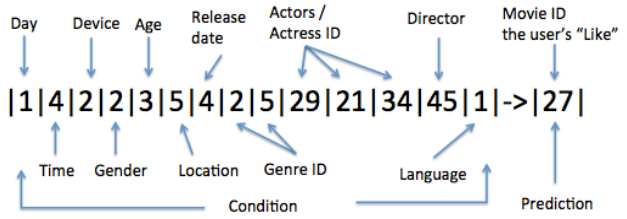


Figure 1. Classifier Representation

The numbers shown in Figure 1 are interpreted according to the following schema:

- Day: [0] = Unknown; [1] = Sunday; [2] = Monday; [3] = Tuesday; [4] = Wednesday; [5] = Thursday; [6] = Friday; [7] = Saturday.
- Time: [0] = Unknown; [1] = 6 AM ~ 8:59 AM; [2] = 9 AM ~ 11:59 AM; [3] = 12 AM ~ 2:59 PM; [4] = 3 PM ~ 6:59 PM; [5] = 7 PM ~ 8:59 PM; [6] = 9 PM ~ 11:59 PM; [7] = 12 PM ~ 2:59 AM; [8] = 3 AM ~ 5:59 AM.
- Type of device: [0] = Unknown; [1] = TV; [2] = Tablet; [3]=Phone.
- Gender: [0] = Unknown; [1] = Male; [2] Female.
- Age: [0] = Unknown; [1] = Below 18; [2] = 18 ~ 29; [3] = 30 ~ 39; [4] = 40 ~ 49; [5] = 50 ~ 59; [6] = 60~69; [7]=70~79; [8]=80 and above.
- Location ID (e.g., [5] = Tokyo).
- Release date: [0] = Unknown; [1] = Before 1970; [2] = 1970 ~ 1979; [3] = 1980 ~ 1989; [4] = 1990 = 1999; [5] = 2000 ~ 2010; [5] = 2010 and after.
- Genres ID (e.g., [2] = Drama; [5] = Romance).
- Actor/Actress ID (e.g., [29] = Leonardo DiCaprio; [21] = Kate Winslet, [34] = Billy Zane).
- Director ID (e.g., [45] James Cameron).
- Language ID (e.g., [1] = English).
- Movie ID (e.g., [27] = Titanic).

Suppose a user likes Titanic (a fact). The system considers the situation when it happened. On what day? What time? What kind of user? What kind of movie? Who are the actors? The classifier shown in Figure 1 can be read this way: "A female teenagers who live in Tokyo who likes 1990s movies AND likes drama and romantic movies AND likes movies starring Leonardo DiCaprio, Kate Winslet, and

 ISBN: 978-1-61208-329-2

Billy Zane AND likes movies directed by James Cameron AND likes English movie WILL enjoy watching Titanic ON a tablet AT night ON Sunday".

Since movies are usually related to multiple genres and have many actors/actresses involved, for each fact the system generates multiple classifiers picking (in our example) two random genres and three random actors/actresses related to the movie. This means that during learning the system will look for which combination of genres and actors/actresses are relevant to the user's profile.

Additionally, for each fact the system also generates classifiers that have "Wildcard" symbols along the input string. This means that during learning the system will look for which parts of the input that are not relevant (the noise) to the user's profile. Consider a more generalized classifier illustrated in Figure 2.

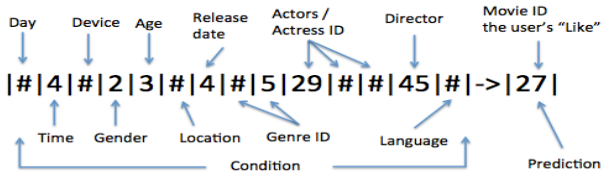


Figure 2. A Generalized Classifier

The classifier illustrated in Figure 2 shows some wildcard symbols ('#' symbols) along the input string. This classifier thinks that the "Day", "Device", "Location", one "Genre", two "Actors/Actresses", and "Language" parts are not relevant. This classifier can be read: "Regardless of the Day, Device and Location, a female teenager who likes 1990s romantic movies AND likes movies starring Leonardo DiCaprio AND likes movies produced by "James Cameron", WILL enjoy watching Titanic at Night".

The more wildcard symbols a classifier has, the more generalized it is. After generating all the classifiers from the initial set of facts, the system will test those classifiers against users in the testing set. During each learning loop, the system will evolve classifiers that make accurate prediction and delete inaccurate classifiers through evolutionary process. After the learning process is finished, the population set should consist the generalized classifiers (but not too generalized) with high accuracy.

The input length for LCS is flexible. If, for example, more profile data are available from the smart devices, it is possible to add those details into the input. If necessary, more metadata about the movies (such as more genres, studio name, music composer, or other metadata) can be added too. The longer the classifier, the more details are taken into consideration.

Obviously, when more details are necessary to be considered, using general decision process, such as decision tree, the decision-making process will end up with a very complicated procedure. LCS simplifies this process by representing the details as strings of possible solution and let the evolutionary process search for the good ones.

C. *Learning*

After the initial population set is generated, the learning phase begins (Figure 3).

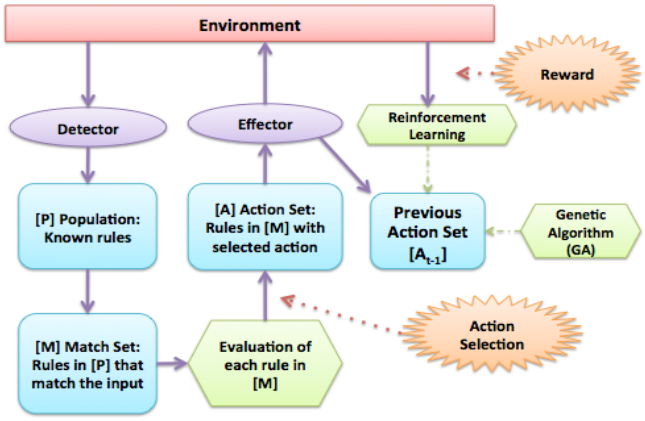


Figure 3. Learning Process

LCS starts by sensing the current environment situation. Suppose in the training set, LCS encounters a user with the situation shown in Figure 4, and LCS has a Population Set [P] of eight Classifiers C shown in Figure 5.

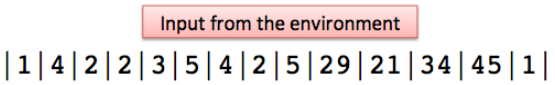


Figure 4. Input sensed from the environment situation

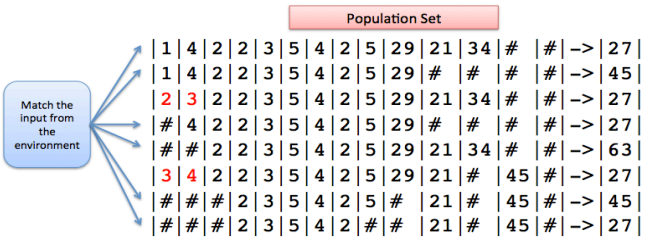


Figure 5. Population Set

From Figure 5, we can see that six classifiers match the input from environment. The first two digits of the other two classifiers do not match the input. LCS will select the six matched classifiers, and put them in a collection called Match Set [M] (Figure 6).

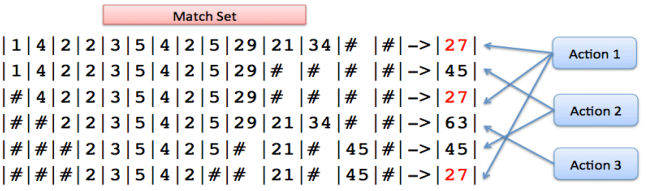


Figure 6. Match Set collected from the Population Set

 ISBN: 978-1-61208-329-2

Figure 6 shows that the six classifiers inside the Match Set predicted three different actions (three movies). For each action, LCS calculate the prediction array *P(A)* using the value of reward prediction (*C.p*) and fitness (*C.f*) of classifier C ∈ [P] (1).

$$P(A) = \frac{\sum_{C.a=a \wedge C \in [M]} C.p \times C.f}{\sum_{C.a=a \wedge C \in [M]} C.f} \quad (1)$$

Essentially, P(A) reflects the average of all reward prediction of classifiers in [M] that advocate action *a*. The algorithm chooses the action that maximizes P(A) (2).

$$A_{\max} = \arg\ \max_A P(A) \quad (2)$$

Suppose in our example above, Action 1 is calculated as the best action. Then, LCS puts the three classifiers in [M] that proposes Action 1 into an Action Set [A].

After LCS applies the action, it receives feedback from the environment whether if its prediction is correct or not. Through RL method [4], the classifiers in [A] are credited with reward *r* as a result of the action performed. Then, the prediction error of each classifier in [A] is updated (3).

$$\varepsilon \leftarrow \varepsilon + \beta\left(|r - p| - \varepsilon\right) \quad (3)$$

where β is the learning rate.

Next, reward prediction of each classifier in [A] is updated (4).

$$p \leftarrow p + \beta\left(r - p\right) \quad (4)$$

Unlike generic LCS, XCS classifiers maintain accuracy parameter that tracks the accuracy of the classifiers\ throughout the learning process. Each rule's relative accuracy is determined by dividing its accuracy the total of the accuracies in [A] (5).

$$\kappa' = \frac{\kappa}{\sum_{C \in [A]} C.\kappa} \quad (5)$$

Finally, fitness *f* is updated with respect to the relative accuracy (6).

$$f \leftarrow f + \beta\left(\kappa' - f\right) \quad (6)$$

The dataset is divided into training set and testing set. The training set is used to train the system to generate a population of classifiers that predict a recommendation for existing items. The testing set acts as new items encountered by the system. This is used to evaluate how well the system is able make predictions for unknown items.

D. Evolution

GA [1] is used to evolve the rules in [A]. GA is triggered when the average time period for classifiers within [A] since the last occurrence of GA is greater than GA's frequency parameter. The GA starts by selecting two "parent" classifiers from [A] with roulette wheel selection (7).

$$p_i = \frac{f}{\sum_{C \in [A]} C.f} \quad (7)$$

Once two parents have been selected, new offspring classifiers are generated by *crossover* and *mutation*. Crossover operation selects a point on the parent classifier strings. All digits beyond that point of one parent classifier are swapped with the digits of another classifier. Finally, the digits of the resulting strings from the crossover are mutated into different acceptable values. This means that two new offspring classifiers are generated maintaining some traits from their parents. The offspring classifiers are inserted into the population set, replacing classifiers with low fitness value. Then, the learning process is repeated again until the population set evolves into a collection of classifiers with relatively high accuracy values.

IV. CONCLUSIONS AND FUTURE WORK

This short paper shows preliminary work on how to apply LCS to recommender system. It shows that LCS can be used to train for, not only predicting a movie a user will enjoy, but also finding the reason why s/he might enjoy it. This is useful to recommend a movie to another user that has similar profile.

LCS is flexible, in a sense that, if the system designer decides to change the input types, the learning algorithm stays the same. S/he may also add more details to the input string if s/he manages to gather more detailed data from the sensors of smart devices.

Ongoing work includes testing with real data, as well as cross-validating the result. Future explorations will include performance analysis, such as training time, as well as comparison to other solutions, such as Naïve Bayes Classifier and k-Nearest Neighbors.

REFERENCES

[1] J. H. Holland, "Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence," Book ISBN 0262581116, A Bradford Book, 1992.

[2] J. Salter, "CinemaScreen recommender agent: combining collaborative and content-based filtering," Intelligent Systems, 2006, pp. 35-41.

[3] R. Mukherjee, G. Jonsdottir, and S. Sen, "MOVIES2GO: An Online Voting Based Movie Recommender System," Proceedings of the fifth international conference on Autonomous agents, 2001, pp. 114-115.

[4] R. S. Sutton, "Reinforcement Learning: An Introduction," Robotica, vol. 17, Issue 2, 1999, pp. 229-235.

[5] P. Symeonidis, and A. Nanopoulos, "MoviExplain: A Recommender System with Explanations," Proceedings of the third ACM conference on Recommender systems, 2009, pp. 317~320.

[6] S. W. Wilson, "ZCS: A Zeroth Level Classifier System," Evolutionary Computation, vol. 2, Issue 1, 1994, pp. 1-18.

[7] S. W. Wilson, "State of XCS classifier system research," Lecture Notes in Artificial Intelligence (LNAI-1813), 2000, pp. 63-81.

[8] M. Irvan, T. Yamada, T. Terano, "Multi-Agent Learning Approcah to Dynamic Security Patrol Routing," Proceedings of SICE Annual Conference, 2011, pp. 875-880.

 ISBN: 978-1-61208-329-2

The Influence of IT Features on M-commerce User Behaviors

Philippe Marchildon and Pierre Hadaya
Department of Management and Technology
ESG-UQÀM
Montréal, Canada
e-mail: marchildon.philippe@courrier.uqam.ca
hadaya.pierre@uqam.ca

Abstract— **M-commerce research suffers from critical shortcomings due to an extensive reliance on TAM. As such, little is known about mobile IT artifacts, their features and their influence on m-commerce user behaviors. Based on evidence which suggests that the use of IT artifact features vary over time and that it is the specific features used at any point in time that determine work outcomes, it is important that we better understand the influence of mobile IT artifact features over time. To do so, we propose to use a different theoretical lens than TAM, to study sensory mobile IT artifact features and to distinguish between stages of the m-commerce adoption process. Accordingly, the literatures on IT artifacts, Kaplan and Kaplan's (1982) framework and mobile IT artifact features are examined to propose a research model, its related hypotheses, and methodological aspects regarding its empirical validation. Finally, the proposed model's anticipated contributions are discussed.**

Keywords-component; M-commerce; IT features; Kaplan and Kaplan's preference framework; longitudinal

I. INTRODUCTION

M-commerce, defined as "the use of mobile information technologies, including the wireless Internet, for communication and coordination within an organization, between an organization and other organizations and/or customers, and for management of the firm" [1, p. 3], is an emerging trend in today's business reality [2]. As with many other IS research topics, the Technology Acceptance Model (TAM) has been extensively used by m-commerce researchers to support their theoretical reasoning. This overwork of TAM has brought unexpected side effects that now undermine further knowledge development [3]. More precisely, to this day, most m-commerce researches have failed to study the different stages, as well as the tangible antecedents of individual behaviors throughout the m-commerce adoption process. Thus, despite the increasing importance of this new way of doing business, little is presently known about mobile IT artifacts (e.g. Internet-enabled PDAs and smart phones), their features and their influence on m-commerce adoption individual behaviors throughout the m-commerce adoption process [4]. As such, despite evidence which suggests that the use of IT artifact features may not only increase but also decrease over time [5], and that it is the specific features in use at any point in time that influence and determine work outcomes achieved trough IT artifacts [6, 7], no sound explanation has yet been provided for variations in the use of mobile IT artifact features. Since the success of m-commerce hinges on mobile IT artifact features, the users' willingness to adopt them, and to engage in activities requiring their usage, a better understanding of their design characteristics and their influence on user behaviors over time is needed [8].

The objective of this study is thus to investigate the influence of mobile IT artifact features on m-commerce users over time. More specifically, the present research aims at answering the following research question: Why does the usage of mobile IT artifact features vary over time? In particular, this study will investigate whether differences exist between features that influence the behavior of m-commerce adopters and those that influence the behavior of post adopters. To do so, a longitudinal stance is proposed which departs from previous m-commerce research by leaving aside the TAM framework and by focusing on a new theoretical lens: Kaplan and Kaplan's preference framework.

The paper proceeds by first examining the literature on IT artifact conceptualizations, Kaplan and Kaplan's [9] preference framework and mobile IT artifact features. Based on these theoretical underpinnings, a research model and its related hypotheses are then developed. This is followed by a discussion that addresses several key research methodological aspects. The paper concludes by presenting the anticipated theoretical and practical contributions of the research.

II. THEORETICAL DEVELOPMENT

A. Conceptualization of the IT Artifact

Early attempts to differentiate artifacts from one another led to the identification of two different sets of distinguishing characteristics: primary and secondary attributes [10]. Primary attributes are those which are "essential to the object or substance and so are inherent in it whether they are perceived or not" whereas, "secondary attributes are those which are perceived by the senses, and so may be differently estimated by different percipients" [11, p. 702]. Subsequently, Griffith and Northcrafth [7] proposed a similar categorization by differentiating between objective and psychological features. As noted by

 ISBN: 978-1-61208-329-2

Downs and Mohr [11] and later on by Griffith and Northcraft [7], because secondary characteristics are better understood as descriptions of users, not of artifacts, and because primary features are highly idiosyncratic, studying artifacts by using either sets of attributes alone leads to critical shortcomings. As such, these authors have suggested that scholars should develop interactive models that combine both primary/objective and secondary/psychological sets of attributes to assess the defining features of IT artifacts and their impacts.

Recognizing the potential of interactive models and anchoring their pioneering work on structuration theory, DeSanctis and Poole [6] were the first to insightfully probe and characterize both artifacts and the work environment within which they are applied. These authors introduced Adaptive Structuration Theory (AST) and a framework anchored on its tenets which assesses the role of IT artifacts from two vantage points: (1) the types of structures that are provided by advanced technologies, and (2) the structures that actually emerge in human action as people interact with these artifacts. DeSanctis and Poole's [6] work not only allows the assessment of each sets of attributes independently but also permits the assessment of interaction effects between these sets of attributes (i.e. the mutual influence of artifacts and social processes) [12]. Drawing from Giddens [13], DeSanctis and Poole [6] first posited the concepts of social structures embedded in technology and social structures in action, and then considered the interplay between them [12]. The concept of social structures embedded in technology is crucial to the characterization of IT artifacts, and includes two dimensions: "structural features" (i.e. specific types of rules and resources, or capabilities, offered by the system) and "spirit" (the general intent with regard to values and goals underlying a given set of structural features) [6].

More recently, Markus and Silver [12] refined DeSanctis and Poole's [6] work in an attempt to address the criticisms made to the concepts of "structural features" and "spirit" as well as to account for the following two observations: (1)"variations in the social structures in technology were seen as encouraging different forms of social action", and (2) "the ways in which people actually used the social structures of technology (i.e. appropriated them) were seen as influencing the outcomes actually observed" [12, p. 612]. More precisely, Markus and Silver noted that people might appropriate a system's features faithfully (i.e. in a manner consistent with the spirit and structural feature design) or unfaithfully, leading to different consequences [5]. As such, Markus and Silver [12] suggested to unpack DeSanctis and Poole's [6] concepts and to redefine them as three new concepts: technical objects, functional affordances, and symbolic expressions. The technical objects concept pertains to the artifacts themselves whereas the functional affordances and symbolic expressions concepts refer to relations between technical objects and users. Specifically, the term "technical objects" refers to artifacts and their components, while the term "functional affordances" refers to the possibilities for goal-oriented action afforded to specified user groups by the technical objects and the term "symbolic expressions" refers to the communicative possibilities of a technical object for a specified user group. A key element in this reconceptualization is the establishment of a link between artifacts and their potential users through functional affordances and symbolic expressions. As a result, the interactions between artifacts and users can now be more clearly defined.

However, despite DeSanctis and Poole's pioneering work, little empirical research has been undertaken to test and validate their theoretical assertions and evaluate the influence of IT artifacts on users. As a result, few schemes are available to provide a sound theoretical grounding for further research. An exception is Rosen and Purinton's [14], empirical study which is based on Kaplan and Kaplan's [9] preference framework, and which theorizes and operationalizes the relationships between Web site features and users. Indeed, it is interesting to note that although Kaplan and Kaplan [9] did not explicitly build their framework on the tenets proposed by either Markus and Silver [12] or DeSanctis and Poole [6], the similarities between the different approaches are striking and provide a sound setting for future research

B. Kaplan and Kaplan's Preference Framework

The "Preference Framework" developed by Kaplan and Kaplan [9] is based on knowledge in psychology, architecture and design. Its basic premises are that artifacts provide users with information (i.e. signs, icons, etc.) and that this information inscribed in artifacts influences theirs behaviors. The influential role of inscribed information in artifacts lies in the assumed informational needs of individuals that are triggered when they interact with artifacts. In other words, Kaplan and Kaplan's [9] "Preference Framework" describes how individuals use the information inscribed in an artifact's design to satisfy their informational needs when interacting with it. As such, the "Preference Framework" is congruent with the tenets proposed by Markus and Silver [12] and DeSanctis and Poole [6] since it recognizes not only the characteristics of the artifacts (i.e. the information inscribed in them) and the characteristics of the users (i.e. their informational needs) but also their mutual interaction by assessing the fulfillment or not of the individual's informational needs. Put differently, by identifying and defining the fulfillment of needs, Kaplan and Kaplan [9] acknowledge the critical role of affordance and symbolism in linking artifacts to individuals.

Kaplan and Kaplan's [9] "Preference Framework" is rooted in a sequence of studies that asked participants to look and assess photographs of physical landscapes and landmarks against a list of items. Through these experiments, the researchers were able to categorize individuals' informational needs when interacting with artifacts along a cognitive and a time dimension, resulting in a 2x2 matrix of informational needs (see figure 1). The cognitive dimension reflects the different types of needs that compose individual informational needs while the time dimension captures the order in which informational

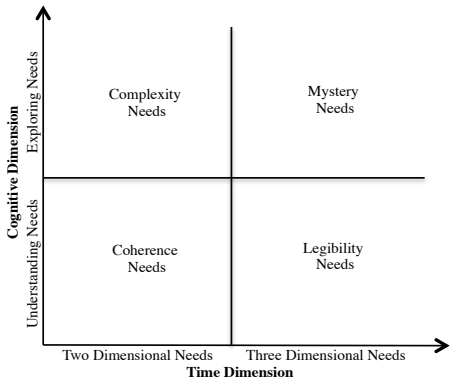


Figure 1. Kaplan and Kaplan's Preference Framework.

needs arise in an individual. More precisely, Kaplan and Kaplan suggest that an artifact can create understanding and exploring individual information needs which arise sequentially according to their assessment of immediate (i.e., two dimensional needs) versus longer-term (i.e., three dimensional needs). As such, four different informational needs exist (i.e., coherence, complexity, legibility and mystery), each representing a certain type of need (understanding vs. exploring) at a certain time (two vs. three dimensional).

Kaplan and Kaplan's preference framework posits that an individual's IT artifact usage behavior stems from the sequential fulfillment or not of his informational needs. As such, the behavior of an individual is first defined by the fulfillment of his immediate needs (i.e. two-dimensional needs) which arise from the instantaneous interaction with an artifact. As seen in table 1, these are coherence and complexity informational needs, which, when fulfilled, allow an individual to proceed to a rapid assessment of an artifact based upon a relatively superficial examination. Coherence informational needs refer to the degree to which the artifact features hang together. Thus, coherence needs can be fulfilled through the redundancy of the artifact's elements and/or textures, where, for example, the coordinated elements of an IT artifact give it a "minimalist" feel since few colors and elements are integrated in its design (e.g., Motorola MOTOPHONE3). On the other hand, complexity informational needs refer to the richness of the artifact features. For example, in a smart phone, the number of buttons or colors on the keyboard, the variety of icons available on the screen, the array of hands-free functionalities, amongst other things, would all contribute to the complexity of the artifact. It is important to note that both coherence and complexity informational needs must be fulfilled first for an individual to follow his course of action when interacting with an artifact.

Subsequently, an individual's usage behavior will be dictated by the fulfillment of longer-term needs (i.e., three-dimensional needs). These are legibility and mystery informational needs. Legibility informational needs refer to whether or not an artifact possesses a memorable component, a landmark facilitating the finding of one's way. This is similar to having a menu bar that is always positioned at the bottom of the screen no matter which application is used in a smartphone or PDA. Mystery informational needs refer to the extent to which an artifact conveys the feeling that many more features can be found. A smartphone could, for example, prompt users to discover new functionalities through certain feed-forward or feedback information.

Kaplan et al. [15] equated the transition an individual makes between immediate and longer-term needs to moving from a two-dimensional space to a three-dimensional space or as the difference between standing at the gate of a garden and actually walking through the garden. Finally, Kaplan and Kaplan's [9] "Preference Framework" implies that individuals will have a preference for artifacts that fulfill all four informational needs, an assumption that has been validated by several researchers who showed that people favor artifacts that answer coherence and legibility informational needs [16], while at the same time accommodating a desire for some complexity [17] and mystery [24].

C. Mobile IT Artifact Features

The notion of features, although well defined at the conceptual level (i.e. the building blocks or component parts of a technology [19]) remains rather elusive at the operational level. Recognizing this conundrum, Griffith [19] suggested that it is only through theoretical anchoring that researchers will be able to rightfully operationalize the concept of artifact features. Therefore, because individuals interact with artifacts through their senses [20] and because this study focuses on the interactions between mobile IT artifacts and their users, the present study defines features at the sensory level. That is, any building block or component directly triggering one of the five human senses (i.e., sight, hearing, smell, touch, and taste) is considered to be a feature.

Mobile IT artifact features have traditionally been investigated along three dimensions: (1) visual, (2) tactile and (3) audio [21], since smell and taste features have yet to be effectively developed and incorporated into m-commerce artifacts. In general, visual features have been the main focus of m-commerce and IS researchers. Findings from these research initiatives underline the significant impact of visual features on user behaviors and perceptions. For example, Karvonen [22] found a relationship between "aesthetic beauty" and e-trust. Rosen and Purinton [14] found that minimalist visual design drew users further into their task and increased perceived artifact efficiency. Furthermore, although early studies which thought that individuals perceived an artifact's visual informational cues in a holistic manner [23], recent findings from Lavie and Tractinsky [24] demonstrate that online users perceive visual informational cues along two sub-dimensions, namely classical aesthetics and expressive aesthetics. The first sub-dimension is associated with clean and orderly design, while the second represents the originality and creativity of the artifact's design, as

 ISBN: 978-1-61208-329-2

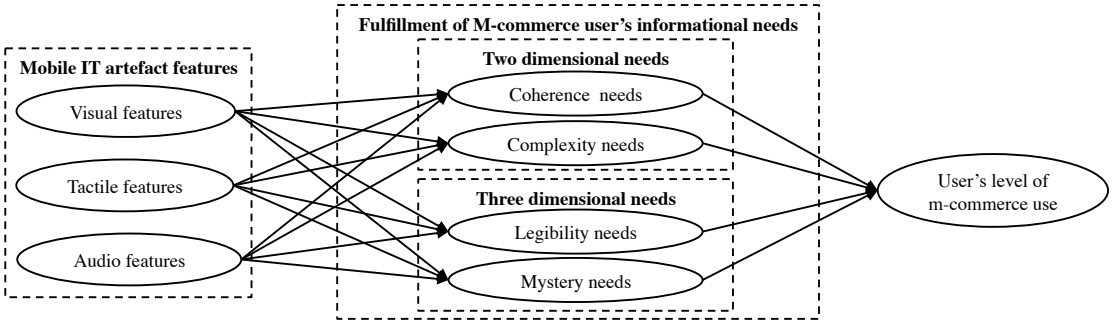


Figure 2. Research Model.

perceived by the users [24]. Altogether, these findings on visual features are in line with Kaplan and Kaplan's [9] framework, which identifies similar perceptual factors in the dimensions of coherence and complexity, and recognizes individual perceptions of artifacts to be multidimensional [25].

In addition, just as Kaplan and Kaplan [9], Tractinsky and Lowengart [25] acknowledged that while these sub-dimensions are distinct, they are not necessarily orthogonal. More precisely, Tractinsky and Lowengart [25] argued that "conceptually, the correlation between the two sub-dimensions reflects a fundamental relation to aesthetic design and perceptions" and that empirical correlations amongst the sub-dimensions "reflect an ecological phenomenon in which capable designers are good at creating balanced designs high on both sub-dimensions, whereas incompetent Web design tends to fail on both". These findings regarding visual features reinforce Kaplan and Kaplan's [9] assertion that individuals will prefer artifacts, which fulfill all four informational needs. However, although the strong emphasis on visual features has brought important insights on these specific characteristics of artifacts, this unbalanced attention has also meant that the IS field has to some extent neglected the study of tactile and audio features. Such features are also important as interaction effects across various features may also have significant impacts on users [19]. Furthermore, insights from the literature on virtual environments suggest that while adding auditory and tactile cues are likely to increase a user's perception in doing a certain task, increasing the level of visual fidelity will not produce similar outcomes [26]. Such findings suggest a potential tradeoff between features and limitations to feature enhancements (i.e., more is not always better). Therefore, a thorough investigation of all mobile IT artifact features and not just their visual characteristics is needed to better understand their role and influence on individuals.

III. CONCEPTUAL FRAMEWORK

A. Research Model

As described above, Markus and Silver [12] have theorized that IT artifacts influence users through their functional affordances and symbolic expressions. On the other hand, Kaplan and Kaplan [9] provide an operationalization of these concepts by defining the linkages between an IT artifact's features and an individual's behaviors through IT artifact informational cues and their fulfillment of informational needs. Based on these ideas, the following research model and related hypotheses are proposed (see figure 2)

B. Hypotheses

1) Two vs. three dimensional informational needs

Individuals are believed to depend on their visual sense for 80% of their external information, and presumably for even more than 80% of their external information when working with a GUI [27]. As such, visual features of mobile IT artifacts, which are likely to fulfill most of the informational needs, are extremely important when individuals are first introduced to the technology. However, such an unbalanced reliance on visual informational cues can heavily tax an individual's sense of vision and render the conveyance of additional information through this specific sense more difficult [28]. Consistent with this idea, Huong et al. [26] found that visual informational cues, although more important then tactile and audio informational cues, were in fact limited in their impact on users due to visual saturation and limited computational power. As such, several researchers have suggested [29], and have empirically validated [28] the idea that other senses should also be put to contribution to overcome the limits of visual informational cues. Results from these studies showed that adding tactile and audio feedback while users interact with IT artifacts did in fact increase their performance. As such, and based on the fact that individuals heavily rely upon visual informational cues, visual features are likely to play a more important role in fulfilling an individual's two-dimensional needs than tactile and audio features. However, since an individual's visual sense is already heavily taxed, leaving limited room for additional informational cues, tactile and audio features are likely to fulfill three-dimensional needs more effectively than visual features. These arguments lead to the first two hypotheses.

Hypothesis 1: The relationship between visual features and the fulfillment of an individual's informational needs will be stronger for two dimensional needs than three

dimensional needs in both adoption and post adoption settings.

Hypothesis 2: The relationship between audio and tactile features and the fulfillment of an individual's informational needs will be stronger for three dimensional needs than two dimensional needs in both adoption and post adoption settings.

2) Adoption vs. post-adoption informational needs

In developing their preference framework, Kaplan and Kaplan [9] recognized that individuals assess new situations and the information available to them in a two-phase, sequential manner. At first, individuals reflect on their immediate and direct perception of the setting's elements. These initial preoccupations are translated into coherence and complexity informational needs that must be fulfilled for the individuals to feel at ease in their novel environment and to follow their course of action. This primary assessment is subsequently followed by a secondary appraisal which emphasizes deeper needs (i.e. legibility and mystery informational needs) [14]. In order words, individuals move from a two-dimensional space, where coherence and complexity informational needs are pre-dominant, to a three-dimensional space, where legibility and mystery informational needs prevail [15].

As such, the sequential manner in which individuals assess a new situation suggests that the importance of each informational need evolves over time and that their influence on individuals varies accordingly, with two-dimensional needs being more important at first, and three-dimensional needs being more important later on. This idea was empirically supported by Rosen and Purinton [14] who used Kaplan and Kaplan's [9] preference framework to assess the quality of website design. More precisely, these authors demonstrated that two-dimensional needs were more important than three-dimensional needs when individuals were first introduced to a new website. Their results showed that users' intentions to revisit a web site after a brief initiation were largely explained by coherence (30.59% of variance explained) and complexity (20.87 % of variance explained) informational needs than legibility (17.8 % of variance explained) and mystery (not significant) informational needs. Thus, two-dimensional needs are likely to be more important for adopters who have limited experience with mobile IT artifacts than for experienced users (i.e., post adopters). Conversely, and again because of differences in experience with using mobile IT artifacts-commerce device, three-dimensional needs are likely to be more important for post adopters than for adopters. These arguments lead to the last two hypotheses.

Hypothesis 3: The relationship between the fulfillment of an individual's two-dimensional needs and his level of use of mobile IT artifact features will be stronger for adopters than for post adopters

Hypothesis 4: The relationship between the fulfillment of an individual's three-dimensional needs and his level of use of mobile IT artifact features will be stronger for post adopters than for adopters.

IV. METHODOLOGY

A. Data Collection

To empirically test the proposed research model we propose to investigate the rollout of an m-commerce solution to a group of pilot users within a company for a one-year period. Two surveys will be administered, one at the beginning of the rollout, after an initial training session, and another at the end of the project. The surveys' instruments will include measures drawn and/or adapted from the literature as well as measures that will be specifically develop for the purpose of this study. At each data collection point, users will be asked about their level of use of mobile IT artifact.. Users will also be interviewed on their informational needs and the features that enable them to fulfill these needs.

B. Statistical Analyses

Structural equation modeling (SEM) will be used to analyze the study data. As such, for each of the two data collection, a two-phase analytical procedure will be employed. In the first phase, a confirmatory factor model (i.e., the measurement model) will be used to measure the fit between the theorized model and observed variables, whereas the results of the measurement model will be used to create a path-analytic model to investigate the relationships hypothesized between the study constructs in the second phase [30]. Subsequently, we'll rely on the procedure proposed by Karahanna et al. [31] to compare the results from the two surveys, identify differences and test the research hypotheses.

V. CONCLUSION

The present research aims to answer the following research question: "Why does the usage of mobile IT artifact features vary over time?" and to validate the premise that individuals use different features at different points in time because the information inscribed in mobile IT artifact features is stable and their informational needs evolve over time. To do so, we propose a research model anchored on Kaplan and Kaplan's [9] preference framework to test the influence of mobile IT artifacts features on the behavior of m-commerce adopters and post adopters. The results obtained are expected to provide important theoretical and practical contributions.

First, this study departs from previous research in the IS field by being one of the few to open the IT artifact black box and to empirically investigate the role of IT artifact features on individual behaviors. As such, the present research, which defines IT artifact features along three dimensions (i.e., visual, tactile, and audio), will improve our understanding of concrete and tangible technology acceptance's antecedents. Second, the longitudinal stance of this study that will permit an

 ISBN: 978-1-61208-329-2

assessment of the varied influence of IT artifact features over time is likely to provide additional insights on the influence of time and IT artifact features on individuals' behaviors. Third, this research relies on a different theoretical lens than the traditionally used TAM framework and as such can contribute to the IS field in validating and providing IS scholars with a new research tool.

From a practical standpoint, the proposed study can help managers to better design and manage m-commerce apparatuses and improve the various outcomes tied to their use. For example, anticipated insights from this study could suggest that training to novice users should focus on visual features while training to experienced users should focus instead on tactile and audio features. Also, if practitioners plan to implement upgrades to the m-commerce devices used in their organizations, insights from this study can inform them on which features to upgrade and at what time in each apparatus' life cycle. Such guidance can be of significant importance as upgrading IT artifact features often entail tradeoffs and represent significant investments in both time and money [26].

REFERENCES

[1] E. Scornavacca and S. J. Barnes, "M-banking services in Japan: A strategic perspective," International Journal of Mobile Communications, vol. 2, 2994, pp. 51-66.

[2] Y.-S. Wang and Y. W. Liao, "The conceptualization and Measurement of m-commerce User Satisfaction, Computers in Human Behavior", vol. 23, 2007, pp. 381-398.

[3] I. Benbasat and H. Barki, "Quo Vadis TAM," Journal of the AIS, vol. 8, 2007, pp. 211-218.

[4] D. Cyr, M. Head and A. Ivanov, "Design aesthetic leading to m-Loyalty in Mobile Commerce," Information & Management, vol. 43, 2006, pp. 950-963.

[5] J. Jasperson, P. E. Carter and R. W. Zmud, "A Comprehensive Conceptualization of Post-Adoptive Behaviors Associated with Information Technology Enabled Work Systems," MIS Quarterly, vol. 29, 2005, pp. 525-557.

[6] G. DeSanctis and M. S. Poole, "Capturing the Complexity in Advanced Technology Use: Adaptive Structuration Theory," Organization Science, vol. 5, 1994, pp. 121-147.

[7] T. L. Griffith and G. B. Northcraft, "Distinguishing Between the Forest and the Trees: Media, Features, and Methodology in Electronic Communication Research," Organization Science, vol. 5, 1994, pp. 272-285.

[8] G. Bruner and A. Kumar, "Explaining consumer acceptance of handheld Internet devices," Journal of Business Research, vol. 58, 2003, pp. 115–120.

[9] S. Kaplan and R. Kaplan, Cognition and environment. New York, NY: Praeger Publishers, 1982.

[10] J. H. Jeans, Physics and Philosophy. Ann Arbor, MI: University of Michigan Press, 1966.

[11] G. W. Downs Jr. and L. B. Mohr, "Conceptual Issues in the Study of Innovation," Administrative Science Quarterly, vol. 21, 1976, pp. 700-714.

[12] L. M. Markus and M. S. Silver, "A Foundation for the Study of IT Effects: A New Look at DeSanctis and Poole's Concepts of Structural Features and Spirit," Journal of the AIS, vol. 9, 2008, pp. 609-632.

[13] A. Giddens, Central Problems in Social Theory. Berkeley, CA: University of California Press, 1979.

[14] D. E. Rosen and E. Purinton, "Website design: viewing the web as a cognitive landscape," Journal of Business Research, vol. 57, 2004, pp. 787–794.

[15] R. Kaplan, S. Kaplan, and R. L. Ryan, With people in mind. Washington, DC: Island Press, 1998.

[16] K. Lynch, The image of the city. Cambridge, England: MIT Press, 1960.

[17] J. F. Wohlwill, "Environmental aesthetics: the environment as a source of affect," In Human behavior and environment, vol. II, I. Altman and J. F. Wohlwill, Eds. New York: Plenum, 1976, pp. 37-86.

[18] R. Kaplan, "Predictors of environmental preference: designers and clients," in Environmental design research, vol. 1., W. F. E. Preiser, Ed. Stroudsburg, Dowden, Hutchinson and Ross, 1973, pp. 265-274.

[19] T. L. Griffith, "Technology Features as Triggers for Sensemaking," Academy of Management Review, vol. 24, 1999, pp. 472-488.

[20] R. Jain, "Experiential Computing," Communication of The ACM, vol. 46, 2003, pp. 48-54.

[21] H. Kiljander and J. Jarnstrom, "User interface styles," in Mobile Usability: How Nokia Changed the Face of the Mobile Phone, C. Lindholm, T. Keinonen, H. Kiljander, Eds. McGraw Hill, 2003, pp.15-44.

[22] K. Karvonen, "The beauty of simplicity," Proc. ACM Conference on Universal Usability, 2000, pp. 85-90.

[23] H. Van der Heijden, "Factors Influencing the Usage of Websites: The Case of a Generic Portal in the Netherlands," Information & Management, vol. 40, 2003, pp. 541-549.

[24] T. Lavie and N. Tractinsky, "Assessing Dimensions of Perceived Visual Aesthetics of Web Sites," International Journal of Human-Computer Studies, vol. 60, 2004, pp. 269-298.

[25] N. Tractinsky and O. Lowengart, "Web-Store Aesthetics in E- Retailing: A Conceptual Framework and Some Theoretical Implications," Academy of Marketing Science Review, vol. 11, 2007, pp. 1-19.

[26] Q. D. Huong, N. Walker, C. Song, A. Kobayashi and L. F. Hodges, "Evaluating the Importance of Multi-sensory Input on Memory and the Sense of Presence in Virtual Environments," IEEE explore, 2008, pp. 1-8.

[27] H. Takao, S. Kaoru, J. Osufi and I. Hiroaki, "Acoustic User Interface (AUI) for auditory displays," Displays, vol. 23, 2002, pp. 65-73.

[28] J. Jacko, K. V. Emery, P. J. Edwards, M. Ashok, L. Barnard, T. Kongnakorn, K. P. Moloney and F. Sainfort, "The effects of multimodal feedback on older adults' task performance given varying levels of computer experience," Behaviour & Information Technology, vol. 23, 2004, pp. 247-264.

[29] G. Fontaine, "The experience of a sense of presence in intercultural and international encounters," Presence: Teleoperators and Virtual Environments, vol. 1, 1992, pp. 482-490.

[30] G. S. Kearns and A. L. Lederer, "A Resource-Based View of Strategic IT Alignment: How Knowledge Sharing Creates Competitive Advantage," Decision Sciences, vol. 34, 2003, pp. 1-29.

[31] E. Karahanna, D. W. Straub and N. L. Chervany, "Information Technology Adoption across Time: A Cross-Sectional Comparison of Pre-Adoption and Post-Adoption Beliefs," MIS Quarterly, vol. 23, 1999, pp. 183-214.

 ISBN: 978-1-61208-329-2

Semantic Agent of Informational Extraction on Big Data Ontological Context

Caio Saraiva Coneglian, Elvis Fusco
Department of Computer Science
UNIVEM –University Center Euripides of Marilia
Marilia, SP - Brazil
caio.coneglian@gmail.com, fusco@univem.edu.br

***Abstract*—The large increase in the production and dissemination of data on the Internet can offer information of high value-added to organizations. This information may be from heterogeneous databases that may not be considered relevant by most systems, e.g., social media data, blogs, and more. If organizations would use such sources, they could build a new management vision known as Competitive Intelligence. In the context of an architecture of Information Retrieval, this research that aims on implementing a semantic extraction agent for the Web environment, allowing information finding, storage, processing and retrieval, such as those from the Big Data context produced by several informational sources on the Internet, serving as a basis for the implementation of information environments for decision support. Using this method, it will be possible to verify that the agent and ontology proposal addresses this part and can play the role of a semantic level of the architecture.**

Keywords-Big Data; semantic web; semantic scrapper; ontology.

I. INTRODUCTION

The massive diffusion of generated data is testing the ability of the most advanced techniques of information storage technological, treatment, processing and analysis. The areas of treatment and information retrieval are being challenged by the volume, variety and velocity of semi-structured and unstructured complex data, offering opportunities for adding value to business-based information providing organizations a deeper and precise knowledge of their business.

Opportunities to add value to the businnes-based information arise due to both the internal and external environment. Hence, there is a need for a new approach to structure Information Technology (IT) companies to transform data into knowledge, which cause far-reaching impact.

To aggregate and use information that are scattered in the internal and external environments of organizations, there is the concept of Competitive Intelligence, which according Fleisher [1], is a process by which organizations gather actionable information about competitors and the competitive environment and, ideally, apply it to their decision-making and planning processes in order to improve their performance.

A proactive informational process leads to a better decision, whether strategic or operational, in order to discover the forces that govern the business, reduce risk and drive the decision maker to act in advance, besides protecting the knowledge generated.

In the current scenario of the information generated in organizational environments, especially in those who have the Internet as a platform, there is data that, due to its characteristics, is classified as Big Data.

In the literature, Big Data is defined as the representation of the progress of human cognitive processes, which generally includes data sets with sizes beyond the capacity of current technology, methods and theories to capture, manage and process the data within a specified time [2]. Gartner [3] defines Big Data as the high volume, high speed and/or high variety of information that require new ways of processing to allow better decision making, new knowledge discovery and process optimization.

In the process of information search for Competitive Intelligence and Big Data robots, data mining on the Internet are used; according to Deters and Adaime [5] robots are systems that collect data from the Web and assemble a database that is processed to increase the speed of information retrieval.

According to Silva [6], the extraction of relevant information can rank a page according to a domain context and also draw information structures them and storing them in databases. To add meaning to the content fetched, the robots are associated with Web search semantic concepts, which let the search through a process of meaning and value, extracting the most relevant information.

The ontology in the philosophical context is defined by Silva [6] as part of the science of being and their relationships; in this sense, the use of ontologies is essential in the development of semantic search robots, being applied in Computer Science and Information Science to enable a search smarter and closer to the functioning of the cognitive process of the user so that data extraction becomes much more relevant.

Thus, an agent presents itself as a solution to retrieve information on the web by semantic means. Currently, the content is organized in a jointly manner, in which syntactic structures do not have semantic data aggregation. In this sense, the role of the agent is to extract the information from the content and use syntactical ontology to achieve semantic relations and apply them to retrieval information.

This research aims to implement a semantic agent for searching on the Web and allowing the retrieval, storage and processing of information, i.e., Big Data from various informational sources on the Internet. Such semantic agent will be the main mechanism for building a computational

 ISBN: 978-1-61208-329-2

architecture that transforms disaggregated information on an analytical environment of strategic, relevant, accurate and usable knowledge to allow managers the access to opportunities and threats in the field of higher education institutions, based on concepts of competitive intelligence. The semantics of the agent will be built using ontological structures.

To achieve this goal, the Semantic Agent will be built using the domain of higher education institution, addressing the problem related to scientific research.

In this paper, the proposed architecture, the test of the ontology and the Semantic Agent are described.

II. Information Retrieval in Big Data

The traditional information systems are unable to cope efficiently with all new data sources and multiple contexts of information that have mainly the Internet as a platform.

Problems are encountered in retrieving, standardizing, storing, processing and usage of information generated by various heterogeneous sources that are the basis for enabling systems for decision support organizations.

In this context, it is questioning whether the computing environments of information actually present in full all relevant information to decision makers in organizations.

In this sense, Beppler [16] proposed a type of architecture of information retrieval. This recovery occurs only by analyzing documents, and removing and storing information, without observing the existing context, e.g., using syntactic analysis.

The solution was proposed to create an architecture for information retrieval in the context of Big Data, as seen in Figure 1.

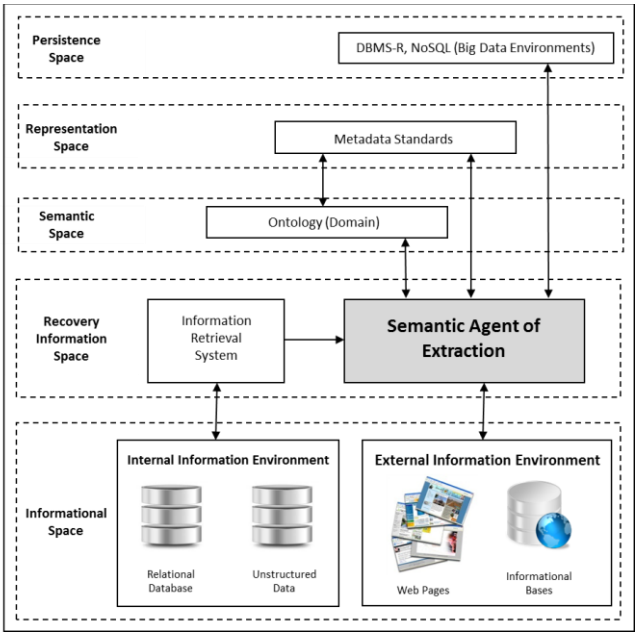


Figure 1. Architecture Context of Semantic Agent of Extraction.

Already Wiesner [17] proposed a semantic solution for this issue, using for making ontologies the recovery of information with a more generic architecture, by this the ontology you can insert semantics during the recovery process.

This architecture was proposed, so that the recovery of the information can be made using the semantic space. The architecture has the agent as the structure of information retrieval and integrates with all elements and layers of the architecture.

This architecture will be used at an institution of higher education, because in this area there is much information not being used several times in competitive intelligence.

This architecture distinguishes itself from others, for doing all the recovery information using the same domain, therefore, only information relating to the problem in accordance with the defined ontology is extracted.

III. Ontology

To set a network of Semantic Web, ontologies have been used frequently [13].

According to Clark [10], an ontology is organized into concept hierarchies, because it cannot reflect optimally specific formalisms; then, it is possible to consider an ontology as the embodiment of the knowledge level.

There are several types of ontologies, as outlined below [11]:

- Upper Ontology: this type of ontology serves to explain what exists in the world. And most are used to represent large knowledge bases;
- Domain Ontology: is a more specific area of an area of knowledge;
- Task Ontology: This ontology serves to solve a specific problem of a domain;
- Heavy-Weight Ontology: these ontologies are much more defined, have well-defined rules, be very careful when conceptualizing the world, and
- Light-Weight Ontology: this type need not be precise as large in the conceptualization.

Noy [8] explains the seven steps that are required to build an ontology, these steps are described below: 1. Determine the domain and scope of the ontology; 2. Consider reusing existing ontologies; 3. Enumerate important terms in the ontology; 4. Define the classes and the class hierarchy; 5. Define the properties of classes—slots; 6. Define the facets of the slots and; 7. Create instances.

The ontology proposal was hatched seeking cover a problem within the domain of higher education institution, which is the issue of scientific research, was used this issue to be able to have a more synthetic ontology, with the focus on extracting value information, why this ontology is not so great; so, one can get a better view about the Semantic Agent, which is the focus of this research.

 ISBN: 978-1-61208-329-2

The focus of the use of ontology in this case is for being the semantics of the agent. The agent will acquire the information from web pages, and from there pass the data by ontology implemented.

IV. SEMANTIC AGENT OF EXTRACTION

The creation of a software agent that aggregates semantically information available on the web in a given domain can bring to a computational platform grants for the creation of an information environment for decision support giving a broader view of the internal and external scenarios of information relevance in organizational management.

In this context, we understand the extreme importance of using agents to extract data through scrapper semantic search with the use of technologies like NoSQL [4] persistence in information processing with characteristics of Big Data, essential in the recovery, storage, processing and use of various types of information generated in these environments of large volume data sets on Competitive Intelligence.

In the context of the architecture presented in Figure 1, this research are dealing the problem of automatic and semantic information extraction of web environments that have as informational sources: web pages, web services and database with the development of the agent semantic of data extraction.

This agent should communicate with internal and external information spaces of Big Data basing their search on ontological rules on a metadata standard to perform the semantic extraction of the domain proposed and supported by other systems in a broader context of Information Retrieval.

From this semantic search, the scrapper comes as a tooling strategy in the search and find the information that really add value to the decision-making process. Inside a huge and massive data structure scattered throughout the web, it is essential that the search engines do not support only syntactic structures of decision in information retrieval, but also in investigations of the use of semantic extraction agents.

The research uses the domain of higher education institutions as a case study to apply the proposed computing platform in the architecture described in Figure 1. For the development of the prototype of the ontology, we used the issues of scientific research within educational institutions, such as notices, grants, funding agencies, search directories, events, and journals, among others.

To elaborate the conceptual notation of ontology, we used Protégé software [14], as shown in Figure 2, which shows the class hierarchy of the ontology. In this figure, the dotted arrows are properties of objects in each class, i.e., when a dotted arrow goes from one class to another, means that the class from which emerged the arrow contains an object of destination class of the arrow.

The agent will act on this proposed ontology that this scenario is called Task Ontology, according Mizoguch [11]; it is an ontology that solves a specific problem within a domain, that is, solves the problem of scientific research within the domain of an institution of higher education.

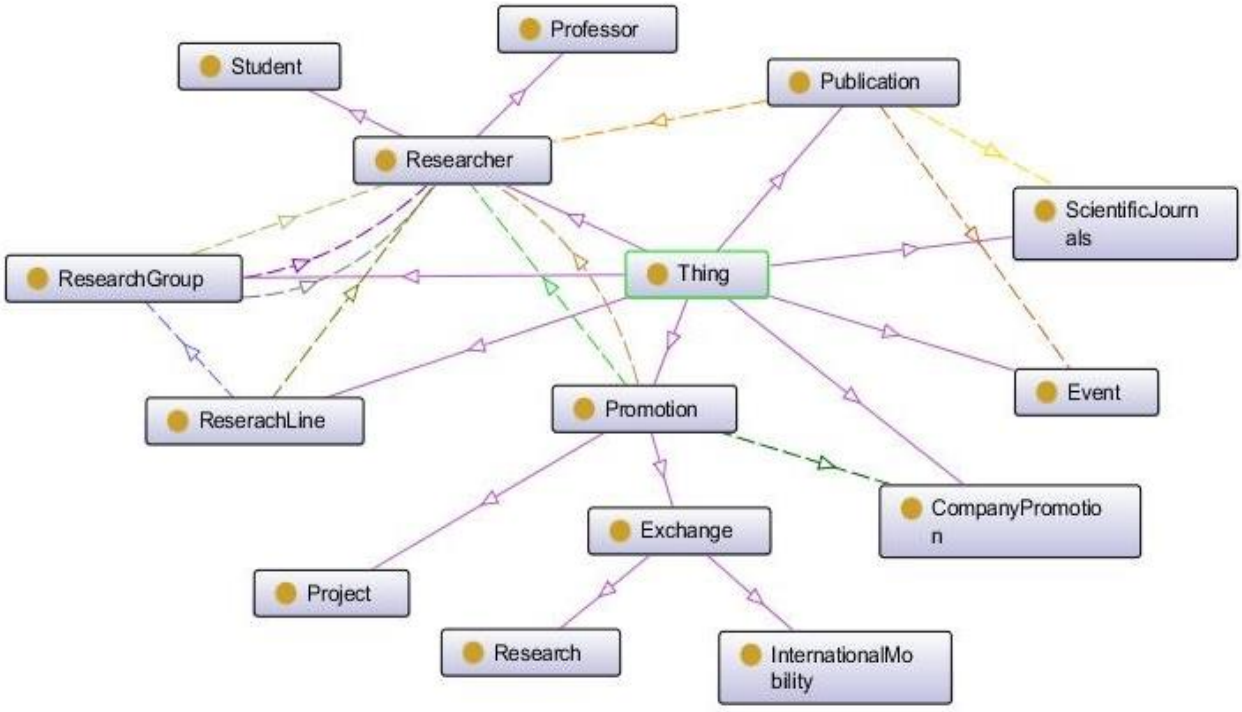


Figure 2. Class Hierarchy of Research Ontology.

 ISBN: 978-1-61208-329-2

V. CONCLUSION AND FUTURE WORK

Having access to information from your business domain is a fundamental requirement for management and decision making in organizations.

An Information Retrieval system has the ability to provide relevant information for accessing Web sites and services, it is necessary the existence of software agents that add semantic information from various informational sources for a specific domain.

In this context, the scrapper semantic search enters as a tooling strategy for searching and finding the information that really add value to the decision-making process, because inside a huge and massive structure of data spread across the Web, is essential that the search engines do not support only in syntactic structures of decision, but also in investigations of the use of semantic extraction agents.

Currently, we are performing the implementation of the ontology, and the integration with semantic extraction agent, that is also being created. This process is nearly being finalized. The agents will be tested for verifying if will be able to extract from there will be tests to check if the agent will be able to extract the information that really will add value to competitive intelligence.

By the development of an agent for semantic extraction, authors envision an effective use of information in the Environments Information Retrieval, an effective use of information in the scenarios of Big Data in the field of Higher Education Institutions will be obtained.

REFERENCES

[1] C. S. Fleisher and D. L. Blenkhorn, “Managing Frontiers in Competitive Intelligence”. 2001. Westport.

[2] “Big data: science in the petabyte era”. Nature 455 (7209): 1. 2008.

[3] Gartner, Douglas and Laney, “The importance of big data: A definition”. 2008.

[4] M. Diana and M. A. Gerosa, “NOSQL na Web 2.0: Um Estudo Comparativo de Bancos Não-Relacionais para Armazenamento de Dados na Web 2.0” (“NoSQL Web 2.0: A Comparative Study of Non-Relational Data Storage Benches for Web 2.0”). São Paulo, 2010.

[5] J. I. Deters and S. F. Adaime, “Um estudo comparativo dos sistemas de busca na web” (“A comparative study of search systems on the web”).

[6] T. M. S. Silva, “Extração De Informação Para Busca Semântica Na Web Baseada Em Ontologias” (“Information Extraction for Semantic Search In Web Based On Ontology”). Florianópolis, 2003. <https://repositorio.ufsc.br/handle/123456789/85791> [retrieved: 03/10/2014].

[7] M. A. A. Mesquita, “Web Semântica E Recuperação Da Informação Na Internet : O Que Esperar Do Futuro?” (“Semantic web and information retrieval on the Internet: What to Expect From the Future?” 2010.

[8] N. F. Noy et al, “Ontology Development 101: A Guide to Creating Your First Ontology”. <http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html> [retrieved: 03/10/2014].

[9] T. R. Gruber, “Towards Principles for a Design of Ontologies Used for Knowledge Sharing”, International Journal of Human and Computer Studies. 1995

[10] D. Clark, “Mad cows, meta-thesaurus and meaning, IEEE Intelligent Systems”. 1999.

[11] R. Mizoguchi, “Tutorial on Ontological Engineering”. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.76.6226&rep=rep1&type=pdf> [retrieved: 03/10/2014]

[12] J. Davies, D. Fensel and F. Van Harmelen, “Towards The Semantic Web: Ontology-Driven Knowledge Management”, John Wiley & Sons Ltd, 2003.

[13] R.A. Falbo, et al., “ODE: Ontology-based software Development Environment”. IX Congreso Argentino de Ciencias de la Computación, p. 1124-1135, La Plata, Argentina, Outubro 2003.

[14] Protégé. Stanford University. <http://protege.stanford.edu/> [retrieved: 03/10/2014].

[15] J.E. Prescott, “The Evolution of Competitive Intelligence”. 1999.

[16] F.D. Bepler, et al. “Uma Arquitetura Para Recuperação De Informação Aplicada Ao Processo De Cooperação Universidade-Empresa” (“An Architecture for Retrieval of Applied Information In Case Of University-Industry Cooperation”) 2005, São Paulo.

[17] K. Wiesner et al. “Recovery Mechanisms for Semantic Web Services” DAIS 2008, LNCS 5053, pp. 100–105, 2008.

 ISBN: 978-1-61208-329-2

Malay Semantic Text Processing Engine

Benjamin Chu Min Xian
MIMOS Berhad
Kuala Lumpur, Malaysia
mx.chu@mimos.my

Liu Qiang
MIMOS Berhad
Kuala Lumpur, Malaysia
qiang.liu@mimos.my

Rohana Mahmud
University of Malaya
Kuala Lumpur, Malaysia
rohanamahmud@um.edu.my

Arun Anand Sadanandan
MIMOS Berhad
Kuala Lumpur, Malaysia
arun.anand@mimos.my

Kow Weng Onn
MIMOS Berhad
Kuala Lumpur, Malaysia
kwonn@mimos.my

Dickson Lukose
MIMOS Berhad
Kuala Lumpur, Malaysia
dickson.lukose@mimos.my

***Abstract*—Semantic Text Understanding is a process that transforms text into conceptual representation. In this paper, we propose a Text Understanding System for Malay Language. The system comprises of two components: Morphology Analyzer and Semantic Text Interpreter. Some initial evaluation experiments were conducted on these components to gain explanatory insights into its performance. All the current text processing systems we reviewed are focused on preliminary algorithms and rules associated to lexical, morphological and syntax analysis. In our paper, we developed an integrated approach for a text understanding system that has the ability to represent the semantics of the text.**

Keywords-Natural Language Processing; Semantic Text Understanding; Morphology Analysis; Semantic Text Interpretation.

I. INTRODUCTION

The development of fast algorithms to understand and exploit the content of a document, and extracting useful information is very critical. In recent years, development in the area of semantic analysis of natural language text has triggered many applications in Text Mining, Summarization, Text Understanding, Information Retrieval and Extraction. Extracting actionable insight from large highly dimensional data sets, and its use for more effective decision-making, has become a pervasive problem across many fields in research and industry.

Extracting meaningful information from natural language text is the essential challenge that needs to be addressed. In developing these systems for main languages (e.g., English), the researchers have addressed several computational linguistic challenges including lexical, morphological, syntax and semantic processing. There are several fundamental challenges to semantic processing. Essentially, an extensive knowledge base is needed to process the text. Moreover, the complexity of defining rules for different languages when designing algorithms need to be addressed [1].

In this paper, our research focus on a Malay Language Text Understanding (MLTU) for standard Malaysian formal language, known as *Bahasa Malaysia (BM)* or the *Malay language*. Although, a wide demand and usage for the Malay language with a population of more than 28 million speakers, text processing systems geared for this language is still lagging behind.

This paper is structured as follows: Section 2 describes the related work on existing text understanding systems for Malay language; Section 3 describes our Semantic Text Processor system; Section 4 evaluates the performance of the system. Finally, Section 5 concludes this paper with a discussion on the overall outcome achieved and future research directions.

II. RELATED WORK

Several existing techniques in the current state-of-the art for text understanding generally aimed at constructing the syntax and semantic structures from texts. The main challenges for opened and natural language text understanding are caused by the ambiguity of natural language. As Malay native speakers, we will easily be able to understand the semantics of the following example sentence.

"Ali melihat Aminah dengan sebuah teleskop dan dia memanggilnya kuat-kuat" [Malay]

 ISBN: 978-1-61208-329-2

"Ali saw Aminah with a telescope and he is calling her loudly" [English translation]

However, the sentence itself for a machine to comprehend the meaning is quite difficult, as it lacks both the background knowledge and issues with the ambiguity of complex linguistic structures. Extracting meaningful information from natural language text is the essential challenge that should be addressed. In the existing systems, several Computational Linguistic challenges have been addressed focusing more on lexical, morphological and syntax analysis while lesser emphasis on semantic processing.

Many previous researchers in Natural Language Processing (NLP) had attempted to develop a Malay Morphology Analyser and Syntax parsers of speech tagger and parsers [2][3][4][5][6][7][8]. However, most works claimed the difficulties in resolving the stemming issues [9][10][11][12].

For example, the affixation method will derive various words that changed their syntactic class category from the original word (i.e., compared to English, which is forming a new word using inflection method; but, usually, the syntactic class category remains the same). For instance, the word makan (verb - purposely) becomes makanan (noun), when adding the suffix 'an'; becomes pemakanan (adjective), when adding circumfirxes 'pe…an', and becomes termakan (verb - unintentionally), when adding prefix 'ter'. Another major method of forming Malay language that is hardly found in other Languages is reduplication method, which can be full-duplication, such as the word kuat-kuat, or the partial duplication, such as lelaki (i.e., laki-laki).

All these characteristics and word formation issues create many problems for morphology analysis in Malay. Although the issues of labeling the morpheme and the dynamic nature of the syntactic category have been highlighted in MALEX [2][3] and MALIM [4], under-stemming and over-stemming problems remain unresolved [9][10][11][12].

All the systems we reviewed above are focused on preliminary algorithms and rules associated to syntax and morphology analysis. None are focused on developing an integrated approach for Malay Semantic Text Understanding. The ability to represent the semantics of the text is the most essential aspect of this approach. In the following section, we will describe the components of our Malay Semantic Text Understanding System.

III. SYSTEM DESCRIPTION

A. *Morphology Analyzer*

In the English morphology analyzer, stemming and lemmatization are the important task to allow the system to identify the root words. In Table I, the English verb for the different tenses may appear in different forms of spelling. For example, the verb 'walk', it will be appended with an affix's' in simple present tense, it spells as 'walks'; in present progressive tense it is appended with an affix 'ing' is appended, it spells as 'walking'; in simple past tense an affix 'ed' is appended and it spells as 'walked'. The verb 'eat' will change its spelling in various forms in different tenses: in simple past tense it spells as 'ate'; in present perfect tense it spells as 'eaten'. The English verbs will be changed in form spelling according to the tense. In the Malay language perspective, there will not be any spelling changes in the word for each grammar tense in Malay language; in the most of situation, an additional word will be added in front of the word to fulfill the grammar tense issue. As we observed, it is possible to perform Malay language analysis without stemming and lemmatization. As mentioned above, we will only be focusing on the Part-Of-Speech (POS) in Malay morphology analyzer in our initial system.

TABLE I. STEMMING AND LEMMATIZATION

English	Malay
walk	berjalan
walks = walk + s	berjalan
walked = walk + ed	telan berjalan
walking = walk + ing	sedang berjalan
eat	makan
ate	sudah makan
eaten	telah makan
beautiful	cantik
beautifully = beautiful + ly	dengan cantik

In the Malay POS module, we use Apache OpenNLP library [13] to perform Malay POS tagging task. The OpenNLP POS tagging module is language dependent and only performs well if the model language matches the language of the input text. Currently, it supports mainly for European languages. The Apache OpenNLP library is a machine learning based toolkit. We need to prepare for the Malay POS annotated corpus to train the OpenNLP POS tagger module for Malay language. In this experiment, we have collected about 2000 Malay sentences. We use of the Malay WordNet [14] to annotate the POS with each token of the sentences and validated by the Malay native speakers. After the corpus is annotated, 80% of the corpus is used for training and 20% of the corpus is used for evaluation. We are able to get very high accurate from the evaluation for Malay POS tagging with the new trained Malay POS module with the known words. Dataset preparation and evaluation results will be elaborated further in details in the following section of the experiment and evaluation in. There are three OpenNLP modules used to perform POS tagging: Sentence Detector, Tokenizer and Part-Of-Speech Tagger.

The OpenNLP Sentence Detector is able to detect punctuation characters to determine the end of a sentence. Malay and English language share the same alphanumeric and punctuation characters. Therefore, it is possible to directly use the existing English sentence module for the Malay language sentence detection task. The sentence

 ISBN: 978-1-61208-329-2

detector can be easily integrated into our application through OpenNLP API. As shown in Fig. 1, the input of the sentence Detector is a text string and the output is an array of Strings, where each string is one sentence.

```
Input
Peningkatan tahap Kepuasan pelanggan dan stakeholder mengenai penyampaian
perkhidmatan kewangan. Pergerakan pesawat udara, kapal dan jalan raya lebih
selamat. Sistem penyampaian perkhidmatan meningkat dan efisyen.

Output
[Peningkatan tahap Kepuasan pelanggan dan stakeholder mengenai penyampaian
perkhidmatan kewangan., Pergerakan pesawat udara, kapal dan jalan raya lebih
selamat., Sistem penyampaian perkhidmatan meningkat dan efisyen.]
```

Figure 1. Sentence Detector Input and Output

The OpenNLP Tokenizer segments the input character sequence into tokens. Tokens are usually words, punctuation, and numbers. The tokenizer module expects an input string, which contains the untokenized text. If possible, one sentence will be best input string for the tokenization module. In this experiment, the input array of the sentences is provided from the output of the Sentence Detector. The sample result is shown in Fig. 2. Tokenizer returns an array of strings where each string is one token.

```
Input
Peningkatan tahap Kepuasan pelanggan dan stakeholder mengenai penyampaian
perkhidmatan kewangan.

Output
[Peningkatan, tahap, Kepuasan, pelanggan, dan, stakeholder, mengenai,
penyampaian, perkhidmatan, kewangan, .]
```

Figure 2. Tokenizer Input and Output

The POS Tagger marks the input tokens with their corresponding POS tag based on the token itself and the context of the token. A token can possibly have multiple POS tags, the POS tagger uses maximum entropy probability model to predict the correct POS tag from the tag set. A tag dictionary is used by the POS tagger to limit the possible tags for a token; this will also increase the POS tagger tagging accuracy and performance. As shown in Fig. 3, the expected input of the POS tagger is a tokenized sentence in the form of string array where each of the strings is a token. The output is a tag array; it contains one POS tag for each token for the input array. The corresponding tag can be found at the same index of the tag array. The final output of the POS tagger will be a sentence where token and tag pairs are concatenated with an underscore, "_".

```
Input
[Peningkatan, tahap, Kepuasan, pelanggan, dan, stakeholder, mengenai,
penyampaian, perkhidmatan, kewangan, .]

Output
[NN, NN, NN, NN, CC, NN, VB, NN, NN, NN, .]

Final Output
Peningkatan_NN tahap_NN Kepuasan_NN pelanggan_NN dan_CC stakeholder_NN
mengenai_VB penyampaian_NN perkhidmatan_NN kewangan_NN ._.
```

Figure 3. POS Tagger Input and Output

B. Semantic Interpreter

For this module, we have extracted the grammatical rules from [19] and we have defined all of these programmatically for each of the thematic roles listed in Table II. Semantic Interpreter will use the rules defined to generate the semantic representation of the sentence. In this case, the semantic representation is in the form of Conceptual Graphs (CG).

For example, we can have a sentence as the input to this module, "*Kawalan ekonomi sepanjang tahun*" which means "*Economy restraint throughout the year*". From the previous module, this sentence will be annotated to produce the conceptual graph, which is shown in Fig. 4, as follows:

Annotated sentence:
Kawalan_NN ekonomi_NN sepanjang_IN tahun_NN

CG:

```
graph1:
        [kawalan]->(objek)->[ekonomi]->(durasi)->[tahun].
```

Figure 4. Simple Conceptual Graph in Malay

As shown in Fig. 4, this is a simple graph representing the meaning of the text. Moreover, we have defined rules to produce nested graphs for several sentence cases as shown below.

Sentence:
Meningkatkan harga barang dan minyak kerana inflasi negara.
English translation:
Increase the price of goods and oil due to the country's inflation

Annotated sentence:
Meningkatkan_VB harga_NN barang_NN dan_CC minyak_NN kerana_CC inflasi_NN negara_NN

CG:

```
g1: [meningkatkan]->(objek)->[harga]-
            {
            (objek)->[barang];
            (objek)->[minyak];
            }
g2: [inflasi]->(objek)->[negara]

g3: [situasi:*:(Penerangan,g1)]->(sebab)->[situasi:*:(Penerangan,g2)].
```

Figure 5. Nested Conceptual Graph in Malay

An example of a nested graph is shown in Fig. 5. In g1, the concept [harga] is the object (objek) of the verb [meningkatkan]. The concept [harga] is linked by the object relation (objek) to both concepts [barang] and [minyak] due to the conjunction in the sentence. Similarly in g2, the concept [inflasi] is linked by the object relation (objek) to the concept [negara]. In g3, a situation described by g2 is caused

 ISBN: 978-1-61208-329-2

by a situation expressed in g1. The relation "caused by" between these two situations is using the Malay thematic role "sebab". Table II shows the complete listing of the thematic roles used in Malay.

TABLE II. THEMATIC ROLES FOR MALAY

Malay	English translation
Pelaku	Agent
Alami	Experiencer
Alat	Instrument
Asal	Origin
Bilangan	Amount
Destinasi	Destination
Deritaan	Patient
Durasi	Duration
Gaya	Manner
Hasil	Result
Kepunyaan	Possession
Kesan	Effector
Manfaat	Beneficiary
Muasal	Matter
Objek	Object
Permulaan	Start
Penyertaan	Accompaniment
Perbandingan	Comparand
Sebab	Because
Sifat	Attribute
Tema	Theme
Tempat	Location
Tujuan	Purpose
Ukuran	Measurement
Waktu	PointinTime
Perhinggaan	Completion
Penafian	Negation
Jalan	Path

IV. EXPERIMENTAL EVALUATION

A. Datasets

In the current state of the art, there is no Malay language POS annotated corpus that is available to train the POS module for Malay language. Many previous attempts have been done to prepare the data manually [3][8][15]. With the unavailability of any Malay POS annotated corpus, data preparation is an important task in this initial research work.

As the first step, the POS data was extracted by utilizing both Malay WordNet [14] and Apertium [16] Malay to Indonesian translation dictionary. The Malay WordNet is a lexical dictionary (currently supports Malaysian and Indonesian). The dictionary comprises of 19,210 synsets, 48,110 senses and 19,460 unique words with POS tag in the Malay WordNet, where all the relations (hypernyms, meronyms etc.) are extracted from WordNet. This project was initiated by Francis Bond from Nanyang Technological University [17]. The project is inspired by Princeton WordNet since there is no lexical dictionary for Malay language. Apertium is a machine translation engine designed to translate closely related languages. The current Apertium engine supports language translation from Indonesian to Malay. In doing so, the engine uses POS information and translation rules for Malay and Indonesian words. We extracted this POS data Apertium, along with Malay WordNet, to build our POS annotation corpus.

In this research work, we collected about 2000 Malay sentences as our dataset. We also created a module to extract and combine the Malay POS data for the Malay WordNet and Apertium. Once Malay POS data dictionary is ready, we created another module to parse and annotate all possible POS for the Malay sentences base on the POS dictionary, as the result some of word may have annotated with multiple POS tag. The final step, native Malay speaker will need to validate and correct the tags for all the Malay sentences. Fig. 6 shows the annotation result for each the steps involved.

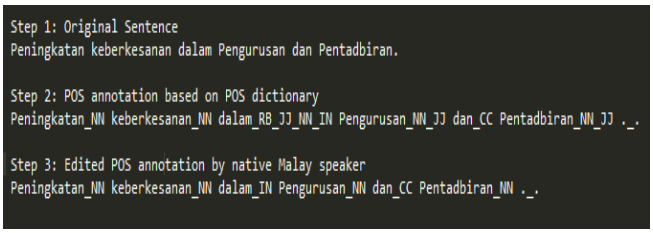


Figure 6. Malay Part-Of-Speech Annotation Sample

During implementation and experiment, 80% of annotated sentences were used for POS module training data; the rest of the 20% were used as evaluation data.

B. Evaluation Results

Based on the methods described above, evaluation has been conducted to determine the accuracy of the two main modules; Morphology Analyzer module and the Semantic Interpreter module.

- *Morphology Analyzer*

The overall accuracy of the POS tagging was calculated as the ratio of correct POS tags found by the system over the total number of POS tags. The accuracy scores along with the corpus size are plotted in Fig. 7. Between Phase 1 and Phase 2, the inconsistencies in the POS annotations were fixed. For example, the Malay word "dan" was annotated as preposition "IN", in some sentences and as conjunction "CC" in other cases. In Phase 3, along with increasing the number of annotated sentences, a Tag dictionary is a word dictionary, which contains specified POS tags for the tokens. This ensured that inappropriate tags were assigned to tokens, which will result in better accuracy. Naturally, increasing the number of annotated sentences resulted in better accuracy, until a plateau was reached, at 2000 annotated sentences.

 ISBN: 978-1-61208-329-2

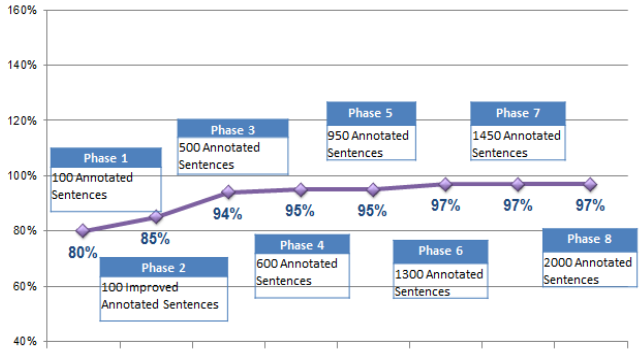


Figure 7. Morphology Analyzer Experimental Results

- *Semantic Interpreter*

In evaluating this module, we have created 70 graphs manually as the gold standard for our benchmark. The results produced by the system were classified as Correct (indicating full match), Partial (indicating a partial match) or Incorrect (indicating incorrect representation). As shown in Fig. 8, the results show that 62 graphs were classified as correct, 7 as incorrect and 1 partial match.

V. DISCUSSION

Upon analyzing the results, it was found that the partial match was due to a missing concept in the knowledge base. Fig. 8 shows the knowledge base is based on the Malay WordNet with over 30,000 concepts. Although the partial match is only 1%, but extending and enriching this knowledge base with more concepts will further improve the interpretation accuracy. One of the reasons behind the incorrect results was found to be the lack of support for anaphora resolution. For example, this is shown in Fig. 9 where the pronoun 'mereka' is not being resolved to the noun 'penduduk'.

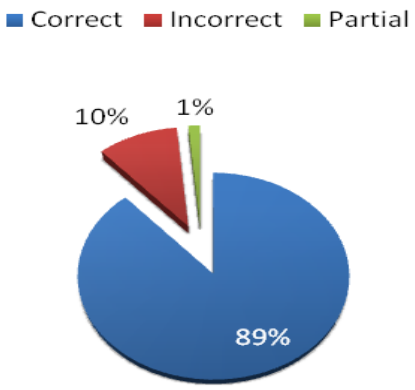


Figure 8. Semantic Interpreter Experimental Results

```
g1:  [penduduk]<-(agen)<-[pergi]->(destinasi)->[bandar]
g2:  [mereka]<-(agen)<-[mencari]->(objek)->[kerja]
g3:  [situasi:*:(Penerangan,g1)]<-(sebab)<-[situasi:*:(Penerangan,g2)]
```

Figure 9. Incorrect representation without anaphora resolution

Anaphora resolution [18][19] is the problem of identifying how contextual entities are referred within a single or several sentences, typically what a pronoun or a noun phrase is referring to. For example, from the sentences *john loves mary* and *he wishes to marry her*, the entity *john* is referred by *he* and *mary* is referred by *her*. Another example in Malay language can be seen in the following sentence. "Penduduk pergi ke bandar kerana mereka mencari kerja" is translated as "the villagers went to the city because they wanted to find a job". Here, the word 'mereka' (they) is referring to a pronoun; therefore it should be resolved to "penduduk" (villagers). As shown in Fig. 10, the correct representation of the graph:

```
g1:  [penduduk:$cc9]<-(agen)<-[pergi]->(destinasi)->[bandar]
g2:  [penduduk:$cc9]<-(agen)<-[mencari]->(objek)->[kerja]
g3:  [situasi:*:(Penerangan,g1)]<-(sebab)<-[situasi:*:(Penerangan,g2)]
```

Figure 10. Correct representation with anaphora resolution

where the reference indicator *$cc9* would denote the coreference.

VI. CONCLUSION AND FUTURE WORK

State-of-the-art-text processing systems for Malay Language are still dealing with problems related to lexical, morphological and syntax analysis. Based on syntax analysis alone, meaning through syntax is still insufficient to explain the comprehension of natural language texts. Therefore, we proposed an integrated approach for Malay Text Understanding, which included both syntax and semantic processing.

In summary, we have developed Morphology Analyzer and Semantic Interpreter components. From a qualitative comparison perspective, we have evaluated both components on how well they can perform (this is quite subjective, and is based on our initial benchmarking exercise).

In future, we plan to enrich our Malay Linguistic knowledge base derived from Malay WordNet with other linguistic resources. We will continue to evaluate both of our components with a large news dataset to improve our semantic rules. Furthermore, we will also explore Coreference Resolution for Malay Language. Coreference Resolution will help to refine the semantic representation produced by resolving all anaphors and cataphors to their intended referents.

 ISBN: 978-1-61208-329-2

REFERENCES

[1] Benjamin Chu, Fadzly Zahari, and Dickson Lukose, Large-Scale Semantic Text Understanding. In Semantic Technology and Knowledge Engineering (STAKE) 2010 Conference Proceedings, Sep. 2010, pp. 28-39.

[2] Knowles, Gerald and Zuraidah Mohd Don, Tagging a corpus of Malay texts and coping with 'syntactic drift'. Proceedings of the corpus linguistics, 2003, pp. 422-428.

[3] Zuraidah Mohd Don, Processing Natural Malay Texts: A Data Driven Approach, TRAMES, Vol 14 (64/59), 2010, pp. 90-103.

[4] Mohd Yunus Sharum, Muhammad Taufik Abdullah, Mohd Nasir Sulaiman, Masrah Azrifah Azmi Murad, and Zaitul Azma Zainon Hamzah, MALIM- A new computational Approach of Malay Morphology, IEEE, Vol 2, 2010, pp. 837-843.

[5] Timothy Baldwin, and Suad Awab, Open Source Corpus Analysis Tool for Malay, Retrieved Nov. 2013, from: https://code.google.com/p/malay-toklem/

[6] Mat Awal, Norsimah Abu Bakar, Kesumawati and Abdul Hamid, Nor Zakiah Jalaluddin, and Nor Hashimah, Morphological Differences between Bahasa Melayu and English: Constraints in Students' Understanding, The Second Biennial International Conference on Teaching and Learning of English in Asia, 2007, pp.1-11.

[7] H. Mohamed, N. Omar Abd, and M. J. Aziz, Statistical Malay Part Of Speech Tagger using Hidden Markov Approach, International Conference on Semantic Technology and Information Retrieval, Putrajaya, June. 2011, pp. 231-236.

[8] Rayner Alfred, Adam Mujat, and Joe Henry Obit, A Rule-based Part Of Speech (RPOS) Tagger for Malay Text Articles; A. Selamat (Eds), ACIIDS 2013, Part 11, LNAI 7803, 2013, pp. 50-59.

[9] Salhana Amad Darwis, Rukaini Abdullah, and Norisma Idris, Exhaustive Affix Stripping and a Malay Word Register to solve stemming errors and ambiguity problem in Malay Stemmers, Malaysian Journal of Computer Sciences, Vol 25, 2012, pp. 213-218.

[10] Bali Ranaivo-Malancon, Computational Analysis of Affixed Words in Malay Language, Unit Terjemahan Melalui Komputer, USM, Technical Report, 2004.

[11] S. A. Fadzli, and A. K. Norsalehen, I. A. Syarilla, H. Hasni, and M. S. S. Dhalila, Simple Rules for Malay Stemmer, The International Conference on Informatics and Applications, Jun. 2012, pp. 28-35.

[12] Y. L. Tan, A Minimally-Supervised Malay Affix Learner, Proceedings of the Class 2003 Senior Conference, Swarthmore, 2003, pp. 55-62.

[13] OpenNLP, Retrieved Nov. 2013, from: http://opennlp.apache.org

[14] Malay Wordnet, Retrieved Dec. 2013, from:
http://wn-sa.sourceforge.net/index.eng.html

[15] Norshuhani Zamin, Alan Oxley, and Zainab Abu Bakar, A Lazy Man's Way to Part-Of-Speech Tagging, Knowledge Management and Acquisition for intellignent system lecture notes in cimputer science, vol 7457, 2012, pp. 106-117.

[16] Apertium a free/open-source machine translation platform, Retrieved Nov. 2013, from: http://www.apertium.org/

[17] Noor Nurril Hirfana, Bte Mohamed, Saquan Suerya, and Bond Francis, Creating the Open Wordnet Bahasa, Proceeding of the 25th Pacific Asia Conference on Language, information and Computation, 2011, pp. 255-264.

[18] Ruslan Mitkov, and Wv. Sb. Wolverhampton, Anaphora Resolution: the State of the Art. Computational Linguistics, 1999, pp. 1-34.

[19] Asmah Haji Omar, Nahu Melayu Mutakhir 5th Edtion. Kuala Lumpur: Dewan Bahasa Pustaka, 2009.

 ISBN: 978-1-61208-329-2

Semantic-based Multilingual Islamic Finance Thesaurus

Aziza Mamadolimova
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: aziza@mimos.my

Nor Azlinayati Abdul Manaf
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: azlinayati.manaf@mimos.my

Jasbeer Singh Atma Singh
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: jasbeer@mimos.my

Farouq Hatem Hamad
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: farouq.hamed@mimos.my

Nurul Aida Osman
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: nurulaida.osman@mimos.my

Khalil Ben Mohamed
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: khalil.ben@mimos.my

Dickson Lukose
Knowledge Technology Department,
Artificial Intelligence Centre
MIMOS Berhad
Kuala Lumpur, Malaysia
e-mail: dickson.lukose@mimos.my

***Abstract*— In this paper, we present an attempt to build a semantic-based multilingual Islamic Finance thesaurus, with the aim of globally standardizing the use of Islamic Finance concepts and providing a rich, semantically sound terminology. We describe a semantic model, which uses international standards such as Simple Knowledge Organization System, a World Wide Web Consortium recommendation designed among others for representation of thesauri, with the aim of enabling easy publication of the thesaurus as part of Linked Data and efficient future use. Finally, we present an Islamic Finance thesaurus collaborative authoring tool, which allows people around the world to contribute in improving the Islamic Finance thesaurus by adding, modifying or deleting concepts, relationships between the concepts or descriptions.**

Keywords-Islamic Finance; Thesaurus; Linked Data; SKOS; RDF

I. INTRODUCTION

Islamic finance and banking institutions worldwide have grown at a remarkable pace for the last three decades. According to a study by the International Monetary Fund (IMF), the number of Islamic finance and banking institutions rose from 75 in 1975 to over 300 in 2005, in more than 75 countries. Total assets worldwide are estimated at $250 billion, and growing at about 15% per annum [1]. Islamic banking and finance concept is also gaining popularity all over the world as highlighted by Ms. Christine Lagarde, director of IMF, who said: "to make (Islamic banking) activities as welcome in Paris as they are in London and elsewhere" [2].

This growth of interest in Islamic Finance and Banking is continuously attracting more stakeholders, but each one uses its own definitions or spellings for core Islamic Finance concepts, creating a lot of ambiguity and misunderstanding within the community. To the best of our knowledge, there exists no standard Islamic Finance thesaurus to refer to; thus, the major objective of this paper is to build such a thesaurus promoting standardization and interoperability in Islamic Finance domain.

Existing financial terminologies are representing knowledge about conventional banking and finance, e.g., the Financial Industry Business Ontology (FIBO) [3]. FIBO

 ISBN: 978-1-61208-329-2

presents knowledge about financial instruments, business entities, market data and corporate actions in a technology neutral format along with formal definitions and defined business relationships. FIBO standardizes the language of conventional financial contracts and promotes unambiguous shared meaning among all participants in the conventional banking and finance world. In this paper, we aim to complement the existing banking and finance ontologies by considering all the terms specifically used in Islamic finance domain, which include Islam-related terms, Islamic contracts legacy, Islamic scholars, etc.

In order to be widely used around the world, the thesaurus must conform to semantic web standards and would eventually be published as part of the Linked Data (LD) [4]. Linked Data provides a network of interlinked structured data from various sources; it contains as per September 2011 more than 31 billion interlinked pieces of information from different domains (e.g., Health, Economy). We built the Islamic Finance thesaurus using Simple Knowledge Organization System (SKOS) [5], a World Wide Web Consortium (W3C) [6] recommendation designed for representing thesauri, taxonomies, or any other type of structured controlled vocabulary. SKOS is part of the semantic web family of standards built upon Resource Description Framework (RDF) [13], and its main objective is to enable easy publication and use of such vocabularies as Linked Data.

The viability of the thesaurus, i.e., its management and continuous enrichment over the years, is a core aspect [7]. Existing well-established thesauri, ontologies or knowledge bases (e.g., AGROVOC [8], Systematized Nomenclature of Medicine Clinical Terms [9], National Cancer Institute [10]) propose management tools in which the collaborative authoring section is primordial. It allows people around the world to contribute making the terminology a success, for example by adding new descriptions of concepts (e.g., translating a term in another language). An example of such a tool is "VocBench" [11], developed to manage and author AGROVOC knowledge base [12]. Thus, the terminology can continuously be refined, until it eventually caters to the needs of stakeholders from the domain. In this paper, we also propose an Islamic Finance thesaurus tool for browsing the thesaurus as well as collaboratively authoring its content, at first locally then worldwide through the LD.

This work is the result of a broader collaboration started in 2011 with the Institute of Islamic Banking and Finance in Malaysia (IIBF). The collaboration includes the semantic representation of Islamic contracts [17], the semantic representation of Islamic finance terms and an application to automatically generate meaningful questions in different modalities. The modalities investigated are multiple choice questions as well as complex modalities using diagrams to represent the question and semantic similarity to assess free-drawn answers. In this paper, we only focus on presenting the semantic representation of Islamic finance terms as well as the developed collaborative authoring tool.

Section 2 recalls the basics of utilized semantic web standards such as RDF and SKOS. In Section 3, we present the semantic model for the Islamic Finance thesaurus and the followed methodology. Section 4 presents a collaborative authoring tool developed to facilitate the management and enrichment of the thesaurus. The prospects of this work are outlined in Section 5.

II. RDF AND SKOS

A. RDF

RDF provides a common framework for expressing information in order to be processed by applications and exchanged between applications without loss of meaning.

RDF can be used to assign attributes and values to resources and to express relationships between resources. It allows computers to know something about a subject and provides a general method to decompose information into pieces. RDF requires a subject (or resource), predicate (or relationship) and object (or value).

B. SKOS

SKOS provides a standard way to represent knowledge organization systems using RDF. SKOS is a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other similar types of controlled vocabularies. SKOS allows concepts to be composed and published on the World Wide Web, linked with data on the Web and integrated into other concept schemes.

The main components of SKOS model are: *Concepts, Labels, Semantic Relationships, Documentary Notes and Concept Schemes*. Below, we present some extracts of the SKOS Primer [5] document explaining briefly the components we are using in the following.

Concepts: Concepts are fundamental elements of the SKOS vocabulary. Concepts are the units of thought—ideas, meanings, or (categories of) objects and events, which underlie many knowledge organization systems [5]. The class *skos:Concept* allows implementers to assert that a given resource is a concept.

Labels: Labels are expressions that are used to refer to the concepts in natural language. SKOS provides three properties to attach labels to conceptual resources: *skos:prefLabel*, *skos:altLabel* and *skos:hiddenLabel*.

Semantic Relationships: In Knowledge Organization System (KOS)'s, semantic relations play a crucial role for defining concepts. The meaning of a concept is defined not just by the natural-language words in its labels but also by its links to other concepts in the vocabulary. Mirroring the fundamental categories of relations that are used in

 ISBN: 978-1-61208-329-2

vocabularies such as thesauri [14], SKOS supplies three standard properties:

- *skos:broader* and *skos:narrower* enable the representation of hierarchical links, such as the relationship between one genre and its more specific species.
- *skos:related* enables the representation of associative (non-hierarchical) links, such as the relationship between one type of event and a category of entities, which typically participate in it.

Documentary Notes: Concepts sometimes can be further defined using human-readable ("informal") documentation, such as scope notes or definitions. SKOS provides a *skos:note* property for general documentation purposes, which is further specialized into *skos:scopeNote*, *skos:definition*, *skos:example*, and *skos:historyNote* to fit more specific types of documentation.

- *skos:scopeNote* supplies some, possibly partial, information about the intended meaning of a concept.
- *skos:definition* supplies a complete explanation of the intended meaning of a concept.
- *skos:example* supplies an example of the use of a concept.
- *skos:historyNote* describes significant changes to the meaning or the form of a concept.

Collections of concepts: SKOS allows the definition of meaningful groupings or "collections" of concepts. In thesaurus terminology these collections are known as "arrays", and the label that groups the terms in the collection is a "node label". To correctly model such concept collection structures, SKOS introduces a *skos:Collection* class. Instances of this class group specific concepts by means of the *skos:member* property.

III. MODELLING OF ISLAMIC FINANCE THESAURUS

The intention of engineering Islamic Finance using semantic technologies is to build a unique and well established thesaurus for all Islamic Finance users. To the best of our knowledge, there exists no similar work in Islamic Finance field. Another advantage of building a thesaurus for Islamic Finance is to enable users around the world not to start their ontology modelling from scratch. It also promotes interoperability between semantic-based applications using the thesaurus.

Modelling of the Islamic Finance Thesaurus begins with the conceptualization of the Islamic Finance concepts and properties, which are represented as resources using RDF/SKOS formalisms. To model the thesaurus we implemented Spiral Modelling [15] technique, which is usually used for ontology modelling. The requirements are provided by IIBF and the knowledge engineers from MIMOS work closely with the domain experts from IIBF to evaluate the facts. Figure 1 shows the iterative process of getting domain knowledge from experts, modelling this knowledge, verifying it with the help of domain expertise and finally engineering it.

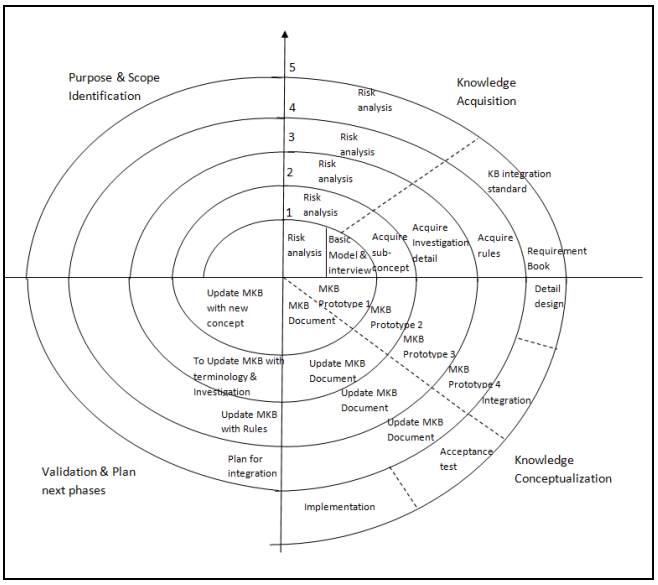


Figure 1. Knowledge Modelling - Spiral Model

We used TopBraidComposer [16] tool for engineering the thesaurus. The SKOS data model of the thesaurus has three concepts at the top level *skos:Collection*, *skos:Concept*, *skos:ConceptScheme*, as shown in Figure 2. All Islamic Finance terms are defined as type of *skos:Concept* class. Islamic Finance concepts that have similar themes are grouped under several categories (e.g., Islamic Sale Contracts), which are defined as type of *skos:Collection*. Finally the whole Islamic Finance thesaurus is defined as type of *skos:ConceptScheme*.

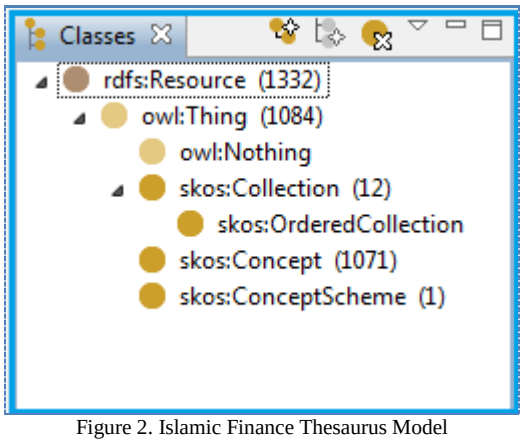


Figure 2. Islamic Finance Thesaurus Model

Each concept is identified by a unique Universal Resource Identifier (URI). Throughout this paper, we will use the prefix “miIFT” to indicate the namespace for

 ISBN: 978-1-61208-329-2

"http://www.mimos.my/IFT#". For example, the concept "http://www.mimos.my/IFT#Ajr" is shown as *miIFT:Ajr*.

Within this thesaurus, there is a total of 1071 concepts and all the terms are related to Islamic Finance in one way or another. For example, there are different kinds of Islamic Finance contracts such as *miIFT:Murabahah*, *miIFT:Ijarah*, *miIFT:Istisna*, *miIFT:Al-Mudharabah*; different types of financial instruments such as *miIFT:Cash*, *miIFT:Check*; different Shariah scholars according to whose opinion certain transaction may or not take place such as *miIFT:Abu_Hanifah*, *miIFT:Abu_Yusuf*, *miIFT:Ahmad_ ibn_ Hanbal*, *miIFT:Ibn_Hazm*.

Figure 3 shows the model of the concepts *miIFT:Ajr* and *miIFT:Ajr-un-kareem* in a graph representation.

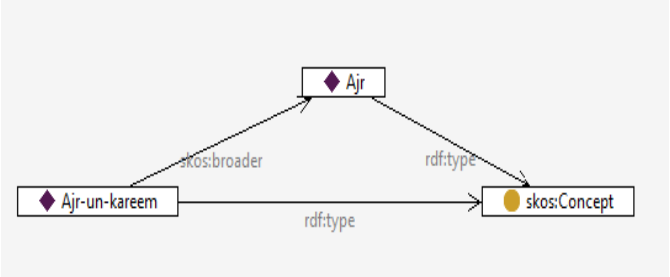


Figure 3. Model of the concepts "*miIFT:Ajr*" and "*miIFT:Ajr-un-kareem*"

Figure 4 shows the modelling of the *miIFT:Murabahah* concept and its related properties. We can see that *miIFT:Murabahah* has broader concept *miIFT:IslamicSaleContract* and it is related to *miIFT:Investment*. All of them are SKOS concepts. It also has several labels (two preferred labels -English and Arabic, two alternate labels and one hidden label), a history note and a definition.

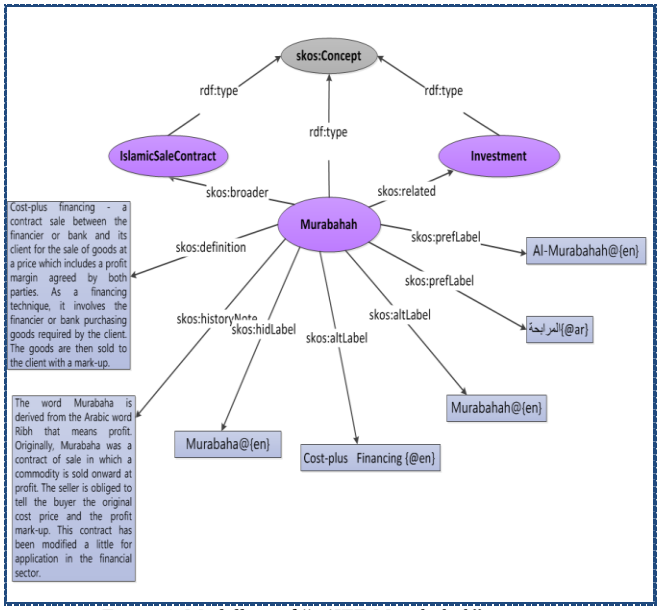


Figure 4. Modelling of "*miIFT:Murabahah*" concept

In the current modelling of the thesaurus, all the terms are in two languages: English and Arabic. But, the system described in the next section is catered to capture any language, because the primary goal of this system is to be able to provide collaborative authoring of its content for users all over the world. Thus one concept can have multiple labels -*skos:prefLabel*, *skos:altLabel*, *skos:definition*, *skos:scopeNote*, in different languages. For example, the word *miIFT:Ajr* has an English preferred label *Wage*{@en} and it has a preferred label in Arabic اجرة{@ar} and اجرت{@ur} in Urdu.

As mentioned before, SKOS data model allows to define other kind of spelling of particular term via the use of *skos:altLabel* (alternate label). The primary purpose for *skos:altLabel* is for synonyms, abbreviations, acronyms and different spellings in different languages. For example, for the organization *miIFT:IslamicConferenceFiqhAcademy* it is possible to have all the following alternate labels:

- *OICFA* (acronym*);*
- *OIC Fiqh Academy* (abbreviation)*;*
- *Organization of the Islamic Conference Fiqh Academy* (official name in English);
- مجمع الفقه الإسلامي المؤتمر منظمة (official name in Arabic).

This approach provides high flexibility at the lexical level.

Another part of our model uses SKOS ability to create various collections, which are designed to group concepts sharing common features together. For instance all Islamic Finance contracts are member of *miIFT:IslamicFinanceContract* collection; all financial instruments are member of *miIFT:FinancialInstrument* collection and all shariah scholars are member of *miIFT:ShariahScholar* collection. We use the property *skos:member* to create these links.

Figure 5 shows that *miIFT:Ijarah*, *miIFT:Salam* and *miIFT:Istisna* are all type of SKOS concept and member of *miIFT:IslamicFinanceContract* collection. Similarly for *miIFT:Cash* and *miIFT:Check*, which are members of the collection *miIFT:FinancialInstrument*.

The model has been reviewed by Islamic finance experts from IIBF and knowledge engineers from MIMOS. The process required three iterations/sprints from April 2013 to September 2013, and IFT model has finally been validated and baselined by IIBF on the 26th of September 2013. The content is still under construction and daily updated by research assistants from IIBF with the support of MIMOS.

IV. COLLABORATIVE AUTHORING TOOL

In this section, we present an Islamic Finance thesaurus collaborative authoring tool (called Islamic Finance

Vocabulary –IFV), which allows people around the world to contribute in improving the Islamic Finance thesaurus by adding, modifying or deleting concepts, relationships between the concepts or descriptions.

IFV is a web-based system, which consists of a set of features to assist users in managing the Islamic Finance Thesaurus. These features include navigation, search, authoring, validation, user management and ontology management, which are essential functionalities for knowledge management activities. IFV front-end was built using Adobe Flex/ActionScript 3 and its back-end was based on Java Spring framework on top of a MIMOS platform called the Semantic Technology Platform (STP).

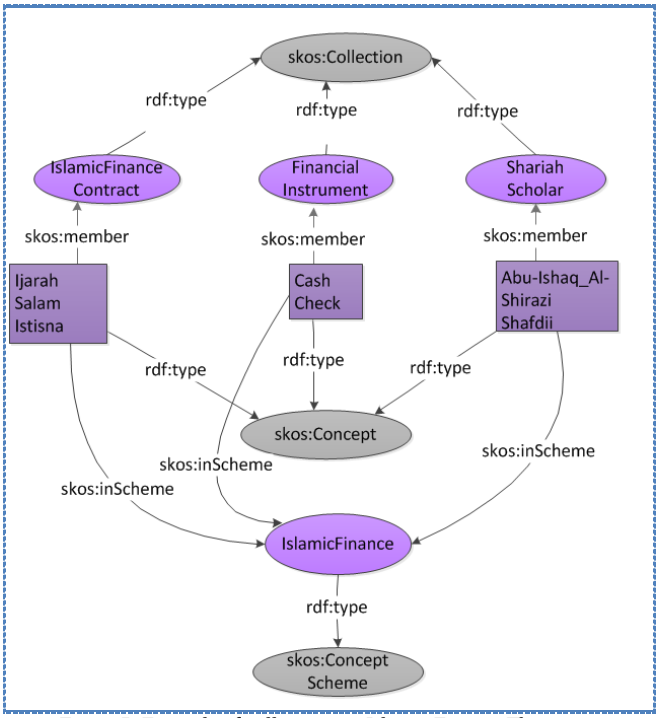


Figure 5. Example of collections in Islamic Finance Thesaurus

Figure 6 shows the IFV overall system architecture. It depicts the web-service components running in the back-end and how they communicate with the Client Application (IFV) GUI. The Service Oriented Architecture (SOA) is adopted as the base technology to ensure scalability. The front-end application is communicating with the back-end components through the core STP web-service components (Delegator, Authenticator and Lookup service). The major function of the Delegator is to distribute all tasks requested by the user from the front-end application to the back-end components and return respective data to front-end application. The front-end application is mainly focused on user interaction and data displaying functionalities.

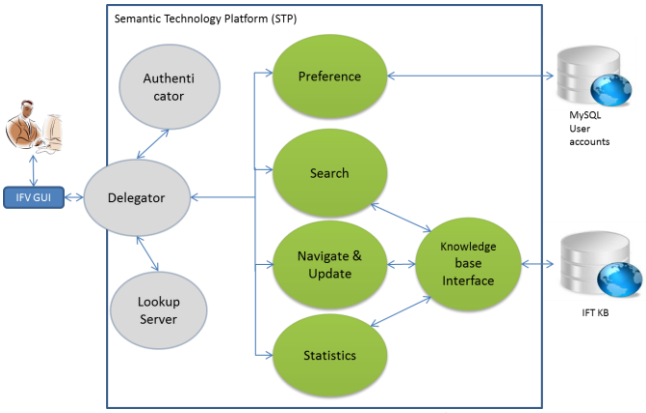


Figure 6. System Architecture of IFV

Apart from the core STP components, the system also includes other web-service components to enable the functional features of IFV, a database and a knowledge base. The features of the components and data sources are described below:

1) IFT knowledge base (IFT KB): contains the Islamic Finance Thesaurus and associated metadata.

2) User account database: contains information about the user details, such as user name, passwords, roles, status etc.

3) Preference: this component is responsible for displaying/modifying the user's profiles and accounts.

4) Search: this component is responsible for receiving queries from IFV GUI and returning the results of the queries.

5) Statistics: this component is responsible for displaying the most discussed, top visited concepts and the top contributors to IFT and their registered places.

6) Navigate and Update: this component is designed for the users to navigate IFT and update its content. The navigation view consists of concepts hierarchy, instances list and details of triples.

The IFV has been reviewed by IIBF on November 2013 and is presently in prototype mode. It is currently in testing phase after the addition of several functionalities such as a search based on Arabic root word.

Figures 8 and 9 are snapshots of the IFV Graphical User Interface. Figure 8 illustrates the landing page.

 ISBN: 978-1-61208-329-2

Figure 8: Home page of IFV system

Figure 9: Navigation view of IFV system

Figure 9 shows the navigation view allowing users to navigate through IFT by selecting specific collections, concepts and instances.

The chart in Figure 7 illustrates the approval workflow when updating IFT. The workflow is as follows, where the validators are the Islamic finance experts from IIBF:

i. The process starts with the user making a suggestion.
ii. It will go to the validator who will check the suggestion.
iii. If the validator needs more clarifications, the system prompts the user for more clarification and it goes back to Step i).
iv. Else, if no clarification is needed, the process will continue to check if the validator agrees with the suggestion or not.
v. If the validator agrees, then the suggestion is approved and committed to the IFT KB and the process ends.
vi. Otherwise if the validator disagrees, the suggestion is rejected and purged, and the process ends.

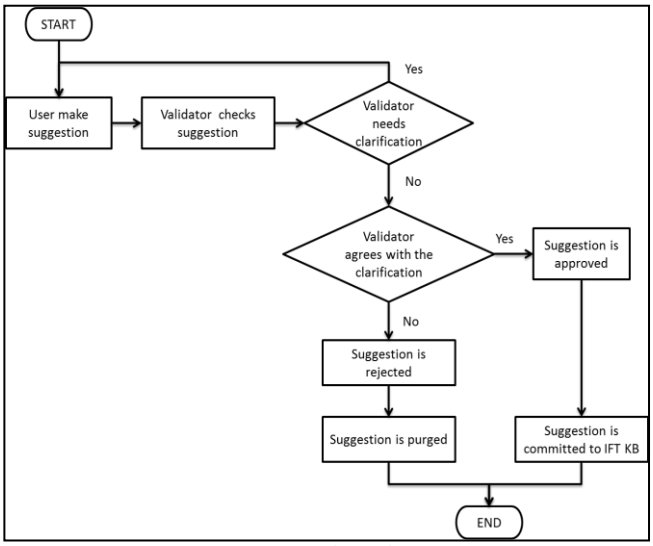


Figure 7: Approval workflow of editing/updating

This workflow has been adopted locally by IIBF, where all the validators are Islamic finance experts from IIBF. We foresee this workflow to be hardly adopted at an international scale, especially if the validators are coming from different Islamic religious groups. A possibility would be to allow the representation of different viewpoints for a single concept, for example by modifying the notion of labels to cover the representation of different school of thoughts, and by categorizing the validators based on their school of thought.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented our work on modelling semantic-based multilingual Islamic Finance Thesaurus. We aim to provide globally standardized use of Islamic Finance concepts as well as providing a rich, semantically sound terminology. To describe Islamic Finance concepts, relations between the concepts and categorization of concepts, we utilized the SKOS data model, a W3C Recommendation designed for representation of thesauri.

We also described a web-based system, the Islamic Finance Vocabulary (IFV), which was developed as a platform to navigate, share and collaboratively author the content of our Islamic Finance Thesaurus among the communities of practice. It serves as a vital tool in the tremendous growth of interest in promoting standardization in Islamic Finance sector.

In the current Islamic Finance thesaurus, most of the information is in the form of annotation. Thus, we plan to transform the representation into an ontology in order to be used for reasoning purpose. We also plan to enhance the multilingual feature of the Islamic Finance thesaurus content to include other languages such as Malay, Mandarin, etc. Finally, to exploit the benefits of the huge Linked Data resources, we want to map the Islamic Finance thesaurus

 ISBN: 978-1-61208-329-2

concepts to other related knowledge bases within the same domain such as Conventional Banking.

ACKNOWLEDGEMENTS

The authors wish to thank the Institute of Islamic Banking and Finance and the Islamic finance experts Dr. Syed Musa Alhabshi, Dr. Hikmatullah Babu Sahib and Prof. Dr. Engku Rabiah Adawiah Engku Ali for their valuable support and advice.

REFERENCES

[1] N. Raphaeli, "Islamic Finance – A Fast-Growing Industry". retrieved March, 2014, from: http://www.aclu.org/files/fbimappingfoia/20111110/ACLURM002789.pdf.

[2] France to Promote Islamic Finance, retrieved March, 2014, from: http://www.expatica.com/fr/news/local_news/France-to-promote-Islamic-finance.html.

[3] D. Newman, "The Financial Industry Business Ontology (FIBO)", Ontology Summit (2013), retrieved March 2014, from: http://ontolog.cim3.net/file/work/OntologySummit2013/2013-05-02_03_OntologySummit2013_Symposium/Keynote-2_OntologySummit2013_Symposium_FIBO-Briefing--DavidNewman_20130502.pdf

[4] Linked Open Data, Retrieved March, 2014, from: http://linkeddata.org/.

[5] SKOS simple knowledge organization system reference. retrieved March, 2014, from: http://www.w3.org/TR/2008/WD-skos-reference-20080125/.

[6] W3C, retrieved March, 2014, from: http://www.w3.org/

[7] W. Schmitz-Esser. "Thesaurus and Beyond: An Advanced Formula for Linguistic Engineering and Information Retrieval", Knowledge Organization, 1999, vol. 26, no. 1, pp. 10-22.

[8] AGROVOC, retrieved March, 2014, from: http://aims.fao.org/standards/agrovoc

[9] SNOMED-CT, retrieved March, 2014, from: http://www.ihtsdo.org/snomed-ct/.

[10] J. Golbeck, G. Fragoso, F. Hartel, J. Hendler, J. Oberthaler, and B. Parsia. "The National Cancer Institute's Thesaurus and Ontology". Journal of Web Semantics (2003), vol. 1, pp. 75-80.

[11] AGROVOC Workbench, retrieved March, 2014, from: http://agrovoc.mimos.my/vocbench/.

[12] D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, and S. Katz, "Reengineering thesauri for new applications: The AGROVOC example" 2004, Journal of Digital Information 4(4).

[13] P. Hayes and B. McBride, "RDF Semantics" (2004), retrieved March, 2014, from: http://www.w3.org/TR/rdf-mt/.

[14] ISO 2788:1986 Documentation - Guidelines for the establishment and development of monolingual thesauri. Second edition. ISO TC 46/SC 9, 1986.

[15] S. Mohamed, S. Mustaffa, and D. Lukose, "Using the Spiral Process Model to Develop a Medical Knowledge Base", The 2nd Semantic Technology and Knowledge Engineering Conference (STAKE 2010), July 2010, pp. 101-114.

[16] Top Braid Composer, retrieved March, 2014, from : http://www.topquadrant.com/products/TB_composer.html.

[17] A. Mamadolimova, N. Ambiah and D. Lukose. "Modeling Islamic finance knowledge for contract compliance in Islamic banking", Proc. 15th international conference on Knowledge-based and intelligent information and engineering systems (KES 2011), Springer-Verlag, Volume Part III , pp. 346-355.

 ISBN: 978-1-61208-329-2

An Experiment in Managing Language Diversity Across Cultures

Amarsanaa Ganbold
School of Information Technology
National University of Mongolia
Ulaanbaatar, Mongolia
amarsanaag@disi.unitn.it

Feroz Farazi
DISI
University of Trento
Trento, Italy
farazi@disi.unitn.it

Fausto Giunchiglia
DISI
University of Trento
Trento, Italy
fausto@disi.unitn.it

***Abstract*—Developing ontologies from scratch appears to be very expensive in terms of cost and time required and often such efforts remain unfinished for decades. Ontology localization through translation seems to be a promising approach towards addressing this issue as it enables the greater reuse of the ontological (backbone) structure. However, during ontology localization, managing language diversity across cultures remains as a challenge that has to be taken into account and dealt with the right level of attention and expertise. In this paper, we report the result of our experiment, performed on approximately 1000 concepts taken from the space ontology originally developed in English, consisted in providing their translation into Mongolian.**

Keywords: Ontology localization, space ontology, space domain, ontology, Semantic Web

I. INTRODUCTION

Building a true, flourishing and successful Semantic Web [1] should involve the participation from all cultures and languages across the world. In the development of the traditional Web, this participation was spontaneous and has been made possible as the necessary tools and resources were available. With the Semantic Web one of the crucial lacks is the capacity to assign precise meaning to words that requires Natural Language Processing (NLP) tools that use Knowledge Base (KB). Still for many languages such resources are not developed at all and for some others what is out there cannot be used effectively as they could not achieve critical mass. However, for English much progress has been made and the WordNet (http://www.princeton.edu) developed at Princeton is one of the well-known and widely used resources in the field. Yet its coverage is often unsatisfactory while dealing with domain specific tasks [2].

Towards solving the issue of the lack of coverage and to gain a critical mass of concepts, some domain ontologies have already been developed. A prominent example is the *space ontology* [3] developed in English with comparatively very large coverage of geo-spatial features and entities around the globe. Domain ontologies can also deal with the specificity of an area of knowledge, for example, by providing relations and attributes specific to the domain. By reducing polysemy (the amount of words with same meaning), they can enable better semantic interoperability.

Ontologies that are developed to perform NLP tasks in one language can hardly be used with their full potential for another. Representing an existing ontology in a new language, taking into account cultural and linguistic diversities, is defined as ontology localization.

In this paper, we describe the development of the space ontology in Mongolian starting from its English counterpart from the Universal Knowledge Core (UKC). Building an ontology without human-level accuracy is a potential obstacle in developing applications (e.g., word sense disambiguation and document classification). Synset base resources (linguistic representation of ontologies) such as WordNet and FinnWordNet [4] are built manually to obtain better quality. Being concerned about the quality and giving utmost importance to it, we followed a manual approach. The contributions of our paper include:

i) The development of an ontology localization methodology that is domain and language independent and seems to achieve very high quality

ii) The development of a methodology for dealing with diversity (e.g., lexical gaps) across cultures and languages

iii) Lessons learned from the execution of the whole process in the generation of the space ontology in Mongolian

The paper is organized as follows. In Section 2, we provide detailed description of the UKC. Section 3 gives an overview of the space ontology. In Section 4, we describe the macro-steps of the translation process. In Section 5, we describe the diversity across English and Mongolian cultures in terms of space related features. Section 6 reports the experimental results, Section 7 discusses the lessons learned and Section 8 describes the related work. In Section 9, we provide the concluding remarks.

II. THE UNIVERSAL KNOWLEDGE CORE

The UKC[3] is a large-scale ontology, under development at the University of Trento which includes hundreds of thousands of concepts (e.g., lake, mountain chain) of the real world entities (e.g., Lake Garda, Alps). It consists of three main components: *domain core*, *concept core* and *natural language core* (See Fig. 1).

As described in [3], the domain core consists of various **domains,** where each of them represents an area of knowledge or field of study that we are interested in or that we are communicating about [5]. In other words, a domain can be a conventional subject of study (e.g., mathematics, physics), an application of pure disciplines (e.g., engineering, mining), the aggregation of such fields (e.g., physical science, social science) or a daily life topic (also called Internet domains, e.g., sport, music). Each domain is

 ISBN: 978-1-61208-329-2

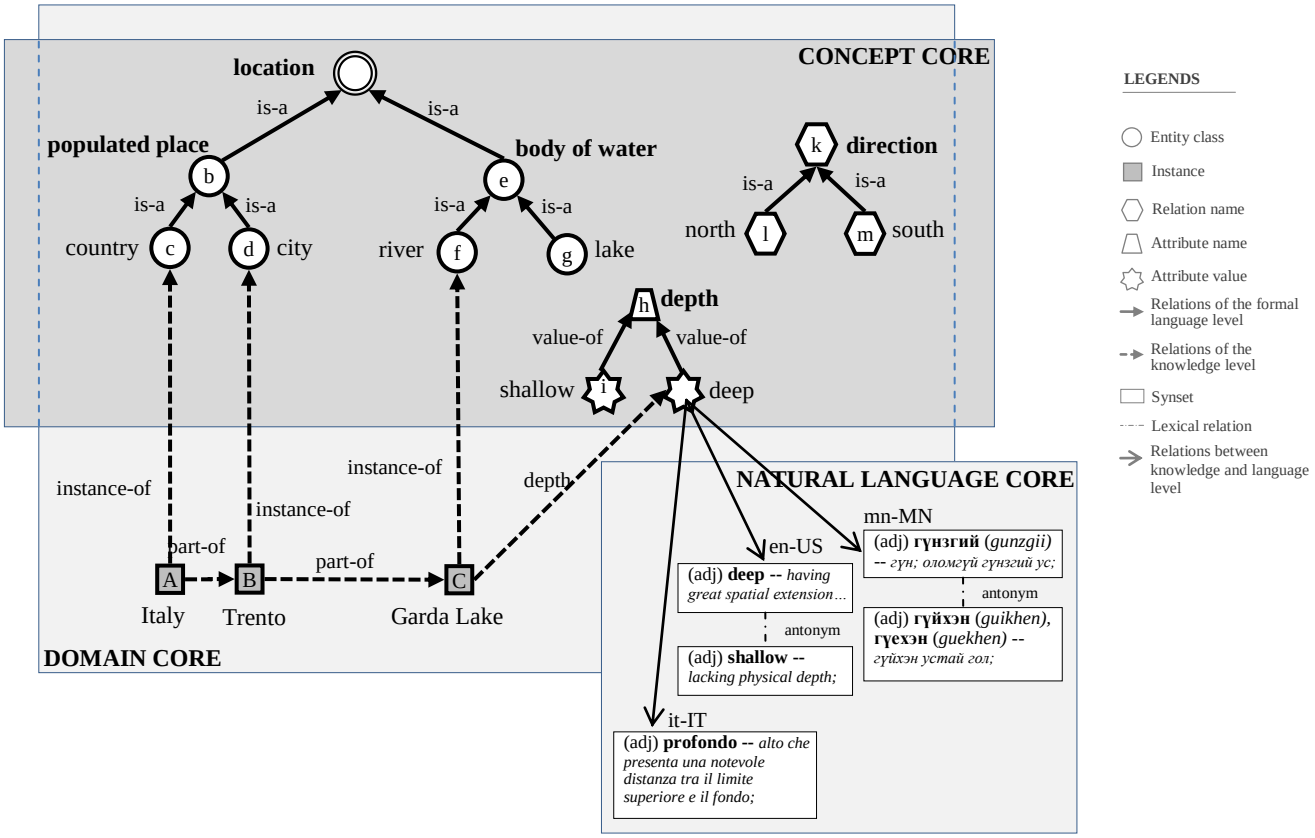


Figure 1. Knowledge Organization in the UKC

organized in **facets,** where a facet can be defined as a hierarchy of homogeneous concepts describing the different aspects of meaning [6]. According to our methodology [7], called DERA, where D stands for Domain, facets are classified into three categories: Entity class (E), Relation (R) and Attribute (A). For example, in the space ontology, country and continent are **entity classes**. **Relations** describe relations between entities; examples of spatial relations are near, above, far etc. An **attribute** is a property of an entity, e.g., depth of a lake.

The concept core consists of concepts and semantic relations between them. The concepts in the concept core form a directed acyclic graph, which provides the terms and the structure from which facets are defined. Entity class, relations and attributes are all codified as concepts. A **concept** is a language independent representation of a set of words (synset) which are synonym of a given word in natural language. For example, country, city, etc. The concept *city* can be represented as *city* in English, *città* (chit'a) in Italian, *xom* (khot) in Mongolian.

The natural language core is built with the complete integration of hierarchically organized synset bases, for instance WordNet and the Italian part of MultiWordNet (http://multiwordnet.fbk.eu). This component consists of words, senses, synsets and exceptional forms. A **word** is the basic lexical unit of the natural language core represented as a lemma. It can be multiword, phrase, collocation, etc. The words in the natural language core provide, for any given language, the translation of the concepts stored in the concept core.

Word senses are organized into four part-of-speeches -- noun, verb, adjective and adverb, one word may have more than one part-of-speech, and synonym word senses with the same part-of-speech are grouped into synset. A **sense** is the meaning of a word. A word can have one or more senses each having a part-of-speech tag. Each sense belongs to only one synset. All senses of a given word are ranked according to most preferred usage. A **synset** is a set of words which share the same meaning. In fact, words in a synset have semantically equivalent relations. Each synset might be accompanied by a gloss consisting of a definition and optionally example sentences.

III. THE SPACE DOMAIN

The space domain [3], [5] is a large-scale geospatial ontology built using the faceted approach. It was developed as the result of the complete integration of GeoNames (http://www.geonames.org) and WordNet. It is also known as space ontology and in this paper, we refer to it with any of these names. It currently consists of nearly 17 facets, around 980 concepts and 8.5 million entities. The ontology (excluding entities) is integrated into the UKC. Some

 ISBN: 978-1-61208-329-2

examples of facet are *land formation* (e.g., mountain, hill), *body of water* (e.g., sea, lake), *administration division* (e.g., state, province) and *facility* (e.g., university, industry).

In Fig. 2, we provide a partial bird's eye view of the whole set of facets. Note that facets are not connected to each other and they do not have concept overlap across or within them.

Fig. 3 shows a small portion of the facet *geological formation* in which the second level represents *natural elevation*, *natural depression* and the level below the *natural elevation* is organized into *oceanic* and *continental elevation*, and so forth.

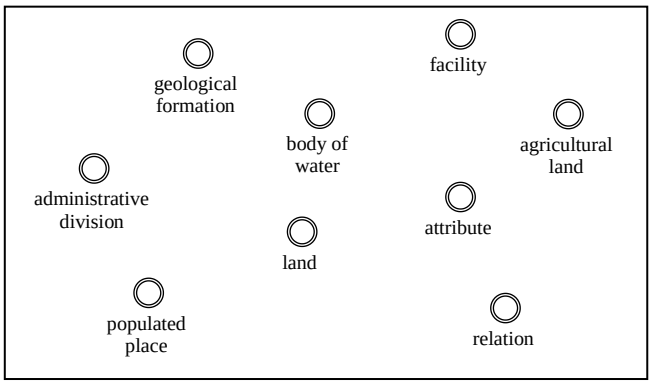


Figure 2. A subset of the facets of the *Space* domain

Note that within a facet with double circled node we distinguish the root concept from the rest of the concepts that are represented with single circle.

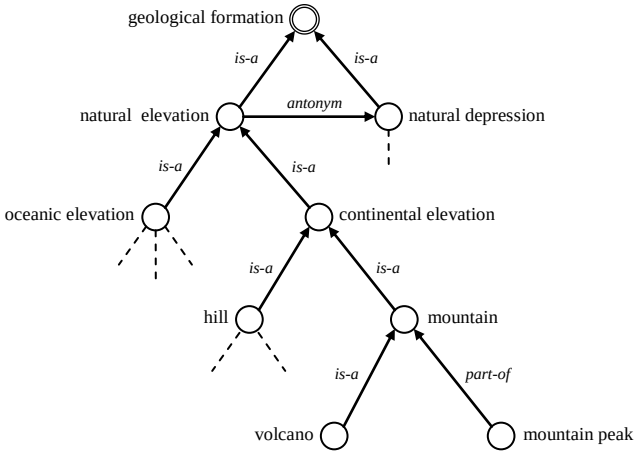


Figure 3. An entity class (E) category facet (partial view)

In the Space domain, the *relation* category contains around 10 facets such as *spatial relation* and *primary outflow*. A partial representation of the *spatial relation* facet is shown in Fig. 4.

The *spatial relation* is the spatial property between geological physical objects or the way in which something is located. Leaf nodes of this facet represent relations between entities. For instance, Mongolia is *south* of Russia and *north* of China. The relation *primary outflow* connects two bodies of water.

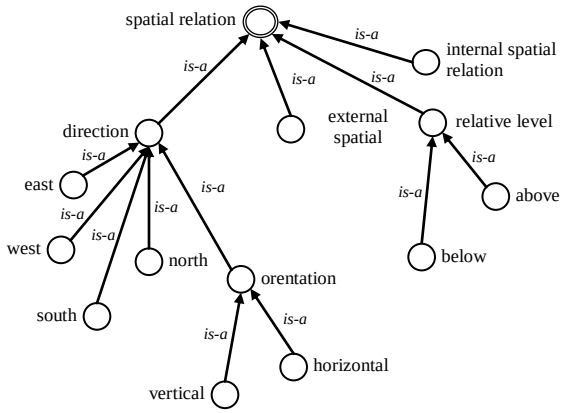


Figure 4. A relation (R) facet (partial view)

Within the domain the *attribute* category consists of around 20 facets such as *rain* and *temperature*.

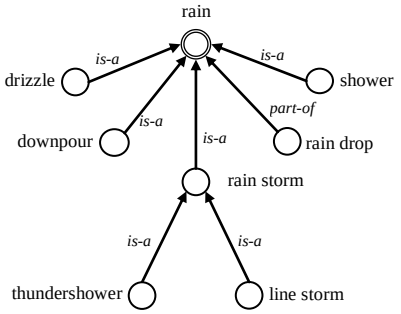


Figure 5. An attribute (A) facet (partial view)

rainstorm, *downpour*, *drizzle* and *shower*. With rain we mean *falling of water in drops from vapor condensed in the atmosphere*. The temperature indicates *the degree of hotness or coldness of an object or environment.*

IV. TRANSLATION APPROACH

The main idea of the translation process is to take the objects of the domain of interest from a source language, in this case English, and to produce the corresponding representation in a target language, e.g., Mongolian in order to update the UKC with translations. The process includes the translation of the synset words and glosses. A direct translation of them is provided whenever possible. However, the world is full of diversity and people of a particular culture might not be aware of some concepts. For instance, Mongolia is a landlocked country, thus some terms (e.g., *dry dock*, *quay*, *pier*, etc.) related to seaport are not known to the community or are rarely used.

In order to provide the most suitable translation for a synset, we follow the macro-steps described below and represented in Fig. 6.

1. A **language translator** takes a synset provided in the source language and gets a clear understanding of its meaning. In case of difficulty, he/she finds the corresponding images or videos of the synset word(s) on the Web to perceive the concept through visualization.

 ISBN: 978-1-61208-329-2

2. The **language translator** provides a suitable translation of the word(s) in the target language. With suitable we mean word, multiword, co-occurrence and phrasal representation as we do not allow a free combination of words as translation of a word. In case of unavailability of the word(s) for the given meaning, the translator can mark it as a lexical gap. However, the translator always provides the translation of the gloss.

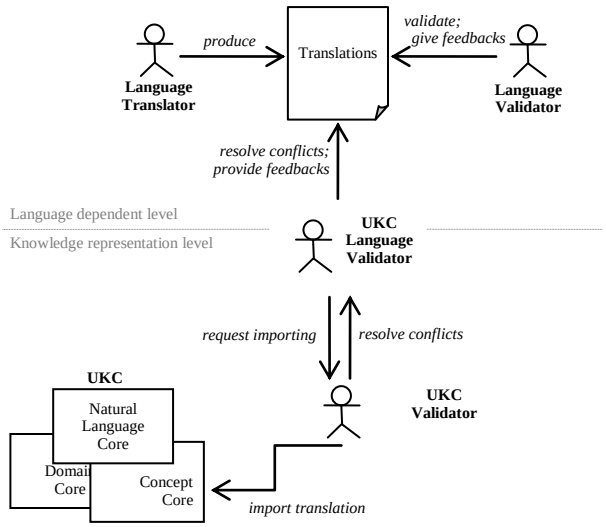


Figure 6. Translation phases of UKC

3. A **language validator** evaluates the translation of the word(s) and the gloss of the synset. In case the concept is marked as a gap, the validator either confirms the gap or suggests a translation for the word(s).
4. Upon receiving feedback on the synset, the **language translator** goes through the comments and updates the translation when necessary. In case of disagreement, the language translator provides comments including mostly the rationale about the disagreement.
5. The **language validator** reevaluates the updated translation. In case of disagreement, the validator generates further feedback and sends it back to the language translator (step 4). Even if after a few iterations a disagreement is not resolved, a second language validator is consulted. If agreed upon, the validation for the given synset is over.
6. A **UKC language validator** takes the validated translation to evaluate its correctness from both the language and UKC perspectives. The validator corrects the mistakes and resolve the issues (if any) communicating with the language validator (if necessary), possibly in a few iterations. Finally, he/she asks a UKC validator for importing the translation to the UKC.
7. The **UKC validator** runs an automatic validation tool to evaluate if the provided input is compliant with the UKC. In case of errors are found, they are corrected with the help of the UKC language validator (if needed) possibly iterating a few times. Once all the issues are resolved, the UKC validator imports the translation to the UKC.

Following these steps we translated the *space ontology* into Mongolian end-to-end, evaluated and finally imported the translations to the UKC. To achieve optimal quality while executing the whole process depicted in Fig. 6, we set the criteria that translators and various validators must possess the competences necessary for the task. The language translator should be a native speaker from the country of origin of the target language with a good command of the source language. The language validator should be a linguist possessing the necessary language competences. The UKC language validator is a native speaker of the target language with knowledge of the UKC. The UKC validator is an expert on the UKC with no specific competence on the language.

From a geographical point of view we expect that, in most cases, the language core will be developed in the countries where that language is spoken, while the UKC is and will be developed centrally. The UKC language validator, whenever possible, should operate centrally where the UKC validator is. This spatial distribution of operations and operators has been designed as an attempt to preserve local diversity and, at the same time, to deal with the need for central coordination required because of existence of a unique, single UKC. The underlying model is that there is a single world, represented by the UKC, and many different views of the world, each represented by a different natural language. The diversity of the world is therefore captured, as it will be described in detail in the next section, in the mapping from the informal natural languages and the unique UKC formal concept language.

V. TYPES OF DIVERSITY

The translation or localization is the adaptation of a piece of knowledge to a particular language and culture [8]. This is nontrivial and linguistic experts might help in this task. Moreover, the localization should be based on the perception of the concepts and entities in the real world within the local communities and not on the literal translation.

A. Concepts

We assume concepts to be universal. However, their representation in natural languages varies. Within the same language a concept might be referred with multiple terms (known as synonymy) and multiple concepts might be referred with the same term (known as polysemy).

The concepts *valley*, *dale* and *hollow* are represented with the same term in Mongolian.

> ***valley*** *– (a long depression in the surface of the land that usually contains a river)*
>
> ***dale*** *– (an open river valley (in a hilly area))*
>
> ***hollow*** *– (a small valley between mountains; "he built himself a cabin in a hollow high up in the Appalachians")*

Moreover, in the UKC dale and hollow are subordinate concepts of valley. In this case, translating them into the

 ISBN: 978-1-61208-329-2

target language increases polysemy. However, we translate them because within the Mongolian culture people can classify their (real world) entities under the specific concept.

Lexical gaps are those concepts that do not have a succinct representation in a given language. However, they can be expressed as a free combination of words [9]. For example, the concept *parish - (the local subdivision of a diocese committed to one pastor)* is a lexical gap in Mongolian. The variation in the concept lexicalization from the source language (S) to the target language (T) is depicted in Fig. 7(a).

As the lexical gap is a feature of the languages, it does happen with all of them. There can be a gap also from the target to source language. For instance, the Mongolian words **бууц** (buuts) and **буйр** (buir) are gaps in English. The word *buuts* can be represented in English as *an area of dried and accumulated manure where a nomadic family was living* and the word *buir* can be represented in English as *a round shaped spot where a nomadic yurt was built*. Note that these words lack a succinct representation in English. Therefore we consider them as gaps. This phenomenon is drawn in Fig. 7(b).

The nomadic lifestyle of Mongolians is the source of these concepts that are not used in the English speaking cultures across the globe.

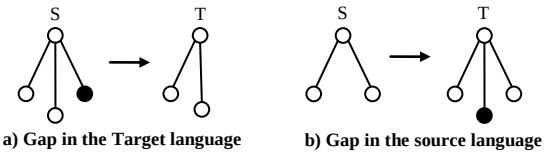


Figure 7. Variations of concept localization

Words pointing to lexical-gap concepts might appear also in the glosses. For instance, the term *piers* appearing in the gloss of *Romanesque architecture* is a lexical gap in Mongolian. In such cases, the translation is produced with a free combination of words.

> *Romanesque architecture – (...characterized by round arches and vaults and by the substitution of **piers** for columns and profuse ornament and arcades)*

B. Senses

In the space ontology, some words have multiple senses that have subtle difference in meaning. For instance, the word fissure has two senses:

> [S1]: *crack, cleft, crevice, **fissure**, scissure – (a long narrow opening)*
>
> [S2]: ***fissure** – (a crack associated with volcanism)*

The two concepts associated with the given word are hyponyms of *continental depression* and they can be represented with the same word(s) in the target language. This phenomenon is shown in Fig. 8(a).

Polysemous words in the source language might correspond to lexical gaps for a subset of senses. For instance, *gorge* has two senses within Space ontology and one of them is a gap as depicted in Fig. 8(b), where 'mn' and 'en' denote Mongolian and English, accordingly.

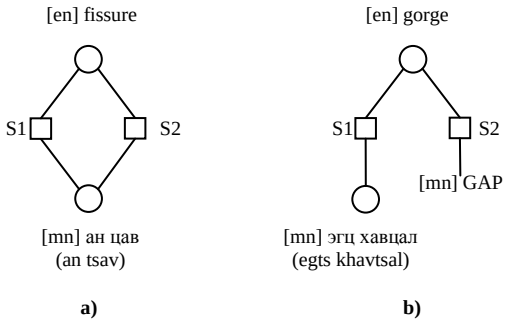


Figure 8. Word sense diversity

C. Synsets

Words in a synset can be directly translated into the target language. However, for some of them there might be a lack of translation. For example, the synset *mountain peak (the top point of a mountain or hill)* has 6 words of which 3 of them lack translation into Mongolian as shown below.

1 peak	→	оргил (ogril)
2 crown		
3 crest		
4 top	→	орой (oroi)
5 tip		
6 summit	→	дээд оргил (deed orgil)

In gloss paraphrasing, some parts of the glosses sometimes are obtained using words with a very close or similar meaning instead of exact translation. Though our first preference is to provide the exact translation, in many cases this could not be achieved. The following example shows a paraphrased translation where the phrase "near a shore" is eliminated from Mongolian version. In this situation, there is no difference between bank and shore in Mongolian language.

> [in English] ***oceanic sandbank** – a submerged bank of sand near a shore, can be exposed at low tide*
>
> [in Mongolian] ***далайн элсэн эрэг*** (gl. oceanic bank of sand) *– шунгаж орсон далайн элсэн эрэг, далайн давалгааны намхан хаялганд үзэгддэг (gl. a submerged sea bank of sand, visible at low tide)*

Example sentences in glosses were also paraphrased or added newly in order to provide a better explanation. For example, well-known place names are often substituted in the target language because famous names within a culture might give better understanding about a concept being translated. The highest mountain peak of the Alps ridge is Mont Blanc that is substituted with Everest as it is known to the most of the people in the East Asian region. Moreover, symbols are kept in their original forms, e.g., measurement unit symbol, pH.

Date and time format, measurement unit and currency were converted into the ones used regionally. For example, 5 inches is converted into 12.7 centimeters because of the pervasive use of MKS system in Mongolia. Note that these types of words appear only in glosses. However, using these types of word might not be suitable as fractions are less intuitive than whole numbers. For example, 3 feet is converted into 0.9144 meters. Such fractions cannot be

 ISBN: 978-1-61208-329-2

TABLE I. LOCALIZATION RESULT OF THE SPACE DOMAIN

Facets	Concepts	Translated	Disagreed words	Disagreed glosses	Translator Identified Gaps	Finally accepted Gaps	Finally Localized Concepts
administrative division	18	18	2	4	0	0	18
agricultural land	19	19	2	1	0	0	19
attribute	85	73	1	23	12	10	75
barren land	7	7	1	0	0	0	7
facility	357	357	54	64	0	2	355
forest	5	5	5	4	0	0	5
geological formation	200	150	73	87	50	52	148
land	15	15	2	3	0	2	13
plain	12	12	0	0	0	3	9
rangeland	8	8	1	4	0	0	8
region	46	44	6	0	2	2	44
relation	54	54	8	32	0	0	54
wetland	8	8	3	1	0	0	8
abandoned facility	16	15	4	1	1	1	15
body of water	116	106	24	17	10	3	113
populated place	13	10	2	1	3	2	11
seat of government	6	4	0	1	2	2	4
Total number of objects	**985**	**905**	**188**	**243**	**80**	**79**	**906**

mapped easily to the real world entities and most often become tedious to remember.

VI. RESULTS

In this Section, we report the results of our experiment. We could translate 91.88% of the concepts of the space ontology into Mongolian and the remaining 8.12% were identified as lexical gaps. In Table I, we report the detailed statistics of the translation task and the obtained results.

In Table I, the number of concepts per facet is shown separately, e.g., administrative division has 18 concepts, agricultural land has 19 concepts and so on. Note that for the sake of space, we group the statistics of all attribute facets as attribute and relational ones under relation.

Language Translators provided Mongolian translation for 905 concepts *Language Validators* provided feedback on each of the produced synset words and glosses separately that help us achieving better quality. The validation procedure identified 188 disagreed words and 243 disagreed glosses. Cases such as disagreements and modifications for improvement were solved in iterations (as many as needed) between the translators and validators until they reached to an agreement. The highest number of iterations was recorded as 4.

Language Validators' evaluation of the lexical gaps revealed that the translators proposed 10 false positives out of 80. We also identified that the translators produced 9 false positive translations of the concepts whereas they are gaps. In the end, we found that there are in total 79 gaps and 906 concept translations being accepted. The *UKC Language validator* and *UKC validator* reported a few (around 5) conflicts which were then solved with little effort. It is worth mentioning that *Language Translators* proposed to add 7 new concepts to the space ontology. This is only initial work and we expect that a few more concepts will be added with the evolution of the space ontology.

VII. LESSONS LEARNED

Assigning word sense rank appears as a difficult task to accomplish since the *Language Translators* contribute the results separately. In the translation work, they were aware of the fact that concepts translated by others might have the same word label. But it remained obscure until the whole translation task was finished. This ranking could be defined once all the concepts are translated. This is a non-trivial task to accomplish because deciding acceptable ranks might require local community agreement or the consultation of high quality linguistic resources that are often insufficient for domain specific tasks in many languages.

Synonymous words within the synsets were often increased after translations were evaluated by the Language Validators. This was the case since Language Translators concentrate in providing the target language correspondence representation of the knowledge objects taken from the source language within a reasonable amount of time. This often results in the postponement of the addition of synsets.

In the cases where an example sentence in a gloss contains a number that has to be converted according to some suitable measurement, we should freely change values and corresponding units since the numbers always give some extra information to provide glosses. For instance, 6000 meters can be changed to 6 km (while value remains same) and 3 kilograms to 3 pounds (while value modifies). Nevertheless, in case of sensitive information found in a gloss, we should exactly convert the number to relevant measurement unit in order to preserve the meaning of the gloss. For example, for understandable measuring unit of the target users 500 feet can be converted into 152.4 meters.

Parts of the glosses that follow the same syntactic pattern in the source language can be translated with little effort. For instance, the gloss part *a facility for [verb]+ing [object]* appeared in around one tenth of the concepts. We repeated the same translation for the part that matched completely. Moreover, we used the translation memory technique which

 ISBN: 978-1-61208-329-2

provides a translation with recurrent structure in the same way as previous translations.

In order to introduce foreign cultures to the community, we can translate lexical gaps as free combination of words. However, this should not always be the case. A first reason is computational: the explicit marking of the lexical gaps could support the KB-based applications in reducing computation time by avoiding the management of (multi)words which will be very rarely or never used. A second, more important reason, is related to the actual existence of a free combination of words capable of capturing, in the mind of a native speaker with no knowledge of the original concept (as it exists in the foreign culture) what the concept actually means, in the real world.

VIII. Related work

MultiWordNet [9] consists of several European language WordNets. It was developed under a model that reuses semantic relations from WordNet as follows: when there are two synsets and a relation holding between them, the same relation is assumed to hold between corresponding synsets in the new language. There is no literal translation in the case of developing Italian version of MultiWordNet of the synsets, words and exceptional forms but the contributors have produced the best possible Italian equivalents according to their skills and experiences in knowledge organization and linguistics. However, a limited number of glosses has been provided, e.g., around 2k in Italian over 33k.

The ontology localization activity described in [10] is an attempt to address the localization and diversity issues. They proposed guidelines and methodologies for enriching ontology with multilingual information. However, we differ from them with respect to the target language and the development approach.

Universal Multilingual Knowledge Base also known as UWN [11] was developed leveraging on the Wikipedia data and linking multilingual terms that are connected to the same page. However, automatically built KB resources often suffer from quality issues, e.g., around 10% of the terms in UWN are attached to the wrong senses, whereas we achieved human-level accuracy.

FinnWordNet [4] was produced from WordNet with the help of professional translators and the output is monitored by bulk validation. While producing the whole WordNet in Finish in 100 days, they traded off the quality for reducing the amount of translation time. Diversity in the languages such as lexical gaps is overlooked in this task.

IX. Conclusion

In this paper, we proposed an approach for generating ontologies through translation from one language into another. This approach was developed to be applied independently of domain and language and to deal with the diversity across the languages. While translating the ontologies, we manage diversity with the identification of diversity features and their presence in a given target language by working together with the linguistic experts and/or native speakers living in the country where it is spoken. We evaluated the effectiveness of the methodology by performing a case study for translating the space ontology into Mongolian. Thanks to the reuse of the ontological backbone structure, we achieved a space ontology in Mongolian that is as high quality as the original one in English. Though manual approach is usually known to be time consuming, adopting this methodology in a crowdsourcing setting can help increase throughput and make this suitable for dealing with large ontologies. Our future plan includes the exploitation of this valuable resource to improve the accuracy of NLP tasks (see [12]) and Concept Search (see [13]) in space domain.

Acknowledgment

The research leading to these results has received funding (partially) from the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement n. 600854 Smart Society: hybrid and diversity-aware collective adaptive systems: where people meet machines to build smarter societies http://www.smart-society-project.eu/. We are thankful to Vincenzo Maltese for his valuable feedback.

References

[1] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web,” Scientific American, vol. 284, no. 5, 2001, pp. 34–43.

[2] F. Giunchiglia, V. Maltese, F. Farazi, and D. Biswanath, “GeoWordNet: a resource for geo-spatial applications,” in ESWC’10, Volume Part I, 2010, no. December 2009, pp. 121–136.

[3] F. Giunchiglia, V. Maltese, and D. Biswanath, “Domains and context: first steps towards managing diversity in knowledge,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 12–13, 2012, pp. 53–63.

[4] K. Lindén and L. Carlson, “FinnWordNet – Finnish WordNet by Translation,” LexicoNordica – Nordic Journal of Lexicography, vol. 17, 2010, pp. 119–140.

[5] F. Giunchiglia, B. Dutta, V. Maltese, and F. Farazi, “A facet-based methodology for the construction of a large-scale geospatial ontology,” Journal on Data Semantics, vol. 1, no. 1, 2012, pp. 57–73.

[6] F. Giunchiglia, B. Dutta, and V. Maltese, “Faceted Lightweight Ontologies,” in in Conceptual Modeling Foundations and Applications, vol. 5600, 2009, pp. 36–51.

[7] F. Giunchiglia, B. Dutta, and V. Maltese, “From Knowledge Organization to Knowledge Representation,” in ISKO UK Conference, 2013, no. June.

[8] M. C. Suárez-Figueroa and A. Gómez-Pérez, “First Attempt towards a Standard Glossary of Ontology Engineering Terminology,” in TKE08, 2008, pp. 1–15.

[9] L. Bentivogli and E. Pianta, “Looking for lexical gaps,” in EURALEX International Congress, 2000, pp. 663-669.

[10] M. Espinoza, E. Montiel-Ponsoda, and A. Gómez-Pérez, “Ontology localization,” in K-CAP ’09, 2009, pp. 33-40.

[11] G. De Melo and G. Weikum, “Towards Universal Multilingual Knowledge Bases,” in Principles, construction, and applications of multilingual wordnets : proceedings of the Fifth Global WordNet Conference, 2010, pp. 149–156.

[12] I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, and X. Huang, “From Web Directories to Ontologies : Natural Language Processing Challenges,” in ISWC’07/ASWC'07, 2007, no. 60673038, pp. 623–636.

[13] F. Giunchiglia, U. Kharkevich, and I. Zaihrayeu, “Concept search,” The Semantic Web Research and Applications, vol. 5554/2009, 2009, pp. 429–444.

 ISBN: 978-1-61208-329-2

Investigating Factors for E-Knowledge Sharing amongst Academic Staff

Hanan Alotaibi
Electrical and Computer Science
University of Southampton
Southampton, UK
hmqa1g09@ecs.soton.ac.uk

Richard Crowder
Electrical and Computer Science
University of Southampton,
Southampton, UK
rmc@ecs.soton.ac.uk

Gary Wills
Electrical and Computer Science
University of Southampton,
Southampton, UK
gbw@ecs.soton.ac.uk

***Abstract*-Knowledge sharing has been considered a significant component of success in Knowledge Management (KM). In the most organizations KM is often inadequate when it comes to knowledge sharing, especially between staff who work in universities. In order to encourage knowledge sharing, it is important to know why/where/when employees choose to contribute or to receive shared knowledge. The purpose of this research is to investigate the factors that affect academics' behavior towards knowledge sharing by using Web technology. A synthesis of factors which already exist in current theory, i.e., the Unified Theory of Acceptance and Use of Technology, as well as other factors which are always explored independently in research studies, are combined to a Knowledge Sharing Technology Model has constructed. The model identifies the key factors that affect the uptake of knowledge technologies in universities.**

Keywords-Knowledge management; Knowledge sharing Technology, Web technology; the Unified Theory of Acceptance and Use of Technology (UTAUT).

I. INTRODUCTION

Over the last few years, the majority of the largest global corporations have Knowledge Management projects to support their development and growth [1]. It is widely recognized that organizations benefit by establishing appropriate knowledge management system to increase their efficiency. The main activities in Knowledge Management [2] are acquiring, sharing, and storing the knowledge. It is recognized that the most crucial activity of all is knowledge sharing since most knowledge is held as tacit knowledge by people [3]. However, Knowledge Management is often inadequate when it comes to knowledge sharing, especially between staffs [4]. Thus, novice staffs are unable to capture valuable information while there is no knowledge sharing mechanism between staff, this can affect staff performance, when tacit knowledge from experts is often lost as the knowledge has not been made explicit codified. So this may result in a poorer employees experience and lower staff achievement.

In the last few decades, the use of technology in supporting Knowledge Management process has been widely recognized, which are sharing and reusing of knowledge and technology represents a highly visible solution while information technology provides direct assistance in the processes of Knowledge Management [5]. Web technology is the most effective technology used in Knowledge Management [6]. Web technology is based on a particular set of technologies enabling users to interact and collaborate with each other in social media: it can be termed the 'Social Web', as it incorporates a strong social component [7]. Sharing knowledge via web technology can be very effective among staffs, who work in universities, such as Wiki or Blogs.

This work considers that application of Knowledge Management to Universities in the Saudi Arabia. The Universities are lacking in management technology system for the academic process. Consequently, tacit knowledge of expert academics is lost, as the knowledge has not been documented. Thus, the novice academics are unable to use useful information, as no knowledge has been shared among academic staff. There are insufficient studies regarding the academics perspective on knowledge sharing technology [8] and their use in Saudi Arabia universities. The majority existing studies are conducted in international commercial organizations. The aim of this research is to investigate factors that influence academics' behavior toward knowledge sharing via web technology.

This paper is structured as follows. Section II provides additional background to the work. Section III describes a conceptual mode that is being used to understand knowledge sharing in Universities in Saudi Arabia. Section IV discusses how this model is to be validated, the paper concludes with a discussion in Section V.

II. BACKGROUND

A. Knowledge Sharing

Knowledge sharing is a mutual relationship between sender, who provides knowledge, and receivers, who are seeking knowledge, exchange of information gained from

 ISBN: 978-1-61208-329-2

experiences, is used to support an individual who is working towards a common goal [9]. Sharing and distributing knowledge is positively linked to Knowledge Management [10] found that knowledge sharing is based on individual behavior, as people do not accept the value of sharing knowledge unless they think it is important. Thus, changing people's behavior is the challenge in Knowledge Management [11] and knowledge sharing behavior is the central process of knowledge management.

Knowledge sharing behavioral is typically affected by certain factors either positively or negatively, hence this research focuses on knowledge sharing technology behavioral factors.

Knowledge sharing behavior is viewed as the degree to which academics actually share their knowledge with their colleagues via Web technology. In practice, knowledge sharing can be considered from two aspects: behavioral and technological.

B. *The Unified Theory of Acceptance and Use of Technology (UTAUT)*

The Unified Theory of Acceptance and Use of Technology (UTAUT) model was defined by Venkatesh et al. [9] and extended the Technology Acceptance Model (TAM), which is the most widely applied model of user acceptance and usage [12]. UTAUT provides a useful tool for managers needing to assess the likelihood of success for new technology [9].

The UTAUT model examined the determinants of user acceptance and usage behavior and found that all contribute to the usage behavior [13]. This research takes advantages of UTAUT to examine staff behavior toward knowledge sharing technology by considering in some factors that will present next section.

III. CONCEPTUAL MODEL

We have synthesized factors affecting knowledge sharing; some of these factors already exist for example UTAUT [14], while other factors, such trust, time, leadership and IT support, that are always explored [15, 16, 17 ,18] Overall, based on researchers reviews and the Unified Theory of Acceptance and Use of Technology (UTAUT), Knowledge Sharing Technology (KST) Model was developed, Figure 1.

A. *Motivation*

The biggest issue facing staff is the difficulties in sharing knowledge because of lack of time for preparing the subject to be presented and the fact that knowledge sharing activities require a high level of effort [19]. In this case, the staff should be encouraged to engage in knowledge sharing activities, and it is also unrealistic to assume that all staff contributes their knowledge. In fact, human beings will offer their knowledge when they expect reward, such as extra payment [20]. Also, in the case of the benefit expected from knowledge sharing, the benefit is not only more payment but also reciprocal benefit.

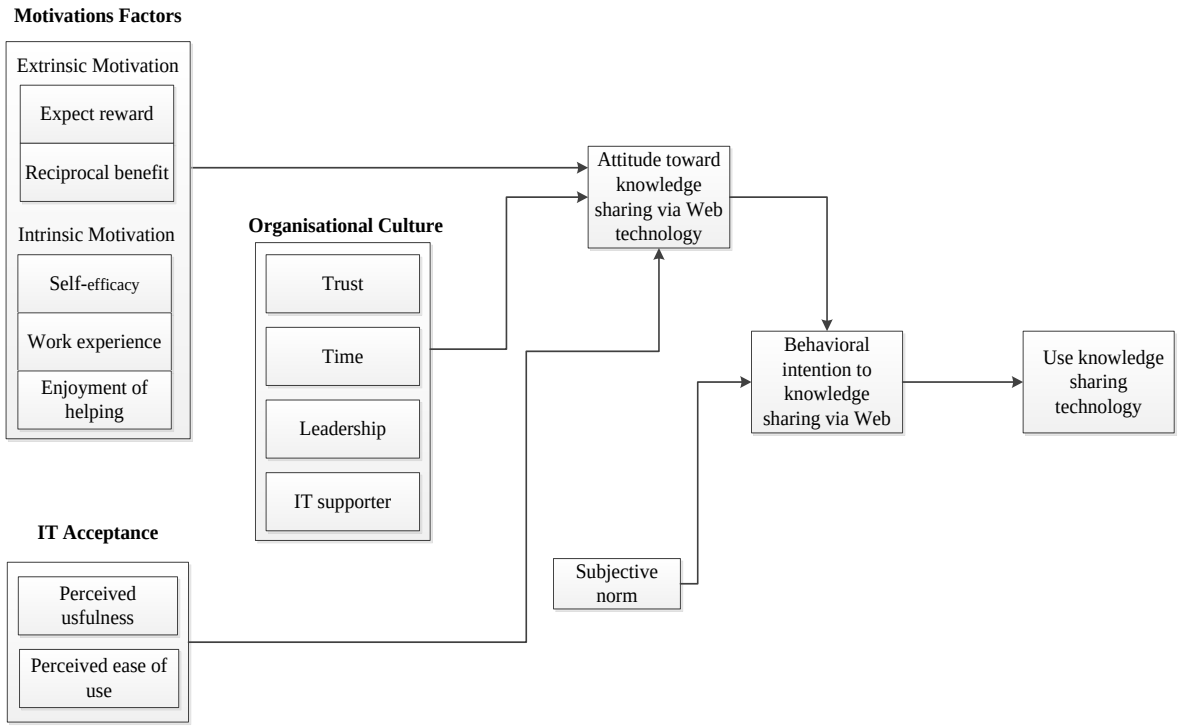


Figure 1 The Knowledge Sharing Technology Model (KST) of the factors that lead to successful adoption.

 ISBN: 978-1-61208-329-2

There are some members of academic staff who contribute their knowledge because of their belief in their self-abilities and competences, and also the belief that their knowledge can help to work improvements; this is a self-efficacy factor.

Furthermore, work experience drives staff towards knowledge sharing in order to obtain satisfaction and pleasure [20]. Some individuals enjoy helping others, especially if they are working in group [21].

According to the UTAUT model, it can be assumed that the employees' attitudes towards knowledge sharing are dependent on intrinsic and extrinsic motivation. Intrinsic motivation is defined as perception that staff will share knowledge because he expect to obtain valuable outcomes, it includes expected reward and reciprocal benefit, whereas, extrinsic motivation defined as staff will share knowledge because the member of staff believes that they have valuable information that should be shared includes self-efficiency, work experience and enjoying helping others.

B. Technology acceptance

In this work we examine two factors that influence individual attitude, perceived usefulness and perceived ease of use, which are the root constructs of UTAUT. Perceived usefulness is defined as the degree to which staff believes that using particular system would be enhanced to share knowledge. Perceived ease is defined as the degree to which staff believes that using particular system would be free of effort to share knowledge of use. They have shown evidence that these factors are strongly correlated with attitude towards the acceptance of information technology. When staff feel the technologies can be used in an easy way, it is more probable that they will present their knowledge. So their attitude to ease of use will affect an individual's knowledge sharing behavior. Also, staffs are more likely to share their knowledge when they feel that they have worthwhile information that is very useful for other staff.

C. Organizational culture

Organizational culture has a relationship with staff's communication and knowledge sharing behavior, as previous research has pointed out [22]. Most of studies [22] have suggested that organizations should create opportunities for employees' interaction and facilitates of knowledge sharing.

Trust has an indirect influence on knowledge sharing, which leads to increased sharing through technology. Trust is defined in this model as the degree of staff 's believe that other members are honest and have valuable and useful knowledge to share. Furthermore, other researchers have examined affect-based trust and cognition-based trust [23], and found that trust has an effect on sharing knowledge when staffs believe other team members are honest.

A further factor is time, Ford and Chan [10] and Ford and Staples [11] examined the influences of time on knowledge share, and found that the most staff unwilling to share their knowledge because of lack of time. Moreover, in [24] claims that time is one of the barriers to knowledge sharing in organizations, as adding information to the system is time consuming. However, the authors' opinion is that knowledge sharing is definitely non-consuming time, once the information is available in the system. Thus, staff can reach the valuable information that has been previously placed in the system more quickly, rather than searching in the other huge sources. Time is defined as the staff believes that sharing knowledge is non- consuming time while information is available on the system.

Furthermore, leadership in a team setting has relationship with knowledge sharing. According to Bain et al. [14], a team's expertise is more highly developed when there is a leader controlling the team in regard to knowledge sharing and moreover, providing a good quality of new ideas and encouraging staff to share their knowledge. So, leadership has an influence on employee's attitude toward knowledge sharing by using technology. We believe that the leadership, which is defined as to encourage employees to share knowledge, has a significant influence in the Saudi organizations' situation.

Knowledge Engineers provide direct assistance in the processes and circumstances to create knowledge [15]. The success of knowledge management is commonly based on implementation of new IT-based systems. Staff codifying and sharing knowledge by a system are required to be familiar with using the system or there is assistance for users who are unfamiliar with IT. In addition, among the fast growing technologies, the changing tools of the system there is continual improvement, so users should be kept up-to-date with new changes.

D. Subjective norm

According to the UTAUT model, the subjective norm is identified as the degree to which a staff member perceives whether social pressure will affect the performance of knowledge sharing technology.

IV. RESEARCH METHODS

This initial study will use both qualitative and quantitative research methods in two phases. In the initial phase we will use in-depth interviews, while the second phase will be conducted by using an online survey.

The interview includes mixed-methods of questions are divided into three categories; knowledge sharing; important of using Web technology and Knowledge sharing via Web. Interviews will be conducted with ten to fifteen expert and novice staffs who work in Saudi Arabian Universities. The interviews will be conducted across a number of disciplines. The purpose of the interviews was to investigate staff'

 ISBN: 978-1-61208-329-2

opinion about knowledge sharing via Web technology and explore other unidentified factors and investigate that requirement toward knowledge sharing among staff and discover the abilities and acceptance of staffs in using Web technology in knowledge sharing purpose.

The survey will conduct after analysing the interviewees' answers. The target subjects will be staff who are practicing at universities in Saudi Arabia. In the survey, we will measure the items that are include the KST model Each item will be measured on a five-point Likert scale [25], ranging from strongly disagree (1) to strongly agree (5). The result of the survey will help to refine the KST model and then confirm it.

V. CONCLUSION

The research objective is to improve knowledge management in Saudi Arabia universities and to facilitate exchanging knowledge between staffs. Therefore, this study explored the effective factors of knowledge sharing using web technology by examining staff's behavior toward knowledge sharing. Using theoretical frameworks UTAUT to underpin the model and using existing research into knowledge management, the model has been constructed by combining, synthesizing and refactoring the factors, which have the most important effects in knowledge sharing using web technologies.

From the authors point of view, the knowledge management systems should be established in the Saudi universities based on all factors that will explore from empirical study, in such a way that they function in a more efficient manner. Exploring the factors will asset knowledge worker to build website for knowledge sharing purpose in Saudi Arabia universities.

ACKNOWLEDGMENTS

The authors thank the Saudi Higher Education for funding Hanan Alotaibi to undertake this research at the University of Southampton.

REFERENCES

[1] J. Brown and P. Duguid, "Knowledge and Organisation: A Social-practice Perspective." Organisation Science, 2001, 198-213.

[2] N. Milton, N. Shadbolt, and H. Cottam , "Hammersley Towards Knowledge Technology for Knowledge Management" Human-Computer Studies, 1999.

[3] I. Nonaka and H. Takeuchi. The knowledge-creating company: How Japanese companies create the dynamics of innovation. USA, Oxford University Press, 1995, pp. 96-104.

[4] H. Lin. "Effect of extrinsic and intrinsic motivation on employee knowledge sharing intentions" Journal of Information Science, 2007, vol. 33 (2), 135-149.

[5] A. Silver, "Where technology and knowledge meet" Journal of business strategy, 2000, pp. 21(6), 28-133.

[6] C. Wagner, "Breaking the knowledge acquisition bottleneck through conversational knowledge management" Information Resources Management Journal, 2006, pp. 23 (3) 70-83.

[7] U. Bojars, G. Breslin, A. Finn, and S. Decker "Using the Semantic Web for linking and reusing data across Web 2.0 communities Web Semantics" Science, Services and Agents on the World Wide Web. 2008.

[8] R.Fullwood, J. Rowley, and R. Delbridge. "Knowledge sharing amongst academics in UK universities" Journal of Knowledge Management, 2013, vol. 17, pp. 123 – 136.

[9] M. Eisenhardt and M. Santos. Knowledge-based view: A new theory2 of strategy? In A. Pettigrew et al. (Eds.). Handbook of Strategy and Management , 2002.

[10] S. Allameh, A. Abedini, J. Pool, and A. Kazemi, " An analysis of factors affecting staffs knowledge sharing in the central library of the University of Isfahan using the extension of Theory of Reasoned Action" International Journal of Human Resource Studies, 2012, vol. 2, p. 1.

[11] R. Ruggles, "The state of notion: knowledge management in practice". California Management Review, 1998, pp. 80-89.

[12] F. Davis, R. Bagozzi, and P. Warshaw,. "User Acceptance of Computer Technology: A Comparison of Two Theoretical Models." Management Science, 1989, vol. (35:8), pp. 982-1003.

[13] K. Bandyopadhyay and K. Fraccastoro "The Effect of Culture on User Acceptance ofInformation Technology" Communications of the Association for Information Systems, 2007, vol 19, Article 23.

[14] V. Venkatesh, M. Morris, G. Davis, and F. Davis F. "User Acceptance of Information Technology toward a Unified View" MIS Quarterly, 2003, vol. 27 (3), pp. 425-478.

[15] D. Ford and Y. Chan. "Knowledge sharing in a multi-cultural setting: A case study." Knowledge Management Research & Practice, 2003, vol.4, pp. 3-16.

[16] D. Ford and D. Staples. "Perceived Value of Knowledge: The potential informer's perception" Knowledge Management Research & Practice, 2006, vol. 4, pp. 3-16.

[17] G. Bain, L. Mann, L. Atkins, and J. Dunning. "R&D Project Leaders: Roles and Responsibilities" Leadership, Management, and Innovation in R&D Project Teams, 2005, pp. 49–70.

[18] B. Bergeron. Essentials of Knowledge Management New Jersey USA, 2003.

[19] K. Husted, S. Michailova, and D. Minbaeva. Knowledge sharing and organizational performances: the role of extrinsic and intrinsic motives. Cairns, Australia. 2005.

[20] H. Lin, "Effect of extrinsic and intrinsic motivation on employee knowledge sharing intentions" Journal of Information Science, 2007, vol. 33 (2), pp. 135-149.

[21] I. Emmerik and I. Jawahar, "Lending a helping hand, provision of helping behavior beyond professional career responsibilities" Career Development International, 2005, vol. 5, p. 10.

[22] I. Nonaka and H. Takeuchi. The knowledge-creating company: How Japanese companies create the dynamics of innovation. USA, Oxford University Press. 1995.

[23] S. Chowdhury, "The role of affect and cognition-based trust in complex knowledge sharing" Journal of Managerial Issues, 2005, vol. 17(3), p. 310−326.

[24] T. Haldin-Herrgard, "Difficulties in diffusion of tacit knowledge in organizations" Journal of Intellectual Capital, 2000, vol. 1, pp. 357-365.

[25] L.Rensis, "A Technique for the Measurement of Attitudes" Archives of Psychology, 1932, 140: 1–55.

 ISBN: 978-1-61208-329-2

A Novel KM Framework for Fostering Creativity and Stimulating Innovation

Amirhossein Roshanzamir, Ahmad Agha Kardan
Department of Computer Engineering and IT
Amirkabir University of Technology
Tehran, Iran
Emails: {amrhssn, aakardan@}aut.ac.ir

***Abstract*—Knowledge Management (KM) is a dynamic system to identify important information, collect it from those who possess it, store it, and finally, share it with those required including employees, customers and other stakeholders. Today, a good KM system consolidates a company's internal expertise with external information by generating and collecting as much as useful information in order to improve processes, customer relations, decision making, employee morale, performance and most importantly revenue and profit. This paper explores the framework of KM by identifying two core phases, i.e., knowledge creation and knowledge usage and then maps each respectively to creativity and innovation. Invention and innovation are extremely dependent on the availability and richness of knowledge; however, the major challenge in many organizations is that KM is focused on maintaining continuity and consistency of capturing that knowledge and publishing it appropriately in order to be used by those who need it. This structure alone treats ineffectively with the creativity and invention process and leave not much spaces for innovation to be stimulated. Meanwhile, there is no solid measurement tool to assess performance and productivity of KM systems. In order to address these challenges, this paper integrates the four traditional steps of KM including Capturing, Storing, Disseminating and Implementing by Planning, Leading and Adapting (PLA) aspect to form a two dimensional model for directing and leading the whole KM process. The new model can potentially foster creativity in the first two steps of capturing and storing and then stimulate innovation in the second two steps of disseminating and implementing. Finally, the paper studies Apple Inc., as one of the most innovative companies in order to illustrate the possible application of this novel model.**

Keywords-knowledge management; creativity; innovation; evaluation; assessment; value creation; Apple, Inc.

I. INTRODUCTION

Knowledge Management (KM) is simply defined as a dynamic system to identify important information, collect, store, and finally share it with those required this information including employees, customers and other stakeholders [1]. KM is an emerging concept and has been around for more than 20 years in terms of growth as a discipline. Meanwhile, in today competitive world and global economy, where its characteristics is described by rapidly evolving technology, shorter product lifecycles and higher rate of new product development, organizations need to foster creativity and innovate their products, services and policies. This approach will enable them to prosper and keep up with highly dynamic environment.

Some anecdotal evidence suggests that KM is more widely accepted within certain industries like the pharmaceutical, energy, aerospace, and manufacturing. These knowledge intensive industries are the leaders in KM organizational adoption as well as creativity and innovation by leveraging new knowledge throughout their organizations, customers and stakeholders. Much is said about the role of KM in supporting innovation within organizations and this is also closely tied in with enhancing the activities of 'knowledge workers' in dynamic organizations such as consulting firms. It is worth remembering that business organizations are neither built for KM, nor for innovation; they are built for profit making and increasing stakeholder's value. This can be achieved by meeting and exceeding what customers perceive as value for price. It is widely recognized that KM can drive and support innovation within organizations, through a wide variety of approaches and techniques which can be embedded within KM frameworks. Bates and Khasawneh [2] suggested that innovation is equated with the adoption and application of new knowledge and practices, including the ability of an organization to adopt or create new ideas and implement these ideas in developing new and improved products, services, and work processes and procedures. Innovation, then, is considered an intangible resource that is very difficult to imitate. However, the main goal of KM in many organizations seems to be focused on improving the management of information and knowledge within and across enterprises [3]. KM in most organization is centered on maintaining continuity and consistency of capturing the knowledge, and publishing it appropriately to be easily and quickly used by those who require it including staff, customers and other stakeholders. This structure lacks measurement and assessment sprite in order to evaluate and improve KM efforts. It can further ineffectively be aligned with organization's goals and fruitlessly may treat with the creativity and invention process. It also leaves not much spaces for stimulating innovation which can be defined as generating drastic change in what customer perceives as value for price.

The general belief is that everything that gets measured can be evaluated, adjusted and then controlled and improved. It is, therefore, the intention of this paper to formulate a novel framework in KM equipped with the basic assessment tools. The model will give readers a new perspective in fostering creativity and stimulating innovation within their organizations in order to create value and deliver benefits. We first review invention and

 ISBN: 978-1-61208-329-2

innovation concept and study major divers of innovation in Section III. Section IV describes invention, innovation and technology. Section V determines incremental and radical innovation. Section VI reviews KM current models. Section VII introduces our two dimensional KM model. Section VIII focuses on idea funnel and integrates it with the new model. Section IX studies Apple Inc. in order to illustrate the possible application of the model. In section X, we emphasize on KM leader role and elaborate on knowledge hierarchy and we sketch out the conclusion and future works in the final section.

II. INVENTION VS. INNOVATION

People normally equate innovation with creativity or invention. However, innovation is different from invention and creativity. Sloane [4] has defined these terms as follows: Creativity is the capability or act of conceiving something original or unusual. Invention is the creation of something that has never been made before and is recognized as the product of some unique insight, while Innovation is the implementation of something new.

Invention is the act of generating a device, process or discovery that is new and useful which reflects extraordinary creativity and can even make a distinct contribution to science advancement. Invention is somehow individual activity focused on internal process of an organization which can potentially result in innovation, if it is properly leaded, managed and finally commercialized. In such a sense innovation is the external manifestation of invention which has a tangible impact. It is about executing and commercializing in order to meet or exceed the customer's needs. Accordingly, creativity and innovation process is divided into five blocks by many scholars, as shown in Fig. 1).

The process begins with idea generation and opportunity recognition which happens when an insight about something new is developed. Once the idea is considered to be of value to either customers or shareholders in form of cost advantage or solution to a problem, then, it must be evaluated by decision makers to address below questions :

- What kind of value does the idea creates ?
- Is there a market for this new value ?
- Is market large enough to justify development ?
- How does the idea fit within organization's strategy ?

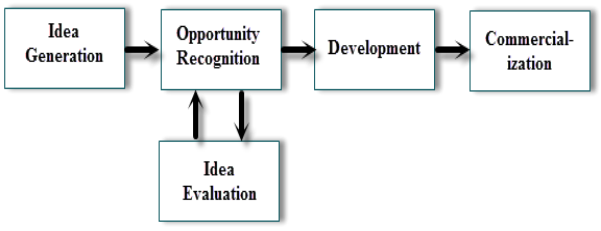


Figure 1. Innovation process adopted from [5].

Ideas that produce affirmative answers to these questions can be considered for development and commercialization stages. Therefore, we can summarize the model to below equation according to Govindarajan [6]:

Innovation = Ideas + Execution
or
Innovation = Creativity + Commercialization

III. MAJOR DRIVER OF INNOVATION

The successful innovations do not come into being by benchmarking and copying. They come into being by overcoming contradictions, limitations, paradigms, by taking another view at the problem and differentiating. . Differentiation can be achieved by a new design, other functions or another view at the customer need [7]. In fact, there are many drivers of innovation which include but not limited to profit and revenue growth, productivity, cost efficiencies, business or organizational model, partnership, route to market or marketing method, employee satisfaction and most importantly new products and services development.

We tend to think of an innovation as a new product; but, we can innovate with a new process, method, business model, partnership, route to market or marketing method too. In fact, every aspect of a business operation is a candidate for innovation. Some of the most powerful innovations we can make, are in business methods and customer services. If we look at companies like Dell, eBay and Amazon, we see that their great innovations were with their business models rather than in new products.

In a broader view, there is a difference between the organization's view of product innovativeness and the customers' view of the same. Firms express product innovativeness by comparing technology product content to competitor offerings and by assessing the degree of technical and marketing resources needed, whereas the customers based their evaluation of innovativeness on their need to alter metal models and behavioral habits. The key drivers and assessment tools like employee satisfaction, productivity and cost efficiency can help organizations to improve the quality, price, image and availability of products and services in order to better serve the customers. Therefore, we can say that the customer is the major driver of innovation and must be at the center of focus in all innovation efforts. Specifically, the major criteria in assessing new ideas should be in involving the customer's real requirements in entire process in order to record his or her feedback in a scorecard. This measurement tool would increase retention rate, increase new customers; reduce complaints and cost of it, reduce response time, increase revenue per customer (new or existing), increase sale volume and increase customer satisfaction [8].

We can think of Apple's iPhone, introduced in June 29, 2007 as of one of the most dynamic example of innovation which, not only changed the mobile industry, but revolutionized people's lives and the way business is done. Here, one can observe the role of putting customers and users at the center of attention and the driver of innovation. The iPhone certainly was not the first smartphone, nor was it the first phone to offer users access to their email and the

 ISBN: 978-1-61208-329-2

Internet. But, it introduced the touch-based user interface, which, like the mouse, changed the way people interact with their devices. This alone together with other user-friendly features made communication more simple, fun, intuitive and interactive for users.

IV. INVENTION, INNOVATION AND TECHNOLOGY

Innovation and technology are not the same, although innovation can be the result of new technology; however, in some cases innovation is based on smart redeployment or combination of existing technologies. Other important issue is that not all invention end in commercial application or useful product, since, simply there might not be market for it or even timing is not correct. For example, Ampex in the US was the company which invented video recording system; yet, JVC of Japan was the one to become successful with the VHS standard. The same is true with Netscape, the first Internet browser, which has not been the most successful one. iPhone as the most successful and amiable smartphone was not certainly the first one on the market either.

Therefore, we can observe that successful innovation must generate higher value for customers in the first step and then sustain this value in long run. Many new and good ideas can be generated within a company; yet, only those which have an internal rate of return that is significantly higher than cost of capital are called innovations. This higher rate can justify resources for stimulating innovation compared to the other alternatives and can also reward for the risks taken with innovation. In brief, a sustainable economic success can lead invention to innovation. To succeed, organizations need to build up their competencies for managing and sustaining both invention and innovation and we can think of this success by multiplying creativity to innovation, if either has a zero score then the success is zero. Apple Inc. is considered as magnificent icon of inventiveness by connecting creativity with technology in order to create value in digital-age economies. The company represented amazing products that directed and transformed seven fields i.e., personal computing, animated movies, music, phones, tablet computing, digital publishing and retail stores. Apple Inc. combined leaps of the imagination with amazing feats of engineering and became the US most admired company.

V. INCREMENTAL AND RADICAL INNOVATION

Innovations can be incremental or radical, based on the nature of knowledge and the amount of knowledge to be acquired and applied.

Likewise, every improvement in products or services can be seen as an incremental innovation which is a kind of solution for problems in current set-up. Incremental innovation exploits existing forms and/or technologies in order improve or reconfigure something that already exists [9]. Most businesses and organizations are good at incremental innovation. A radical innovation, in contrast, is a departure from existing technologies or methods which in many occasions creates new and emerging market [9]. A radical innovation demands an entirely new approach to do or make things. As such it is often risky and challenging and requires more time and budget. Most large organizations are not so good at radical innovation.

The four types of innovation on the basis of the nature of knowledge and the amount of knowledge to applied and acquired are plotted in Figure 2.

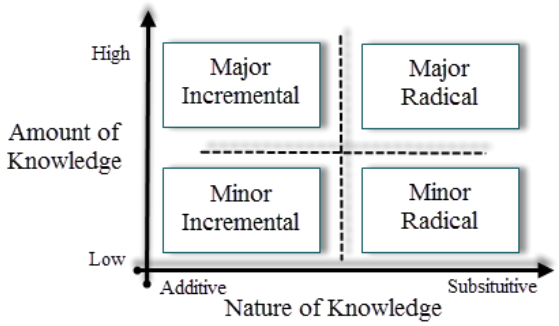


Figure 2. Innovation plot adopted from [9].

A. Major Incremental Innovation

The innovation is a major incremental if the nature of knowledge is additive that is to say there is a new thing but it is in sequence with the existing findings and the amount of knowledge that is to be acquired and applied is high [9].

An incremental innovation would build upon the existing knowledge and resources and enhance the competence of an organization. Mobile commerce, which is in sequence of developing and enhancing e-commerce can be considered in this innovation category.

B. Minor Incremental Innovation

The innovation is minor incremental if the nature of knowledge is additive that is to say there is a new thing but it is in sequence with the existing findings and the amount of knowledge that is to be acquired and applied is low [9].

An incremental innovation would involve modest technological changes to improve the existing products and/or services and sustain their competitiveness. Adding more features on existing 2010 Microsoft Office and updating its bug fixing is considered to be in this innovation category.

C. Minor Radical Innovation

The innovation is minor radical, if there is a new thing but it is in sequence with existing findings and the amount of knowledge that is to be acquired or applied is low [9]. When an organization first published its website for e-commerce and online sales that is considered a minor radical innovation for the organization since, many other websites of the same nature are already exist. Yet, the nature of sales and marketing for this particular organization is radically changed.

 ISBN: 978-1-61208-329-2

D. Major Radical Innovation

A radical innovation, on the other hand, requires completely new knowledge and/or resources and will be, therefore, competence destroying [9]. A radical innovation involves large technological advancements, rendering the existing products, rendering the existing products non-competitive and obsolete [9]. As in many field, a radical innovation undermines and abolishes established products and services by passing of time. For example, music downloads over the internet as a radical innovation would make music CD obsolete. Digital photography has also eroded demand for traditional photography film. Founded in 1880 by George Eastman, Kodak was one of the America's most notable company, which established market for camera film and dominated the field afterwards. However, the company has been struggling for years to adapt to an increasingly digital world before filing for bankruptcy protection on January 2012.

The big winners often are not the companies that obtain new technologies and use them to enhance existing products; rather they are the companies that understand how those technologies can be used to create better customer experiences than existing applications do and the biggest winners will be companies that learn to systematically produce one technology epiphany after another [10]. Drucker [28] said that 'Every organization must prepare for the abandonment of everything it does.'; this statement brings us to the fact that those companies who cannot keep up with innovation and technology trend will be soon out of the market. The change in the mobile phone market caused by iPhone has most severely affected Nokia and Sony Ericsson which used to sell quality and affordable feature phones. Apple dared to be different and innovator by offering unique features. While more expensive than many alternatives, iPhone is wildly popular with low return rates and high user satisfaction levels.

To sum up, it is presumed that creativity should be incorporated in daily work of the organization where everybody is encouraged and rewarded to generate new ideas again and again. The aim should be to build up a climate where innovation is stimulated and sustained in both dimensions, i.e., the amount of knowledge and the nature of knowledge as a culture.

VI. KM CURRENT MODELS

KM has fueled the creative process of many companies like Google, Microsoft and Apple Inc. as it is an intricate weaving of knowledge by collecting, storing, sharing and finally putting it into practice. KM activities must have a conceptual framework to operate within in order to ensure that they will be coordinated and produce the expected KM benefits [1].

Many KM models represent a holistic and comprehensive perspective (i.e., they are comprehensive and take into consideration people, process, organization and technology dimensions [1]. We review the most known models here.

Weick [23] proposed a theory of sense making to describe how chaos is transformed into sensible and orderly processes in an organization through the shared interpretation of individuals. He claims sense making consists of four integrated processes, i.e., ecological change, enactment, selection and retention. Nonaka and Takeuchi [24] studied how knowledge is produced, used, and diffused within an organization and how such knowledge is contributing to the diffusion of innovation. Wiig [23] focuses on the three conditions that need to be present for an organization to conduct its business successfully: it must have a business (products and services) and customers for them, it must have resources (people, capital, facilities), and it must have the ability to act. von Krogh and Roos [26] distinguished between individual knowledge and social knowledge and took an epistemological approach to managing organizational knowledge: the organizational epistemology KM model. Boisot [22] distinguished information from data and emphasized that effective movement of information goods is very much dependent on senders and receivers sharing the context and same coding scheme or language. Choo [25] has described a model of knowledge management that stresses sense making, knowledge creation, and decision making. The model focuses on how information elements are selected and subsequently fed into organizational actions.

Bennet and Bennet [21] described a complex adaptive system approach to KM and believed that the organization can be viewed as a system which is composed of living subsystems that combine, interact, and coevolve to provide the capabilities of an advanced, intelligent, technological, and sociological adaptive enterprise.

Despres and Chauvel [16] suggested that four dimensions cut across KM field:

- Time: referring to a linear and simplified representation of cognitive process, including the (a) mapping, (b) acquisition, (c) codification, (d) storage, (e) application and (f) transformation of knowledge or its elements.
- Type: referring to tacit and explicit knowledge
- Level: referring to different levels of social aggregation.
- Context: referring sense-making, in that no knowledge element has any meaning outside of a given context.

Meanwhile, they concluded that seven major clusters of activity are currently active in KM and the majority of behaviors and practices associated with KM may be located in this classification.

- Business intelligence;
- Benchmarking;
- Data warehousing;
- Groupware/virtual teaming;
- Communities of Practice;
- Innovation/synergies, Creativity, and
- Learning/Competencies/Employee Development.

 ISBN: 978-1-61208-329-2

VII. INTRODUCING PLA MODEL

In fast-moving world, most organizations rely on their ability to consistently deliver new and improved products and services to their audiences. KM managers have a pivotal role to play in helping their firms to become more innovative [11]. Indeed, creativity and innovation are at the cutting edge of KM, although there is generally lengthy time span between development of the new knowledge and its transformation into commercially viable products and services. Most company's innovation efforts start with ideas and brainstorming sessions which are nothing more than a one dimensional approach. More importantly, it ignores the organizational capabilities and lacks the assessment tool whilst reducing innovation chance of success.

The basic model of KM, on the other hand, lacks the measurement and assessment spirit in order to analyze, adjust and improve KM outcome. The model is too general to align with organization's goals and treats ineffectively with the creativity and invention process. Therefore, this KM approach leaves not much space for fostering creativity and stimulating innovation. As the saying goes everything that get measured can be evaluated, adjusted and then controlled and improved. There is an old saying that what gets measured gets managed.

In today's global business organizations need to integrate the measurement and management of company's tangible assets with the assessment of knowledge assets. On the other hand, Invention and Innovation are extremely dependent on the availability and richness of knowledge; therefore, we developed a new approach for planning, leading and adapting new ideas and integrate it into KM classical model, as shown in Fig 3. We name the model PLA (Planning, Leading, Adapting), which enables us to have a closed loop. The model tries to measure and assess the whole process in order to foster creativity in the first two steps of capturing and storing and then stimulate innovation in the second two steps of disseminating and implementing.

	CREATIVITY		INNOVATION	
Planning Leading Adapting	Capture and Generate	Store and Structure	Analyze and Disseminate	Implement and Extent

Figure 3. KM Framework for fostering creativity and stimulating innovation.

Planning: Refers to the policies, methods and logic formulated for steps 1 through 4 of classic KM model in order to meet and exceed current and future customer's perceived value. We need to accurately describe the whole KM model and shall consider where to begin, how to structure, which people should participate and how they should be trained (assembling team). In this phase, we delineate the objectives and boundaries of our KM system in terms of scope, time and budget. This is mainly focused on defining our customers and their requirements as well. The goals and/or objectives of KM system are then set based on the customer's current requirements and future expectations. At this stage, as shown in Fig. 4, we form a brainstorming team of creative thinkers in the community

Leading: Refers to the extent to which and how well the company execute steps 1 through 4 by creating the required environment and sustaining invention and innovation capacity and capabilities in order to implement policies and methods formulated in planning stage. It also involves defining the indicators to monitor and measure success level for each step of classic KM model and then quantify it. To measure KM success, we first develop a data collection plan and document the steps we intend to quantify. We will then select our metrics and key performance indicators and then conduct data collection and measure the indicators. At this stage, as shown in Fig. 4, we allow new ideas to be evaluated by selected groups of community and/or potential customers.

Adapting: Reflects on the assessment of planning and leading processes to ensure the right modifications for effective execution of KM steps in order to adapt the whole approach for engaging and responding to customers and stakeholders. This step focuses on the indicators to reduce the gap between the current performance of system versus the desired goal. KM systems need to find new ways for doing things better, cheaper and faster. Adapting also ensures that performance improvement remains at the desired level. We institutionalize this by modifying policies, procedures and incentive system. We can plot these modifications on a small scale to determine their sustainability and then implement them on a wider scale. In this stage as shown in Fig. 4, we build another team of selected experts to put ideas into practice.

VIII. IDEA FUNNEL AND PLA

The process from idea generation to market place is challenging and demands systematic assessment since many ideas - if not most - are either technically unfeasible, too expensive to implement or simply not appealing to the customers. The idea funnel is a metaphor to eliminate unpromising ideas at early stages in order to avoid wasting time and resources. The funnel has wide mouth into which all ideas are poured and a few of them pass towards marketplace while the funnel narrows by the criteria which are already defined. While aggregating PLA approach into idea funnel, we can observe that planning is bold at the entrance of funnel which demands most of time and resources should be spent in formulating a plan at the beginning. Then, the ideas are leaded and screened according to the criteria set by the organization and pass through for evaluating and finally adapting to the customers' or potential customers' requirements. The diagram highlights the fact that as ideas pass through the funnel the center of attention will be shifted from planning to leading and then adapting. The same is true in KM framework of Fig. 2 in which more attention to be given to planning and leading in the first two steps of KM whereas, in the second two steps the emphasize is on adapting.

 ISBN: 978-1-61208-329-2

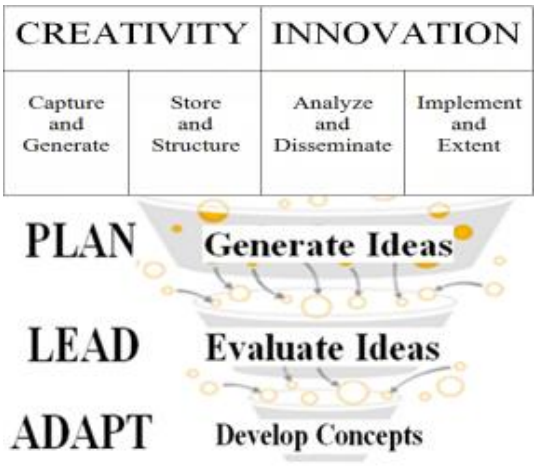


Figure 4. Idea Funnel integrated with PLA approach.

IX. CASE STUDY - APPLE INC.

Apple Inc., incorporated on 1977 by Steve Jobs and Steve Wozniak, designs, develops, and sells consumer electronics, including computer software, personal computers, tablets and mobile phones. Apple achieved widespread success with iPhone, iPod Touch and iPad products, which introduced innovations in mobile phones, portable music players and personal computers respectively. Apple closed at a record share price of $665.15 on August 2012 reported by CNBC and became one of the most profitable companies on the Earth. In the same month, it had a market capitalization of $622.98 billion which was the highest nominal market capitalization ever reached by a publicly traded company. Apple's success story and astronomical growth stems from the mastery of several areas including, but not limited to, creativity, innovation, supply chain management, knowledge management, and operations.

We believe that KM is the core competency of Apple Inc., which is manifested in its corporate culture and engaging in the innovation value processes to capitalize on new market space and providing an understanding of predictive markets, finally translates into amazing products and the brand loyalty of customers. Apple's culture and knowledge-sharing initiative, implanted by its great pioneer late Steve Jobs when he created Apple University to assure continuity of his vision. He hired Yale and Harvard academics to teach the company's history to Apple executives, and he commissioned internal and external "case studies" to prevent managers from repeating strategic errors [12]. This initiative is fairly new and is not easy to measure its influence on the company's innovation and success story, yet it proves the commitment of Steve Jobs to manage knowledge flow within the company.

Contrary to current business trends toward transparency and flatter hierarchies, Apple Inc. has fiercely encouraged secretiveness, silos and a start-up mentality and Apple's marketing campaigns resists imparting information and actively seeks to protect it [13]. Despite all these and the lack of first hand data of what exactly is happening inside Apple Inc., we observed the external manifestation of Apple innovation and focus on its strengths in knowledge management and try to evaluate this with our proposed KM model in this section.

A. PLANNING in Apple

Apple had formulated solid polices, methods and logic to create and collect knowledge through different source including staff, customers and stakeholders. It further constraint its secrets so there have been limited leaks whereas data were distributed among design teams, selected suppliers and premium customers for further studies. Apple also had the ability to constantly change its structure and goals with new information in order to meet the customers' requirements by targeting six industries including personal computing (with Apple computers and laptops), animated movies (through Pixar, which pioneered computer animation), music (through the iTunes Store and the iPod), phones (with iPhone), tablet computing (with iPad), digital publishing and retail store (by opening Apple own physical store).

B. LEADING in Apple

In leading stage, Apple stimulated creativity and fostered invention and innovation capacity and capabilities and came up with the grand vision that the personal computer should become a "digital hub" for managing all of a user's music, videos, photos, and content. Apple, thus, got into the personal-device business by designing and selling collection of finest products in respective six industries that are iTunes music store in April 28, 2003, which had over 40 billion downloads before Xcode introduced. In January 2007, Apple introduced the iPhone and sold 37.04 million in Q1 this year alone. Apple TV, iPod touch, and iPod classic introduced as well. iPhone 3G in 2009, iPad in 2010, which became the number one selling tablet to date by having half of the market share. In 2008, MacBook Air introduced, worlds slimmest computer at the time iPhone 3G introduced. In January 2010, Apple introduce the iPad, the number one selling tablet to date which still owns more than half the market share. Apple also set up its own direct-to-consumer product distribution service, first with the online Apple Store, which handled $12 million in sales in its first month of operation, and then with Apple retail stores in high-end, prized locations. Steve Jobs brought his famous attention to detail to every design aspect of Apple's stores, including the shelving, flooring and lighting [14]. In 2010, Apple came up with the successor strategy—the "hub" would move to the cloud—and Apple began building a huge server farm so that all a user's content could be uploaded, and then, seamlessly, synced to other personal devices [12].

C. ADAPTING in Apple

Apple constantly assessed the former stages to ensure the right modifications for effective execution of KM steps in order to adapt the whole approach for engaging and responding to customers' and stakeholders. Apple would never release any new product unless its designers and engineers had successfully answered his favorite question,

 ISBN: 978-1-61208-329-2

"Will this help the purchaser?" [14] Apple also ensures continuous improvement in overall performance in delivering the right goods and/or services to the customers. This happened when Apple introduced the original iMac. The device was quite useful for managing a user's photos and videos, but it was left behind when dealing with music. People with PCs were downloading and swapping music and then ripping and burning their own CDs and the iMac's slot drive couldn't burn CDs [12]. Apple created an integrated system that would transform the music industry instead of upgrading the iMac's CD drive. The result was the combination of iTunes, the iTunes Store, and the iPod, which allowed users to buy, share, manage, store, and play music better than they could with any other devices [12]. After the iPod became a huge success, Apple explored all possible reaction of competitors and learned that mobile phone maker might add music player to their handsets. So, Steve Jobs cannibalized iPod sales by creating the iPhone. "If we don't cannibalize ourselves, someone else will," he said [12].

X. KNOWLEDGE HIERARCHY

Reviewing Apple case study pointed out an important role of its founder, Steve Jobs that, we believe, acted as a KM director and leader to nurture creativity and foster innovation. Visionary leaders as Goleman [19] discusses can see the far-flung consequences of local decisions and imagine how the choices they make today will play out in the future. Jobs was the one who saw the commercial potential of many innovations in computer, music and mobile industry, well ahead of anyone else. That is why he has been described as the "Father of the Digital Revolution" [17], or a "Master of Innovation" [18]. The case also illustrated how Jobs created an environment where a clear vision of KM leader challenged the people to deliver and break out of their traditional thinking patterns. The other important issue as indicated by Dugan and Kaigham [20] is that breakthrough innovations, by their very nature, do not lend themselves to consensus. Most of innovative products developed by Apple Inc. dazzled and jumped off the page because Jobs himself played a significant role as KM leader and director who had visibility into and the authority to define and select projects while reallocating and reprioritizing resources.

As such, we add the 5th level in Data-Information-Knowledge-Wisdom hierarchy (DIKW), which was introduced by Russell Ackoff in his address accepting the presidency of the International Society for General Systems Research in 1989 [29]. Perspicacity as Indicated in Table I. is the gift of seeing and understanding people, things, or situation intelligently far ahead of others and setting trends to reshape specific sector and push value to the relevant part of the eco-system.

TABLE I. REVISED KNOWLEDGE HIERARCHY

TYPE	Definition	Purpose
PERSPICACITY	The gift to see and understand people, things, or situations intelligently and setting trends.	Prediction, Intuition, Sixth Sense, Inspiration
WISDOM	Knowledge of what is true or right coupled with good sense, judgment and expertise.	Decision Making, Creation, Innovation
KNOWLEDGE	Information in context to make it insightful and relevant for human action.	Production, Development, Improvement,
INFORMATION	Data placed into a form that is accessible, timely and accurate.	Storing , Accessing
DATA	Raw facts, figures and records contained in a system.	Capturing, Processing

XI. CONCLUSION AND FUTURE WORK

This paper has explored the convergence of knowledge management and creativity and innovation. The proposed framework and model tries to promote the continuous quest of the business community to describe two of the most important resources of sustaining and developing the business of a company - creativity and innovation. Developing and bringing to market innovative products ahead of competitors can generate various benefits in economic, preemptive, technological and behavioral factors. [15].

We argue that the proposed model builds new insights into the role of knowledge management systems in knowledge-intensive organizations for fostering creativity and stimulating innovation. We further highlighted an important role of KM director and leader who employs perspicacity to see, understand and recognize things and situations that are beyond the realm of normal expectations in science and technology to guide and direct innovation trends. We believe that implementing a successful KM system to foster creativity and stimulate innovation in every organization requires adopting a multidisciplinary perspective, encompassing issues of strategy, structure, systems and human resource management. This requires more detailed analysis on successful cases in order to develop KPIs to bridge the gap between goals and results in KM which is beyond the scope of the present paper. We hope that the idea of formulating a two dimensional model for directing and leading KM process will generate interest for further research in this area.

REFERENCES

[1] K. Dalkir "Knowledge Management in Theory and Practice" The MIT Press, 2011.

[2] R. Bates and S. Khasawneh "Organizational learning culture, learning transfer climate and perceived innovation in Jordanian organizations" International Journal of Training and Development, vol., 9, pp. 96 – 109, 2005.

 ISBN: 978-1-61208-329-2

[3] M. Grauer, U. Müller, D. Metz, S. Karadgi, and W. Schäfer "About an Architecture for Integrated Content-Based Enterprise Search" The Third International Conference on Information, Process, and Knowledge Management, pp. 48 – 54, 2011.

[4] P. Sloane "The Innovative Leader: How to Inspire Your Team and Drive Creativity" Kogan Page, 2007.

[5] Harvard Business Essentials "Managing Creativity and Innovation" Harvard Business Review, 2003.

[6] V. Govindarajan "The Other Side of Innovation: Solving the Execution Challenge" Harvard Business Review Press, 2010.

[7] J. K. Sturiak "Innovations and knowledge management" Human Systems Management 29, pp. 51 – 63, 2010.

[8] C. Hannabarger, B. Buchman, P. Economy "Balanced Scorecard Strategy for Dummies" Wiley Publishing Inc., 2007.

[9] N. Mundra, K. Gulati, and R. Vashisth "Achieving Competitive Advantage Through Knowledge Management and Innovation: Empirical Evidences from the Indian IT Sector" The IUP Journal of Knowledge Management, vol. 9, no.2, pp. 7 – 25, April 2011.

[10] R. Verganti "Designing breakthrough products" Harvard Business Review, pp. 114 – 120, October 2011.

[11] C.V. Winkelen and W. Jordan "Building the capability to be innovative" KM Review, vol. 11, issue 3, pp. 8 – 13, July/August 2008.

[12] W. Isaacson, "The real leadership lessons of Steve Jobs" Harvard Business Review, pp. 92 – 102, April 2012.

[13] A. Lashinsky "Inside Apple How America's Most Admired – and Secretive – Company Really Works" Business Plus, 2012.

[14] J. Elliot, L. William and W.L. Simon "The Steve Jobs Way Leadership for a New Generation" Vanguard Press, 2011.

[15] C. E. Castro and K. A. Desender "Analyzing Porter's Ideas: Horizontal Differentiation and Product Innovation", The IUP Journal of Knowledge Management, vol. 8, no. 3, 2010.

[16] C. Despres and D. Chauvel "A Thematic Analysis of the Thinking in Knowledge Management". In: Charles Despres and Daniele Chauvel (Eds.), Knowledge Horizons: The Present and the Promise of Knowledge Management. Butterworth-Heinemann, 2000.

[17] "Steve Jobs: Father of the Digital Revolution". People and Lifestyle. October 14, 2012. Archived from the original on July 15, 2012. Retrieved February 18, 2014.

[18] "Steve Jobs: Master of Innovation". SUCCESS. May 4, 2010. Archived from the original on September 16, 2012. Retrieved February 18, 2014.

[19] Daniel Goleman "The focused leader" Harvard Business Review, pp. 50 – 60, December 2013.

[20] R. E. Dugan and K.J. Gabriel "Putting the Breakthrough Back into Innovation" Harvard Business Review, pp. 73 – 84, October 2013.

[21] A. Bennet, and D. Bennet "Organizational survival in the new world: The intelligent complex adaptive system. A new theory of the firm. Burlington" Elsevier Science, 2004.

[22] M. Boisot "Knowledge assets" Oxford University Press, 1998.

[23] K. Weick, "Making sense of the organization" Blackwell Publishing, 2001.

[24] I. Nonaka and H. Takeuchi "The knowledge-creating company: How Japanese companies create the dynamics of innovation" Oxford University Press, 1995.

[25] C. Choo "The knowing organization" Oxford University Press, 1998.

[26] G. Von Krogh and J. Roos. "Organizational epistemology" St. Martin's Press, 1995.

[27] K. Wiig,"Knowledge management foundations: Thinking about thinking. How people and organizations create, represent and use knowledge" Schema Press, 1993.

[28] P. F. Drucker "Classic Drucker: Wisdom from Peter Drucker from the Pages of Harvard Business Review" March 1, 2006.

[29] R. L. Ackoff "From Data to Wisdom" Journal of Applied Systems Analysis, 16 (1), pp. 3– 9, 1989

 ISBN: 978-1-61208-329-2

An Implementation Tool for the Expertise Model using CommonKADS Methodology

Titah Mawloud, Mouss Mohamed Djamel, Aitouche Samia
Laboratory of automatics and manufacturing, Industrial engineering department, University Hadj Lakhdar
Batna – Algeria
{t.mawloud@yahoo.fr, d_mouss@yahoo.fr, samiaaitouche@yahoo.fr}

***Abstract*— Our work is a part of the manufacturing monitoring systems, using the model of knowledge creation for the realization of industrial diagnosis dependability aspects. Knowledge capital has an important role in organizations, particularly in the industrial sector based on knowledge. The aim of this work consists in outsourcing tacit knowledge into explicit knowledge at the thermal power plant of Jijel city in Algeria. For our analysis, we used the methodology ''CommonKADS'' of knowledge acquisition, which is standard for the development of knowledge-based systems in Europe; but, the weak points are (i) the lack of an implementation tool for this method, (ii) weak modeling language CML (Conceptual Modeling Language), because it is a semi-formal language, and (iii) lack of inference (the role of knowledge-based). Therefore, we proposed the expert system generator G2 as a computerized model of expertise for this methodology; it is a highly efficient development assistant of knowledge-based systems. This comes from the fact that it contains a natural and formal language. It is structured and allows the definition of all the elements of the methodology CommonKADS, it offers possibilities more than an inference engine. The studied thermal power plant is using an online monitoring system; it makes the detection of signs that show abnormalities using alarms. We have proposed a knowledge-based system that follows the detection to diagnose in real-time the process that ensures good continuity of production and availability of inputs, and results in quality of monitoring equipment and rapid diagnosis. Saved expertise should allow a better fit of interventions. Our contribution is in the conduct and support the diagnosis of a production system. The proposal is a tool for implementing CommonKADS, based on improvement of its weaknesses.**

Keywords-knowledge acquisition; CommonKADS; industrial diagnostic model of expertise; language CML; G2; knowledge-based system.

I. INTRODUCTION

The knowledge management is an important need, whether a company is conscious or not. It should allow locating and making visible the tacit knowledge of experts, to be able to keep, access and update, and disseminate best use of knowledge. Engineering knowledge is not simply a means of extracting expert knowledge, but it includes methods and techniques of knowledge acquisition, modeling, representation and use [1]. We chose the CommonKADS methodology [2] because it provides a framework for modeling the knowledge level. The issue is that there is a conflict between former expert without a degree and new graduate recruited employee in our companies in Algeria. Experts resolve problems rapidly using their know-how acquired over the years of accumulated experience and, the graduate employees aren't reactive and make the time to explore guides, plans, etc. The latter cannot take advantage of experts because they do not have the same background. To break these barriers, the company should externalize the tacit knowledge hidden in the minds of experts; this is the aim of our work. In the next sections, a short review of methods of knowledge management systems shows the differences between them, then, a description of CommonKADS method and its weaknesses are presented, jointly with proposals. The ameliorated CommonKADS is applied to thermal power plant to give best results of externalization of tacit knowledge.

II. METHODS OF DEVELOPMENT OF KNOWLEDGE SYSTEMS

The pioneer in the methods of knowledge capitalization is SKANDIA [13], introduced by the Swedish insurance company SKANDIA. Its strategy is to focus on human resources and their capacity to innovate and bring wealth to a business. CYGMA is a method dedicated to profession memory, in the framework of a design task, while REX and MKSM are methods which do not focus on a kind of corporate memory and do not restrict to a kind of task [12]. REX relies on the building of pieces of experience, stemming from several kinds of sources (human, documents, databases); such pieces can be retrieved in answer to natural language request. MKSM [12] takes inspiration form the complex system theory for offering a theoretical analysis of organization knowledge, considered as a complex system. The modeling phases proposed by MKSM are close to CommonKADS notions. All three methods were applied to several industrial applications. Criteria for comparing them more precisely could be: (i) the complexity level of the method application, (ii) the kind of corporate memory it enables to build, (iii) the kind of task it restricts to, (iv) the number and features of effective applications built with them, and (v) evaluation of such applications by their end-users.

III. THE METHOD COMMONKADS

This methodology is one of the results of the ESPRIT projects KADS-I and KADS-II [12]. It relies on the premise that knowledge sharing is based on the communication of knowledge and recreation. Therefore, knowledge management means sharing knowledge among multiple

 ISBN: 978-1-61208-329-2

individuals. The primary objective of the method is to assist in the knowledge modeling of an expert or group of experts in order to make a decision support knowledge-based system. CommonKADS uses more of the three categories listed above, six models to analyze the knowledge: organization, task, agent, communication, knowledge and design.

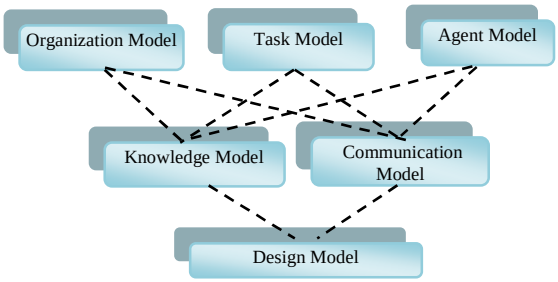


Figure 1. CommonKADS models.

- Organization model: It supports the analysis of major features of an organization, in order to discover problems and opportunities for knowledge systems.
- Task model: It analyzes the global task layout, its inputs and outputs, preconditions and performance criteria, as well as needed resources and competences.
- Agent model: Agents are executors of tasks. An agent can be human, an information system, etc.
- Communication model: It models communicative transaction between the agents involved in a task.
- Knowledge model: Its purpose is to explain in details the types and structures of the knowledge used in performing a task.
- Design model: The CommonKADS models together constitute the knowledge system.

There are several works using CommonKADs method; Recordel [9] found that CommonKADS provides a good starting point for modeling multi-agent systems as they are made to create knowledge-based systems. Therefore, extensions for CommonKADS [9] have been proposed for modeling multi-agent systems, as CoMoMAS and MAS-CommonKADS. The combination of CommonKADS with System Dynamics [10] provide effectiveness in fostering learning and transferring knowledge since such combination, integrates all important elements of an organisation's strategy and operations. CommonKADS was used by Zhang [11] to develop a learner model to give a user advice based on his knowledge to help the teacher and the learner in their tasks. In the next sections, we will illustrate the use of CommonKADS to save expertise of experts to share and to reuse it, to minimize professional mistakes, knowing dangers and risks in thermal power, and to mitigate or even better inhibit conflicts between experts and new graduate employees.

IV. APPLICATION OF COMMONKADS TO AN ALGERIAN THERMAL POWER PLANT

We applied CommonKADS to an Algerian thermal power plant. Organizational models, tasks, agents are shown in successsive sections. An extract of data sheme is presented in the class diagram in UML language [12] (Fig. 2). The principle of monitoring and diagnosis is illustrated in the diagram activity in UML language (Fig. 3).

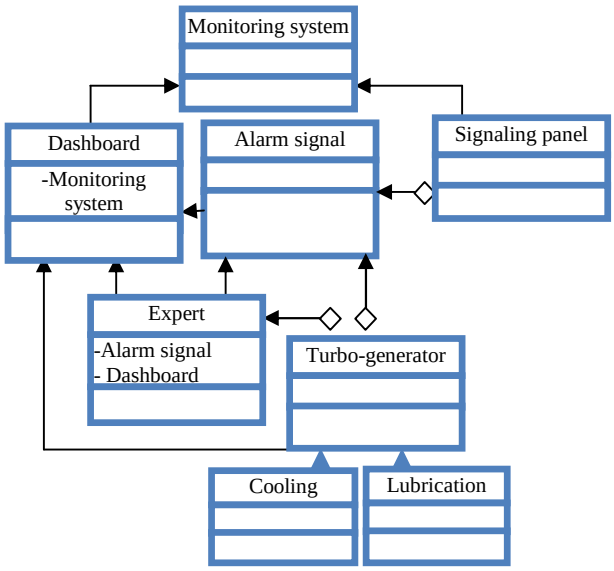


Figure 2. Class diagram of monitoring system.

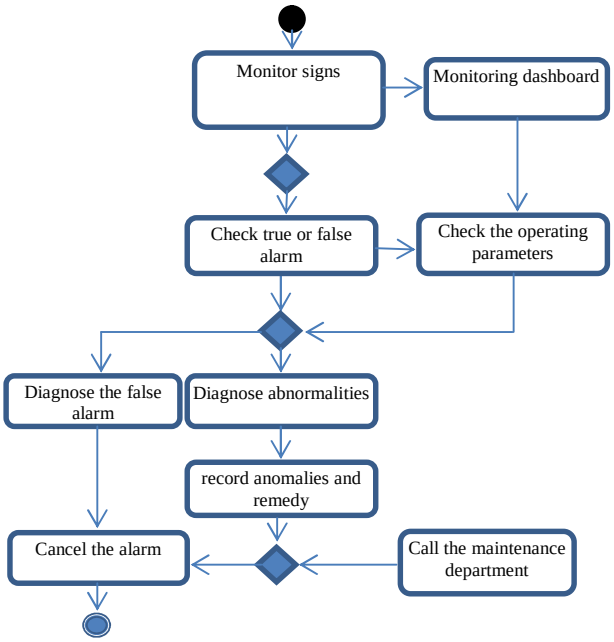


Figure 3. Activity diagram quarter production service of thermal power.

A. Organizational model

Identifying problems in the organization and solution-oriented knowledge opportunities are the first steps. During the next decade, the working age population will begin to decline when experts retire. Exporting know-how (tacit knowledge) is critical to the future of a company expertise.

TABLE I. KNOWLEDGE PROBLEMS AND THEIR PROPOSED SOLUTIONS

Problems and opportunities	✓ Lack of coordination between the division operation and maintenance division ✓ Lack of knowledge sharing between managers of company (experts) and new operating engineers. ✓ Response time in case of abnormality is very slow, which causes downtimes.
Organizational context	Mission ✓ Produces electric power of 630 MW. ✓ Ensures good continuity of production and the availability of means of production. ✓ Monitor the economic parameters and improve equipment performance. ✓ Training employees under the responsibility of the expert
Solutions	Externalization of tacit knowledge into explicit knowledge through direct interviews with experts in order to build a knowledge-based system, based on experience to ensure the transition to generations of younger workers.

B. The knowledge model

The knowledge model proposed in the CommonKADS methodology allows specifying the types, structures and roles of knowledge. This model contains three kinds of knowledge, namely, knowledge of the domain, inference, and task.

a) The knowledge domain

The S-lubrication concept has attributes which can take values. For each attribute, we define a type-of-value (value-type), such as the type-sealing system. It is a symbolic variable and takes two values (sealed or unsealed) (Fig. 4).

```
CML2 CONCEPT S-lubrication;
Description
"Lubrication system to power the turbine bearings and alternator''
ATTRIBUTES:
D-Huil: real;
T-Huil: real;
Pr-Huil: real;
Res-huil: real;
Pr-Eau-ref: real;
Fuit: symbol;
State operation: symbol;
Sealing system: symbol;
END CONCEPT S-lubrication;
```

Figure 4. The concept S-lubrication in CML2 language.

```
CML2 value-type ::= sealing-system;
Type: VARIABLE;
value-List: {sealed, unsealed}
End value-type [sealing-system];
```

Figure 5. Value-sealing-system in CML2 language.

```
CML2 BINARY-RELATION group-turbo-altrnat-alarm-signal;
DESCRIPTION:
"The relationship between lubrication system and alarm signals";
ARGUMENT-1: groupe-turbo-alternat;
CARDINALITY; 1;
ARGUMENT-2: Alarm signals;
CARDINALITY; 1..*;
END BINARY-RELATION conduccion-maquinaria;
```

Figure 6. The relationship between group-turbine generator and alarm signals in CML2 language.

b) The task knowledge

The second step in building the knowledge model is knowledge of the task; therefore, the task identification is very important. The task will support the knowledge-based system of the diagnosis failure. The knowledge model of the task has to define the task and the method to achieve it.

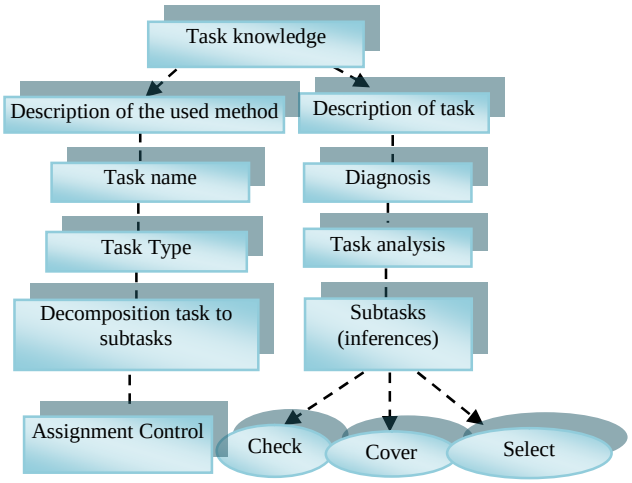


Figure 7. Knowledge model of the task.

c) The knowledge inference

Knowledge inference in a knowledge model describes the inferences, which is the lowest level of a functional decomposition. The last step of building a knowledge model is a description of each inference. Figure 8 shows the inference structure for fault diagnosis task. Inferences

proposed in this model of knowledge are: Check, Cover, Select.

```
CML2        INFERENCE :: = Check;
                    ROLES:
                INPUT: Alarm message;
            OUTPUT: true alarm, false alarm;
                    STATIC:
            Rules to check the alarm;
SPECIFICATION: "The entrance is an alarm signal in the form of a
message that indicates a fault in the system. The output is a message "
true" or "false alarm". »
END INFERENCE Check;
```

Figure 8. Description of the selected inference in CML2 language.

```
CML2        INFERENCE:: = cover;
                    ROLES:
            INPUT: Alarm message = true alarm;
        OUTPUT: All probable causes defined by the expert;
                    STATIC:
            Procedure to cover causes;
SPECIFICATION: "The entrance is an alarm condition with" true"
alarm. The set of hypotheses output (the likely causes of failure)."
END INFERENCE Check;
```

Figure 9. Description of cover inference in CML2.

```
CML2    INFERENCE:: = selected;
                    ROLES:
            INPUT: probable causes;
        OUTPUT: the most probable cause;
                    STATIC:
            Procedure to select the cause;
SPECIFICATION: “Admission is assumptions. The output is the cause
of failure”
END INFERENCE selected;
```

Figure 10. Description of the selected inference in CML2 language.

V. AN IMPLEMENTATION TOOL FOR THE EXPERTISE MODELUSING THE COMMONKADS METHODOLOGY

A. CML2 language

CML2 (Conceptual Modeling Language) is a semi-formal language and specific model of knowledge used by CommonKADS method.

B. Presentation of G2

G2 is a generator of high performance expert systems development assistance; it is used to support many applications involving various techniques of artificial intelligence: Diagnosis, alarm filtering optimization control and supervision.

C. Specific language of G2

The natural language of G2 is a formal and structured language; a developer can express instructions with familiar terms and syntax, because G2 is close to English, which is a benefit for a developer. The natural language of G2 offers:

- An interactive text editor to edit instructions for rules, procedures.
- An interactive graphics editor with:
  - Icons objects
  - Curves, plans, tables, tools
  - Buttons, dialog boxes
  - Message, etc.

D. The domain knowledge

Knowledge of the domain contains a domain schema, which describes schematically the types of knowledge and information to build a knowledge-based system.

1) Domain schema

A schema of concepts, attributes, types of values, relationships between concepts, types of rules and relations between values is defined.

2) Comparative study between CML2 and G2

A comparison is elaborated, according to some criteria, between modeling language of CommonKADS which is CML2 and the language of the generator of expert systems G2, which we used for implementing our human expert system after externalization of tacit knowledge from experience of experts of the thermal power plant.

a) Concepts

The object-oriented concept is the basis of development in G2. The object can represent something physical like a pump, a valve or something abstract like an event, a task, a message, etc.

An object class defines the properties and behavior of objects (attributes icon, etc.). G2 contains several classes that can be defined and inherited by the classes defined by the user. G2 in any class that inherits from a class above should contain all the attributes of the parent class.

Figure 11 presents the notion of concept in the language CML2 of CommonKADS method and the specific language of G2:

CML2	G2 Notes	ok
concept = Concept degasser;	Authors	Pks (July 6th 2011 12:41 p.m.)
	Item configuration	none
super-type-of: Centrale	Class name	degasser
[**disjoint**: **yes** | **no** ;]	Direct superior classes	power plant
[**complete**: **yes** | **no** ;]	Instance configuration	none
[**sub-type-of**: Concept , ... ;]	Change	none
[**has-parts**: has-part+]	Menu option	A final menu choice
[**part-of**: Concept , ... ;]	Class inheritance path	none
[**viewpoints**: viewpoint+]	Inherited attributes	none
[attributes]	Attribute initializations	none
[axioms]	Attributs display	inherited
end concept [Concept ;] .	stubs	inherited
	Icon description	inherited

Figure 11. Presentation of the concept of language and language-specific CML2 and G2.

 ISBN: 978-1-61208-329-2

b) Attributes of concepts

The attributes are the characteristics of an object. Each execution of an application under G2 is based on the behavior of objects and defined object classes.

CML2

CONCEPT degasser;
ATTRIBUTES:
Pr-deg:real;
T-deg:real;
N-deg:real;
T-deg:real;
END CONCEPT degasser;

G2

Pr-deg is given by a pression, initially is given by a pression;
T-deg is given by a temperature, initially is given by a temperature;

Figure 12. The attributes CML2 and G2.

c) The type of values

In G2, types of current values of attributes are more interesting than in CML2, and easier to be handled.

These values are recorded directly in the table object.

CML2

value-type::=**value-type** *Value-type*;
[terminology]
[**type**: **nominal** | **ordinal** ;]
< **Value-list**: { *Value* , ... } >
| < **value-specification**: primitive-type | "Text" > ;

G2

E1, E2: symbol;
t-deg, p-deg, n-deg: quantity;

Figure 13. The types of values CML2 and G2.

d) Relations between concepts

G2 has two ways to define relationships between objects: connections and relationships, connections are used to represent a physical connection between objects. Relations are only created at runtime and have not a graphical representation, and they have no attributes. They can be specified one by one, one to many, or many to many.

CML2

binary-relation::=**binary-relation** *Relation*;
[terminology]
[**sub-type-of**: *Relation* , ... ;]
[**inverse**: *Relation*;]
argument-1: argument
argument-2: argument
[relation-type]
[attributes]
[axioms]
end binary-relation [*Relation* ;].

G2 Notes	Ok
First class	Groupe-turbo-alter
Second class	Paneau-signalisation
Relation name	In-communication-with
Inverse of relation	none
Type of relation	One-to-many
Relation is	no

Figure 14. Types of relationship CML2 and G2.

E. The inference knowledge

G2 is an inference engine developing an object referring to the rules associated to this object, uses backward chaining to find values and forward chaining rules if a value is received.

CML2

Inference::=**inference** *Inference*;
[terminology]
[**operation-type**: *Name*;]
roles:
input: *Dynamic-knowledge-role*; **output**: *Dynamic-knowledge-role* , ... ;
[**static**: *Static-knowledge-role* , ...;]
[specification]
end inference [*Inference*;].

G2

If the temperature of deg…
When ….
Unconditionally…
Initially…….

Figure 15. Presentation inference CML2 and G2.

F. Communication model

The communication model allows conceptual independent modeling of interactions between different agents involved in a task. The agent could be an expert, operator or a system of monitoring. Figure 16 illustrates communication between different agents at quarter production service.

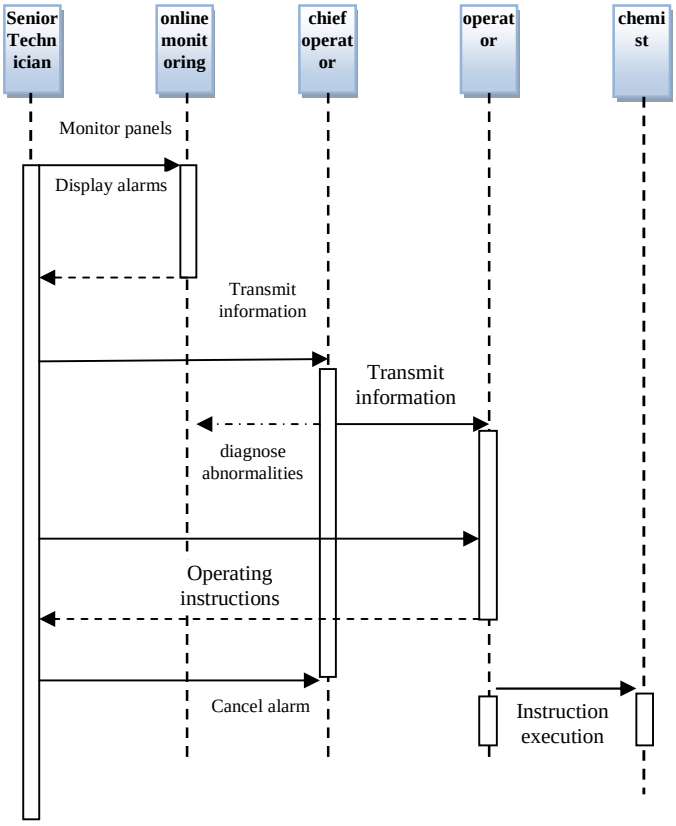


Figure 16. Communication model of agents.

VI. IMPLEMENTING KNOWLEDGE MODEL BY G2

A. The domain knowledge in G2

Domain objects are defined by icons. Each class object can have its own icon with, in this case the superclass is the central class, all elements inherit the characteristics of the superclass.

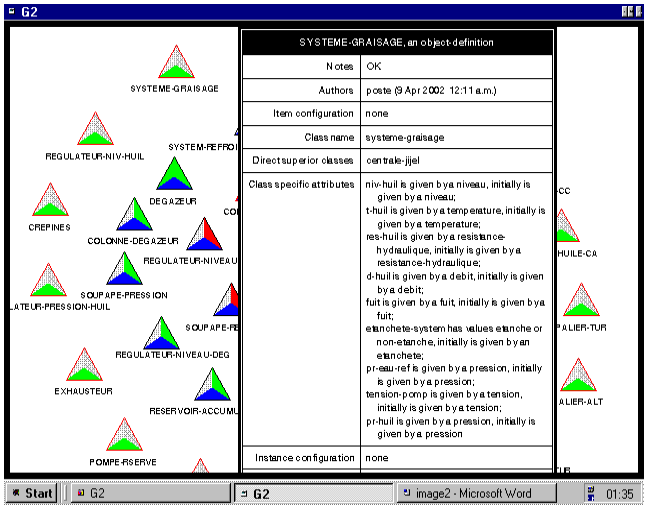


Figure 17. The attributes of the lubrication system in G2.

B. Variables

All objects of the same class have the same general structure, using the following variables. These variables are recorded directly in the object table, and can have a real-time representation.

Figure 18 refers to the alarm messages lubrication system.

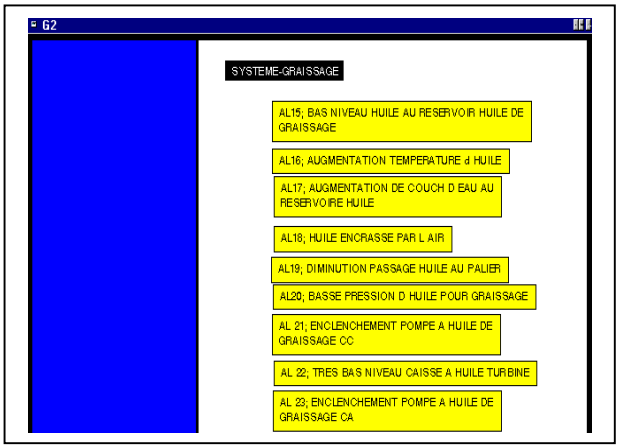


Figure 18. Variables used in the knowledge-based system in G2.

C. The rules used to check alarm

These rules for inference check if alarmis true or false.

D. The procedures

G2 contains a procedural programming language; it provides procedures to perform sequential actions. These procedures for inference identify probable causes and appear as a message.

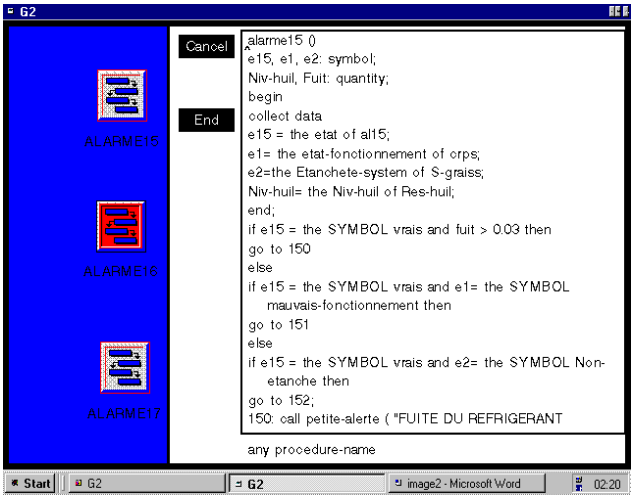


Figure 19. The procedures used for inference as identified in G2.

E. The results of the identified inference

The results of the inference are suggesting probable causes of a failure identified by the expert, as a message understood by the operating personnel.

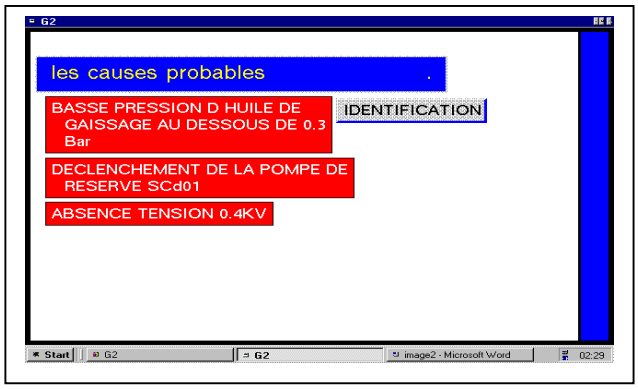


Figure 20. The output of the inference in G2.

VII. CONCLUSION

We developed a knowledge-based system under CommonKADS methodology for the application at the thermal power. The knowledge model specifies requirements knowledge/reasoning system knowledge base to implement. Then, we presented a tool that combines domain knowledge (concepts, attributes, relationships, values, variables) and knowledge inference (G2 contains an inference engine, rules, procedures, formulas, methods).

 ISBN: 978-1-61208-329-2

We found that CommonKADS is structured and offers a systematic development of based-knowledge systems, via many facilities in knowledge modeling. It is easy to understand its configuration, and ensures reusability. Otherwise, CommonKADS presents weaknesses, such as difficulties in the acquisition phase of knowledge, use of a semi-formal language CML2, which we replaced in our work with the natural language of the generator of expert systems G2.

We proposed to use the structured and natural language of G2 to define all the elements of the CommonKADS method for extracting better knowledge of an application, without using CML language which is a language with a semi-formal complexity in reasoning rules.

This broadens the scope of use of the method and builds a knowledge-based system more sophisticated.

We believe that this methodology will be used to keep formalization of the panel memory; achieving system diagnosis aid may be concretely useful to promote the sharing of knowledge between experts and all operating agents and manage the operational know-how of expert's field.

Our work emerges from several perspectives:

- The application of the methodology CommonKADS in real time.
- The generation of this method for all operating tasks and maintenance.
- The use of model expertise to optimize the process of preventive and predictive maintenance.
- Building a book of knowledge that provides a complete memory reproducing the know-how and skills of experts is useful for a company. Our solutions could be used by other manufacturing systems; cement and cylinder manufacturing, etc.

REFERENCES

[1] S. S. Bhandari, N. Chakpitak, K. Meksamoot and T. Chandarasupsang, "Knowledge Based Model for Power Transformer Life Cycle Management Using Knowledge Engineering", World Academy of Science, Engineering and Technology 72, 2010.

[2] A. Th. Schreiber, J. M. Akkermans, A. A. Anjewierden, R. de Hoog, N. R. Shadbolt, W. Van de Velde and B. J. Wielinga, "Knowledge Engineering and Management the CommonKADS Methodology", 1999, MIT Press.

[3] C. R. Guillermo, "Management of technological innovation and knowledge: synergy between theory and TRIZ Case-Based Reasoning, Application in process engineering and industrial systems", Ph.D. thesis, Graduate School: Systems Specialty: Industrial Systems, 2006, Toulouse, France.

[4] C. Dominguès, S. Christophe, and L. Jolivet, "Operational knowledge for the automatic design of legends cards", proc. extraction et gestion des connaissances (EGC), 2009, Hammamet, Tunisia,.

[5] S. V. Marinho and C. Cagnin,"Using a combination of the Commonkads and system dynamics methodologies to make operational the transition between the definition of a join innovation strategy and its implementation and management", proc. 4th International Seville Conference on future-oriented technology Analysis, 2011, Seville, Spain.

[6] S. Bruaux, G. Kassel, and G. Morel, "Critical study of the CommonKADS method application timing calculation codes", proc. francophone days of knowledge engineering: IC, 2003.

[7] M. D. Mouss, , "Diagnosis and conduct production systems approach to knowledge-based system", Ph.D thesis, 2005, industrial engineering department, university Hadj Lakhdar, Batna, Algeria.

[8] S. Aitouche, , "Development of an e-knowledge management system to ameliorate performance of a manufacturing system", Ph.D thesis, 2013, industrial engineering department, University Hadj Lakhdar Batna, Algeria.

[9] P. M. Ricordel, "Development and deployment of multi-agent systems vowels", Ph.D. thesis, 2001, laboratory Leibniz, doctoral school, mathematics, science and information technology.

[10] Y. F. Zhang, 2010, "Learner modeling in the context of a computing environment for human learning offering personalized advice", master thesis in computing, department of computer science and software engineering, Laval Quebec university.

[11] L. Edvinsson, M. Malone, "Le capital immatériel de l'entreprise Identification, mesure; management", commented by Mazar, 1999, Maxima edition.

[12] M. Titah, "Externalization of tacit knowledge into explicit knowledge", master thesis, 2013, industrial engineering department, university Hadj Lakhdar Batna, Algeria.

 ISBN: 978-1-61208-329-2

A hybrid method to develop a knowledge management system

Aitouche Samia, Mouss Mohamed Djamel, Mouss Kinza,
Laboratory of automatics and manufacturing, Industrial engineering department, University Hadj Lakhdar – Batna – Algeria
samiaaitouche@yahoo.fr, d_mouss@yahoo.fr, kinzmouss@yahoo.fr

Kaanit Abdelghafour, Boutarfa Youcef and Rezki Djamil
Laboratory of automatics and manufacturing, Industrial engineering department, University Hadj Lakhdar – Batna – Algeria
k_abdelghafour@yahoo.fr, y_boutarfa@yahoo.fr, d_rezki@yahoo.fr

Abstract—**The question is the understandability and usability of a method to develop a knowledge system. A method could be the most pertinent, but it might not have a guideline to allow its appliance. CICM (Comprehensive Intellectual Capital Model) is an algorithmic guide to develop a knowledge system. To make it more pertinent, we compared it to GERAM (Generalized Enterprise Reference Architecture and Methodology) requirements, knowing that GERAM is an ISO standard (ISO 15704) of engineering of enterprise. We deducted strengths and weaknesses of CICM. We proposed ACICM Model which is an improved CICM Model, contributing to its change, to become closer and for some criteria to go beyond to the requirements of GERAM. ACICM is more convivial and pragmatic, so more chosen by designers of knowledge management systems. Based on weaknesses of the method of performance scorecard SKANDIA, we proposed an adaptation by enrichment of its set of indicators to be more suitable to developing a knowledge management scorecard, to give (ASKANDIA).** A **hybridization of ACICM, ASKANDIA and business intelligence leaded to propose SKACICM method. We applied SKACICM on a cement company to develop e-knowledge management system ameliorating performance by a rate of 26%.**

Keywords-knowledge management; business intelligence; CICM Model; ACICM; SKANDIA; ASKANDIA; SKACICM.

I. INTRODUCTION

The emerging field of knowledge management addresses the broad processes of locating, organizing, transferring and more efficiently using information and expertise within an organization. New market forces and infrastructure changes have prompted an interest in knowledge management.

Knowledge management is the name of a concept in which an enterprise consciously and comprehensively gathers, organizes, shares, and analyzes its knowledge in terms of resources, documents, and people skills. The benefits of knowledge management are focused on improving decision making, reducing cost of employee training, and increasing versatility of the workforce. The aim is to collect explicit and tacit knowledge to develop an e-knowledge management system to ameliorate the performance of manufacturing system. First, a succinct literature review of the most recent works in knowledge management. We tried making an easy usable method (method with clear guidelines) more pertinent. We improved existing methods using proposed matamodels to propose a new method; we used to develop a knowledge management system. Finally, we discussed the contributions of the new system.

II. RECENT WORKS IN KNOWLEDGE MANAGEMENT

The main objective of Mehralian [5] is to develop and prioritize the most important indicators of intellectual capital in knowledge-based industries. The highest ratios are particularly positive work environment, the ratio of investment in R&D (research and development) and the number of R&D projects in the structural capital, while considering the relational capital. In his work, Lin [6] implies that, in addition to the direction of the market using innovative practices, the high-tech industry should focus more on market knowledge and knowledge management to customers. In contrary, we focused on capitalization of existing knowledge in the company especially the externalization of the tacit ones from the minds of company experts to share them and to improve tangible and intangible efficiency of the company. Khalid et al. [12] found that technology reshapes human behavior and it is the key of implementing knowledge management. In our proposed method, cultural change is necessary to ensure the success of implementing knowledge management. CommonKADS is a method we applied in thermal power plant [13] to externalize tacit knowledge, it presented insufficiency in its modeling language and the absence of implementing tool. We applied SKANDIA in cylinder manufacture [14], to develop a dashboard for human resources of cylinder manufacture. The indicators of SKANDIA were insufficient to cover all managerial aspects of the manufacture. Therefore, we adapted SKANDIA, adding a set of important indicators.

III. GERAM (GENERALIZED ENTERPRISE REFERENCE ARCHITECTURE AND METHODOLOGY)

GERAM has been developed by the IFAC/IFIP3 task force [8]. formed at the IFAC World Congress in 1990. It is a framework for the unification of various methods of several domains to support enterprise engineering and integration. It comprises methods from industrial engineering, control engineering, and information systems. It represents a framework rather than a modeling method comprising process models and modeling languages. The GERAM components describe requirements concerning

 ISBN: 978-1-61208-329-2

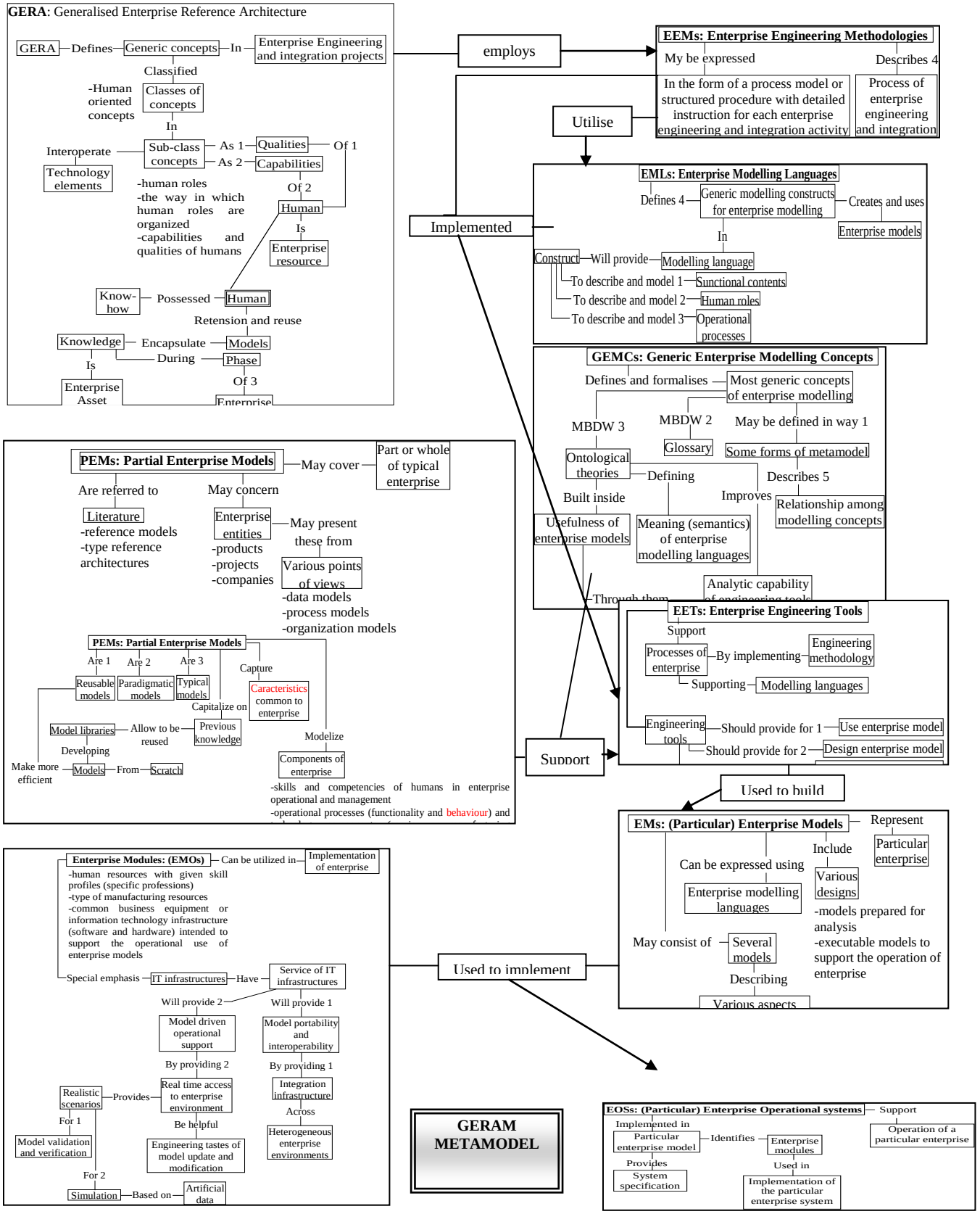


Figure 1. Proposed metamodel of requirements and their relationship of GERAM reference

 ISBN: 978-1-61208-329-2

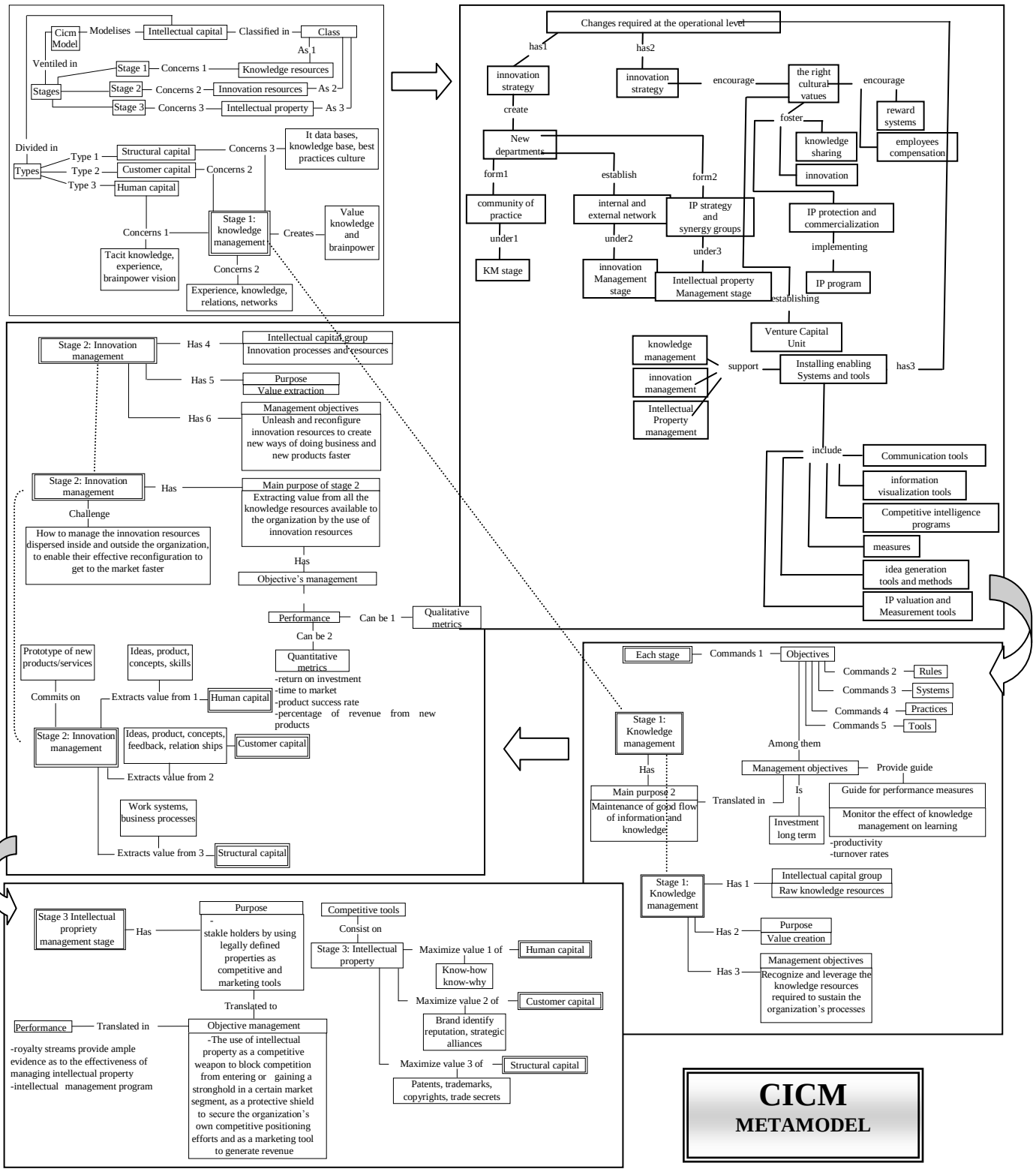


Figure 2. Proposed metamodel of the three stages of CICM model

reference architecture, modeling language, process model, tools, and enterprise models. We proposed an original metamodel (Fig. 1) defining the role and goals of GERAM.

A metamodel is the analysis, construction and development of the frames, rules, constraints, models and theories applicable and useful for modeling a predefined class of problems. As its name implies, this concept applies the notions of meta and modeling.

IV. CICM MODEL

The CICM model was developed by Al-Ali [9] to provide a comprehensive framework for the management of intellectual capital regardless of its function in the business cycle, and whether it is a resource, a process or a product. The CICM model is based on the idea that creating value from intellectual capital follows the same business process or cycle like other tangible resources and assets. It provides a framework that enables management to understand the relationship between the various disciplines and approaches that pertain to intellectual capital management, understand the various practices and services that developed under the banner of intellectual capital management and know where each of these practices fit under the three stages (knowledge management, innovation management and intellectual property management). Based on its weakenesses deduced from its comparison to GERAM requirements, we ameliorated it to give ACICM model. To perform this comparison of GERAM and CICM, metamodels are proposed in Fig. 1 and Fig. 2. This comparison consists to check whether CICM model satisfies GERAM requirements

V. AMELIORATION OF CICM MODEL AGAINST GERAM REQUIREMENTS

In Table 1, the ameliorations of CICM model are proposed, based on confrontation of metamodels of GERAM requirements and CICM model. The relation between the metamodels is wheather the CICM metamodel is included in the GERAM one.

TABLE I. AMELIORATIONS OF CICM MODEL AGAINST REQUIREMENTS OF GERAM

GERAM requirement	Conformity of CICM Model to GERAM requirement	Amelioration of CICM Model to give Ameliorated CICM Model (ACICM)
Metamodel	Not provided	Proposal of metamodels; GERAM and CICM ones
glossary	Not provided	Proposal of glossary containing all semantics used in proposed ontology and metamodels
Semantics and ontological foundation	Not provided	Proposal of an ontology which will be presented in subsequent work
Human centered modeling language	CICM is by definition human centered model for management of intellectual capital	We integrate Management of human resources in designing and accompanying the change involved by the new manner
Relationship between level of description and level of machine orientation for a selection of modeling language	Not specified modeling language for description	UML (unified modeling language) is recommended to guarantee this link. We propose for UML at its turn a new diagram, we called expertise diagram, which the role is to structure and save the expertise of experts and share it with the whole of enterprise
Kinetic representation (phases)	More Dynamic than static, very structured in kinetic phases	Proposal of detailed metamodels for each process and for each step of the process of the three stages, concerning components or structure needed for the accomplishment of the processes and steps example the step 3 (Fig. 5) of the process 3, of the knowledge management stage. It's possible recursively to perform it for all the guide step by step of CICM model
Model can be shared, reused and kept up to date	The time line isn't explicit in but implicit in saving knowledge, experience and share it and transform it in skills	We proposed a new concept which is a master plan for knowledge management with a timeline and eventual costs and value added to the enterprise. This plan should be prepared before applying the Ameliorated CICM.
Maintenability of models in different views in case of change	The difficulty in identifying which step to revise	The proposed three levels of abstraction will make revising easy, we aren't obliged to revise all levels of abstraction, only the physical, or logical one except if the change is radical, we revise the conceptual level
Performance and to_be_state views	Performance view is insufficient and not explicit	Adapted SKANDIA (SKANDIA) proposed offers this view in the three stages of ameliorated CICM (ACICM); knowledge management stage, innovation management stage and intellectual property stage.
Linkability between views	Use organization, information, functional, and resource views	We proposed a fifth new view; the knowledge view and create diagrams for it.

VI. ADAPTATION OF SKANDIA (ASKANDIA)

Skandia Navigator is an early model to analyze intangible assets, pioneered by Edvinsson [11] at Skandia enterprise. It comprises a holistic view of performance and goal achievement with respect to intangibles. It's built of 5 focus areas or perspectives, each capturing different areas of interest: financial focus, customer focus, human focus, process focus and renewal and development focus. Fig. 3 shows globally, the adaptation of SKANDIA to be a quantitative method of controlling performance by enrichment of its book performance with new set of performance indicators to be up to date and to be more convivial to world business. In process 2 of SKANDIA, we proposed to use value explorer. For the third stage of intellectual capital management, which is intellectual property stage, we proposed a suitable method; intellectual asset valuation.

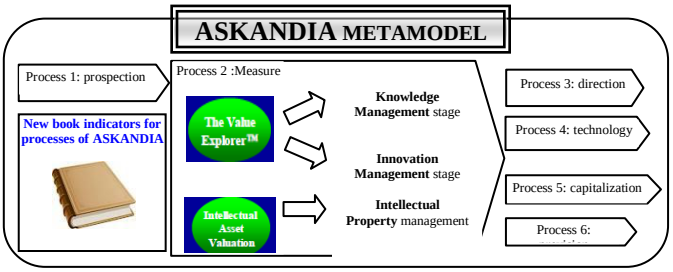


Figure 3. Proposed metamodel of ASKANDIA

VII. HYBRIDIZATION OF ACICM, ASKANDIA AND BUSINESS INTELLIGENCE TO GIVE SKACICM

Fig. 5 shows the principle of the hybridization proposed of two ameliorated methods, ASKANDIA and ACICM, described in the precedent sections, that consists in the use of ACICM to develop the three stages of a system of knowledge management (knowledge management, innovation management and intellectual property management), and performance evaluation using ASKANDIA. A system of business intelligence was integrated to our developed system to allow to the company to be all the time competitive and resistant to environment change. Fig. 4, from the left to the right, explains the processes and their steps in the knowledge management

 ISBN: 978-1-61208-329-2

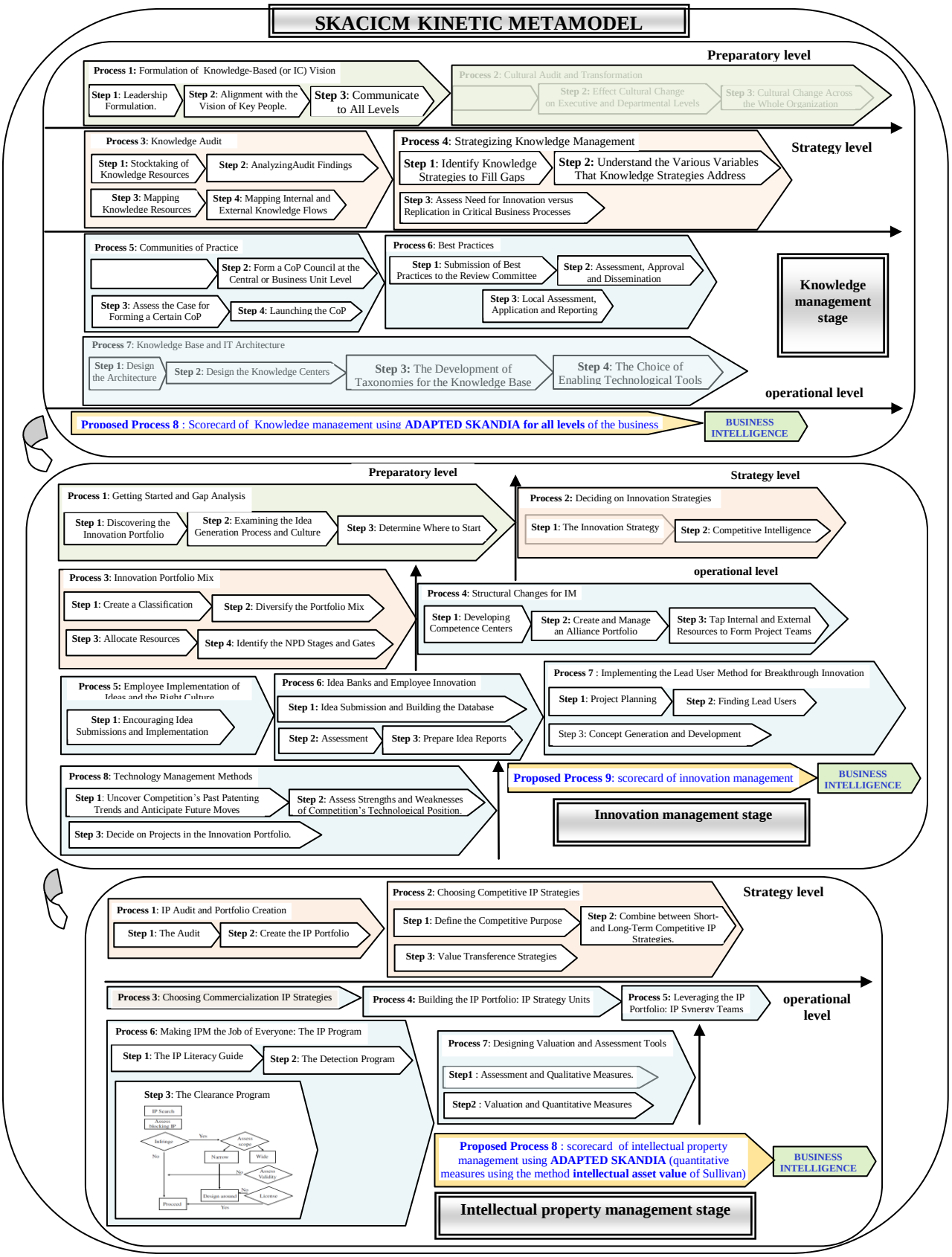


Figure 4. SKACICM kinetic Metamodel

stage. It's constituted by seven processes; formulation of knowledge-based vision, cultural audit and transformations to knowledge audit, which an important step is consisting to mapping knowledge resources, mapping internal and external knowledge flows. Next, it consists to find strategies to fill gaps of knowledge, defining communities of practice and best practices. Finally, implementation of the knowledge base on information technology architecture after its design.

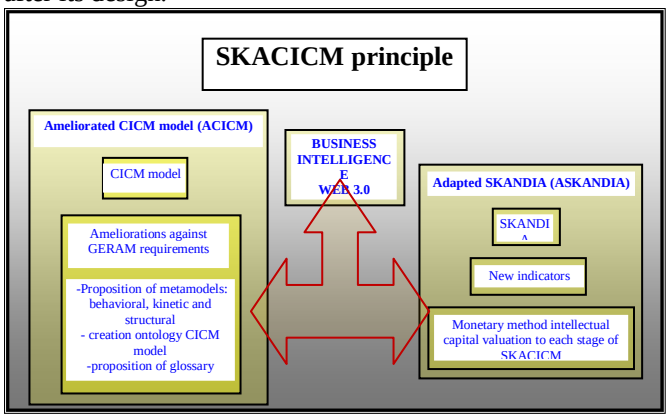


Figure 5. SKACICM principle

VIII. DEVELOPMENT OF KNOWLEDGE MANAGEMENT AND INNOVATION SYSTEM USING THE PROPOSED METHOD; SKACICM

The stages of SKACICM method was applied in cement company SCIMAT to develop a flexible tool to gather, share knowledge between the members of the communities of practice. Fig. 6 is an extract from database scheme; it shows a part of complexity of the structure of information and knowledge needed for the knowledge management.

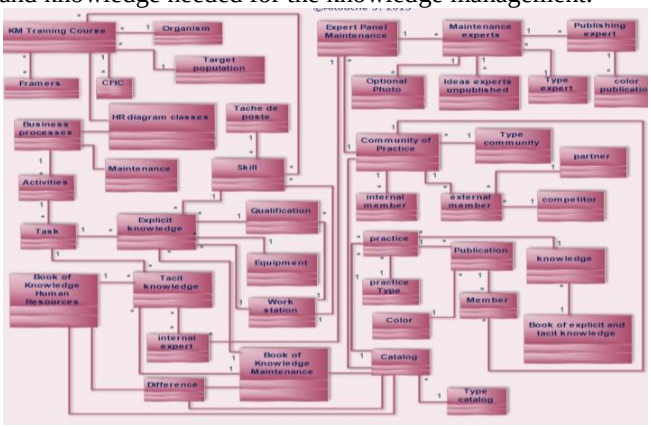


Figure 6. Extract from database scheme

Five communities of practice were created; maintenance, human resources, SCIMAT community which is the sum of all communities of the company. Two extended communities were created; SCIMAT with its partners, SCIMAT with University. Further, a special community will be created, namely, SCIMAT with its competitors. Difficulties are expected when creating this community because of the nature of the relationship between them, and the conflict of their interests.

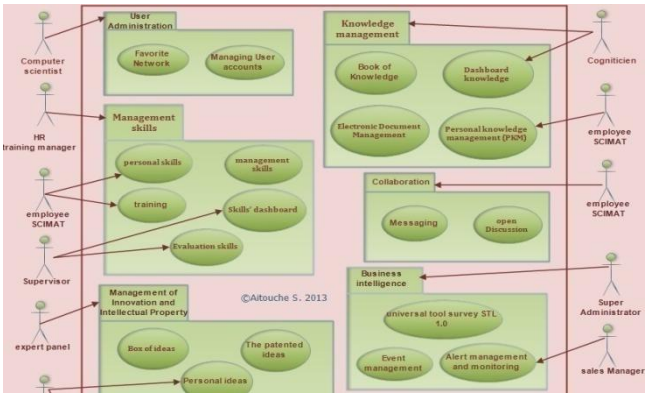


Figure 7. Functionalities of developed knowledge management system

Among the functionalities of Knowledge management system (Fig. 7); learning module, where the members of concerned community can consult knowledge books. We proposed knowledge books, in maintenance and human resources, containing the explicit knowledge and the tacit knowledge externalized from experts of the company via brainstorming, brainwritings, questionnaires and interviews.

The module "Electronic Document Management" allows the findings of electronic archived documents, plans; this makes it easy the research of documents without going to document center to retrieve documents wasting in sometimes hours. The module "business intelligence" allows the managements of new events, alerts and surveys keep company up to date. The module "collaboration and innovation" allows a communication, sharing of knowledge and ameliorates the creativity of the members of community. The novelties are proposed by a member of community, reviewed by a panel of experts and published if the new idea is accepted. The published idea should be applied by the other members of community and voted. If the vote gives more than 60% of acceptation, it becomes good practice and if it gives more than 75%, it becomes best practice. Then, we add the practice to the catalog of good and best practices (Fig. 8), and so on. This approach will improve innovation and bring wealth to the company.

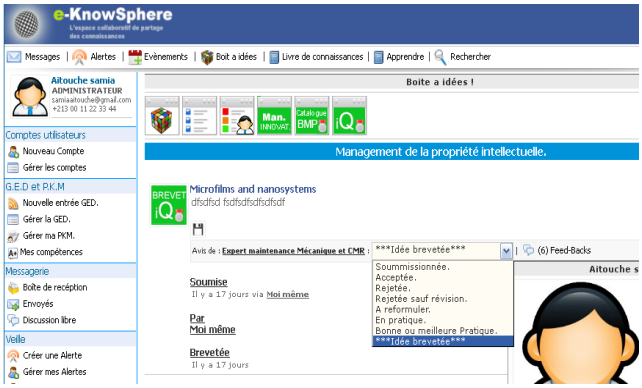


Figure 8. Knowledge and innovation management

 ISBN: 978-1-61208-329-2

IX. RETURN OF EXPERIENCE OF DEVELOPPED KNOWLEDGE MANAGEMENT SYSTEM

A. *Return of experience of professionals, extra studied company SCIMAT*

A satisfaction questionnaire was used for the professionals extra the company SCIMAT; 72% of users are satisfied by the functionalities of the developed system. 51% of them consider that it is directly usable by other manufacturing system without arrangements. Modules to improve according to the questionnaire are the skill management and the electronic management of documents.

B. *Return of experience at the studied company SCIMAT*

At SCIMAT, 83% of users are satisfied by completeness of the developed system. 61% of them consider that it is directly usable by other manufacturing systems without arrangements. Module to improve according to the questionnaire is the security of the system, which we should strengthen. The users of company are more satisfied because of their participation and validation during the development of the system of e-knowledge management.

X. CONCLUSIONS AND LIMITS

The analysis performed in this paper has shown that the metamodels are helpful to designer to refer to need easily. This comes from the readability of the metamodels. Furthermore, the ontologies and glossaries are which we expect to develop for ACICM. It could not satisfy the all requirements of GERAM because ACICM model is an enterprise engineering methodology which is only a component of GERAM and not a complete Enterprise architecture and methodology, certain requirements couldn't be analyzed in this work. A detailed analysis will be done based on detailed metamodels, in further work. All of these stages processes and steps seem long to apply, in contrary; they are very useful for the understandability of SKACICM.

Contributions of the developped knowledge management system are summarised in:

- It provides the ability to manage its knowledge, ideas and personal skills.
- Stored knowledge is a competitive advantage that improves performance in terms of quality and profitability. It allows company improvement performance of 26.82% after its operationalization.
- A business intelligence system allows internal and external audit to prepare company for contingencies.

For supporting the change, and accompany the developed system, we recommand to the company:

- Integrating the competency-based approach because it encourages the training according the new needs.
- Recruitment of a knowledge engineer.
- Set the new recruits lining with experts.
- Appeal to retirees if necessary to train new employees
- Encourage reflective practitioners because they are experts of the future.

As perspectives of our work, we foresee apply SKACICM in more industrial systems to argue its strengths, and at the studied company we foresee:

- Make the extension to other departments.
- Integrate existing expert systems.
- Improve the generator polls for automatic statistical interpretation.
- Quantification of intellectual capital for concrete gratification.

The developed knowledge management is still to complete by the other communities of the company, create their knowledge books and practices' catalogs.

REFERENCES

[1] S. Aitouche, "Designing a software platform for support of decision in a disrupted environment", Master's thesis, 2009, Department of industrial engineering, University Hadj Lakhdar, Batna, Algeria.

[2] S. Aitouche, A. Kaanit, K.N. Mouss, "Proposed decision support in a disturbed environment using the method GIMSI", proc. of IEEE international conference on computer science and automation engineering (CSAE), 2011, Shanghai, China.

[3] S. Aitouche, A. Kaanit, and K.N. Mouss, "Taking in consideration of disruptions for the analysis and design of a production system", proc. of international conference of systems and information processing (ICSIP), 2011, Guelma, Algeria.

[4] S. Aitouche et al., "Comparative study, based on metamodels, of methods for controlling performance", IJCSI international journal of computer science issues, 2012, vol. 9, Issue 3, no 2, pp. 1-9

[5] S. Aitouche, "Developpement of e-knowledge management sytem for amelioration of performance of production system", doctoral thesis, 2013, industrial engineering department, university Hadj Lakhdar, Batna, Algeria.

[6] G. Mehralian, H.R. Rasekh, P. Akhavan and A.R. Ghatari, "Hiérarchisation des indicateurs du capital intellectuel dans les industries du savoir", International Journal of Information Management, 2013, 33 (1), pp. 209-216.

[7] R.J. Lin, R.H. Che, C. Y. Ting, "transformation of knowledge management into innovation in high-tech industry", industrial management and data systems, 2012, 112 (1), pp. 42-63.

[8] IFIP–IFAC Task Force, , GERAM: version 1.6.3, ISO WD15704, 1999.

[9] N. Al-Ali, "comprehensive intellectual capital management, Step-by-Step", 2003, John Wiley & Sons, USA.

[10] O. Noran, "An analysis of the Zachman framework for enterprise architecture from the GERAM perspective", annual reviews in control, 2003, 27, pp. 163-183.

[11] L. Edvinsson, M. Malone, "Le capital immatériel de l'entreprise Identification, mesure; management", commented by Mazars, 1999, edition Maxima.

[12] H. Haron, K. Khalid, "Structurational analysis of knowledge management technology in oil and gas industry", proc. international conference on research and innovation in information systems, 2011, art. no. 6125678

[13] M. Titah, "Externalisation of tacit knowledge into explicit knowledge", Master's thesis, 2013, industrial engineering department, university Hadj Lakhdar, Batna, Algeria.

[14] N. Mebarki, "Elaboration of a dashboard of social capital using the method SKANDIA", master's thesis, 2012, industrial engineering department, university Hadj Lakhdar, Batna, Algeria.

 ISBN: 978-1-61208-329-2

Learner Satisfaction of e-Learning in Workplace

Case of Oil Company in Middle East

Muhammad Al-Qahtani
Saudi Aramco
Dhahran, Saudi Arabia
qahtms1b@aramco.com

Mansour Al-Qahtani
School of Information System
University of Western Sydney
Sydney, Australia
16605842@student.uws.edu.au

Hatim Al-Misehal
Al-Misehal Company
Dammam, Saudi Arabia
hatim.almisehal@gmail.com

***Abstract*--The goal of this paper is to identify factors that affect e-Learners' satisfaction in a large corporation in Saudi Arabia. By adopting an organisational/technological perspective to the e-Learning system, the existing body of literature on e-Learning was reviewed and a suitable theoretical model was selected as an initial theoretical framework characterising various underlying factors for the e-Learners' satisfaction in today's workplace environments in general. Through four semi-structured interviews with employees selected from the case study organisation the current study attempts to identify additional factors that may be relevant to the large Middle Eastern corporations. This, in turn, will facilitate development of a survey questionnaire that crosses distinctively different cultures, which constitutes the author's future study.**

Keywords—e-Learning; workplace learning; organizational learning; technology acceptance model.

I. INTRODUCTION

A. Background

Over the past decade, there has been a sharp global trend in the adoption of e-Learning systems by multinational corporations. E-Learning is now considered as a popular approach to learning and teaching in the workplace due to its flexibility and ease of access, just-in-time delivery, and cost-effectiveness [1], as well as ease of access, consistency, and customer value [2].

Workplace learning is defined by Ellinger [3] as "the processes, means, and activities in the workplace by which employees learn from basic skills to high technology and management practices that are immediately applicable to their jobs, duties, and roles". Workplace learning, in general, and e-Learning at workplace, in particular, is a rapidly expanding sector in many of the Middle Eastern countries. Based on a recent study, most large companies in Saudi Arabia have a long history of adopting e-Learning in their workplace [4]. The top management is well informed about e-Learning and its possibilities, and they are looking for ways to enhance their initiatives with the latest that is available in the world. Based on a recent study, the Saudi e-Learning market has been growing at 33% annually during the period 2008 – 2012 [4]. Based on other studies, the projected growth rate for the entire Middle East during the period of 2009-2014 is just 8%, making it one of the slower growth regions in the world. This may indicate that the Saudi Arabia is one of the fastest growing countries in the Middle East for embracing e-Learning methods for workplaces [4].

The current study applies an existing e-Learning satisfaction model to a workplace environment within a large petroleum company in the Middle East in order to: (i) evaluate the applicability of the model to the selected industry, and (ii) provide insights into additional underlying industry- and/or culture-specific factors that affect e-Learners in workplace environments that have not been explored in previous studies.

The ultimate aim of the study is to gain a better understanding of the above factors in workplace environments, the latter being severely under-studied. In other words, by assessing applicability of various success factors in e-Learning environments in higher education to the workplace environments the current study aims to develop a specialised theoretical framework for e-Learning satisfaction in the workplace as perceived by the adult learners. To achieve the above goals, the study adopts qualitative research methodology by conducting various semi-structured interviews with various stakeholders. Then, the results are analysed using thematic analysis in order to both evaluate the applicability of various underlying factors, as well as exploring additional factors that may affect e-Learners' satisfaction in the workplace in a large Middle Eastern corporation.

B. Case Study

The case study organization is a large oil company in the Middle East with main activities being crude oil production and petroleum refining. The company has almost 52,000 employees and the e-Learning environment of the study include Area Information Technology (IT) department with 3,000 user-learners. Arabic language development is a key requirement of the company's e-Learning program in the company, although many younger generations of employees who are graduates from the universities in western countries

 ISBN: 978-1-61208-329-2

are comfortable with just English language courses. A distinct cultural mix exists within the organisation as a result of the existence of a significant number of American and British citizens as well as young Saudi graduates from western universities at senior levels of company management. As a result of such cultural mix, one major objective of the company's workplace e-Learning program is to eliminate communication barriers throughout the organisation while, at the same time, hierarchy is to be respected and things may move slowly when it comes to decision making.

II. THEORETICAL MODEL

Historically, several theoretical models have been proposed for explaining individuals' attitudes and acceptance of information systems in general. The majority of those models are extensions of the original works, such as the *theory of reasoned action* [5], the *theory of planned behaviour* [6], and *technology acceptance model* [7], and have generalised these theories to the e-Learning environment by adopting an organisational approach a combination of organisational and technological approaches to the e-Learning. On the other hand, researchers in the fields of social psychology and information systems have identified different sets of factors for the success of e-Learning and e-Learners' satisfaction in higher education. A list of the major factors and brief explanation for each is presented in Table 1.

TABLE 1: FACTORS AFFECTING E-LEARNER SATISFACTION (ADOPTED FROM [8]

Factors	**Source**
Perceived usefulness and ease of use, course flexibility, interaction	[9][10]
Motivation, learner's attitude towards technology, computer anxiety, self-efficacy, instructor teaching style and technology quality	[11]
Course flexibility, perceived usefulness and ease of use, instructor's experience, timeliness and interaction	[12]
Learner's gender, age, initial computer skills, learning style, course quality, timeliness and interaction	[13]
Learner's initial computer skills, initial knowledge of e-Learning and courses, age, course design, instructor's timeliness, assessment methods and interaction	[14]
Motivation and interaction	[15]

A study by Sun et al. [16] provides an integrated e-Learning model based on the technology acceptance model aimed at learners in higher education with six categorical dimensions including student, teacher, course, technology, system design, and environmental dimension. The current study applies a modified version of the above model to the domain of e-Learning in the *workplace* while addressing specific issues and differences between the higher education and workplace environments. The above model is presented in Figure 1, along with the hypotheses of the study. On the other hand, some previous studies have provided insights into the differences between higher education and workplace environments, and indicate that one major difference between the two environments is the nature and motivation of the learners in the two environments [17]. The current study has incorporated these differences when designing interview questions and analysing the results.

III. RESEARCH METHODOLOGY

This study adopts a positivist qualitative research method for the following two reasons: (i) a reasonable size of literature already exists in the domain of e-Learning in higher education that can initially be used for 'exploring' relevant and specialised constructs in workplace environments (hence, adopting a positivist perspective), and (ii) there are limited prior studies that investigate e-Learning in workplaces (hence, the choice of interview method).

Interview questions were developed using the theoretical model in Figure 1 in order to assess the relevance of those factors to the workplace environments. By analysing the interview results, we aim to develop a specialised model that explains e-Learning effectiveness in workplace environments in large Middle Eastern organisations. The results from these interviews provide guidelines for developing a comprehensive set of survey questions incorporating large number of respondents for further validation of the model. However, the latter is subject of future work.

 ISBN: 978-1-61208-329-2

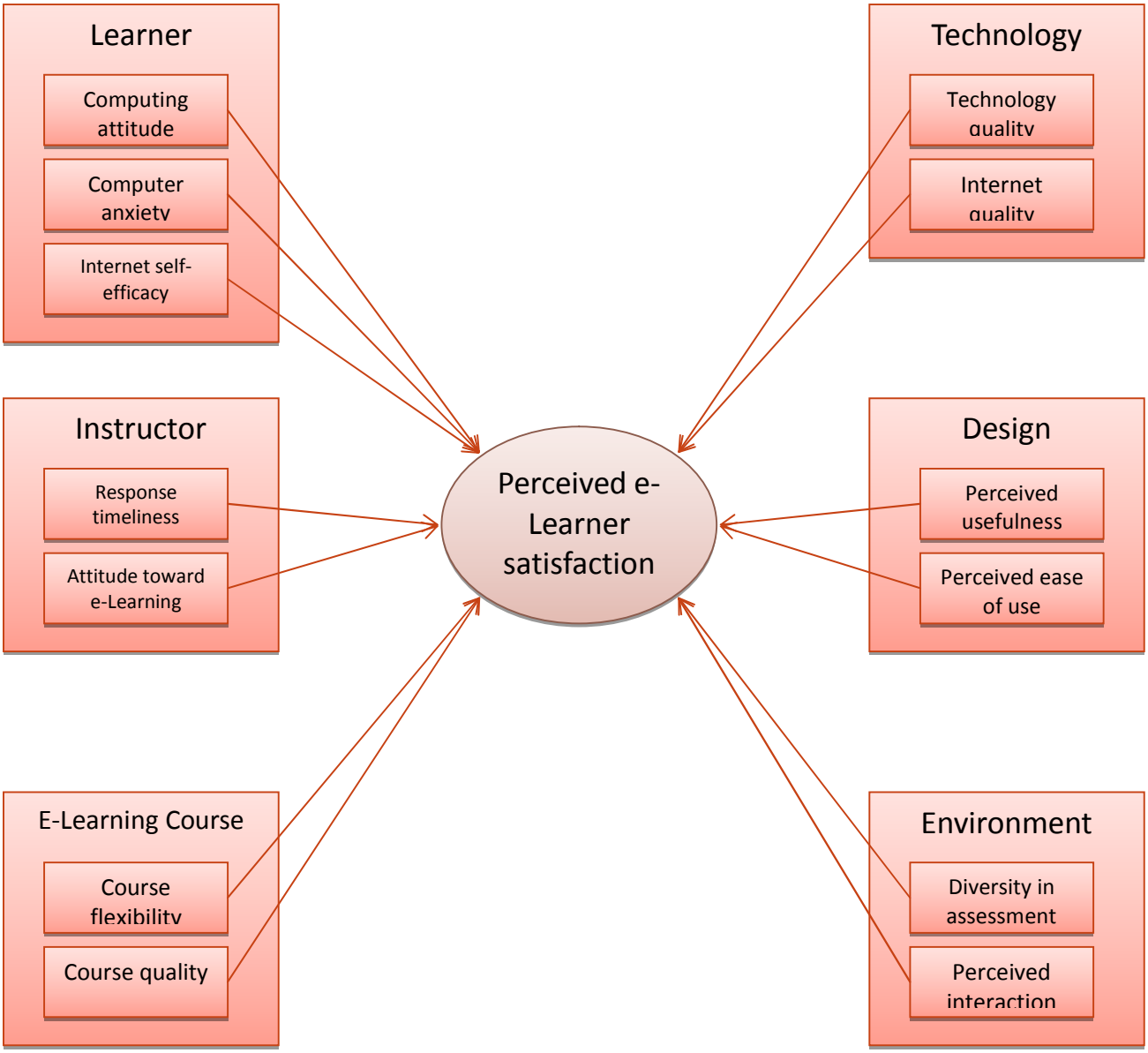


Figure 1. Proposed Conceptual Model (adopted from [16])

IV. INTERVIEW QUESTIONS

Four interviewees were selected purposefully each representing sample from various major user groups within the e-Learning organisation with adequate knowledge about the program. Each interview took almost 30 minutes. The major aim of these interviews was to assist in the development of a large survey questionnaire instrument for a future study by identifying various hidden factors and measures that collectively define e-Learning satisfaction in the case study organisation. More specifically, the interview questions were developed to achieve the following goals:

- Understanding the background of the company's e-Learning system by applying the theoretical model of the study. Such understanding was achieved by learning about various e-Learning environmental dimensions specified in the theoretical model of Figure 1.
- Confirming/assessing suitability of the above model, and
- Discovering possible additional factors that may be specific to the large corporations in the Middle East.

A summary of the interview questions is shown in the following paragraphs. These questions were used to guide the interviewee towards an understanding of underlying factors that may affect learner's satisfaction, as described by the learners and their management. Detailed responses are not reported in this paper. However, an analysis of the findings is provided in the next section.

 ISBN: 978-1-61208-329-2

Question 1: What type of computer attitude do you expect employees must have towards successful completion of this e-Learning course? Do they have to love computers and feel easy and comfortable about it OR they can still succeed even if they are not very comfortable with computers in general? Does such fact make any difference in the success of employees? Please explain your reasons in some details.

Question 2: Can you answer the above question again in relation to the knowledge of the Internet? Do you expect learners/employees to be very knowledgeable about the Internet? Why and why not? What is your minimum set of criteria for the learner's knowledge of the Internet for successful completion of the course?

Question 3: Compared with the traditional face-to-face learning, in what ways do you believe that taking the e-Learning course will help learners to improve their work and efficiency? In other words, explain the advantages and disadvantage of each method in terms of working efficiently: which method is more useful and why?

Question 4: Compared with the traditional face-to-face learning, in what ways do you believe that taking the e-Learning course will help learners to progress in their career more quickly? In other words, explain the advantages and disadvantage of each method in terms of career progress of the learners. Which method is more useful and why?

Question 5: Do you believe e-Learning course motivates learners to learn better? Why and why not?

Question 6: In what ways do you think the e-Learning course might improve the way learners learn? Explain

Question 7: How do you think the e-Learning course can be provided in more useful way?

Question 8: Do you think that there is any productivity gains in using the e-Learning courses compared to the traditional courses? Justify your answer.

A summary of the results, in the form of the most frequent responses (as interpreted by the researcher) is provided in Table II below:

TABLE II: A SUMMARY OF INTERVIEW RESPONSES

Question	Response
1	- Not a major problem today. Everyone here knows about computers and Internet - People's knowledge differs; it should be flexible in that so no one gets bored (also related to Q3, 'motivation').
2	<same as above>
3	- available when users need to complete a task - available when users need it: on-demand - available in small chunks (15-20 minutes) - it should be tailor for individuals; one size does not fit all, otherwise some will lose motivation (past experience)
4	- It must bring learning to people not people to learning - must have visible effects on my career/promotion
5, 6, 7	- Local language: Here, very few are comfortable with just English language courses. - Adults like socialization & require emotional engagement in activities. There should be opportunities for people to communicate, collaborate and share experience - Use of mobile phones is widespread. Most people carry two phones. However, data access is costly. - In the past, sometimes strictly observing hierarchy has had negative effects on me
8	- If it brings various (small/large) groups together, productivity will increase through experience-sharing - It must be directly related to my work activities

V. ANALYSIS OF RESULTS

The current study is the first phase of a large research project that consists of a preliminary qualitative study (the current study) and a comprehensive quantitative study (the future study) in order to enhance effectiveness of e-Learning methods when applied to large workplace environments. The goal of the interviews in the current study is to develop further insights into the design of the main quantitative study. Based on such objective, the interview results are analysed in this section.

One major finding from the interviews was that no instructor was normally present in the case study organisation; therefore, there were no interactions between the learner and instructor. As a result, no questions could be asked about the *Instructor's Response Timeliness*, *Instructor's Attitude Towards e-Learning* and *Perceived level of Interaction* factors that appear in the proposed theoretical model originally developed for the Higher Education.

Furthermore, findings indicated that their e-Learning system did not provide any interaction and communication functionalities among e-Learners, indicating that the current study was unable to investigate the relationship between the level of interaction among the e-Learners and the perceived e-Learner satisfaction in the workplace. However, the

 ISBN: 978-1-61208-329-2

interviewees provided insightful comments on this matter. They suggested that one of the main motivations for the case study organisation to adopt e-Learning method is the need for “managing employee’s e-Learning rather than just promoting rich learning experiences” as currently the e-Learning courses are offered with little interaction and communication functions. Confirming the above suggestion, many researchers have already suggested that interactive instructional design (highly communicative) is an essential factor for learning satisfaction and success [13][18], therefore it is recommended that organisations implement, utilise and promote interaction and communication functions in their e-Learning courses in order to enhance the life-long learning experience in the workplace.

Overall, respondents identified a variety of measures for measuring various environmental dimensions shown in Figure 1 that can be utilised when designing the survey questions of the future main study. For example, the answers to Question 1 implied that, while some employees may still prefer ‘paper work’, almost all of them recognised that the use of computers will certainly facilitate their task. Such recognition of the benefits that automated workflows may ultimately be useful to for them to overcome the short-term difficulties of learning about computers (e.g., anxiety of working with computers, etc.). Similar interpretations of other interview results will provide useful hints for developing the main survey questionnaire instrument as was one of the main intentions of the current study.

VI. CONCLUSION AND FUTURE RESEARCH

Among other things, one major objective of the current study was to act as a pilot study for developing a survey questionnaire instrument for a large sample across multiple organisational cultures. Most e-Learning courses are designed in Western cultures, whereas the largest and fastest-growing countries originate from eastern cultures such as China, Japan, India, and now the Middle East. With globalisation increasing and cross-cultural exchanges accelerating in the recent years, a future similar study that incorporates various organisational and national cultures would facilitate success of e-Learning in workplaces. We intend to extend the current study to other major corporations in various other cultures in order to explore culture-specific factors for the success of e-Learning programs in today’s organisations.

REFERENCES

[1] Wang, M. (2011) Integrating organizational, social, and individual perspectives in Web 2.0-based workplace e-learning. *Information Systems Frontiers* 13, pp. 191-205.

[2] Rainer, R.K., Turban, E. and Potter, R.E. (2007) *Introduction to Information Systems: Supporting and Transforming Business*, Hoboken, NJ., John Wiley and Sons.

[3] Ellinger, A. D. (2005). Contextual factors influencing informal learning in a workplace setting: the case of "Reinventing Itself Company". Human Resource Development Quarterly, 16(3), 389–415.

[4] Garg, A. (2012) Upside Learning Solutions, Punakar Complex, Survey No-117, Maharashtra, India, http://www.upsidelearning.com/blog/index.php/2010/03/29/workplace-elearning-in-saudi-arabia-first-impressions/ (last viewed 28th September 2012).

[5] Ajzen, I. and Fishbein, M. (1980). *Understanding attitudes and predicting social behaviour*, Prentice-Hall, Englewood Cliffs,NJ.

[6] Ajzen, I. (1991). The theory of planned behavior.*Organizational Behaviour and Human Decision Processe*s, 50, pp. 179–211.

[7] Davis, F.D., Bagozzi, R.P. and Warshaw, P.R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science*, 35(8), pp. 982–1002.

[8] Daneshgar, F., Van Toorn, C., Ramburuth, P and Hsu, J., (2010), An Investigation into e-Learner Satisfaction in the Workplace: An Australian Experience, *The International Journal of Learning*, V. 7(12), pp 29-44.

9] Arbaugh, J.B. (2002) Managing the on-line classroom.A study of technological andbehavioral characteristics of web-based MBA courses.*Journal of High Technology Management Research*, 13 (2), pp. 203-223.

[10] Arbaugh, J.B. (2000) Virtual classroom characteristics and student satisfaction with internet-based MBA courses.*Journal of Management Education*, 24(1), pp. 32-54.

[11] Piccoli, G., Ahmad, R. and Ives, B. (2001) Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic it skills training. *MIS Quarterly: Management Information Systems*, 25 (4), pp. 401-

[12] Arbaugh, J.B. and Duray, R. (2002) Technological and Structural Characteristics, Student Learning and Satisfaction with Web-based Courses: An Exploratory Study of Two On-line MBA Programs. *Management Learning*, 33 (3), pp. 331-347.

[13] Hong, K.S. (2002) Relationships between students' and instructional variables with satisfaction and learning from a Web-based course.*Internet and Higher Education*, 5(3), pp. 267-281.

[14] Thurmond, V.A., Wambach, K. and Connors, H.R. (2002) Evaluation of student satisfaction: determining the impact of a web-based environment by controlling for student characteristics. *The American Journal of Distance Education*, 16 (3), pp. 169-189.

[15] Kanuka, H. and Nocente, N. (2003) Exploring the effects of personality type on perceived satisfaction with web-based learning in continuing professional development. *Decision Education*, 24 (2), pp. 227-245.

[16] Sun, P.C., Tsai, R.J., Finger, G., Chen, Y.Y. &Yeh, D. (2008). What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction.*Computers and Education*, 50(4), pp. 1183-1202

[17] Daneshgar, F. and Van Toorn, C. (2009) e-Learning in Workplace versus e-Learning in Higher Education. *Australian Educational Computing*, 24(1), pp. 16-22.

[18] Nahl, D. (1993) Communication dynamics of a live interactive television system for distance education. *Journal of Education for Library and Information Science*, 34 (3), pp. 200-217

 ISBN: 978-1-61208-329-2

Toward a Crowdsourcing Platform for Knowledge Base Construction

Kazuhiro Kuwabara and Naoki Ohta
College of Information Science and Engineering
Ritsumeikan University
Kusatsu, Japan
emails: {kuwabara@is.ritsumei.ac.jp, n-ohta@fc.ritsumei.ac.jp}

***Abstract*—This paper proposes an approach to construct a knowledge base using crowdsourcing where the knowledge base is represented as linked data. With the crowdsourcing concept, the contents of a knowledge base are accumulated. We represent the process of knowledge base construction as a workflow. The ontology of the knowledge base's target domain is utilized in creating and executing workflows. Applications to the construction of knowledge bases in the domains of e-learning content and multilingual frequently asked questions (FAQs) are described as examples. We discuss our proposed approach from the viewpoint of a crowdsourcing platform that facilitates the use of the crowdsourcing concept to construct a knowledge base, and show how the domain ontology can be made use of.**

Keywords-crowdsourcing; linked data; ontology; knowledge base; e-learning.

I. INTRODUCTION

Crowdsourcing is attracting much attention as an approach to exploit the power of many people [3], [5], [16], [21]. For example, Wikipedia was basically created by volunteers on the Internet. Such web sites as Amazon Mechanical Turk provide crowdsourcing services.

In this paper, we propose an approach that exploits the crowdsourcing concept in knowledge base construction. With typical crowdsourcing, a given job is divided into smaller independent tasks. When such a division is difficult, coordinating the execution of tasks becomes an issue. For example, constructing a knowledge base in a company or in a group requires many tasks that involve creating a piece of knowledge. Since these tasks are not independent from each other, the crowdsourcing approach is not easily applied.

CrowdForge was proposed as a framework for applying the crowdsourcing concept to such complex tasks as writing a magazine article from a scientific journal paper [9]. In addition, TurKit, a scripting language, was proposed to specify the crowdsourcing's control flow as a script program [13]. However, it remains unclear how they can be applied to the construction of a knowledge base using domain ontology.

Here, we focus on an approach that utilizes the domain ontology to partition a task into sub-tasks and integrate their results in the construction of knowledge bases. Based on this approach, our framework can be customized to suit a particular domain for which a knowledge base is constructed by providing the domain dependent ontology. We assume that the contents of a knowledge base are represented as linked data [1], whose basic idea was proposed in the area of the semantic web to create the so-called Web of Data. Since linked data are inherently web-based, they are compatible with crowdsourcing.

This paper is structured as follows. The next section describes a crowdsourcing platform based on our proposed approach. Section III shows example scenarios using our proposed framework, and Section IV discusses related works and the features of the proposed framework. The final section concludes this paper.

II. CROWDSOURCING PLATFORM

A. Overview

The target of our proposed crowdsourcing platform is the construction of a knowledge base in a specific domain. We assume that the ontology of the target domain is provided, which is utilized to customize the platform.

The workflow of the knowledge base construction must maintain the quality of the knowledge base. The workflow is basically comprised of the division into tasks and the aggregations of the task results. The workflow's execution is monitored so that system administrators can grasp the progress of the knowledge base construction and intervene if necessary.

In addition, we consider a case where a human task and a program-based service coexist. For example, to construct a multilingual knowledge base, the contents of the knowledge base must be translated. Since many machine translation services are available, we can use such a service or a human translation service. From the viewpoint of knowledge base construction workflow, it is preferable that both the human and machine translation services be treated with the same programming interface. To achieve this, we introduce the concept of a software agent. Each task's interface is defined as an interaction with a software agent. If an individual task is to be executed by a human, then the software agent acts as an intelligent user interface to the human user. If a particular task is executed by a web service, the software agent acts as a wrapper function to that web service.

This resembles the orchestration of such web services as Business Process Execution Language (BPEL), which is extended so that a human task can be incorporated in the orchestration of services (BPEL4People [6]). BPEL4People focuses on tasks that can only be executed by humans, such as decision authorizations. In contrast, our tasks can be interchangeably executed either by a human or a web service to increase workflow flexibility.

Users of the proposed platform include not only contrib-

 ISBN: 978-1-61208-329-2

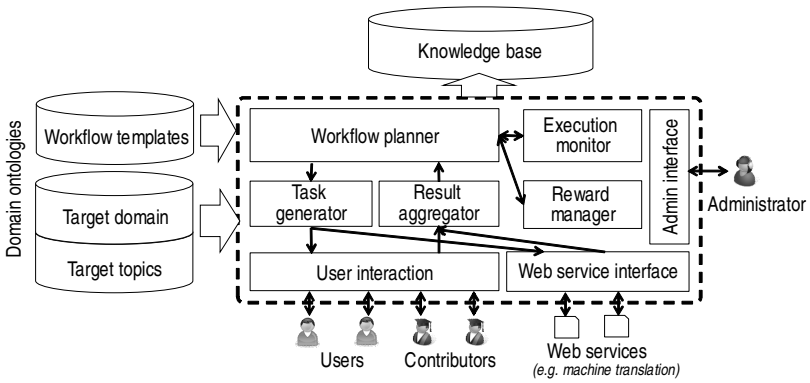


Figure 1. Overview of proposed crowdsourcing platform

utors to the knowledge base construction but also guest users who may just make comments or revision requests. In addition to them, we also assume administrator-type users who are regarded as owners of the job of constructing a knowledge base. We provide a function to monitor the progress of the knowledge base construction.

How to determine the rewards given to contributors and/or users is another issue. In the case of Wikipedia, volunteers basically contribute their time and knowledge to it. For Amazon Mechanical Turk, monetary compensation is standard, and determining rewards is important. To handle this reward issue, we define a reward manager in the platform.

Fig. 1 shows an overview of the proposed platform that incorporates the above functions. In addition, we borrow the idea of the blackboard model, a classical distributed problem solving model [7]. We focus on one that separates data and goal blackboards (Fig. 2).

The blackboard model consists of a blackboard and various knowledge sources. Each knowledge source is supposed to work on the data posted on the blackboard and write its results on it. The data on the blackboard define the level of abstraction, For example, in a typical blackboard application that interprets sensor signals, the lowest data level is the sensor's output signals, and the highest data level is the result of their interpretation. The knowledge source is assumed to process the sensor's output signals, and the results are written on the blackboard. Then, another knowledge source for a higher level works on the results of the lower level. The blackboard model with a separate goal blackboard was proposed to more easily control the execution of knowledge sources [12].

In our proposed crowdsourcing platform, a task is posted to the goal (task) blackboard, and the task's result is written on the data blackboard. The difference with the original blackboard model is that two kinds of tasks are considered. One further divides a task into sub-tasks and posts the generated sub-tasks on the task blackboard. The other executes the task itself and writes its results on the data blackboard.

B. *Workflow*

We use a workflow template to facilitate making a workflow. The following is the basic workflow of a knowledge base construction. The job owner posts a task to solicit a piece of knowledge about a particular topic. For example, if the target of the knowledge base is e-learning content, a posted task might create an exercise in the target domain. To maintain the quality of the created knowledge base, the contents must be checked and revised. We represent such a workflow with the template shown in Fig. 3. When the specific task to be executed is identified, the workflow template is instantiated and executed.

Consider another example of creating multilingual e-learning contents. The domain ontology can be divided into layers. The upper layer ontology describes the vocabulary that is common to e-learning content creation. The multilingual aspect is also described as a workflow for translating the content using a human or a machine service. The target domain

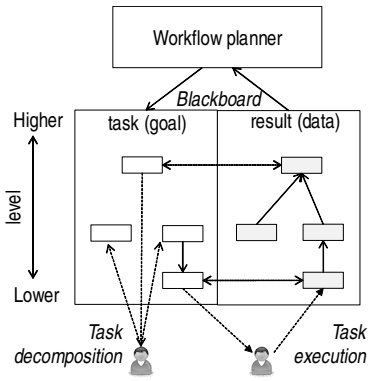


Figure 2. Using a blackboard metaphor

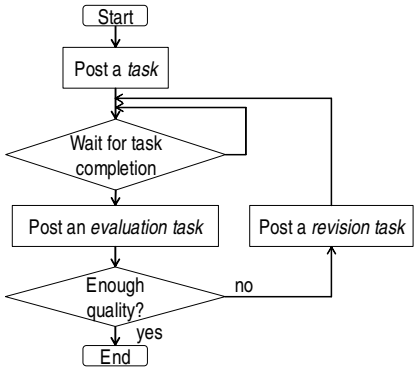


Figure 3. Workflow template for revising task results

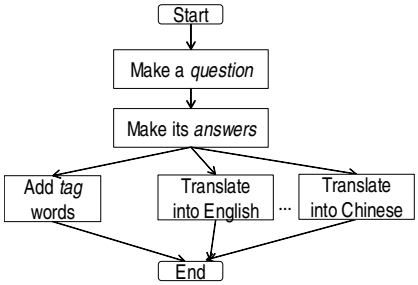


Figure 4. Workflow template for multilingual FAQs

of the e-learning environment is described as a lower layer of the domain ontology. For example, if the target e-learning content is about *artificial intelligence*, the topics that must be covered in this domain, such as *search algorithm*, will be described in the lower domain ontology. Tasks to create content for each topic are made using this domain ontology. As a result of the generated task, another task is also created for checking the generated contents.

III. Example Applications

Next, we consider two example applications: one creates a knowledge base of multilingual frequently asked questions (FAQs) in a domain of rental apartments [8], and the other creates content in an e-learning environment.

A. Multilingual FAQ system

In this application, the knowledge base contains questions that are often asked by international students living in Japan and answers in four languages. This application is intended to provide useful information to international students. Currently, the FAQs are stored in a linked data format. It is preferable that more FAQs are collected and stored in the system. Thus, it is necessary to provide a way to add to the FAQs an entry that consists of a question and its possible answers. This job can be divided into three sub-tasks: 1) adding a question, 2) adding its answer, and 3) adding translations. Its workflow template can be represented as shown in Fig. 4.

To cover as wide a domain area as possible, the job owner may want to make a task that obtains a question regarding a specific topic. In such a case, the workflow is instantiated to solicit a question and its answers in the specified topic and finally their translations in the target languages.

B. Content creation in the e-learning environment

As another example, we consider a task that creates content in an e-learning environment. In the proposed framework, a knowledge base is implemented as linked data. Thus, its contents are represented using a Resource Description Framework (RDF) [17]. In a RDF, information is represented as a set of triples, each of which consists of a subject, a predicate, and an object. In the following, we create a task that adds new exercises as e-learning content.

Assume that the target domain of the e-learning environment is *search algorithms*. An example domain ontology that describes the relationships among search algorithms can be represented, as shown in Fig. 5. The ontology itself is also represented in the RDF format.

A typical exercise can be represented, as shown in Fig. 6.

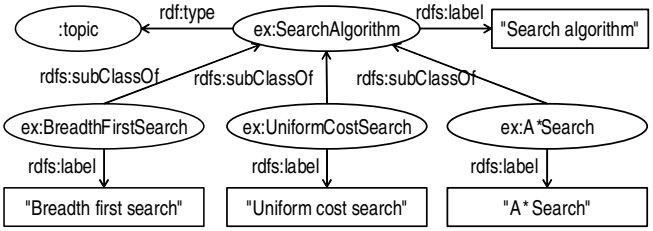


Figure 5. Example domain ontology

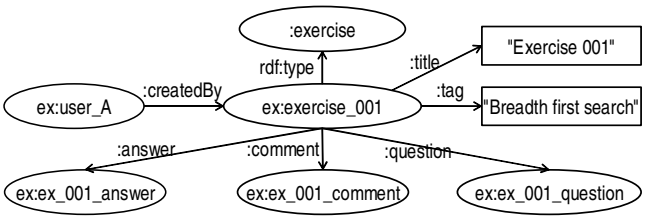


Figure 6. Example structure of exercise

This example represents an exercise that is related with *Breadth First Search*, as specified by its tag.

Next, consider a situation where some topics are not covered by existing exercises. First, such topics are specifically searched for using a SPARQL [19] query to the RDF database that stores the e-learning content. Then, if such a topic is found, a task is generated to create an exercise to cover it. Additionally, a task to revise the created exercise is also generated. These steps are determined by dividing a task of *making an exercise* into subtasks according to the workflow template (Fig. 7). The divided subtasks are executed in turn. The subtasks include a task executed by a machine (finding a topic, in this example) and a human intelligence task (making an exercise itself).

IV. Related Works and Discussion

A. Workflow execution

In crowdsourcing, a job is basically decomposed into small independent tasks that are assigned to individuals. Task decomposition itself can be performed by crowdsourcing. For example, a tool called *Turkomatic* collaboratively makes workflows with crowd workers and job requesters [11]. A method for dynamically controlling a crowdsourcing process is also proposed using the model of active rules [2]. In the proposed

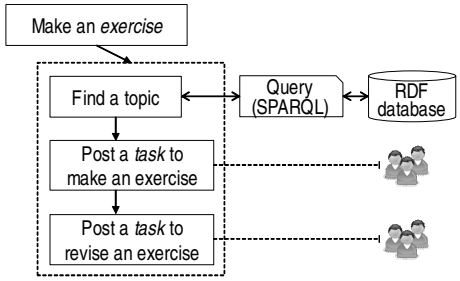


Figure 7. Dividing a task of making an exercise

platform, based on the blackboard model, we introduce two blackboards: one holds tasks and another separately holds task results. We plan to opportunistically execute task decomposition and task execution to adapt changes during a long job execution. This will be necessary for jobs like knowledge base construction that require incremental and possibly never-ending processes.

The crowdsourcing approach is used to build an ontology (for example, [4]), and we also apply crowdsourcing to revise the ontology of the target domain itself. By combining the process of the domain ontology revision and knowledge base construction, we can achieve greater flexibility.

As for a workflow's execution model, handling an exception that may occur explicitly is preferable. Human task execution is often error-prone and may fail. We plan to introduce a *meta-level* workflow to handle task failures so that the main workflow can be created by focusing on the execution of the task itself without paying attention to exception handling.

B. Distribution of rewards

A mechanism must be designed that rewards users for their contributions to the knowledge base. When a monetary reward is involved, determining the value for each task becomes an issue. To solve this problem, we will apply the concept of cooperative games [14]. Some methods have been proposed to distribute rewards in Internet environments where such malicious manipulations as false names or collusion are possible [15], [20]. It is a future issue to implement a mechanism that ensures the incentive of crowd workers, as a function of the reward manager in the proposed platform.

C. Monitoring knowledge base construction

We must also monitor how content is accumulated in knowledge base construction. A visualization mechanism is helpful for such a purpose. For example, CrowdScape [18] controls the *crowd* by visualizing the behaviors of workers. As for workflow execution, CrowdWeaver [10] visually manages complex crowd work.

In our proposed framework, we plan to exploit the domain ontology to visualize the progress of knowledge base content accumulation. For example, we plan to design a function to show how many topics are covered in the target domain by clarifying the correspondence between the contents and the topics described in the domain ontology. In this way, it is possible to provide a job owner with a feedback on the progress of the crowdsourced works. The job owner can intervene in the knowledge base construction, if necessary, to ensure the quality of the knowledge base.

V. Conclusion and Future Work

This paper described a crowdsourcing platform for the construction of a knowledge base and discussed its required functions. The domain ontology of the target domain of the knowledge base is utilized in preparing workflows and executing them. Adopting the blackboard metaphor allows the workflows to be executed opportunistically.

Currently, we are designing our platform as a web application, and are writing workflow templates for constructing knowledge bases in the example domains discussed in this paper. We plan to evaluate the proposed approach using this platform from the viewpoint of making use of the domain ontology.

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data - the story so far," International Journal on Semantic Web and Information Systems, vol. 5, no. 3, pp. 1–22, 2009.

[2] A. Bozzon, M. Brambilla, S. Ceri, and A. Mauri, "Reactive crowdsourcing," Proc. the 22nd Int. Conf. on World Wide Web (WWW '13), May 2013, pp. 153–164.

[3] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," Comm. ACM, vol. 54, no. 4, pp. 86–96, 2011.

[4] K. Eckert et al., "Crowdsourcing the assembly of concept hierarchies," Proc. of the 10th Annual Joint Conf. on Digital Libraries (JCDL'10), June 2010, pp. 139–148.

[5] J. Howe, Crowdsourcing: Why the Power of the Crowd Is Driving the Future of Business. Crown Business, 2009.

[6] D. Ings, et al. (eds), "WS-BPEL extension for people (BPEL4People) specification version 1.1," [Online] Available: http://docs.oasis-open.org/bpel4people/bpel4people-1.1.html, Aug. 2010 (accessed Jan. 15, 2014).

[7] V. Jagannathan, R. Dodhiawala, and L. S. Baum (eds), Blackboard Architectures and Applications. Academic Press, 1989.

[8] S. Kinomura and K. Kuwabara, "Developing a multilingual application using linked data: A case study," Computational Collective Intelligence. Technologies and Applications (ICCCI 2013), Lecture Notes in Computer Science, Springer, 2013, vol. 8083, pp. 120–129.

[9] A. Kittur, B. Smus, S. Khamkar, and R. E. Kraut, "CrowdForge: crowdsourcing complex work," Proc. of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST '11), Oct. 2011, pp. 43–52.

[10] A. Kittur, S. Khamkar, P. André, and R. Kraut, "CrowdWeaver: Visually managing complex crowd work," Proc. of the ACM 2012 Conf. on Computer Supported Cooperative Work (CSCW '12), Feb. 2012, pp. 1033–1036.

[11] A. Kulkarni, M. Can, and B. Hartmann, "Collaboratively crowdsourcing workflows with Turkomatic," Proc. of the ACM 2012 Conf. on Computer Supported Cooperative Work (CSCW '12), Feb. 2012, pp. 1003–1012.

[12] V. R. Lesser and D. G. Corkill, "The distributed vehicle monitoring testbed: A tool for investigating distributed problem solving network," AI Magazine, vol. 4, no. 3, pp. 15–33, 1983.

[13] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller, "TurKit: human computation algorithms on Mechanical Turk," Proc. of the 23nd Annual ACM Symposium on User Interface Software and Technology (UIST '10), Oct. 2010, pp. 57–66.

[14] J. von Neumann and O. Morgenstern, Theory of Games and Economic Behavior. Princeton University Press, 1944.

[15] N. Ohta, V. Conitzer, Y. Satoh, A. Iwasaki, and M. Yokoo, "Anonymity-proof shapley value: Extending shapley value for coalitional games in open environments," Proc. of the 7th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2008), May 2008, pp. 927–934.

[16] J. Pedersen, et al., "Conceptual foundations of crowdsourcing: A review of IS research," 46th Hawaii Int. Conf. on System Sciences (HICSS), Jan. 2013, pp. 579–588.

[17] RDF Working Group, W3C, "Resource Description Framework (RDF)," [Online] Available: http://www.w3.org/RDF/, Feb. 2004 (accessed Jan. 15, 2014).

[18] J. Rzeszotarski and A. Kittur, "CrowdScape: interactively visualizing user behavior and output," Proc. of the 25th Annual ACM Symposium on User Interface Software and Technology (UIST '12), Oct. 2012, pp. 55–62.

[19] SPARQL Working Group, W3C, "SPARQL 1.1 Overview," [Online] Available: http://www.w3.org/TR/sparql11-overview/, Mar. 2013 (accessed Jan. 15, 2014).

[20] M. Yokoo, V. Conitzer, T. Sandholm, N. Ohta, and A. Iwasaki, "Coalitional games in open anonymous environments," Proc. of the 20th National Conf. on Artificial Intelligence (AAAI), July 2005, pp. 509–514.

[21] M.-C. Yuen, I. King, and K.-S. Leung, "A survey of crowdsourcing systems," 2011 IEEE Third Int. Conf. on Privacy, Security, Risk and Trust (PASSAT), and 2011 IEEE Third Int. Conf. on Social Computing (SocialCom), Oct. 2011, pp. 766–773.

 ISBN: 978-1-61208-329-2

Center for Scientific and Technical Information – Library Services for Business and Science at Wroclaw Univeristy of Technology

Anna Walek
Wroclaw University of Technology
Wroclaw, Poland
anna.walek@pwr.edu.pl

Abstract **- Wroclaw University of Technology is situated in Lower Silesia – the dynamically developing region of Poland. Focusing on adopting its own offer to the market needs had been selected as a strategy. Due to that, a synergy effect has been achieved with the development of segments strategic for the region. The University is strongly oriented to cooperation with the economy and industry. One of the key initiatives was establishment within the University structure the Center for Scientific and Technical Information on January 1, 2014. This is the unit responsible for collecting and providing scientific and technical information for the needs of performing scientific research and supporting didactics, as well as coordinating cooperation with the economy and technology transfer. Within the structure of the Center, the Traditional and Electronic Libraries were established, providing the library-information services and creating also the digital library, knowledge repository and the data base for scientific achievements. Besides the library resources, data bases and electronic periodicals, the Center makes available the patent and standardizing information, as well as the information on new technologies both, for the needs of scientific society and industry representatives. Tasks of the Center involve, among others, also collecting knowledge on technologies developed at the University, conducted projects, available scientific equipment and informing on opportunities for commercialization of the research results. In the paper, the genesis of formation of the Center for Scientific and Technical Information and its basic tasks directed in particular at cooperation of science with economy and industry have been presented. Changes in the library and information system of the University and establishment of the Center reveal the new role of libraries and centers of scientific and technical information.**

Keywords-digital libraries; Wroclaw University of Technology; open repository; electronic resources; library services; technology transfer; knowledge commercialization.

I. INTRODUCTION

Wroclaw University of Technology is one of the best technical universities in Poland. For many years it has been on top of Polish university rankings [1]. It was founded in 1946 in Wroclaw. The same year Wroclaw University of Technology Library was established. Ever since then, it has been the largest library in Lower Silesia region collecting and circulating materials in the area of technical and exact sciences. In 1970s, Wroclaw University of Technology Library became a pioneer library as far as automating and computerizing library and information processes and implementing new information technologies are taken into account. In that period, the Library conducted scientific and research work in the above areas. Thus, it developed pioneer solutions nationwide in terms of registration, collecting and circulating scientific-research information [2].

In 1990s, the first Polish Integrated Library System (APIN) was created and implemented here, too. Its role was to provide services for all library processes. Moreover, in 2004, one of the first digital libraries in Poland was founded here as well [3].

Activities in the area of digitization, development and providing access to modern electronic resources of information has been limited for many years due to space shortage and lack of appropriate equipment. Rooms which have been occupied by the Library in a historic building from early 20th century made it impossible to develop comprehensive patron services in the area of providing electronic resources of information.

II. BIBLIOTECH PROJECT

For many years, the University authorities and scientific circles were aware of the necessity of building a modern scientific library at Wroclaw University of Technology. Eventually, the university managed to obtain EU funds for building it and in 2007 within Operational Program Innovative Economy. The funds covered the costs of constructing and fitting a modern building of "Library of Exact and Technical Sciences for the purposes of Innovative Economy" (Bibliotech) [4]. Bibliotech building was finished and handed over to the university by the contractor in October 2013.

Specific nature of Wroclaw University of Technology as a technical university which closely cooperates with the industry and develops a number of solutions and technologies implemented in the industry defines yet another, additional role of the library. It is a place of collecting and providing access to scientific and technical information for students and academic teachers. Moreover, it's another role is to serve business entities and innovative entrepreneurs from the region of Lower Silesia.

 ISBN: 978-1-61208-329-2

New library, already at the stage of planning, was meant to be a modern, electronic unit. Electronic resources, such as e-books, e-journals and databases both shared and developed in modern and well-equipped laboratories were supposed to be the basis for its activities. It was decided that the Electronic Library will become the foundation for the university library and information system. Its activities will be supported by other university units, which collect and circulate standardization and patent information, as well as information on new technologies aimed in particular at business entities cooperating with the university within knowledge commercialization. All the above assumptions have been included in Bibliotech project.

It is an ambition of Wroclaw University of Technology Library to once more become a leader and a nationwide pioneer as far as implementing new technologies and modern IT solutions are considered. More than 10 modern buildings of scientific libraries have been built in Poland within the last several years. However, to a great extent, they moved their previous tasks, collections and services to new premises only. Within the Bibliotech project, we decided to move a step forward. We brought together in one place all library services, units of scientific and technical information, standardization, patent, new technologies information office, computer and multimedia laboratories, group work rooms, individual projects laboratories, and above all, new research laboratories working on data collection, warehousing, processing and sharing. All of the above will constitute a unit supporting research work in the areas of exact and technical sciences and improving access to scientific achievements by entrepreneurs.

III. THE CENTER FOR SCIENTIFIC AND TECHNICAL INFORMATION

On January 1st 2014, the university launched the Center for Scientific and Technical Information of Wroclaw University of Technology. It is a unit serving the whole university performing scientific, research and service-oriented tasks. The new unit includes departments included within the project and located in Bibliotech premises (Electronic Library, digital laboratory, Regional Patent Information Center, New Technologies Information Office and IT research laboratories) [4]. Furthermore, it includes also other units dispersed within the campus, yet providing similar services for the sake of academic environment (e.g. Department Libraries). The aim of this procedure was gathering within one structure the previously dispersed units, restructuring them and reintegrating their activities as well as improving their functioning. At present, the Center is being organized. The new Bibliotech building is being provided with new generation equipment and cutting-edge data center.

A. Library Services

The university library-information system in the form of Traditional Library and e-Library operates within the structure of the Center for Scientific and Technical Information.

Traditional Library is aimed at students in particular. Its main goal is to circulate printed sources (journals, books, academic text books) and providing basic information regarding library collection and electronic resources.

The e-Library is the basis for the university library-information system. Within the structure of the e-Library there will be units responsible for collecting and circulating electronic resources, testing and organizing access to databases and scientific information services, as well as providing comprehensive information services for users representing scientific and business environment. The Center, along with the Libraries, will also serve as an information center, where highly-qualified information brokers will collect, verify and prepare sets or resources in various areas for the sake of users. Within the e-Library, current activities connected with registering scientific achievements will be performed (bibliographic database – DONA). Furthermore, the Laboratory of Scientometrics will be performing research in the area of scientific achievements analysis and citation analysis.

The Center continuously expands its offer of electronic sources, data bases, standards and patent data bases to meet the needs of present and future users. In 2013, the then Main Library (the present Center) was providing access to over 90 thousand titles of electronic books, over 45 thousand of periodicals and 92 specialist data bases. About 37 thousand of registered users are making use of those resources [5]. The specialist data bases and electronic periodicals provided by the Center, serve preparing the comprehensive information services, thematic sets and elaborations.

The services are offered both, for workers, doctoral students and students of Wroclaw University of Technology, as well as other interested parties, such as economic entities and industry representatives.

The Center is a key venture for the University development which defines library as the institution linking various functional structures related to scientific and research activity of the University workers.

B. Knowledge Commercialization and Technology Transfer

Within frames of the Center operate also the units dedicated for cooperation of science with the economy such as the Center for Scientific Cooperation with Economy. Its activity is focused on tasks supporting and initiating undertakings of all types in cooperation with representatives of business environment. The Center is running also a Contact Point for Technology Transfer. By creating a network of mutual relations with business and industry representatives, it identifies individual needs of enterprises in innovation, enabling that way development of solutions facilitating functioning of mechanisms of knowledge commercialization and widely understood cooperation of the University with businesses [6].

 ISBN: 978-1-61208-329-2

An important agenda in the structure of the Center is also a Department of Intellectual Property and Patent Information, where the Regional Center of Patent Information incorporated into the European network PATLIB is operating. It is also consulting intellectual property issues, and provides legal services for the commercialization process.

Guidelines of the European Commission [7][8] and Polish national and regional strategies [9] underline the importance of cooperation between science and the economy. The aim is to introduce innovative solutions to enterprises, and on the other hand, directing the scientific research performed in universities to the market needs. In the elaboration „Program assumptions of the conference Science for Business, Business for Science" [10] of 2013, Andrzej Rabczenko says about necessary changes to be introduced by universities to meet the challenge. Most of the assumptions have already been implemented in the Wroclaw University of Technology just due to activity of the Center and its particular units. Rabczenko underlines the necessity of establishing the Knowledge Center, and within its structure the Knowledge Transfer Office. Units of the Center responsible for cooperation between science and business have already conducted actions defined as necessary. They stimulate research aimed at results commercialization, associate individual laboratories with enterprises and participate in evaluating the commercialization potential of projects [10]. The Center operates two-ways – on the one hand, it is observing results of scientific research and is taking care for their dissemination including finding specific applications, and on the other hand supports relations with interested enterprises by placing orders with the University for the specific solutions [6].

C. Knowledge Repository and Digital Library

A new unit established within the Center is the Knowledge Repository. Wroclaw University of Technology is open towards the idea of Open Access. This is expressed by publishing an open mandate and supporting the Berlin Declaration. As a result, the necessity to create a multi-functional platform to collect and archive university scientific achievements, to promote research done at Wroclaw University of Technology and to support research results transfer between industry appeared. The system developed in the Library currently exists in a beta version. It allows for archiving and searching scientific publications of employees, PhD students and degree students of Wroclaw University of Technology. It will also enable to personalize thematic search and full-text search. Moreover, it will include an expert search, helping to find and contact an author specializing in a given research area and doing particular research projects. Author profiles deposited in the Repository will include a scientific resume, list of publications, citation information, grant and project information. The Repository will be closely integrated with other databases developed at the University with the aim to support research commercialization, developed by the Center for Science and Business Cooperation among others: the laboratories, project database, key equipment database [11].

Within the Center there are laboratories of Lower Silesian Digital Library and the Laboratory of Digitization Methods. Lower Silesian Digital Library was founded in 2004 as The Digital Library of Wroclaw University of Technology. In 2006, on the basis of a contract with Ossolinski National Institute and other higher education institutions from Wroclaw and Lower Silesia, the library was renamed and a Consortium of Lower Silesian Digital Library was created. The Consortium constitutes the common, regional electronic resources [12]. For nearly 10 years, Wroclaw University of Technology Library and the Consortium have been scanning library resources using their own, not always professional, scanners. They also outsourced digitization services from other companies. In Lower Silesian Digital Library laboratory, as well so far have been only 2 scanners. One Minolta PS 7000 scanner is scanning in grayscale and a flat Flex 50i full color scanner is scanning using the so-called cold light. For modern laboratories of the Lower Silesian Digital Library in Bibliotech building, the cutting-edge digitization equipment has been bought. Twelve scanners and posts to take photographs of resources (using 3D technology, among others) will be used to scan library and archive collections of Wroclaw University of Technology and other Consortium Libraries, research project documents, doctoral dissertation, diploma papers and other documents which are created at the University. Digitization laboratories will provide services for other companies and institutions. The variety of digitization equipment will allow for digitization of cultural heritage resources, large scale maps and 3D objects. It will also enable the Library to perform mass digitization of current resources, doctoral dissertations, documents and academic text books, which will be partly accessible through the Lower Silesian Digital Library or the Repository.

D. Laboratories

A group of modern research laboratories has been provided within the structure of the Center:

- Laboratory of Remote Access to Digital Resources – conducts research related to development and implementation of new technologies, among others for safe and fast sharing of library resources, as well as access to digital repositories through computer network (including the wireless one).
- Laboratory of Protection and Safety of Digital Repositories – conducts research related to development and implementation of technologies (methods, equipment, procedures) for providing secure and confidential storing of digital data, as well as widely understood protection of digital repositories.
- Laboratory of Data Bases and Warehouses – is focused on research of data mining technology use,

 ISBN: 978-1-61208-329-2

among others for defining relations between user attributes and type of knowledge sought by him.

- Laboratory of Extraction and Acquisition of Knowledge – is designated for conducting research concerning development and implementation of new technologies for extraction of knowledge from digital resources, determining methods of collecting knowledge and application of knowledge extraction methods in e-learning.
- Laboratory of Exploration and Analysis of Digital Resources – conducting research concerning development and implementation of technologies (methods, procedures, applications) for providing control over plagiarism, as well as creating intelligent systems of searching for information in network and Internet systems.
- Laboratory of e-Learning Technology – aimed at development and implementation of new technologies in e-learning for the needs of university, companies and governmental sector.
- Laboratory of Utilitarian Quality of Information Systems – the Laboratory is designated for testing the quality in use (usability) of information systems. According to standards proposed in the ISO 9241 and ISO 9126 Standards the quality in use is being considered in the aspects of effectiveness and satisfaction. The Laboratory will enable application of tests with users both, in the subjective evaluation mode and recording of the selected physiological parameters during the tests. Innovative and unique in the world equipment for automatic recognition of human emotional behavior during operation of information systems will allow for inclusion of the laboratory to the world research trends in that field.
- Laboratory of Tyfloinformatics – conducting innovative in the country research over modern methods and techniques of making knowledge available to blind and partially sighted persons.
- Research and Development Laboratory for Multimedia – the reference laboratory designated, among others, for self-testing of various tele-information equipments in the functionality and suitability in creating wideband networks and services.
- Laboratory of Service Oriented Systems – the laboratory will develop and implement test methodology of service quality offered within SOA architecture (Service Oriented Architecture), participate in development of Polish standards of software quality and methodology of their use in practice, as well as creating own standards of product certification [4].

The Laboratories are to be run by scientific workers of the University, among others from the Department of Computer Science and Management and Department of Electronics. They will offer services and applications suitable for implementation by various economic entities

IV. CONCLUSION

The Center for Scientific and Technical Information of Wroclaw University of Technology has gone into development phase. Particular departments are gradually moving to the modern building. Tender procedures for the purchase of computer and digitization equipment, as well as research laboratories equipment are in progress. Official opening of the Center is planned in autumn 2014. It will then become the most modern academic and library institution this type in Poland serving collection, management and circulation purposes of knowledge produced at Wroclaw University of Technology.

At the basis of establishment the Center of Knowledge and Technical Information lay the need for cumulating in one place the whole knowledge and information both, acquired and produced at Wroclaw University of Technology. The Bibliotech building, where the key agendas of the Center are located, is to be a place where the scientific workers, doctoral students and students may come seeking for information sources for the conducted scientific research, as well as those looking for opportunities of commercialization of the innovative solutions developed by them. This is also to be a friendly place for entrepreneur, who is seeking in the university the innovative technology, a possibility of cooperation or a specific knowledge or information. So far, the tasks performed by the individual units of the Center were scattered and therefore the quality of that type services was insufficient. Project of the Center for Scientific and Technical Information redefines the role of library in the university of technology. The library of future integrates access to various type resources and provides services tailored to the user needs. Moreover, it is not only the unit disseminating knowledge but also cresting it

REFERENCES

[1] Perspektywy University Ranking, http://www.perspektywy.pl/portal/index.php?option=com_content&view=article&id=724:uczelnie-akademickie&catid=93&Itemid=230 [retrieved: 10 February, 2014].

[2] Biblioteka Politechniki Wrocławskiej 1946-2011 - historia, działalność, organizacja, Henryk Szarski (red.), Jadwiga Wojtczak (red.), Wrocław 2011, http://www.dbc.wroc.pl/publication/14819[retrieved: 15 February, 2014].

[3] H. Szarski, "Komputeryzacja w Bibliotece Głównej i OINT Politechniki Wrocławskiej. Stan obecny i kierunki zmian", Przegląd Biblioteczny, nr 3/4, 1991, pp. 305–315.

[4] Studium Wykonalności projektu Biblioteka Nauk Ścisłych I Technicznych na potrzeby Innowacyjnej Gospodarki, Wroclaw 2010.

[5] Działalność Systemu Biblioteczno-Informacyjnego Politechniki Wrocławskiej w roku akademickim 2012/2013, Raport Seria: U nr 226, unpublished.

[6] Opracowanie modelu współpracy Nauki z Gospodarką w Politechnice Wrocławskiej, Wrocław: Politechnika Wrocławska, 2014, unpublished.

[7] University Business Dialogue: a new partnership for the modernisation of Europe's universities P7_TA(2010)0187, European

 ISBN: 978-1-61208-329-2

Parliament resolution of 20 May 2010 on university-business dialogue: a new partnership for the modernisation of Europe's universities (2009/2099(INI)) 2011/C 161 E/15. http://ec.europa.eu/prelex/detail_dossier_real.cfm?CL=en&DosId=202297 [retrieved: 12 February, 2014].

[8] Communication from the Commision to the European Parliament, the Council, the European Economic and Social Committee and the Committee of Regions Entrepreneurship, 2020 Action Plan, Reigniting the entrepreneurial spirit in Europe COM(2012) 795 final, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2012:0795:FIN:en:PDF [retrieved: 12 February, 2014].

[9] Strategia Rozwoju Województwa Dolnośląskiego 2020, Wrocław 2013.

[10] A. Rabczenko, Założenia programowe konferencji Nauka dla Biznesu, Biznes dla Nauki, 10.12.2013, unpublished.

[11] A. Wałek, Wroclaw University of Technology Knowledge Repository – project objectives, in press.

[12] A. Wałek, Biblioteki cyfrowe na platformie dLibra, Warszawa: Stowarzyszenie Bibliotekarzy Polskich, 2009.

 ISBN: 978-1-61208-329-2

Towards Quality Driven Schema Integration Process Tasks

Peter Bellström
Information Systems
Karlstad University
Karlstad, Sweden
e-mail: Peter.Bellstrom@kau.se

Christian Kop
Institute for Applied Informatics
Alpen-Adria Universität Klagenfurt
Klagenfurt, Austria
e-mail: chris@ifit.uni-klu.ac.at

***Abstract*—If the structure of information within several departments of an organization has to be integrated, the process of integration must meet quality criteria. In this paper, we address quality in the schema integration process, more specifically, quality driven schema integration process tasks. Therefore we searched the literature for the best practices used for conceptual modeling as such and applied these to integration tasks. We address in particular three tasks within the integration process that should improve the quality of the integrated schema when used with best quality practices. Within each best practice we emphasize the use of knowledge repositories to support the process of creating a high quality schema. The three tasks are: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction and introducing inter-schema properties to improve and clarify dependencies.**

Keywords-Information Management; Organizational Information; Schema Integration; Schema Integration Process; Schema Quality

I. INTRODUCTION

Schema integration has a long research tradition. Nevertheless, it is still ongoing and many tasks of the schema integration process are needed at all times because schemata are not built from scratch anymore. There are a great many schemata available on the Web. Furthermore, if enterprises merge, also the schemata (e.g., enterprise and business process models) available in the enterprises must be merged. Last but not least, if enterprises use available Web Services, then it might be good to know the business process model and at least match the business process models and data models to check the compliance of the Web Service models with the respective enterprise models.

A good quality of results in such contexts is very important. Literature on quality mainly focuses on the quality of the product (i.e., the model). The criteria a model must meet in order to have a certain quality are specified. To achieve this quality, the process and the improvement of process tasks must be considered.

The aim of this paper is to provide a description of what can be done in the integration process of static schemata in order to get a good, integrated model. Since an integrated model is a model too, we analyzed the literature with a focus on static modeling, and on the kind of process tasks that lead to a model with better quality. Then we applied the strategies to the tasks that have to be done in the integration process. Particularly, we addressed three tasks within the integration process that when used with best quality practices should improve the quality of the integrated schema. The three tasks are: choosing the right integration strategy, choosing the right conflict resolution methods for the chosen level of abstraction and introducing inter-schema properties to improve and clarify dependencies.

Since the paper covers schema integration, the integration process and the quality of the integrated schema, and the process, this paper is structured as follows. In Section 2 we give an overview on integration approaches and quality of schemata. In Section 3, we describe the integration process. Section 4 focuses on some best practices for improving schema quality. In Section 5, we describe the influence of best practices mentioned in Section 4 on three of the tasks mentioned in Section 3. The paper closes with a summary and a brief outline of our future project.

II. RELATED WORK

A. Integration

There is a long research history on several aspects of integration. A first substantial study of integration was made by Batini and Lenzerini [2] in the mid-80s. In another work by Batini et al. [3], other approaches on integration were summarized. In the following years, other integration approaches focusing on several aspects of the integration problem were published.

Larsen et al. [24] used attribute equivalence as the most basic concept to explain the integration of structural schemata. Savarese et al. [33] presented operators for deciding on the similarity or dissimilarity of schema construct. On the basis of defined assertions, Johannesson [21] proposed a method to detect equivalent schemata and to automatically integrate two schemata. Bhargava and Beyer [4] concentrated on the automatic detection of naming conflicts. Further algorithms for structural schema integration can be found in Geller et al. [19]. García-Solaco et al. [18], integrated semantically enriched database schemata. Dai [12] presented an object oriented framework for the integration of heterogeneous databases. Metais et al. [28] introduced linguistic knowledge for the integration step. For relationships, for instance, verbs can name relationships. Knowledge of the verbs and their linguistic semantic roles support the integration. Ram and Ramesh [31] described a blackboard architecture for schema integration of existing

 ISBN: 978-1-61208-329-2

databases. With this system, knowledge from designers and end users who feed the system is shared. The impact of similarity measures for schema matching and data integration is discussed in Spaccapietra and Parent [35]. Frank and Eder [16] described the integration of state charts object oriented models. The work of Cheng and Wang [11] is based on the formalization of state chart constructs. Stumptner et al. [36] proposed a meta-class framework on which integration should be based. Raut [32] gave an overview of business process integration. Fan et al. [14] proposed OWL-S ontologies as a support for business process integration. Lee et al. [26] described the integration of use cases on the basis of petri net models. Finally, Winter et al. [37] used a behavior tree approach for integrating requirements.

B. Schema Quality

A great deal of work has also been written on the quality of conceptual schemata (models). Although quality is a feature of a product or artifact (e.g., a schema), it is also necessary to think about the quality of the process of generating the product to support the quality of the product.

Batini et al. [1] listed eight schema quality characteristics. Lindland et al. [27] proposed a framework consisting of the three dimensions: "syntax", "semantic" and "pragmatics". The syntax-dimension reflects the vocabulary and grammar (i.e., meta-model) of a schema. The semantic dimension relates the used terms and notions to the domain context. The chosen notions modeled by modeling elements must be legal and relevant in the domain, and they must be relevant and legal to the purpose for which the schema has been built. Finally, the pragmatic dimension is achieved if the audience can understand and follow the schema.

Moody [30] concluded that there is still a need for standards, which are also accepted by the industry.

In Moody and Shanks [29], the authors focused on process quality for the development of data schemata (ER diagrams). Their approach was evaluated in a large Australian bank. In the empirical study, it was also important, that the quality was checked throughout the schema development process. In particular, quality-checking was not only made at the end of a phase but before, during and after the schema development phases. Furthermore it turned out that an information architect, who checks the model with respect to enterprise terms can support quality.

In Cherfi et al. [10] the authors presented a framework of four quality characteristics for the ER modeling language.

Becker et al. [5] described the "Guidelines of Modeling (GoM)". Six principles of modeling are introduced in this framework: correctness, relevance, economic efficiency, clarity, comparability, and systematic design. These principles can be seen as general strategic and objective definitions for modeling. Based on these goals, the concluded modeling process consisted of the following steps: goal definition, construction of an overall navigation and structural framework, modeling as such, and completion and consolidation.

With the **se**miotic **qual**ity framework (SEQUAL), Krogstie [23] explains quality of models with model externalization, goals of modeling, modeling domain, explicit knowledge of social actors, interpretation of the social actors and technical actors as well as with languages extension.

C. Summary of the Literature

We adopted the integration process as described in Batini et al. [3] since this is a well-established process. They divided the integration process into four phases: pre-integration, comparison of the schemata, conforming the schemata and merging and restructuring. In Section 3, we describe this process in more detail.

In Section 4, we continue the description about schema quality according to some selected best practices out of the list of schema quality approaches. We have chosen these approaches since they have been shown in practice to improve schema quality. In Section 5, we will then take specific best practices and combine them with three tasks of the integration process steps described in Section 3.

III. INTEGRATION PROCESS

This section should be viewed as a reference point for the following sections in which we describe and discuss best practices in the schema integration process. The integration process starts with a set of schemata, often referred to as views. These views are integrated in order to evolve the global schema. The schema evolvement takes place in four phases proposed by Batini et al. [3]. The output of one phase is used as the input of the next phase.

A. Pre-Integration

Several tasks should be carried out in this phase. Song [34] mentioned that: translating all schemata to the chosen modeling language, checking for differences and similarities in each schema and selecting the integration strategy are all tasks to be performed in pre-integration. Three additional tasks to perform in pre-integration were proposed in Bellström and Vöhringer [7] as follows: schema element name adoption, schema element disambiguation and introduction of missing relationships.

The output from this phase is a set of revised schemata, the definitions of schema elements and the chosen integration strategy.

B. Comparison of the Schemata

This phase has been researched a great deal and has been called an important [34] and difficult phase [13][25]. Several authors [3][22][34] assigned the following tasks to this phase: recognition of name conflicts, recognition of structural conflicts and recognition of inter-schema properties.

The output from this phase is a description of schema element similarities and a description of differences and a description of inter-schema properties.

C. Conforming the Schemata

Also conforming the schemata has received some attention by other researches. For instance, Lee and Ling

 ISBN: 978-1-61208-329-2

[25] called it the most critical phase and Spaccapietra and Parent [35] the key issue in schema integration.

In conforming the schemata, the recognized similarities and differences are resolved by adjusting the input schemata.

The recognized inter-schema properties are also used in this phase. However, its full value is shown in merging and restructuring.

The output of this phase is a set of revised schemata.

D. Merging and Restructuring

The first task performed in this phase is to merge the revised input schemata into one global intermediate schema. The intermediate schema is then restructured e.g., detected inter-schema properties are introduced to semantically enrich the schema. Furthermore, schema elements that are truly redundant are recognized and removed from the schema. Merging the schemata as well as restructuring the schemata results in a new intermediate schema.

Before the integrated schema is handed over to the developers implementing the information system, the schema is again analyzed, meaning that the schema is checked and verified according to several quality criteria [1][3] and/or quality factors [29].

The result of this phase should be a high quality schema that can be passed on to the following phases in which the information system is implemented.

IV. SOME BEST PRACTICES REGARDING SCHEMA QUALITY

Both the Guidelines of Modeling (GoM) [5] and the quality factors explained in Moody and Shanks [29] focus on: improving quality of the modeling process and quality of the resulting product (i.e., the conceptual model).

Both frameworks are a good basis for understanding the quality of the conceptual modeling integration process. The Guidelines of Modeling are a more strategic framework for covering all aspects of enterprise models (e.g., data, organization, processes, and behavior). The work in Moody and Shanks [29] focuses on data schemata more specifically ER data models.

Because of its more operational focus, we adopted the following practices from Moody and Shanks [29] for the integration process in order to fulfill the quality factors and improve the quality of the modeling:

- Introducing a specific kind of stakeholder – the information architect
- Introducing continuous quality checks and reviews.

As well as the general practices:

- Stakeholder participation
- Introducing naming conventions, standards, etc.

We will adopt these practices for the integration process as well.

The information architect (in [29] called data administrator) is a person that was introduced to review a schema with respect to the other data schemata (models) existing in the enterprise.

According to Moody and Shanks [29], who proposed continuous checks and reviews for schema development, reviews must not only be made at the end, but also before and during a development step. Such reviews should support the total quality management aim that the quality checks and reviews should not detect errors, but prevent errors.

The participation of different kind of stakeholders is a successful technique used in Information Systems and Enterprise Engineering. Since the schemata (models) represent the knowledge of ideas of people with different backgrounds, it is necessary that different stakeholders are involved.

The introduction of an information architect also implies the usage and management of standards (e.g., what a schema should look like syntactically, which terms are used and preferred to other terms, etc.).

V. APPLYING BEST PRACTICES TO INTEGRATION TASKS

In general the best practice of "continuous improvement" is a driver for the whole integration process. Although quality is usually considered in or even after the last step of schema integration, we will follow the principle of introducing quality as early as possible here. Therefore we will focus on tasks needed in earlier steps. We will relate them to the best practices in order to improve them. These tasks are: *choosing the right integration strategy*, *choosing the right conflict resolution methods for the chosen level of abstraction* and *introducing inter-model properties to improve and clarify dependencies*. The first is a task that has to be done during pre-integration. The second and the third tasks are at least executed during the 2nd and 3rd steps.

A. Choosing the Right Integration Strategy

In Batini et al. [3], several strategies are proposed for integrating end-user schemata (views). They distinguish between binary and n-ary integration strategies. Among binary strategies a ladder strategy [2] or a balance strategy [3] can be chosen. In the ladder strategy, the stakeholders start with two views. They integrate these two views. Afterwards the first integrated schema is compared and matched with another view, and so on. In the balanced strategy, two views are integrated in an intermediate schema. This intermediate schema is integrated with other intermediate schemata until the global schema is reached. The n-ary strategies are the one-shot strategy (a global schema is generated at once from all views) and the iterative strategy. The iterative strategy uses one shot strategies only to produce intermediate schemata. These schemata are then integrated with each other (two or more). Integrated schema can also be integrated with views. The iterative strategy can be seen as a mixture of the previous three strategies.

1) Continuous Checks and Reviews

For continuous checks and reviews, the integration strategy must prove enough definite points of inspections.

A one shot strategy can be excluded as a good strategy by applying this best practice. Otherwise, it would mean that a global schema exists without any intermediate results. If intermediate results are missing, then it is impossible to identify definite review milestones. Following the best practice of continued improvement given in literature, an iterative, and balanced or ladder strategy should be applied.

 ISBN: 978-1-61208-329-2

Doing so each time, an intermediate schema is generated, this intermediate schema can be reviewed.

It cannot be determined which of the other three strategies should be chosen since all these strategies have intermediate points where schemata can be reviewed before or during integration. The choice between a balanced, a ladder, or an iterative strategy, is a pragmatic decision of available time for the integration and other environmental factors.

2) Information architect, stakeholder participation and standards

Since integration is part of modeling, an information architect, stakeholder involvement, and standards are also necessary for integration.

The information architect has to assure that a certain intermediate schema as well as the views already integrated is in compliance with existing schemata in the enterprise. Stakeholders check the semantic correctness and completeness with respect to a certain examined section represented by the views (schemata) or intermediate schemata. For both the information architect and stakeholder involvement, strategies that have more intermediate points for discussions and reviews (i.e., ladder, balanced, iterative strategy) are more supportive.

Standards help to check if the schema is syntactically correct and if terms are used in compliance with the enterprise. It is therefore necessary that standards are used. Standards equally drive all the four strategies (one shot, ladder, balanced and iterative). Knowledge repositories, such as stemmers and lemmatizers, could be used to facilitate the task of checking that terms are used in a correct way. Drawing tools might also aid in the modeling process and be used to check that the schema is syntactically correct.

B. Choosing the Right Conflict Resolution Methods for the Chosen Level of Abstraction

In the phase comparison of the schemata two schemata are compared for the purpose of finding similarities as well as differences, often more generally referred to as conflicts. In the phase that follows, conforming the schemata, the conflicts are resolved. However, the same resolution methods are often proposed (and used) for implementation-neutral schemata and implementation-dependent schemata. Using different conflict resolution methods for different levels of abstraction is very important since an implementation-neutral schema is often used in the earlier phases of information systems development while an implementation-dependent schema in the later phases is close to programming and technical issues.

The purpose of the schema under design may also vary. Boman et al. [9] address this in their four schema purposes as follows: "A schema can serve at least four different purposes. First, it can be used for clarifying the language used in an organisation. Secondly, it can be used for making explicit the rules that prevail in an organisation, which helps to criticise them and possibly to draw up new rules. Thirdly, a schema can be useful for reviewing existing information systems. Fourthly, a schema can be used for developing a new information system" (p. 122).

One way of combining the mentioned two levels of abstraction with the four purposes stressed by Boman et al. [9] might be as follows. First, clarifying the language is closely related to the implementation-neutral level since then the designers are interested in concepts and connections between concepts rather the implementation-dependent issues and trying to reduce the number of concepts and connections [6]. Secondly, making explicit the rules is also closely connected to the implementation-neutral level since rules must be expressed so that all stakeholders understand the rules and therefore also can criticize them. Thirdly, using a schema for reviewing an already existing information system is closely related to the implementation-dependent level since the schema describes an already implemented information system. Finally, using a schema during the development of a new information system refers to both levels of abstraction. This is motivated since the designers might use different schemata during the development of the information system. The designers might also use different modeling languages dependent on phase and focus in the information systems development process. If choosing the right conflict resolution methods for the chosen level of abstraction are ignored the integrated schema might not only suffer semantic loss but also being hard to understand.

1) Continuous Checks and Reviews

Having designed the schemata on the chosen level of abstraction and in comparison of the schemata recognized the conflicts between two schemata, it is important that in conforming the schemata the right conflict resolution methods are used. However, this is not always the case. Therefore, while applying the best practice of continued checks and reviews, it is important to check that the right conflict resolution methods have been chosen for the current level of abstraction. If the wrong conflict resolution method has been introduced, it should not only be recognized during continuous checks and reviews but also changed to the right one. This should in the end contribute to an integrated schema with high quality since an additional check and review has been conducted. For instance, if during the comparison of the schemata, we recognize a synonym conflict (e.g., article in schema 1 and product in schema 2), it should during conforming the schemata be resolved. However, if the schemata are designed on an implementation- neutral level it is important that all concept names and dependencies are kept as long as possible since they might be of importance for one or several stakeholders. We should therefore not rename of one or both concept names, which is one of the most ordinary proposed resolution methods for a synonym conflict, but instead introduce a resolution method that keeps both concept names. One way to fulfill this could be to introduce mutual inheritance dependency described as A and B are synonyms if and only if A inherits B and B inherits A [20].

2) Information architect, stakeholder participation and standards

While doing schema integration, it is important that both the information architect as well as the stakeholders are very much involved. By involving these actors several of the mentioned pitfalls should be recognized and addressed as

early as possible in the integration process (the current iteration cycle) and not included into the global integrated schema. This is the case since it is the stakeholder and the information architect that possess the knowledge of how their concepts should be named and which concepts should be connected to each other. However, the information architect also has to take into account already existing data schemata within the enterprise and therefore should have a holistic perspective. A stakeholder might instead focus on integrating a schema of a specific department.

Naming conventions, standards and ontologies, so called knowledge repositories, might also exist in the enterprise that need to be taken into account in the integration process. However, it is important that these naming conventions and standards do not restrict the naming of concepts which impoverish the language used in the schema but instead are used as a tool to facilitate the integration process. Therefore standards should not enforce the usage of one concept name but instead give guidelines on how concepts names should be used such as name concepts in singular.

C. Introducing Inter-Schema Properties to Improve and Clarify Dependencies

Another task in comparison of the schema is the recognition of inter-schema properties. An inter-schema property is not really a conflict, but instead it describes a specific link between two concepts. The two most common inter-schema properties described in the literature are hypernym-hyponym dependencies (often referred to as "is-a") and holonym-meronym dependencies (often referred to as "part-of"). When an inter-schema property has been recognized it is documented and passed to the next phase in the schema integration process in which it is used. However, its full value is shown in the last phase of the schema integration process where the inter-schema properties are used as guidance while merging and restructuring the global integrated schema. Introducing inter-schema properties in the schema integration process is of great importance since an inter-schema property has a clear meaning and should therefore also be used not only to clarify and improve a specific meaning between two concepts but also to reduce the number of concepts in the integrated schema if possible. However, reducing the number of concepts should be done carefully. Deleting a concept might reduce the quality of the integrated schema instead of improving its quality. In the worst case, it violates the completeness quality factor addressed in [29].

Finally, it should be noted that a holonym-meronym dependency might be of two types: aggregation and composition in which composition is the stronger.

1) Continuous Checks and Reviews

In the comparison of the schemata, the binary strategy (or n-ary iterative) is used while recognizing similarities and differences, e.g., inter-schema properties, between two schemata. When an inter-schema property has been recognized, it should be documented and passed on to the following phases in the integration process. At the end, the inter-schema property should not only in merging and restructuring be treated as a source of semantic improvement but also be used as guidance, a knowledge repository, while merging and restructuring the integrated schema.

However, since an inter-schema property is used in at least two phases in the integration process, it is substantially important that the inter-schema property is used in a right way and not instead polluting the input schemata and/or the integrated schema. An even worse scenario could be that the inter-schema property is used in a wrong way causing semantic errors. Applying the best practices of continuous checks and reviews is therefore of great importance to improving not only the quality of the integrated schema as such but also to verifying that the inter-schema property is used in a correct way.

For instance, if we in comparing the schemata have recognized not only a hypernym-hyponym dependency between concept A and B in schema 1 but also a hypernym-hyponym dependency between concept B and A in schema 2, problems might later on be introduced into the integrated schema. The inter-schema dependencies are documented and passed on to the following phase in which the schemata are adjusted to solve the recognized conflicts and inter-schema properties. Having done that, the schemata (and some extra information resources) are passed to the last phase in which the schemata are integrated. In the worst case, both hypernym-hyponym dependencies described above are introduced to the integrated schema causing what is sometimes called reverse subset relationship [1] or cyclic generalization [34]. However, applying the best practice of continuous checks and reviews, this problem should be recognized and resolved in the current iteration cycle and not be left to later iterations in the integration process.

2) Information architect, stakeholder participation and standards

Introducing inter-schema properties should result in a semantic richer schema since the inter-schema properties should have a clear meaning compared with, for instance, the association dependency with or without specified cardinality. However, introducing new schema constituents might also result in new problems and errors. Involving information architect as well as stakeholders are also of great importance, since these actors possess the knowledge of their specific domain. However, the information architect has to take into account the schemata already existing within the enterprise and make sure that these match the new schema being developed. On the other hand, a stakeholder from one department might instead only focus on his/her part of the schema (model) and therefore argue for his/her point of view in the integration process.

Finally, naming conventions, standards as well as ontologies, so called knowledge repositories, might also exist within the enterprise. Ontology, or even domain ontology, might for instance be useful when deciding how to resolve the cyclic generalization dependency. This is the case since a description on how concept A and concept B are dependent might be stated in ontology.

 ISBN: 978-1-61208-329-2

VI. CONCLUSION AND FUTURE WORK

In this paper, we have addressed schema quality within the schema integration process. In doing so, we have focused on four best practices of quality improvement given in the literature and three specific integration tasks that should increase the quality of the schema being designed. The four best practices addressed are: *continuous checks and reviews*, *information architect*, *stakeholder participation* and *standards*. The three integration tasks addressed are: *choosing the right integration strategy*, *choosing the right conflict resolution methods for the chosen level of abstraction* and *introducing inter-schema properties to improve and clarify dependencies*. Within each integration task we have also addressed how knowledge repositories might be used to aid in the process of producing a high quality schema.

To conclude (see also Table I), the four best practices used for conceptual modeling if addressed in connection to schema integration can improve the three mentioned tasks and hence the integration process. Continuous checks and reviews, information architect and stake holder participation can be drivers for choosing the right integration strategy. Standards do not have an influence on this task. Continuous checks and reviews, standards, information architect and stakeholder participation are essential in the conflict resolution task. The more conflicts are checked and resolved the better. The more the stakeholders and the information architect are involved, the more conflicts can be resolved. Standards support this task as long as they do not restrict the enterprise specific naming of concepts.

For the inter-schema property introduction, which is used in at least two phases of the integration process, continuous checks and reviews can help verify that the inter-schema property is used in the correct way. Stakeholders and the information architect are the ones who possess the domain knowledge and can thus support the aim to get a semantically richer schema with clear meanings. Standards and ontologies are useful to support the detection of inter-schema properties.

In the long run these improved tasks contribute to a high quality integrated schema.

In future, we will continue our work on identifying particular best practices for quality improvement for other tasks of the integration process. Specifically, we will look at other tasks of the phases (e.g., recognition of name conflicts and structural conflicts, merging the revised schema). We will also investigate the process from the perspective of aspects of quality (e.g., the SEQUAL views – physical, empirical, syntactical, semantic quality).

TABLE I. BEST PRACTICES AND INTEGRATION TASKS

Best Practice	Choosing the Right Integration Strategy	Choosing the Right Conflict Resolution Methods for the Chosen Level of Abstraction	Introducing Inter-Schema Properties to Improve and Clarify Dependencies
Continuous Checks and Reviews	are facilitated by the ladder, balanced and iterative integration strategy.	are the enablers to verify that the schemata illustrate the chosen level of abstraction during the whole integration process.	are the enablers to verify that the inter-schema properties are used in a correct way during the whole integration process.
Information Architect	checks that the schemata are in compliance with existing enterprise schemata.	checks that the choosen conflict resolution methods are in compliance with existing enterprise schemata.	checks that the introduced inter-schema properties are in compliance with existing enterprise schemata.
Stakeholder Participation	is the enabler to check the semantic correctness and completeness of the schemata.	is the enabler to check that choosen conflict resolution methods are semantically correct and that the schema is complete.	is the enabler to check that the introduced inter-schema properties are semantically correct and that the schema is complete.
Standards	help in the process of checking that the schemata are syntactically correct and that terms are used in compliance with the enterprise schemata	help in the process of introducing the correct resolution method for not only naming conflicts but also structural conflicts.	help in the process of introducing the correct inter-schema propterty and help in the process of introducing the inter-schema property in a correct way.

REFERENCES

[1] C. Batini, S. Ceri, and S. B. Navathe, Conceptual Database Design an Entity Relationship Approach. Redwood City: Benjamin/Cummings Publishing Company, 1992.

[2] C. Batini and M. Lenzerini, "A Methodology for Data Schema Integration in the Entity-Relationship Model," IEEE Transactions on Software Engineering, vol. 10 (6), 1984, pp. 650-664.

[3] C. Batini, M. Lenzerini, and S. B. Navathe, "A Compartive Analysis of Methodologies for Database Schema Integration," ACM Computing Surveys, 1986, vol. 18 (4), pp. 323-364.

[4] H. K. Bhargava and R. M. Beyer, "Automated Detection of Naming Conflicts in Schema Integration: Experiments with Quiddities," Proceedings of the 25th Hawaii International Conference on System Sciences, IEEE Press, 1992, pp. 300-310.

[5] J. Becker, M. Rosemann, C. von Uthman, "Guidelines of Business Process Modeling," W. van der Aalst et al. (Eds.):

Business Process Management, LNCS 1806, Springer-Verlag Berlin Heidelberg, 2000, pp. 30-49.

[6] P. Bellström, "On the Problem of Semantic Loss in View Integration," in Information Systems Development Challenges in Practice, Theory, and Education. Volume 2, C. Barry, K. Conboy, M. Lang, G. Wojtkowski, and W. Wojtkowski, Eds. New York: Springer, 2009, pp. 963-974.

[7] P. Bellström and J. Vöhringer, "A Semi-Automatic Method for Matching Schema Elements in the Integration of Structural Pre-Design Schemata," International Journal on Advances in Intelligent Systems. vol. 4 (3 & 4), 2011, pp. 410-422.

[8] P. Bellström and J. Vöhringer, "Towards the Automation of Modeling Language Independent Schema Integration," International Conference on Information, Process, and Knowledge Management (eKNOW 2009), IEEE Press, 2009, pp. 110-115.

[9] M. Boman, Jr. J. A. Bubenko, P. Johannesson, and B. Wangler, Conceptual Modelling. London: Prentice Hall, 1997.

[10] S. S. Cherfi, J. Akoka, and I. Comyn-Wattiau, "Perceived vs. Measured Quality of Conceptual Schemas: An Experimental Comparision," Proceedings of Tutorials, Posters, Panels and Industrial Contributionof the Twenty-Sixth International Conference on Conceptual Modeling (ER 2007), vol. 83, 2007, pp. 185-190.

[11] B. H. C. Cheng and E. Y. Wang, "Formalizing and Integrating the Dynamic Model for Object Oriented Modeling," IEEE Transactions on Software Engineering, vol. 28 (8), 2002, pp. 747-762.

[12] H. Dai, "An Object-Oriented Approach to Schema Integration and Data Mining in Multiple Databases," Proceedings on the Technology of Object-Oriented Languages (TOOLS), IEEE Press, 1997, pp. 294-303.

[13] L. Ekenberg and P. Johannesson, "A Formal Basis for Dynamic Schema Integration," Conceptual Modeling – ER'96. LNCS 1157, 1996, pp. 211-226.

[14] S. Fan, L. Zhang, and Z. Sung, Z., "An Ontology Based Method for Business Process Integration," International Conference on Interoperability for Enterprise Software and Applications in China, IEEE Press, 2008, pp. 135-139

[15] G. Fliedl, C. Kop, H. C. Mayr, W. Mayerthaler, and C. Winkler, "Linguistically based requirements engineering – The NIBA project," Data & Knowledge Engineering, vol. 35 (2), 2000, pp. 111-120.

[16] H. Frank and J. Eder, "Towards an Automatic Integration of Statecharts," International Conference on Conceptual Modeling (ER 1999), 1999, pp. 430-444.

[17] A. Gal, "Interpreting Similarity Measures: Bridging the Gap between Schema Matching and Data Integration," Data Engineering Workshop of ICDEW 2008, IEEE Press, 2008, pp. 278-285.

[18] M. García-Solaco, F. Saltor, and M. Castellanos, "A Structure Based Schema Integration Methodology," Proceedings of the Eleventh International Conference on Data Engineering, IEEE Press, 1995, pp. 505-512.

[19] J. Geller, A. Mehta, Y. Perl, E. Neuhold, and A. Sheth, "Algorithms for Structural Schema Integration," Proceedings of the Second International Conference on Systems Integration (ICSI'92), IEEE Press, 1992, pp. 604-614.

[20] R. Gustas, Semantic and Pragmatic Dependencies of Information Systems. Kaunas: Kaunas Technologija, 1997.

[21] P. Johannesson, "A Logical Basis for Schema Integration," Third International Workshop on Research Issues on Data Engineering (RIDE-IMS'93), IEEE Press, 1993, pp. 86-95.

[22] P. Johannesson, Schema Integration, Schema Translation, and Interoperability in Federated Information Systems. Dissertation. Stockholm University & Royal Institute of Technology No. 93-010-DSV, 1993.

[23] J. Krogstie, Model based Development and Evolution of Information Systems – A Quality Approach. London: Springer, 2012.

[24] J. A. Larson, S. B. Navathe, and R. Elmasri, "A Theory of Attribute Equivalence in Databases with Application to Schema Integration," Transactions on Software Engineering, vol. 15 (4), 1989, pp. 449-463.

[25] M. L. Lee and T. W. Ling, "A Methodology for Structural Conflict Resolution in the Integration of Entity-Relationship Schemas," Knowledge and Information Systems, vol. 5 (2), 2003, pp. 225-247.

[26] W. J. Lee, S. D. Cha, and Y. R. Kwon, "Integration and Analysis of Use Cases Using Modular Petri Nets in Requirements Engineering," IEEE Transaction of Software Engineering, vol. 24 (12), 1998, 1115-1130.

[27] O. L. Lindland, G. Sindre, and A. Solvberg, "Understanding Quality in Conceptual Modeling," IEEE Software, vol. 11 (2), 1994, pp. 42-49.

[28] E. Métais, Z. Kedad, I. Comyn-Wattiau, and M. Bouzeghoub, "Using Linguistic Knowledge in View Integration: Toward a Third Generation of Tools," Data & Knowledge Engineering, vol. 23 (1), 1997, pp. 59-78.

[29] D. L. Moody and G. G. Shanks, "Improving the Quality of Data Models: Empirical Validation of a Quality Management Framework," Information Systems Journal, vol. 28 (2), 2003, pp. 619-650.

[30] D. L. Moody, "Theoretical and Practical Issues in Evaluating the Quality of Conceptual Models: Current State and Future Directions," Data & Knowledge Engineering, vol. 55 (3), 2005, pp. 243-276.

[31] S. Ram and V. Ramesh, "A Blackboard-Based Cooperative System for Schema Integration," IEEE Expert, 1995, vol. 10 (3), pp. 56-62.

[32] A. Raut, "Enterprise Business Process Integration," Conference on Convergent Technologies for Asia-Pacific Region, IEEE Press, 2003, pp. 1549-1553.

[33] A. Savasere, A. Sheth, and S. Gala, "On Applying Classification to Schema Integration," Proceedings of the First International Workshop on Interoperability in Multidatabase Systems (IMS'91), IEEE Press, 1991, pp. 258-261.

[34] W. Song, Schema Integration – Principles, Methods, and Applications. Dissertation. Stockholm: Stockholm University & The Royal Institute of Technology No. 95-019, 1995.

[35] S. Spaccapietra and C. Parent, "View Integration: a Step Forward in Solving Structural Conflicts," IEEE Transactions on Knowledge and Data Engineering, vol. 6 (2), 1994, pp. 258-274.

[36] M. Stumptner, M. Schrefl, and G. Grossmann, "On the Road to Behavior-Based Integration," Proceedings of the 1st APCCM Conference, 2004, pp. 15-22.

[37] K. Winter, I. J. Hayes, and R. Colvin, "Integrating Requirements: The Behavior Tree Philosophy," 8th IEEE International Conference on Conference on Software Engineering and Formal Methods (SEFM), IEEE Press, 2010, pp. 41-50.

 ISBN: 978-1-61208-329-2

Uncovering File Relationships using Association Mining and Topic Modeling

Namita Dave, Delmar Davis, Karen Potts, Hazeline U. Asuncion
Computing and Software Systems
University of Washington, Bothell
Bothell, WA, USA
{namitad, davisdb1, pottsk2, hazeline}@u.washington.edu

***Abstract*—Software maintenance tasks, such as feature enhancements and bug fixes, require familiarity with the entire software system. A modification task could become very time consuming if there is no prior knowledge of the system. Association mining has been used to identify the files that frequently change together in a software repository, and this information can aid a software engineer locate relevant files for a maintenance task. However, association mining techniques are limited to the amount of project history stored in a software repository. We address this difficulty by using a technique that combines association mining with topic modeling, referred to as Frequent Pattern Growth with Latent Dirichlet Allocation (FP-LDA). Topic modeling aims to uncover file relationships by learning semantic topics from source files. We validated our technique via case studies on two open source projects. Our results indicate that topic modeling can increase the effectiveness of association mining in uncovering the file relationships.**

Keywords-Association mining; Topic Modeling; Software Engineering

I. INTRODUCTION

Software maintenance is the largest cost contributor in the lifecycle of a product [1]. This may be due to an engineer's unfamiliarity with the software to modify, requiring more time to understand the source code [2]. Maintenance tasks also become more difficult as the complexity of the code increases and as code degradation occurs over time due to patches and workarounds [3].

Techniques to find related source code files include static and dynamic analysis, recommendation systems, and code search techniques. Static analysis techniques, more specifically, dependency analysis, provide file relationships based on call graphs [4]. Dynamic analysis tools, meanwhile, are able to identify relationships between files based on execution traces [5]. These techniques, however, are generally language-specific. Recommendation systems, meanwhile, provide possible files of interest based on a developer's past activities, textual similarity, check-in records, or email records [6, 7]. These systems generally use information retrieval techniques along with user context to provide files of interest. Code search techniques find related code based on syntactic or structural matches [8].

Association mining (AM) is another technique to find related files. AM uncovers relationship between files based on files that are modified together in the past. This technique generates rules, which specify which files are frequently changed together. Unlike the other techniques, AM is not specific to the programming language used or restricted to syntactic or structural matches of a query.

Most commonly used algorithms for association mining are Apriori [9] and Frequent Pattern Growth (FP-Growth) [10]. While these algorithms provide some level of accuracy, they are highly dependent on the project history. If there are not enough modifications in the software project, or if the modifications are sparse throughout the software system, there are fewer chances that AM will result in correct rules.

Meanwhile, machine learning techniques such as Latent Dirichlet Allocation (LDA), allows us to detect relationships between files based on semantic similarity. LDA is an unsupervised statistical approach for learning semantic topics from a set of documents [11]. It is a fully automated approach that does not require training labels. It only requires a set of documents and number of topics to learn.

Thus, we aim to address the challenges of AM by combining AM with LDA. Our technique, Frequent Pattern Growth with Latent Dirichlet Allocation (FP-LDA), allows us to achieve better recall results than solely using AM. By combining these two techniques, we are able to overcome the limitations of each technique. LDA allows us to find file associations even with limited modification history. AM, meanwhile, allows us to find associations among files where semantic similarities may not be readily apparent.

The contributions of this research paper are as follows: (1) combination of AM and topic modeling to find file relationships in a software project, and (2) case studies on two open source projects. We also created a set of tools that automates the entire process—from pre-processing the data to querying related files.

The rest of the paper is organized as follows. The next two sections provide background on our combined approach. Section 2 covers AM and Section 3 covers topic modeling. In Section 4, we present our combined approach, FP-LDA. We then validate our approach in Section 5. Related work is discussed in Section 6. We conclude with future work.

II. ASSOCIATION MINING

A. Background

Association rule mining is a method to discover patterns in large data sets. Initially, it was used in Market Basket Analysis to find how items bought by customers are related [12]. Rules are mined from the dataset, such as "Customers who bought item A also bought item B". In the case of software projects, rules such as "Developers who modified

 ISBN: 978-1-61208-329-2

file A also modified file B" are mined [13]. In order to mine these rules, patterns must be analyzed in the dataset.

We now discuss the main concepts of AM [9], as applied to software development.

Let $I = \{i_1, i_2, \ldots\ldots i_m\}$ (Eq.1) represent total set of items. In this paper, the files in the repository are items. T represents a set of transactions $T = \{t_1, t_2, \ldots\ldots t_n\}$ (Eq. 2), which are in the software repository being mined. Each transaction t is a set of items such that $t \subseteq I$. In this paper, t represents one atomic commit.

Given the set of transactions T (see Eq.2), the goal of AM is to find all the association rules that have support and confidence greater than the user specified threshold values. An itemset is a collection of items. The support is defined as the fraction of transactions that contain the itemset and from which the rule is derived. The confidence denotes the strength of a rule. An association rule is represented as

$X \rightarrow Y$ [support = 20%, confidence = 80%] (Eq.3)

In this notation, itemset X is called the antecedent and itemset Y is called consequent such that $X, Y \subseteq I$. Both antecedent and consequent comprise of one or more items. Assume that both X and Y consist of one file each namely x and y respectively. Then, this rule says that in 20% of the check-in transactions, both x and y files are modified and the transactions which changed file x also changed file y 80% of the time.

The threshold support value specified by the user is called minimum support. This is an important element that makes AM practical. It reduces the search space by limiting the number of rules generated. The threshold confidence value specified by the user is the minimum confidence [14].

B. Selection of Association Mining Technique

Two commonly used association-mining techniques are Apriori algorithm [9] and Frequent Pattern Growth Algorithm, or FP-Growth [10].

Apriori is a classic algorithm for learning association rules over transactional databases [9]. The essential idea behind Apriori algorithm is that it iteratively generates candidate itemsets of length (k + 1) from frequent itemsets of length k and then tests their corresponding frequency in the database. Apriori is not efficient when used with large data sets as generation of candidate item sets and support counting is very expensive, as confirmed in [15].

FP-Growth is a faster and scalable approach to mine a complete set of frequent patterns by pattern fragment growth using a compact prefix tree structure for storing transaction dataset [10]. In the first step, it creates a compact Frequent Pattern tree to encode the database. The construction of an FP-tree begins with pre-processing the input data with an initial scan of the database to count support for single items. The single items that do not meet the threshold support values are eliminated. The database is then scanned for the second time to produce an initial FP-tree. The second step runs a depth first recursive procedure to mine the FP-tree for frequent itemsets with increasing cardinality. The FP tree stores a single item at each node. The root node of an FP tree is empty. The path from root to a node in the FP tree is a subset of transactions database. The items in the path are in decreasing order of support. In the second step, the algorithm examines conditional-pattern base for each itemset starting with length 1 and then constructs its own conditional FP-tree. Unlike Apriori algorithm, it avoids generating expensive candidate itemsets. Each conditional FP-tree is recursively mined to generate frequent itemsets. The algorithm uses divide and conquer approach to decompose mining task into smaller tasks of mining the confined conditional databases. Interested readers can refer to [10] for more information.

C. Limitations

AM is useful in finding patterns in the data that satisfy minimum support and minimum confidence constraints. However, some researchers have shown that AM often results in redundant and unimportant rules. A drawback is that it is difficult to eliminate insignificant rules [16].

In this research, the number of association rules generated depends on the amount of modification history of a project. Also, there is a possibility that not all modules or files may be changed during a software maintenance phase. This can affect the number of rules generated.

III. TOPIC MODELING

A. Background

LDA is an unsupervised statistical approach for learning semantic topics from a set of documents [11]. Since it is an unsupervised machine learning technique, no training labels are necessary. This is a fully automated approach that only requires a set of documents and the number of topics to learn. Here are some concepts used in LDA:

- A word is a basic unit of discrete data.
- A document is characterized by a vector of word counts.
- A corpus has a total of W words in its vocabulary.
- D documents placed side by side, gives W x D matrix of counts.
- A topic is a probability distribution over W words.
- Each document is associated with a probability distribution over T topics.

LDA is a generative Bayesian topic model for a corpus of documents. The basic concept behind LDA is that it discovers topics from a collection of documents [11]. Then it learns a distribution over words for each topic. To obtain a semantic interpretation of a topic, we simply examine the highest-probability words in that topic. For example, if a topic has high probability words "window", "dialog", "height", "width", "button", we can infer the topic to be related to the user interface of the software. Lastly, it defines each document as a probabilistic mixture of these topics. Each document can belong to multiple topics. Additional details regarding LDA's generative process are in [11].

As we discuss in the next section, we use LDA to determine possible relationships between source code files through their topic distributions. Each source code file equates to a document in LDA.

 ISBN: 978-1-61208-329-2

B. Selection of Topic Modeling Technique

Topic modeling algorithms generally fall under two categories: sampling-based and variational methods [11]. Sampling-based algorithms collect samples to approximate the posterior with an empirical distribution. Variational methods, meanwhile, use a parameterized family of distributions and then find the member of the family that is closest to the posterior. In this paper, we use a fast version of Collapsed Variational Inference (CVB0) for LDA [17], which has been shown to be among the fastest and most accurate methods for learning topic models.

C. Limitations

LDA has generally been applied to unstructured text [18]. Meanwhile, source code is a highly structured text that has a limited range of semantic concepts. The results are also subject to parameters used in LDA. As a result, researchers have examined ways to fine-tune the parameters [19].

We processed the source code prior to running LDA such that reserved words are removed and only semantically meaningful words are used. Our pre-processing technique is similar to the pre-processing technique described here [20].

IV. COMBINED APPROACH

Our technique, FP-LDA, aims to uncover possible file relationships regardless of the amount of project history available and regardless of the programming language used. (Some knowledge of the language used in the source code is needed to eliminate language-reserved words from the source code. See the next section for more details.) FP-LDA consists of the following steps: (1) data extraction and pre-processing, (2) association data mining, (3) topic modeling, and (4) result querying. Fig. 1 shows a high level process of our technique. Each layer in the figure corresponds to each of these steps. All the processes represented by a rectangle have been implemented.

A. Data Extraction and Pre-processing

Pre-processing version history for data mining. This first step involves extracting the version history of an open source project and preparing the data to be fed as input to the mining algorithm (Step A2). We created a tool that accesses the version history of the project, processes it, and stores the history data in MySQL database. For projects using Subversion (SVN), we used SVNKit application programming interfaces (APIs) [21] to access the version history of the project. SVNKit is an open source Java-based SVN library. For projects using Git, we used JavaGit [22] API to access the version history.

Data pre-processing is an important step in that it removes all unwanted data that may impact data mining (A2). In our technique, our goal is to create a generic pre-processing step to support different open source projects. Thus, we used the following conditions when determining the type of transactions to include in our AM. Similar to [13], we do not include transactions with more than one hundred files since these transactions may contribute to noise. Such commits may be due to specialized tasks, such as formatting all source code files and then checking-in all files together. We also removed transactions that do not assist in identifying relationships between files, such as single file commits, non-source code commits (e.g., graphic files), and commits of deleted files. The remaining valid transactions are then stored in a database (A3). This is the dataset that will be analyzed by the mining algorithm. We then transform this dataset into a file format that conforms to expected format of the mining algorithm (A4).

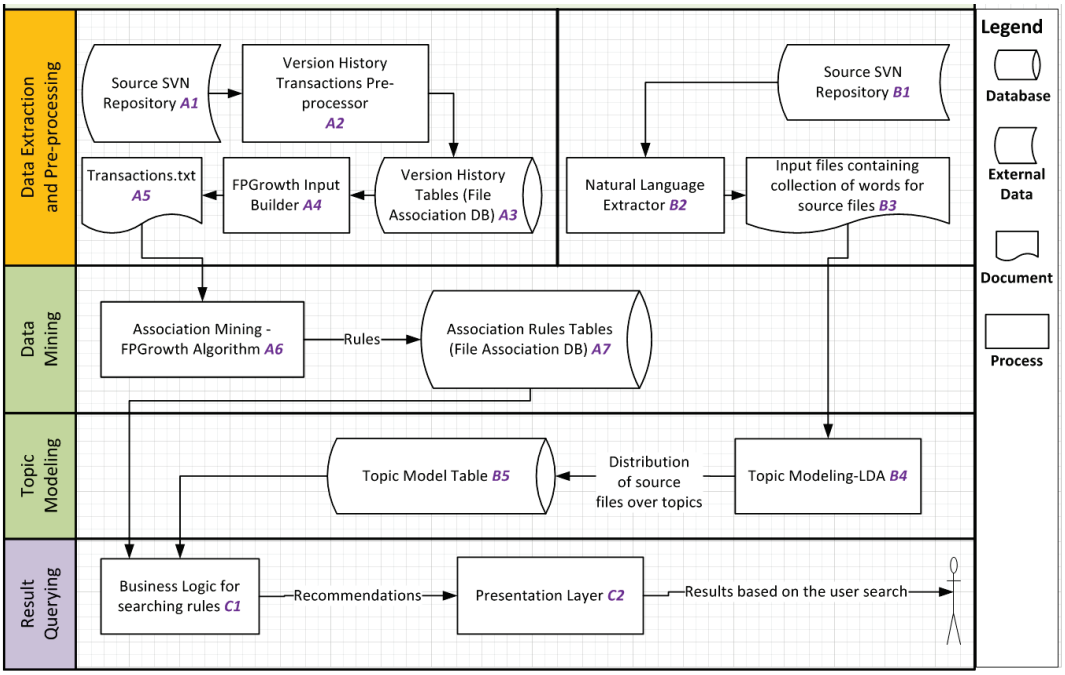


Figure 1. FP-LDA data flow to find file dependencies

Pre-processing source files for LDA. While the mining algorithm examines the entire commit history, we use topic modeling to extract topics from the latest version of the source code. We extracted from the source code semantically meaningful text, such as comments, identifier names and string literals. These words provide clues on the purpose or functionality of the code (B2).

To extract these words, we run each file through a tokenizer. The tokenizer aids in splitting words with underscore or in camel case to obtain the name of objects or variables. We also specified a set of stop words that are programing language-reserved words, commonly occurring words (e.g., the, was), and common terms in a software project. We also removed words like "get" and "set" since source files contain methods that start with these words. This requires some knowledge of the programming language syntax. Another option is to generate the Abstract Syntax Tree using tools like ANTLR [23] to support multiple languages. The generated tree can then be explored to extract the comments and identifiers inside the source code.

B. Association Data Mining

Once the data is preprocessed, we run the data mining algorithm (A6). We used Frequent Pattern Growth (FP-Growth) algorithm for AM, more specifically, the Liverpool University Computer Science – Knowledge Discovery in Data (LUCS-KDD) implementation of FP-Growth. This Java implementation uses tree structures for AM [24]. In our current case study, we restrict the length of the frequent itemsets generated to 2. (In the future, we will use the rules with more than one item in antecedent and consequent to uncover more complex relationships. For example, which files change given that two input files are modified.) Since we currently store the frequent 2-itemsets in database, we can compute the union of consequents using an SQL query. We store the generated frequent itemsets in a database (A7).

C. Topic Modeling

Once the source files are pre-processed, we extract semantic topics using LDA (B4). We used the CVB0 implementation of LDA [17]. Our implementation of LDA has the following parameters: number of topics and number of iterations. Number of topics is the number of topics we specify. The greater the number of topics, the more fine-grained will be the generated topics. Number of iterations is the number of times the algorithm will run. The higher number of iterations increases the likelihood that the topics will converge. For our case studies, we observed that the topics converged by 5000 iterations.

D. Result Querying

The last step is to query the results of both the rules generated from AM and the document relationship to topics (C1). We assume that the user is aware of at least one file that has to be modified for a given modification task. This file is used as the input. The output will show all the files that are recommended or predicted to change along with the input file.

V. VALIDATION

In this section, we discuss how we assessed our technique. We cover the setup of our case study, the results, and the limitations of our approach.

A. Case Study Setup

We conducted a case study on two open source projects: ArgoUML and EclipseFP. ArgoUML is a UML editor that also performs model checks [25]. This project uses the SVN repository, has around 6600 files, 14,000 commit transactions, and has a modification history of more than ten years. The second project we used is EclipseFP [26]. This is an Eclipse plugin for Haskel programming. This project uses Git version control system. This project consists of around 2000 files, has 1,796 commit transactions, and has a modification history of eight years. We selected these projects because these are active projects.

To measure the effectiveness of our approach, we used precision and recall. Precision measures the conciseness of a recommendations provided by the approach. Recall measures how many relevant recommendations are made by using this approach. We followed the same approach as used by Ying et al [13]. In this case study, we have assumed that developer is aware of at least one file for a given modification task. Therefore, we specified only one file f_s for generating recommendations for a modification task m. As explained in [13], the precision $precision(m, f_s)$ of a recommendation $recom(f_s)$ is the fraction of files that are predicted correctly and are part of the solution $f_{sol}(m)$ for the modification task m. The recall $recall(m, f_s)$ of a recommendation $recom(f_s)$ is the fraction of files recommended out of $f_{sol}(m)$.

For example, let us consider a modification task that requires changing files {a, b, c, d}. In addition, let us assume that the recommendations obtained for file b using our approach are files {a, c}. In this case, the precision for file b in this modification task is 100% as the approach recommended correct files. The recall value for file b for same modification task is 66.67% because the approach could predict only two files {a, c} out of {a, c, d}.

In order to determine the effectiveness of our prediction algorithm, we generated FP-Growth rules using 90% of the commit transactions. We then calculated the precision and recall rates of the generated rules on the remaining 10% of the commit transactions. We split the dataset based on time, since this simulates actual practice. Then, we compared these precision and recall rates with the precision and recall rates of FP-LDA.

The parameters we used are as follows. Minimum support for AM for ArgoUML is 10 and EclipseFP is 15. Confidence value for both projects is 25%. The number of topics for LDA was 20 topics. We measured the precision and recall of FP-LDA approach for 10%, 40% and 75% topic distribution values.

 ISBN: 978-1-61208-329-2

B. Results

The FP-growth resulted in 401 rules for ArgoUML and 42 for EclipseFP. The average precision and recall values obtained for ArgoUML with just FP-Growth are 0.48 and 0.06 respectively. The average precision and recall values for EclipseFP using AM are 0.32 and 0.13 respectively. FP-LDA results in lower precision and higher recall using the topic distribution cutoffs. Fig. 2 shows these values for both projects at various distribution cutoffs for topic modeling.

C. Discussion

The calculation of precision and recall gives a general understanding of how the approach fares in finding relationships. We assumed that each of these transactions was a task presented to a developer. For each file in the test transaction, we calculated precision and recall values to see if the tool can predict the remaining files.

We see that average precision reduces with the use of FP-LDA. This is due to the fact that the number of total recommendations increases due to topic modeling. However, a higher recall value shows that there is an increase in the number of relevant files predicted. This proves that number of correct recommendations increases with LDA. Although the overall precision is lower with combined approach, the utility of the approach lies in the fact that a developer needs to search only the set of recommended files, and not the entire source code base. Moreover, since we solely base our precision and recall on actual check-in records in the latter 10% of the history record, it is entirely possible that two files are related, but they may not have been checked-in together within this subset of the data. Thus, one can consider our precision number as a lower bound (e.g., FP-LDA precision for EclipseFP is at least 32% for 20 topics).

In addition to calculating precision and recall, we examined certain transactions to see how the combined FP-LDA fares over using AM alone. For example, in the ArgoUML dataset, AM fails to predict any relationship between files CrSingletonViolatedMissingStaticAttr.java and CrConsiderFacade.java (see Fig. 3). However, topic-modeling results show that these files are related. Upon analyzing these files, we found these two files have an indirect inheritance relationship. Similarly, in EclipseFP dataset, the classes DeltaVisitor.java and FullBuildVisitor.java implement interfaces IResourceDeltaVisitor.java and IResourceVisitor.java respectively. IResourceDeltaVisitor extends from IResourceVisitor.java. Topic Model shows that these files are related. However, since there was only one transaction where these files were changed together, AM does not show these files as related. These examples illustrate that FP-LDA technique can correctly identify related files.

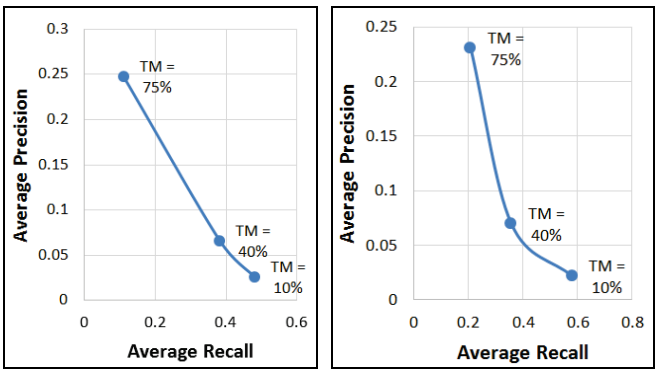


Figure 2. Precision and recall for ArgoUML, minimum support = 10 (left) and EclipseFP (right) minimum support of 15

D. Limitations of the study

Our precision and recall numbers may be subject to the specific datasets we selected. However, even though these datasets are different (e.g., the ratio between transactions and number of files is much greater in ArgoUML than in EclipseFP), we still observe a general increase in recall rates when using FP-LDA.

The number of topics used in LDA may also affect precision and recall rates. We ran our technique using different topic numbers and observed that the smaller the topic number, the higher the recall rates and the lower precision rates are generated. The "right" number of topics differs from one dataset to another.

VI. Related Work

Our work is most closely related to previous work in mining frequently changed files from a software repository [13, 27]. We used AM as other software engineering researchers have used this technique in the past. We build on top of this existing work and examine the benefits of combining AM with topic modeling. While others have used collaborative filtering [28], we use topic modeling, which is a probabilistic version of matrix factorization over the word-document matrix. In this paper, we use topic modeling to analyze the semantic content of source code and commit comments. In previous work, we have used topic modeling to identify associations between various software files and architecture components [29]. In the future, we plan to use topic modeling to identify associations between files and authors. Our work is also related to other techniques that seek to identify relationships between software files, such as recommendation systems, code search techniques, and dependency analysis.

Recommendation systems for software engineering may also recommend files for modification. Not all recommendation systems use association rule mining, but eRose a plugin for Eclipse does [7]. The common factor among all recommendation systems for software engineering is that they rely on the user's context in order to provide recommendations. While recommendation systems may help find related files in source code, the issue of user context is outside of the scope of our work.

 ISBN: 978-1-61208-329-2

Trans#	Name of File	P-noTM	R-noTM	P-TM75	R-TM75
17831	CrConsiderSingleton.java	0	0	0.13	0.75
17831	InitPatternCritics.java	0	0	0	0
17831	CrSingletonViolatedOnlyPrivateConstructors.java	0	0	0.13	0.75
17831	CrSingletonViolatedMissingStaticAttr.java	0	0	0.13	0.75
17831	CrConsiderFacade.java	0	0	0.13	0.75

Figure 3. FP-LDA is able to predict more files that are related to each other (P-TM75 and R-TM75) than FP-Growth alone (P-noTM, R-noTM)

Code search techniques may also be used to find source files that are related to one another. These techniques have their roots in traditional information retrieval methods [30]. An equivalency study was undertaken to the compare various IR methods in the area of traceability recovery [31]. The results of this study showed that while Latent Semantic Indexing (LSI), Jensen-Shannon (JS), and Vector Space Model (VSM) provided higher accuracy in identifying related files, LDA was able to capture associations, which the other methods could not. Recent work in code search has been performed to enhance the accuracy of these methods by allowing the user to specify both the syntactic and semantic properties of a search [30]. Code search techniques, however, fall short in finding relationships between project files, which are not semantically or syntactically related. Meanwhile, our technique finds these relationships based on the change history of the project and semantic relationship.

Dependency analysis tools may be used to find relationships between source files based on call graphs [4]. By making use of the project histories, we can mine relationships between any files that are checked-in together, as opposed to simply analyzing the code structure. As discussed previously, we also have the ability to find relationships between source code written in different languages. Most importantly, this approach helps to detect cross cutting concerns in which there may be a relationship between two files, but no relationship in a call-graph. For example, a project created for multiple operating systems may contain two source files, which accomplish the same task, but have no relationship in the calling tree. In this case, dependency analysis cannot detect these relationships, but our approach can because of the semantic similarity between files.

VII. Conclusion and Future Work

In this paper, we used AM and topic modeling together to assist developers in software maintenance task. These techniques were used to uncover the source file dependencies within a software project. We applied AM on version history of a project to find files that frequently change together. We complemented this technique by using topic modeling on the source code documents. We showed that using topic modeling could uncover file dependencies that are not captured due to lack of version history for those files. Our evaluation indicates that this combination of techniques increases recall rates by more than double, based on the open source projects we analyzed.

In the future, we would like to explore various options that can measure the usefulness of this approach. We also plan to examine other means to pre-process our data (e.g., aggregating the transactions based on time interval and committer to obtain a logical grouping of transactions). Finally, we plan to analyze more open source projects as well as conduct user studies to determine whether our approach reduces the time required for impact analysis or any maintenance task.

Acknowledgment

We thank Arthur U. Asuncion for his insights on LDA and providing the CVB0 implementation of LDA. We also thank Eamon Maguire for his assistance in extracting version histories and running the mining algorithm. This material is based upon work supported by the National Science Foundation under Grant No. CCF-1218266. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] S. S. Yau and J. S. Collofello, “Some stability measures for software maintenance,” Trans. on Software Engineering (TSE), vol. SE-6, Nov. 1980, pp. 545–552, doi:10.1109/TSE.1980.234503.

[2] A. J. Ko, B. A. Myers, M. Coblenz, and H. H. Aung, “An exploratory study of how developers seek, relate, and collect relevant information during software maintenance tasks,” TSE, vol. 32, Dec 2006, pp. 971–987, doi:10.1109/TSE.2006.116.

[3] R. N. Taylor, N. Medvidovic, and E. Dashofy, Software Architecture: Foundations, Theory, and Practice. John Wiley & Sons, 2010.

[4] M. Sharp and A. Rountev, “Static analysis of object references in RMI-based Java software,” in Proc of the Int’l Conf on Software Maintenance (ICSM), Sep. 2005, pp. 101–110, doi:10.1109/ICSM.2005.84.

[5] M. Eaddy, A. V. Aho, G. Antoniol, and Y. G. Gueheneuc, “Cerberus: Tracing requirements to source code using information retrieval, dynamic analysis, and program analysis,” in Proc of the 16th Int’l Conf on Program Comprehension (ICPC), Jun. 2008, pp. 53–62, doi:10.1109/ICPC.2008.39.

[6] D. Cubranic, G. C. Murphy, J. Singer, and S. Booth Kellogg, “Hipikat: a project memory for software development,” Trans. on Software Engineering (TSE), vol. 31, Jun. 2005, pp. 446–465, doi:10.1109/TSE.2005.71.

[7] M. P. Robillard, R. J. Walker, and T. Zimmermann, “Recommendation systems for software engineering,” IEEE Software, vol. 27, Jul-Aug. 2010, pp. 80–86, doi:10.1109/MS.2009.161.

[8] S. Bajracharya, J. Ossher, and C. V. Lopes, “Sourcerer - an infrastructure for large-scale collection and analysis of open-source code,” Science of Computer Programming, vol. 79, Jan. 2014, pp. 241–259, doi:10.1016/j.scico.2012.04.008.

 ISBN: 978-1-61208-329-2

[9] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules in large databases,” in Proc of 20th Int’l Conf on Very Large Data Bases, Sep. 1994, pp. 487–499.

[10] J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in Proc of the 2000 Int’l Conf on Mgmt of Data, May 2000, pp. 1–12, doi:10.1145/342009.335372.

[11] D. M. Blei, “Probabilistic topic models,” Comunications of the ACM, vol. 55, Apr 2012, pp. 77–84, doi:10.1145/2133806.2133826.

[12] R. Agrawal, T. Imielinski, and A. Swami, “Mining association rules between sets of items in large databases,” SIGMOD Rec., vol. 22, Jun. 1993, pp. 207–216, doi:10.1145/170036.170072.

[13] A. T. T. Ying, G. C. Murphy, R. Ng, and M. C. Chu-Carroll, “Predicting source code changes by mining change history,” TSE, vol. 30, Sep. 2004, pp. 574–586, doi:10.1109/TSE.2004.52.

[14] B. Liu, W. Hsu, and Y. Ma, “Mining association rules with multiple minimum supports,” in Proc of the Fifth ACM SIGKDD Int’l Conf on Knowledge Discovery and Data Mining, Aug. 1999, pp. 337–341, doi:10.1145/312129.312274.

[15] J. Pei et al., “Mining sequential patterns by pattern-growth: the PrefixSpan approach,” Trans. on Knowledge and Data Engineering, vol. 16, Nov. 2004, pp. 1424–1440, doi:10.1109/TKDE.2004.77.

[16] B. Liu, W. Hsu, and Y. Ma, “Identifying non-actionable association rules,” in Proc of Int’l Conf on Knowledge Discovery and Data Mining, Aug. 2001, pp. 329–334, doi:10.1145/502512.502560.

[17] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh, “On smoothing and inference for topic models,” in Proc of Conf. on Uncertainty in Artificial Intelligence, Jun. 2009, pp. 27–34.

[18] B. Gretarsson et al., “TopicNets: Visual analysis of large text corpora with topic modeling,” Trans. on Intelligent Systems and Technology, vol. 3, Feb. 2012, pp. 23:1–23:26, doi:10.1145/2089094.2089099.

[19] A. Panichella et al., “How to effectively use topic models for software engineering tasks? an approach based on genetic algorithms,” in Proc of the Int’l Conf on Software Engineering (ICSE), May 2013, pp. 522–531.

[20] T. Savage, B. Dit, M. Gethers, and D. Poshyvanyk, “TopicXP: Exploring topics in source code using latent dirichlet allocation,” in Proc of the Int’l Conf on Software Maintenance (ICSM), Sep. 2010, pp. 1–6, doi:10.1109/ICSM.2010.5609654.

[21] TMate Software, “SVNKit.” http://svnkit.com/, retrieved: Jan. 2014.

[22] “JavaGit.” http://javagit.sourceforge.net/, retrieved: Jan. 2014.

[23] “ANTLR.” http://www.antlr.org/, retrieved:Jan. 2014.

[24] F. Coenen, G. Goulbourne, and P. Leng, “Tree structures for mining association rules,” Data Mining and Knowledge Discovery, vol. 8, Jan. 2004, pp. 25–51, doi:10.1023/B:DAMI.0000005257.93780.3b.

[25] CollabNet, “ArgoUML.” http://argouml.tigris.org/, retrieved:Jan. 2014.

[26] “EclipseFP, the Haskell plug-in for Eclipse.” http://eclipsefp.github.io/, retrieved: Jan. 2014.

[27] T. Zimmermann, A. Zeller, P. Weissgerber, and S. Diehl, “Mining version histories to guide software changes,” TSE, vol. 31, Jun. 2005, pp. 429–445, doi:10.1109/TSE.2005.72.

[28] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in Proc of the 10th International Conference on World Wide Web, May 2001, pp. 285–295, doi:10.1145/371920.372071.

[29] H. U. Asuncion, A. U. Asuncion, and R. N. Taylor, “Software traceability with topic modeling,” in Proc of ICSE, vol. 1, May 2010, pp. 95–104, doi:10.1145/1806799.1806817.

[30] S. P. Reiss, “Semantics-based code search,” in Proc of ICSE, May 2009, pp. 243–253, doi:10.1109/ICSE.2009.5070525.

[31] R. Oliveto, M. Gethers, D. Poshyvanyk, and A. De Lucia, “On the equivalence of information retrieval methods for automated traceability link recovery,” in Proc of ICPC, Jun-Jul. 2010, pp. 68–71, doi:10.1109/ICPC.2010.20.

 ISBN: 978-1-61208-329-2

Extracting Representative Words of a Topic Determined by Latent Dirichlet Allocation

Toshiaki Funatsu
Graduate School of Information Science
and Electrical Engineering,
Kyushu University
funatsu@nlp.inf.kyushu-u.ac.jp

Yoichi Tomiura
Faculty of Information Science and Electrical Engineering
Kyushu University,
tom@inf.kyushu-u.ac.jp

Emi Ishita
Research and Development Division,
Kyushu University Library,
Kyushu University
Ishita.emi.982@m.kyushu-u.ac.jp

Kosuke Furusawa
Graduate School of Information Science
and Electrical Engineering,
Kyushu University
furusawa@nlp.inf.kyushu-u.ac.jp

***Abstract*—Determining the topic of a document is necessary to understand the content of the document efficiently. Latent Dirichlet Allocation (LDA) is a method of analyzing topics. In LDA, a topic is treated as an unobservable variable to establish a probabilistic distribution of words. We can interpret the topic with a list of words that appear with high probability in the topic. This method works well when determining a topic included in many documents having a variety of contents. However, it is difficult to interpret the topic just using conventional LDA when determining the topic in a set of article abstracts found by a keyword search, because their contents are limited and similar. We propose a method to estimate representative words of each topic from an LDA result. Experimental results show that our method provides better information for interpreting a topic than LDA does.**

Keywords-LDA; topic analysis; Gibbs sampling.

I. INTRODUCTION

Web search engines are very widely used. Users are able to access different information resources easily using keywords for a search. Academic information retrieval systems have also become common and popular. As academic research disciplines have become more specific or more interdisciplinary, users who search related documents need narrow or focused topics. However, a keyword search is often not able to address this need. When users use very specific words as search terms, they generally obtain only a few search results. On the other hand, when they use general words as search terms, they obtain many search results. In this case, it is time-consuming to select relevant documents from search results.

Therefore, the following retrieval support system is useful when a user searches academic papers related to narrow or focused topics: (1) the user retrieves academic papers with generalized keywords, (2) the system does a topic analysis of the abstracts found in the search and presents some information about their topics to the user, (3) the user chooses a particular topic among them, and (4) the system narrows down the search results to academic papers that mainly contain that topic. Some methods perform a keyword article search using the feedback of the latent topic [2] or search with a novel topic model that organizes articles using the author information [3].

Latent Dirichlet Allocation (LDA) is a well-known method for topic analysis. In LDA, a topic is treated as a latent variable for determining probabilities of words. The user is able to understand a topic based on a list of words that appear with high probability in the topic. However, when a keyword search yields results with similar content, it may be difficult to understand a topic with the word list presented by LDA. The word list contains many unnecessary words for expressing a topic. Then, we consider that there are two types of words in the list. One is a word expressing the content of a topic, and the other is a word attendant to the first type of word. We call the first type a representative word of a topic. In this paper, we propose a method for identifying representative words of a topic from the word list acquired by LDA to help the user to understand the topic.

Our method first constructs a set of documents for each

 ISBN: 978-1-61208-329-2

topic that contains only words that LDA assigns the topic to, and next identifies a representative word for each topic and each document. We assume that the representative word of a document generates the other words in that document. We use Gibbs sampling to identify the representative word of each document. The higher the probability that the word *w* represents a document of topic *t*, the more representative *w* is of *t*.

In Section II, we discuss some related studies and explain the model underlying our method. In Section III, we propose the model of our method. In Section IV, we discuss an experiment that compares the results of LDA and to those of our method.

II. RELATED STUDIES

LDA is a generative probabilistic model of a corpus [1]. LDA assumes that each document has a probability distribution over topics and each topic has a probability distribution over words. Its generative process for a document in a corpus is as follows:

For each word in the document,

a) choose a topic *t* according to the probability distribution over topics that the document has;

b) choose a word *w* according to the probability distribution over words that the topic *t* has.

Blei et al. estimate the parameters using the variational Bayesian method [1]. Griffiths et al. analyze topics in a document based on LDA, but they use Gibbs sampling in parameter estimation [4].

We define our notation as follows:

M : number of documents,

$w_n^{(m)}$: the n-th word in the m-th document,

$\mathbf{w}^{(m)} = (w_1^{(m)}, w_2^{(m)}, \cdots, w_{N_m}^{(m)})$: the m-th document,

$\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(M)})$: set (sequence) of documents,

$z_n^{(m)}$: latent variable expressing a topic to be assigned to the word $w_n^{(m)}$,

$\mathbf{z}^{(m)} = (z_1^{(m)}, z_2^{(m)}, \cdots, z_{N_m}^{(m)})$,

$\mathbf{z} = (\mathbf{z}^{(1)}, \mathbf{z}^{(2)}, \cdots, \mathbf{z}^{(M)})$,

K : number of topics,

$\theta_t^{(m)}$: probability of words with topic t in the m-th document,

$\theta^{(m)} = (\theta_1^{(m)}, \theta_2^{(m)}, \cdots, \theta_K^{(m)})$,

$\theta = (\theta^{(1)}, \theta^{(2)}, \cdots, \theta^{(M)})$,

V : number of words (by type),

$\phi_w^{(t)}$: occurrence probability of word w from topic t,

$\phi^{(t)} = (\phi_1^{(t)}, \phi_2^{(t)}, \cdots, \phi_V^{(t)})$,

$\phi = (\phi^{(1)}, \phi^{(2)}, \cdots, \phi^{(K)})$.

The joint probability of w and z in LDA is expressed as

$$p(\mathbf{w}, \mathbf{z} \mid \theta, \phi) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} p(z_n^{(m)} \mid \theta^{(m)}) p(w_n^{(m)} \mid z_n^{(m)}, \phi) . \quad (1)$$

$$p(t \mid \theta^{(m)}) = \theta_t^{(m)}, \ p(w \mid t, \phi) = \phi_w^{(t)}$$

The prior distribution of $\theta^{(m)}$ is the dimensionality K-1 of the Dirichlet distribution with parameter α :

$$p(\theta^{(m)} \mid \alpha) = \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \prod_{k=1}^{K} \theta_k^{(m)} . \quad (2)$$

The prior distribution of $\phi(t)$ is the dimensionality V-1 of the Dirichlet distribution with parameter β :

$$p(\phi^{(t)} \mid \beta) = \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \prod_{w=1}^{V} \phi_w^{(t)} . \quad (3)$$

The probability of $\mathbf{z}$ given the set of documents $\mathbf{w}$ and the hyper parameters α and β is obtained via

$$p(\mathbf{z} \mid \mathbf{w}, \alpha, \beta) \propto \prod_{m=1}^{M} \frac{\Gamma(K\alpha)}{\Gamma(\alpha)^K} \frac{\prod_{k=1}^{K} \Gamma(n_{DZ}(m,k;\mathbf{z}) + \alpha)}{\Gamma(n_{DZ}(m,*;\mathbf{z}) + K\alpha)} \times \prod_{k=1}^{K} \frac{\Gamma(V\beta)}{\Gamma(\beta)^V} \frac{\prod_{w=1}^{V} \Gamma(n_{ZW}(k,w;\mathbf{w},\mathbf{z}) + \beta)}{\Gamma(n_{ZW}(k,*;\mathbf{w},\mathbf{z}) + V\beta)} , \quad (4)$$

where $n_{DZ}(m, k ; \mathbf{z})$ is the number of times a word from the m-th document is assigned to topic k in $\mathbf{z}$, $n_{ZW}(k, w ; \mathbf{w}, \mathbf{z})$ is the number of times word w is assigned to topic k in ($\mathbf{w}$, $\mathbf{z}$), and $n_{DZ}(m, * ; \mathbf{z})$ and $n_{ZW}(k, * ; \mathbf{w}, \mathbf{z})$ are

$$n_{DZ}(m,*;\mathbf{z}) = \sum_{k=1}^{K} n_{DZ}(m,k;\mathbf{z}) = N_m ,$$
$$n_{ZW}(k,*;\mathbf{w},\mathbf{z}) = \sum_{w=1}^{V} n_{ZW}(k,w;\mathbf{w},\mathbf{z}) . \quad (5)$$

In our study, we use the results of topic analysis to estimate representative words for each topic through Gibbs sampling [4].

Blei et al. [5] studied a method for extracting significant multi-word expressions for a topic from the results of topic analysis using the procedure in [1]. Our approach is different from this. We provide representative words of a topic as useful information for understanding the topic. We do not analyze multi-word expressions, but simply treat multi-word expressions in Wikipedia entries as single words in the preprocessing in our experiment, because multi-word expressions help users to understand topics. Our approach of estimating representative words of a topic can be applied to the results of topic analysis

 ISBN: 978-1-61208-329-2

with the procedure of [5].

III. OUR PROPOSED METHOD

LDA works well for a document set that is large and has a variety of contents. It is also necessary to be able to predict topics contained in a document set to some extent so as to identify a topic from a list of words that appear with high probability in the topic. However, for a set of abstracts obtained by a keyword search, it may be difficult to identify a topic with the word list presented by LDA.

TABLE I. REPRESENTATIVE WORD SETS MATCHED TO TOPICS BY LDA

Topic	Representative Word Set Express Topic
1	retrieval model information framework space task theory problem vector process method concept question similarity modeling
2	ir retrieval text paper language issue indexing evaluation xml image processing area application research discussion
3	system information user retrieval study paper process interaction time management knowledge result tag case performance
4	method datum algorithm structure information problem feature data number analysis classification technique music network combination
5	document query collection term retrieval concept approach relevance context result technique feedback performance analysis ir
6	library computer system use access information service storage science index technology labor description resource program
7	search web information user engine approach result need content domain page use ontology interest strategy
8	information retrieval research field development multimedia application technique technology machine researcher type book tool form
9	ir word experiment retrieval work evaluation text function term performance trec measure set system graph
10	database author protocol scheme problem pir record server report privacy communication software requirement file number

This is because such a document set has technical and similar contents.

Table I shows the results of LDA topic analysis for an abstract set consisting of 525 academic papers found by the query "information retrieval" on Cute.Search (the academic search service at Kyushu University). One can see that it is difficult to determine what each topic is.

Hence, we propose a method for estimating the representative words of each topic from the results of LDA topic analysis of an abstract set obtained by a keyword search. Our method consists of the following three components:

a) Improving LDA [4]:
We improve the algorithm so as to calculate the semi-optimum solution z maximizing $p(\mathbf{z} \mid \mathbf{w}, \alpha, \beta)$.

b) Deleting unnecessary words that occur in many topics, and generating a document set for each topic.

c) Estimating representative words from a document set for each topic.

A. *Improving LDA*

Griffiths et al. estimate θ and ϕ using the s-th result of sample $\mathbf{z}$ (where s is large enough) [4]. It does not matter actually if we are only interested in θ and ϕ, and if both the document size and document number are large. In the proposed method, we construct a document set of each topic using $\mathbf{z}$. This makes it a problem using the s-th result of sample $\mathbf{z}$. Therefore, we improve the algorithm so as to get the semi-optimal solution $\mathbf{z}$ that maximizes (4) among s samples. We call the obtained $\mathbf{z}$ the "suboptimal topic assignment."

B. *Deleting Unnecessary Words and Constructing a Document Set of Each Topic*

We calculate the idiosyncrasy of each word for a topic and remove words that have low idiosyncrasy. Specifically, we calculate the entropy of a word. We remove words that seem to be ineffective for topic expression by setting a threshold for entropy.

The entropy of word w is given as

$$E(w) = -\sum_{t=1}^{K} p(t \mid w) \log p(t \mid w) , \quad (6)$$

where $p(t \mid w)$ is the maximum likelihood estimate by suboptimal topic assignment z as follows:

$$p(t \mid w) = \frac{n_{ZW}(t, w; \mathbf{w}, \mathbf{z})}{\sum_{t=1}^{K} n_{ZW}(t, w; \mathbf{w}, \mathbf{z})} . \quad (7)$$

The word that has the lowest idiosyncrasy is the word w that satisfies

 ISBN: 978-1-61208-329-2

$$p(t \mid w) = \frac{1}{K} \quad , \tag{8}$$

for every topic t. The entropy of this word is $\log_2 K$. Then, we consider a word w as unnecessary and remove it if w satisfies

$$E(w) > \kappa \log_2 K \ . \tag{9}$$

Now we set κ to 0.25 for a preliminary experiment.

The document set of each topic t (=1, 2, …, K),

$$\mathbf{w}[t] = (\mathbf{w}^{(1)}[t], \mathbf{w}^{(2)}[t], \cdots, \mathbf{w}^{(M_t)}[t]) \ ,$$

is constructed from the results (**w**, **z**) of LDA topic analysis as follows:

a) Set m =1 and i =1.

b) Seek the following word set (word sequence):

$$\{w_n^{(m)} \mid z_n^{(m)} = t, \text{and } E(w_n^{(m)}) \le \kappa \log_2 K\} \ .$$

If the number of elements (words) in this set is over L, then set as follows:

$$w^{(i)}[t] = \{w_n^{(m)} \mid z_n^{(m)} = t, \text{and } E(w_n^{(m)}) \le \kappa \log_2 K\}$$

and $i \leftarrow i$+1.

c) If m equals M, the construction process is finished. Otherwise, $m \leftarrow m$ +1 and repeat step (b).

We do not replace pronouns with their antecedents when constructing input data for LDA. Then, a word that appears frequently is not always important for a certain topic. A word that is referred by a pronoun is sometimes important, which is why we delete redundant words in a document. We also delete any document that has less than L words from the document set of a topic, because it seems difficult to estimate a representative word of such a document. There would be noise for estimating a representative word. Now, we set L = 4.

C. *Estimating Representative Words for a topic*

We estimate the representativeness of each word for topic t from the document set of t (document sequence):

$$\mathbf{w} = (\mathbf{w}^{(1)}, \mathbf{w}^{(2)}, \cdots, \mathbf{w}^{(M)}) \ .$$

This is constructed with the method of the preceding paragraph. We omit t from this because the following process is executed for each topic. In addition, the number of documents M, the identification of each word and number of words (by type) V are also set for each topic t. In the same way, parameters introduced in the following model are also set for each topic t.

In our model, the document $\mathbf{w}^{(m)}$ is generated in the following two steps: 1) a word x is generated as a representative word, and 2) x generates the other words in the document. The probability of generating x as a representative word is denoted by η_x, the probability of w occurring in the document whose representative word is x is denoted by $\xi_1^{(x,\,w)}$, and $\xi_0^{(x,\,w)}$ is 1 - $\xi_1^{(x,\,w)}$. Here, the probability of generating x ($\in \mathbf{w}^{(m)}$) as a representative word and generating the other words in $\mathbf{w}^{(m)}$ from x is expressed as follows:

$$\begin{aligned} & p(\mathbf{w}^{(m)}, x \mid \eta, \xi) \\ & = \eta_x \times \left(\prod_{\substack{w=1 \\ w \notin \mathbf{w}^{(m)}, w \neq x}}^{V} \xi_0^{(x,w)} \right) \times \left(\prod_{\substack{w=1 \\ w \in \mathbf{w}^{(m)}, w \neq x}}^{V} \xi_1^{(x,w)} \right). \end{aligned} \tag{10}$$

The prior distribution of η = (η_1, η_2, ... , η_V) is the dimensionality V-1 of the Dirichlet distribution with parameter γ:

$$p(\eta \mid \gamma) = \frac{\Gamma(V\gamma)}{\Gamma(\gamma)^V} \prod_{x=1}^{V} \{\eta_x\}^{\gamma-1} \ . \tag{11}$$

The prior distribution of ($\xi_0^{(x,\,w)}$, $\xi_1^{(x,\,w)}$) is the beta distribution (one-dimensional Dirichlet distribution) with parameter δ:

$$p(\xi_0^{(x,w)}, \xi_1^{(x,w)} \mid \delta) = \frac{\Gamma(2\delta)}{\Gamma(\delta)^2} \{\xi_0^{(x,w)}\}^{\delta-1} \{\xi_1^{(x,w)}\}^{\delta-1} \ . \tag{12}$$

The representative word of document $w^{(m)}$ is denoted by $x^{(m)}$, and the set of representative words for all documents is denoted by **x**:

$$\mathbf{x} = (x^{(1)}, x^{(2)}, \cdots, x^{(M)}) \ .$$

We define two counters as follows. $n_R(x \ ; \ \mathbf{x})$ is the number of times x has been selected as a representative word in **x**, and $n_C(x,w \ ; \ \mathbf{w}, \mathbf{x})$ is the number of elements in the set:

$$\{m \mid x^{(m)} = x, \text{and } w(\neq x) \in \mathbf{w}^{(m)}\} \ .$$

(In other words, n_C is the number of documents that have x as a representative word and contains word w)

The probability of occurrence (**w**, **x**) is

 ISBN: 978-1-61208-329-2

$$p(\mathbf{w},\mathbf{x}\mid\eta,\xi) = \left(\prod_{x=1}^{V}\{\eta_x\}^{n_R(x;\mathbf{x})}\right) \times\left(\prod_{x=1}^{V}\prod_{\substack{w=1\\ w\neq x}}^{V}\{\xi(x,w)=0\}^{n_R(x;\mathbf{x})-n_C(x,w;\mathbf{w},\mathbf{x})}\{\xi(x,w)=1\}^{n_C(x,w;\mathbf{w},\mathbf{x})}\right). \quad (13)$$

Then, we obtain the conditional probability of $x^{(m)} = x$ given the representative words of $\mathbf{w}$ without $\mathbf{w}^{(m)}$ via

$$p(\mathbf{w},\mathbf{x}/x^{(m)},x^{(m)}=x\mid\gamma,\delta) \propto\left(\frac{n_R(x;\mathbf{x}/x^{(m)})+\gamma}{(M-1)+V*\gamma}\right) \times\left(\prod_{\substack{w\notin\mathbf{w}^{(m)}\\ w\neq x}}^{V}\frac{n_R(x;\mathbf{x}/x^{(m)})-n_C(x,w;\mathbf{w},\mathbf{x}/x^{(m)})+\delta}{n_R(x;\mathbf{x}/x^{(m)})+2\delta}\right) \times\left(\prod_{\substack{w\in\mathbf{w}^{(m)}\\ w\neq x}}^{V}\frac{n_C(x,w;\mathbf{w},\mathbf{x}/x^{(m)})+\delta}{n_R(x;\mathbf{x}/x^{(m)})+2\delta}\right), \quad (14)$$

where $\mathrm{x}/x^{(m)}$ means representative words of $\mathbf{w}$ without $\mathbf{w}^{(m)}$.

We estimate η using Gibbs sampling as follows. $E[\eta_x \mid \mathbf{w}, \mathbf{x}]$, the expectation value of η_x from the posterior distribution of η_x given $\mathbf{w}$ and its representative words $\mathbf{x}$, is calculated according to

$$E[\eta_x \mid \mathbf{w},\mathbf{x}] = \frac{n_R(x;\mathbf{x})+\gamma}{M+V\gamma}. \quad (15)$$

$E[\eta_x \mid \mathbf{w}]$, the expectation value of η_x from the posterior distribution of η_x given the document set $\mathbf{w}$, is found from

$$\begin{aligned} E[\eta_x\mid\mathbf{w}] &= \int \eta_x p(\eta,\xi\mid\mathbf{w})d\eta d\xi \\ &= \int \eta_x \frac{\sum_{\mathbf{x}} p(\mathbf{w},\mathbf{x},\eta,\xi)}{p(\mathbf{w})}d\eta d\xi \\ &= \sum_{\mathbf{x}}\frac{p(\mathbf{w},\mathbf{x})}{p(\mathbf{w})}\int \eta_x \frac{p(\mathbf{w},\mathbf{x},\eta,\xi)}{p(\mathbf{w},\mathbf{x})}d\eta d\xi \\ &= \sum_{\mathbf{x}} p(\mathbf{x}\mid\mathbf{w})E[\eta_x\mid\mathbf{w},\mathbf{x}] \end{aligned}. \quad (16)$$

Let $\mathbf{x}(S_0+1)$, $\mathbf{x}(S_0+2)$, ..., $\mathbf{x}(S_0+S)$ be the sequence of representative words obtained by Gibbs sampling from (S_0+1) to (S_0+S) rounds. Then, $E[\eta_x \mid \mathbf{w}]$ is approximated by (15), (16), and the law of large numbers:

$$E[\eta_x\mid\mathbf{w}] = \frac{1}{S}\sum_{s=S_0+1}^{S_0+S}\frac{n_R(x;\mathbf{x}(s))+\gamma}{M+V\gamma}. \quad (17)$$

Our method presents a list of representative words w with high probability η_w for each topic. Table II shows the results of the topic analysis performed by our method for the same dataset as in Table I.

TABLE II. SAMPLE TOPICS FROM OUR METHOD

Topic	Representative Word Set Express Topic
1	Model space method largesystems findings determine andtajikistan cells researcheshave methodsin efficiency were applied unwanted subjected
2	Processing issue indexing image conference participant format storey forseveral child andperformance ai articolo nostril name
3	System behavior difference interaction medium sinceinformation characterization control recovery management ehrlich gate eigenvector completion agent
4	Datum algorithm method structure value deviation omit market mechanical acceptance complete avenue stemmer between decision
5	document query collection factor temperament preference ohio occupation chicago feedback department finder lsa formalism proposition
6	computer library index access organization storage university control labor science classroom rs skill instruction subscales
7	search web interface dei indexdocuments iv onthe day north request ofwordnet keywordstoindexing print collaboration tapas
8	field part multimedia tool portland researcher roll machine illustration film discovery sidebar facilitarne hypertext diffuse
9	recall sense trec function word effect component weight class investigation iss efficacy combination isss thesaurus
10	database author notice report fax communication general american rule horizon hole analogue correlation radiation the

IV. EXPERIMENT

We performed an experiment to compare the method of [1] and our method. We prepared 20 queries and collected about 500 to 1500 Japanese abstracts from the article database CiNii for each query. We did topic analysis for the collected abstracts using LDA with 10 topics, and then estimated representative words of each topic using our method. We set α and β of LDA's meta-parameters to 2.0

 ISBN: 978-1-61208-329-2

and 0.1 respectively. We set γ and δ of our model's meta-parameters to 0.05 and 0.1, respectively.

We evaluated the analysis results with each method as follows. 1) Four students studying the areas of electrical engineering and computer science evaluated the results. We divided them into two groups of two students. The four evaluators are denoted by a1, a2, b1 and b2. For each query, we assigned methods to the students so that the method that a1 and a2 evaluated was different from the method that b1 and b2 evaluated. For every query, we replaced the methods being evaluated. As a result, each student evaluated the results for 10 queries using each method. 2) We evaluated the analysis result for method M and query Q. The evaluators were given the word list (15 words) for every topic determined by method M and the 10 abstracts randomly selected from the search results by query Q. For each abstract *a*, the evaluator selected three topics that seemed to be included in *a* using his sense based on the word lists of topics, and we scored the size of the intersection between selected topics and the following set to *a*:

$$\{t \mid \theta_t^{(a)} \geq \text{the third highest value in } \theta_1^{(a)}, \theta_2^{(a)}, \cdots, \theta_{10}^{(a)}\}.$$

As a result, the score for an abstract is from 0 to 3. The score for query Q is the sum of the scores of each evaluator in a group for every abstract in a set retrieved by Q. As a result, a score for query Q (that is, an abstract set retrieved by Q) is from 0 to 60.

The results of the evaluation are in Table III. The scores for our method are higher than those for LDA for most queries (No. 1, 4–7, 10–13, and 16–20), and the average of the scores for all abstract sets for our method is 1.3 points higher than for LDA. However, this is not a very big increase. We assume that users use the topic analysis to narrow down the results of a keyword search about their own field or related fields. The evaluators were unfamiliar with some of the prepared queries. For the abstract sets retrieved by familiar queries (No. 1, 4–10, 13, 16, 19, and 20 in Table III), the average score for our method is 2.52 points higher than for LDA.

TABLE III. RESULTS OF EXPERIMENT

No.	Query	LDA	Our method
1	“Natural language”	23	25
2	“Translation”	25	21
3	“Medical treatment”	26	24
4	“Light, Energy”	25	27
5	“Ion, Electricity”	14	19
6	“Sensor, Measurement”	19	20
7	“Energy, Environment”	20	25
8	“Electric power, Supply”	19	17
9	“Retrieval, Support”	15	17
10	“Radio wave, Transmission”	19	20
11	“Concrete”	20	21
12	“Fluid mechanics”	13	14
13	“Quantum”	13	20
14	“Plasma”	22	22
15	“Nuclear fusion”	20	17
16	“Sensing”	24	25
17	“Project management”	18	19
18	“Aviation, Cosmos”	21	23
19	“Artificial intelligence”	20	23
20	“Communication network”	16	19

V. CONCLUSION AND FUTURE WORK

The representative word lists generated by our method does not contain some unnecessary words that are contained in word lists generated by LDA, but there are many non-content words and general terms in our lists. Our goal is to make LDA analysis more intelligible. We cannot expect a very big improvement in expression of topic contents when LDA analysis is not good. In this work, the meta parameters α and β in the LDA were set to 2 and 0.1, respectively. In future work, we will explore better values of the meta parameters and compare the results for LDA and our method. In addition, we will evaluate the effect of filtering words in LDA word lists using entropy and explore a better entropy threshold.

VI. REFERENCE

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, Latent Dirichlet Allocation, Journal of Machine Learning Research, 3, January, 2003, pp. 993-1022.

[2] D. Andrzejewski and D. Buttler, Latent Topic Feedback for Information Retrieval, Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, August 21-24, 2011, pp. 600-608.

[3] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, Citation Author Topic Model in Expert Search, Proceedings of the 23rd International Conference on Computational Linguistics: Posters, August 23-27, 2010, pp. 1265-1273.

[4] T. L. Griffiths and M. Steyvers, Finding scientific topics, PNAS, vol. 101, 2004, pp. 5228-5235.

[5] D. M. Blei and J. D. Lafferty, Visualizing Topics with Multi-Word Expressions, Technical Report, arXiv:0907, 1013vl [stat.ML], 2009.

 ISBN: 978-1-61208-329-2

Discovery of Precursors of Serious Damage by Disaster Context Library with Cross-field Agents

Taizo Miyachi[1], Gulbanu Buribayeva[1]
[1] School of Information Science and Technology, Tokai University, Hiratsuka, Japan
e-mail: miyachi@keyaki.cc.u-tokai.jp, banu.b@hotmail.com

Saiko Iga[2], Takashi Furuhata[3] and Tseveenbolor Davaa[3]
[2] Keio Research Institute of SFC, Fujisawa, Japan
[3]University of Utah, USA
e-mail: sai@s05.itscom.net, Takashi.furuhata@utah.edu
Tseveenbolor.davaa@utah.edu

Abstract— **Unpredictable scale of disasters (e.g., earthquake and Tsunami, etc.) has caused tremendous number of deaths all over the world. Citizens need to study effective basic actions and adaptive actions for the sudden change of disaster contexts in order to survive. By referring the case of The Great East Japan Earthquake in 2011, we propose disaster context library "d-Library" that shows citizens an eight layered disaster contexts framework, allowing them to easily recognize important layers of knowledge to survive. They can also study latest scientific knowledge and learn how to discover precursors of various types of disasters and their serious damages. We also discuss a methodology to avoid "catastrophic occurrence forgetting" used to remind important problems and to ensure safety after the evacuation. We also discuss experimental results of decision making assistance and checking the safety assist function.**

Keywords-disaster context; cross-field agent; catastrophic occurrence forgetting; warning picture; risk grade

I. INTRODUCTION

The Great East Japan Earthquake in 2011 and Tsunamis [1] killed about 20,000 citizens. Moreover, many people were irradiated [2] and about 300,000 residents have lost home town by explosions of Fukushima nuclear power plant and occurred some nuclear leakage. From the experiences of such serious damages, it became clear that it is very difficult for ordinary citizens to understand separate scientific knowledge over many academic fields, such as seismology, physics, tsunami science, psychology [3], geology, unpredictable accidents by human behaviors, and invisible radioactive contaminants in a limited short time in the disaster. They could not also adapt such knowledge against serious dangers by the great earthquakes, giant tsunamis and radioactive contaminants. They should evacuate indoor, outdoor, and indoor respectively. Therefore, citizens need different types of organized knowledge. They also need information distribution systems, and prior learning systems in order to quickly start evacuation and survive from the great disasters. Public organizations already have a variety of information sharing systems for disaster prevention [4] and have built mutual assist society among many cities and companies both in Japan and in the world. We should enhance the disaster prevention system from the view point of human centered design. The citizen can find disaster contexts and some precursors of the great disasters. We propose an eight layered context-oriented disaster library "d-Library" as enhancing parts of the disaster information systems in order to reduce the serious damage [5] not only by separate scientific knowledge, but also by coping with human nature and ability, social behaviors, and collaboration in the mutual assist societies, etc. New technologies, such as enhancing collaboration by social networks [6] and Big Data [3] analysis should be also introduced into the library.

The citizens should ensure the safety at the place of refuge when they complete evacuation utilizing such disaster contexts since many people were irradiated in a place of refuge. We also discuss safety check functions for avoiding Catastrophic Occurrence Forgetting (COF) [7] that is caused by mass-media reports of a deluge of remarkable occurrences. We also discuss experimental results of decision making assist and checking safety assist function.

II. DISASTER CONTEXTS IN MANY ACADEMIC FIELDS AND CATASTROPHIC OCCURRENCE FORGETTING

The Seismological Association of Japan (SAJ) could not predict not only "Hanshin-Awaji great earthquake Mj7.3" but also "Great East Japan Earthquake Mw9.0" at all. Even famous ten meters height beachside fences in Taro-cho in Tohoku district could not guard the Taro village from giant tsunamis in The Great East Japan Earthquake. Whole village was washed out by the giant tsunamis. The myth of safe nuclear power plant has been broken by explosions of Fukushima nuclear power plants.

On the other hands, about 80 percent of the 184 students that were on their way home from Kamaishi Primary School could safely evacuated by the quick actions based on the disaster contexts of "Tsunami Tendenko" [8].

(1) Disaster context and safe evacuation

"Disaster Context" includes context of occurrences in a disaster. Sharing disaster contexts enables citizens to avoid serious damage and adaptively evacuate from unpredictable danger. Disaster contexts are also very important for detecting disaster precursor of serious damage in order to reduce the damage. Disaster precursors can also be acquired in "technology assessment" [9].

Disaster context mainly consists of layer ID and six sub-contexts. (c1) Preparation, (c2) Environment, (c3) Action, (c4) Information, (c5) Psychology, and (c6) Safety check. Sub-context consists of "sc-attribute," "discription," "state," and "risk grade of context." "state" also includes

 ISBN: 978-1-61208-329-2

"concealed." This means the necessity of careful check of safety. Relationships between sub-contexts are useful for checking the safety.

Major assists for citizens by disaster contexts are classified into eleven categories.

(a1) Know a "current situation" in typical contexts
(a2) Avoid "invisible/unpredictable danger," such as radioactive contaminants and tsunamis from hill-side
(a3) Find "precursors of serious damages" like a warning
(a4) Find "preparation stratagem" based on a pair of disaster contexts and serious damages
(a5) Find a "slighted danger" that causes the death like fast ebb tsunami of 20 cm depth
(a6) Know "how to make evacuation leaders" like "be the first in "Tsunami Tendenko"" [8]
(a7) Know "the reason of missing a chance of survive" like swallowed persons in a car park in a shelter
(a8) Know "the reason of seizing a chance of survive" like catching at a branch at the head of a great tsunami
(a9) Know "the reason of death by same actions" like evacuation by car near a few bridges [3]
(a10) Mind the change of degree of dangers
(a11) Find danger and ensure safety

A disaster context should include at least one description in six kinds of descriptions.

(i) Dangerous situations with a photo for the rescue
(ii) Change of dangers with photos or video
(iii) Serious damage and its change with the causes.
(iv) Embedded dangerous phenomenon
(v) Wrong decisions and serious results with photos
(vi) Success stories in evacuation

(2) Catastrophic Occurrence Forgetting

People should remind both ensuring their safety at a new place of refuge and reconsidering pending problems when they complete safe evacuation according to disaster contexts. Awareness of disaster contexts in an important occurrence to be observed is sometimes reduced by the catastrophic occurrences. We call these phenomena "Catastrophic Occurrence Forgetting (COF)." COF often lets citizens forget the important occurrence to survive. This has caused serious damages in emergency time. There are mainly four categories of catastrophic occurrences to make human forget disasters. (i) Lost of property, (ii) Encounter with what victims expect to forget, (iii) Catastrophe Forgetting (CF) [7] by a torrent of remarkable occurrences and big news, (iv) Sever life-style in separated family in a stricken society. (i) Lost property are (a) physical property, such as house and car, (b) family, such as child and parents, (c) community and neighborhood, (d) home town, (e) pleasant lifestyle, (d) economic base, such as field, farm, ground, non-radioactive fishes etc. Some occurrences in human's memory are forgotten or wiped out by COF.

(ii) Encounter with what victims expect to forget.

Human instinctively forgets abnormal experiences in order to recover from Post-Traumatic Stress Disorder (PTSD) etc. A victim tries to forget the image of disasters, for example as a tsunami, people rolled off upper body and lower body by a swirl of strong undertows, public hall swallowed by tsunami and an expressionless face of the wife who were waving her hands on the top of a big tsunami [10].

(iii) Catastrophe Forgetting (CF).

Japanese mass-media continually reported a torrent of remarkable happenings, such as large fires, explosions of gas tanks, broken buildings, stopped trains, accidents in highway tunnels, North Korean nuclear missiles, avian influenza and tsunamis just after The Great East Japan Earthquake. A victim had to dispose broken stuffs and furniture and tidy up the rooms with fallen bookshelves and fallen products just after the great earthquakes. S(he) often forgot passed, but important occurrences [7] and reduced their awareness against tsunamis. We call this "Catastrophe Forgetting (CF)" in disaster psychology.

Example 1. Forgetting the accidents in nuclear power plants. All Japanese were surprised at the scenes such that tsunami swallowed a whole town including lots of cars and all houses, and the fire spread over a town etc. Japanese citizens could not help reducing the awareness of the accidents in Fukushima nuclear power plants because of CF although a BBC news caster suspected the serious accidents in Fukushima nuclear power plants and French government prepared free airplanes for the French people in Japan.

Example 2. Forgetting the certification of the safety in a new safe place at the arrival time. A family has evacuated to a safer place in Iitate village that locates about 60km away from the nuclear power plant. A father forgot to certify the safety of Iitate village at the arrival time. All the families were irradiated by strong radiation. Young women may not be able to have babies.

III. Discovery of precursor of serious damage by d-Library with Cross-field agent

Early evacuation in a safe direction in safe environments is the best evacuation. Early discovery of disaster precursor is very important for such safe evacuation.

Citizens need different types of organized knowledge and information distribution systems in order to easily understand disaster contexts, quickly start reasonable evacuation in a short time period and survive from the great disasters. Root and transfer routes of a disaster context decide the reliability of the disaster context and facilitate the easy understanding.

From the experiences of serious damages and after effect of the earthquakes, we propose "eight layered disaster context" over multiple academic fields and disaster context oriented disaster library "d-Library" (Fig. 1) [11]; so, as to enhance the disaster information systems in local governments. The aims of the eight layers are facilitating the reduction of serious damage not only by separate scientific knowledge, but also by coping with human nature and ability, social behaviors, and collaboration in the mutual assist societies, etc. Introduction of new technologies, such as enhancing collaborations by social networks and Big Data analysis are also the aims.

The eight layers are (a) Collective unconsciousness, (c) General Knowledge in home country, (d) Good ideas in the other areas/foreign countries, (e) Big Data analysis, (f)

 ISBN: 978-1-61208-329-2

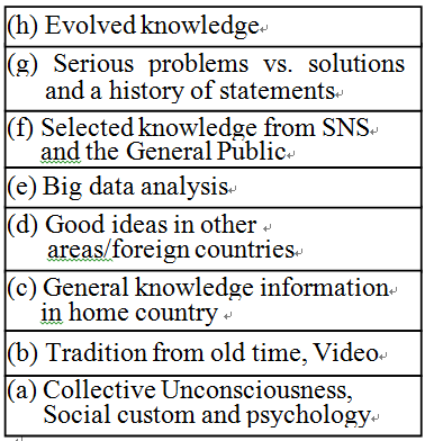


Figure 1. Eight layered disaster context base "d-Library." Problem/solutions and history of statements are kept in (g)

Selected knowledge from SNS and the general public, (g) Serious problems with solutions and a history of statements (PS), and (h) Evolved knowledge (EK). Ordinary citizens can understand a target disaster context and the reason of effectiveness of practical actions in each layer. Big Data analysis often shows unpredictable important causes against human's bias. Separate academic knowledge with some conditions belongs to layer (c) "General knowledge information in home country" in case of the great disasters. For example "a nuclear power plant is safe from the point of killing a person" belongs to layer (c). Effective knowledge in any disasters belongs to layer (h) "Evolved knowledge."

The Japanese government and the local governments have information sharing system for disaster prevention. d-Library should be introduced in the information sharing system.

d-Library consists of two kinds of feature (Fig. 2): (i) Additional database in the information sharing system, and (ii) a report from users and the mutual assist society in disaster time.

(i) Additional database in information sharing systems of a government. Tables in the additional database include "description of context with key words," "raw information in text, picture, video, multimedia, and AR type of multimedia data, etc.," and "URL: link to information in the internet, such as URL of information" since copy right of the information is owned by the other person in the world.

(ii) A report from users and the mutual assist society in disaster time. A contributor can easily make an article with a name of original contributor in a blog site and contribute it with the URL of the article to a bulletin board or a mail auto-delivery system of the information sharing system. Even older people can easily acquire the latest information in real time since the mail auto-delivery system enables push type of information distribution. Administrators of the certified layer like layer (h) should check articles and deliver them to the registered people.

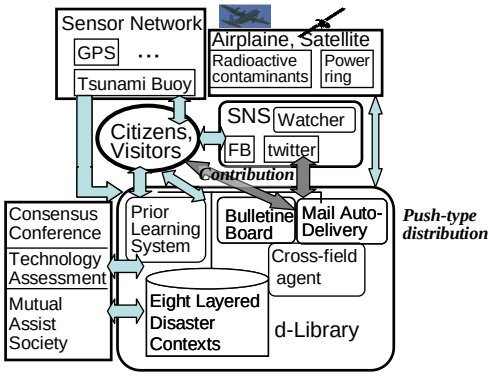


Figure 2. The citizens share eight layered disaster contexts in real time through bulletine board or mail auto-delivery

Dangerous areas with disaster contexts should also be opened for the citizen since citizens can neither watch physical phenomenon nor understand serious risk in them.

We propose "cross-field agent" in disaster context library (d-Library) in order to safely guide the citizens to safe places and reconstruct home town with pleasant lifestyle.

Cross-field agents provide the citizens with the eleven assists by useful disaster contexts and practical actions. Useful disaster contexts enable citizens to find clues of practical actions and a right decision to survive.

Example 2. There are a lot of small towns in saw tooth coastline in the east Japan. Disaster contexts have constraints in human actions by the small plains, a few main streets and a few large bridges [3]. The bridges caused the traffic jams. Residents should not use automobiles in case of giant tsunami. They might adaptively drive to the hillside and walk up to a higher ground when they notice the jams.

Cross-field agent provides citizens with useful disaster contexts for either a great disaster or a serious accident and allows them to timely find a safe way of evacuation corresponding to the change of disaster environments. Cross-field agent shows the user the tsunami warning with a picture of evacuation in order to make an immediate decision to leave the house and go to a higher ground in case of giant tsunamis (Fig. 3). The resident can know the serious situation that a friend near beachside has already left his/her house to survive when s/he accesses the disaster web-site. The picture also shows the resident that a friend is running up the stairs in an evacuation route. The resident reconsiders the decision of leaving the house although COF would happen in himself/herself.

"Risk Grade" of a disaster context should be decided in six levels. (rg0) safe, (rg1) safe with some dangers, (rg3) dangerous staying in a safe place, (rg4) evacuation to a place of refuge, (rg5) evacuation to a place of secondary refuge/ dangerous staying in a dangerous place, and (rg6) evacuation to the outside of a place of secondary refuge. We should check current risk grade and the history of it. The grade and the history contribute to the decision making.

Cross-field agent can become a reliable disaster guide in emergency time. It reminds the important problems with solutions if COF would force citizens to forget the

 ISBN: 978-1-61208-329-2

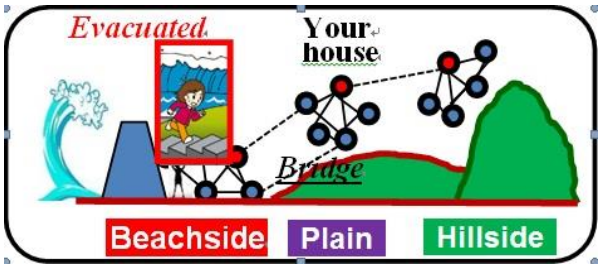


Figure 3. A picture of tsunami warning shows that a friend in the beachside has already left the house

important problems. The field-agent also gives citizen awareness and a chance to ensure the safety that s/he selected. Well experienced people and good watchers could find the danger and its solution with maps that shows the serious risks. Cross-field agents also give citizen a chance to acquire dynamic change of situations and to ensure the safety that citizen selected.

IV. PRACTICE OF DECISION MAKING AND CHECKING SAFETY

4.1 Practicing decision making to survive

During the Great East Japan Earthquake in 2011, about 70 percent of residents did not evacuate because of Normalcy Bias [11], although they knew the tsunami warning of the government in text and reading out by an announcer. We tested the efficiency of tsunami warning with a picture. A user of d-Library can acquire the real-time information by one pushing "Where is people?" icon in a window of d-Library. Field agent shows whether the residents in a zone have started evacuation to a safe place or not (Fig. 3, 5) based on the actions of personal examiners or community/area examiners that a user of d-Library chosen. The user can also know which zone the other residents started moving to a safe place.

We built a prototype of safety check system by web base system and tested whether a subject feel the danger of tsunami such that s/he has to leave his/her house.

Conditions C0. The subject lives near a beachside. The tsunami warnings have already been issued. Resident A lives in a house that locates between the house of the subject and the beachside. Resident B lives in a house that locates on higher ground than the ground of the subject's house. The house of resident B locates farer from the beachside than that of resident A.

Devices. (1) TV, radio and mobile phone, (2) Cross-field agent stand next the subject

Picture (p1). A friend near the beachside has already evacuated to survive (Fig. 3).

Case 1. The tsunami warnings in text were shown on a screen of TV and an announcer of TV and radio tells the residents the tsunami warnings.

Case 2. A subject accessed a disaster web-site. S/he watches the picture (p1) of the tsunami warning and hears the voice guide of the tsunami warning.

Subjects:

Group 1. 16 persons between 23 and 65 years old.
Group 2. 45 persons between 18 and 22 years old.

Test 1. Difference of tsunami warnings between in text with reading out and in picture.

Operations O2.

(1) A subject hears an explanation of the conditions.

(2) A subject hears an explanation of case 2 and watch a picture of evacuation.

(3) A subject hears an explanation of case1 and reminds the scene of the tsunami warning on a screen of TV and voice reading of sentences in the warning. Japanese often watch such TV screen in daily life.

Question 1. "Which case do you feel a mortal danger in?"

Answer 1. (Fig.4)

Group 1. Case 1: 1, Case 2: 15, both cases: 0 (person)

Group 2. Case 1: 2, Case 2: 41, both cases: 2 (person)

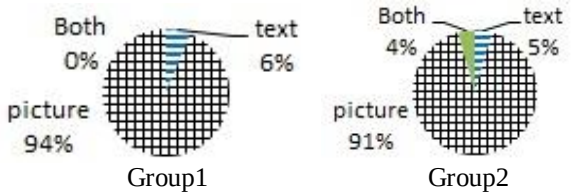


Figure 4. Effective warning in text or picture

Interview 1. Why do you feel the mortal danger in the case1 that you chose?

Answers in Case 1.

Group 1. (a1) One aged person trusts the voice reading out of the warning by an announcer.

Group 2. (a2) Two subjects trust the warnings of JMA and could feel the mortal danger from the tsunami warning on TV.

Answers in Case 2.

Group 1. (a3) 15 subjects felt the mortal danger in the picture since they were accustomed to watch the warnings of JMA and cannot feel the mortal danger in the warnings on TV etc.

Group 2.

(a4) 41 subjects answered as same as (a3).

(a5) A picture showed an image of concrete scene of evacuation by the other resident.

(a6) A picture let me know that there was a person that had already evacuated.

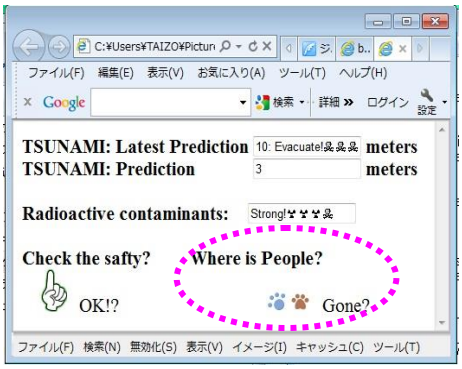
(a7) Most subjects were accustomed to see the tsunami warning and have never been attacked by the great tsunami in their life for more than 18 years.

(a8) A woman did not trust the warning and tried to check actions of the neighbors before the evacuation.

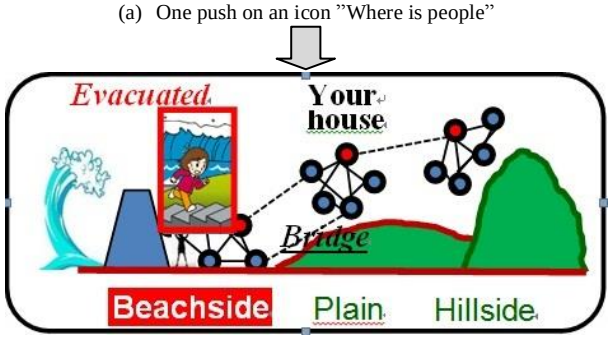
(a9) Most subjects could make a concrete image of evacuation from the picture. The picture produced a strong motivation in subject's mind in case 2. All subjects decided to immediately leave his/her house to survive.

Answers in the both cases. (a10) Two subjects basically evacuate anytime when they hear the word of "tsunami," since they lives near the beachside and the plain near the beachside is small. They could guard their lives against any kinds of tsunamis.

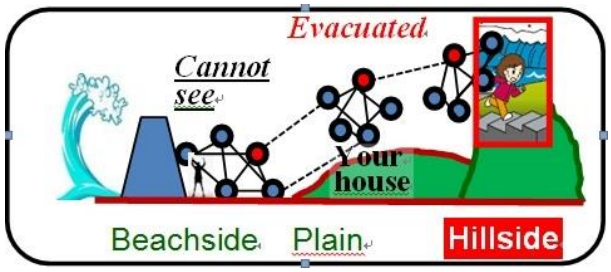
 ISBN: 978-1-61208-329-2



(a) One push on an icon "Where is people"



(p1) A friend in the beachside has already left the house



(p2) A friend in the hillside has already left the house

Figure 5. A user can know the latest action of the other residents by only one push on "Where is People?"

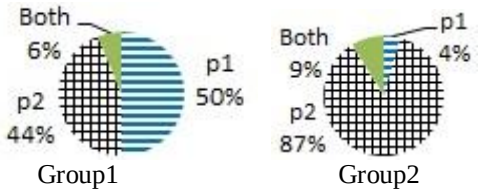


Figure 6. Useful pictures and embedded danger

Test2. Difference between tsunami warning from a beachside friend and that from a hillside friend.
Pictures. (Fig.5)
(p1) A picture shows that a friend near the beachside has already evacuated to survive.
(p2) A picture shows that a friend near the hillside has already evacuated to survive.
Operations O2.
(1) A subject hears situations of the tsunami warning with a picture.
(2) A subject watched two pictures for the tsunami warnings.
Question 2. "Which picture do you feel more mortal danger?"
Answer 1. (Fig.6)
Group 1. p 1: 8, p 2: 7, both p1 and P2: 1 (person)
Group 2. p 1: 2, p 2: 39, both p1 and P2: 4 (person)
Interview1. Why did you feel more mortal danger in the picture that you chose?
Answers for interview 1.
Group 1. Eight subjects thought that residents in the beachside would be the first persons that felt the danger.
(a12) Seven subjects thought that residents in the hillside could find some dangers on a higher ground. They could find the black wall of the great tsunamis coming.
(a13) One subject thought that any tsunami is dangerous.
Group 2. (a13) Only two subjects thought as same as a11.
(a14) 39 subjects thought as same as a12. Some of them thought that it would be too late for residents in the beachside to evacuate. They would be swallowed by the great tsunami.
Discussion. We confirmed that no serious damage for 37 years has made most subjects be accustomed to the tsunami warning and feel no mortal danger in the tsunami warning. We found that a good picture could distribute the effective warning with dangerous situations and contexts in Japan. Japanese subjects could not only easily extract evacuation actions by a resident and make an image of reasonable situation for the surroundings, but also embedded dangerous situations. All subjects except one subject decided to immediate evacuation by both pictures p1 and p2. A female subject in 40's answered that she would watch the surroundings and reconsidered the evacuation. To our surprise, most young subjects in group2 could not only understand the pictures, but also analyze the embedded danger from actions by a friend in the hillside instead of the actions by a friend in the beachside (Fig.6). We should prepare two kinds of good pictures for both young people and ordinary citizen.

4.2 Application of practicing safety check in d-library

Objective analysis of the serious damages by great disasters should be supplied citizens since citizens cannot afford to find the useful information in such an emergency time for example, The Great East Japan Earthquake in 2011, citizens experiencing earthquakes, Tsunami, even nuclear leakage. We proposed "cross-field agent" that shows awareness of checking the safety for citizens in order to just push the icon of easy safety check "E-Safe". E-Safe icon allows a user to acquire objective information and many kinds of risk map like radioactive contaminants map of the Energy Department of State in the U.S. from the outside of the interested parties or from foreign countries by one push of an icon (Fig. 7). The tsunami warning changed the prediction of height of tsunami twice "from 3 meters to 6 meters" and "from 6 meters to more than 10 meters." A citizen should also be given awareness of these two changes

 ISBN: 978-1-61208-329-2

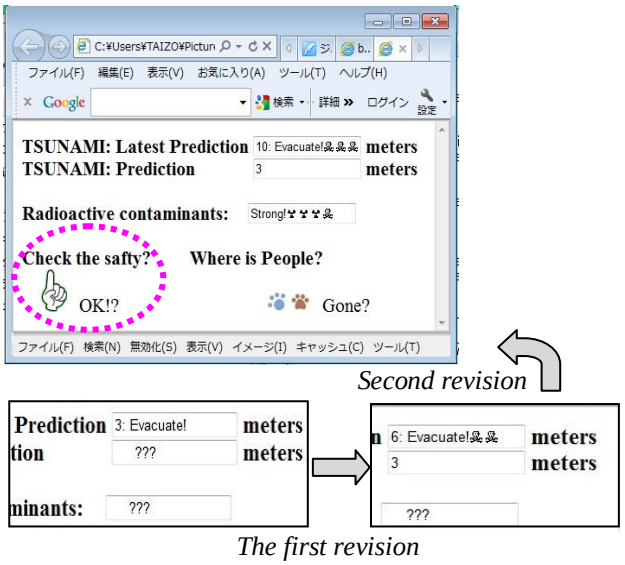


Figure 7. One push of an icon "Check the safety" and two revisions of the tsunami warning

by field-agent in d-Library and only push the button for the check. A user of d-Library could safely evacuate from the great tsunami. S/he also could know the latest right information and objective information like the map of radioactive contaminants from foreign countries etc. A citizen could avoid the potential strong nuclear leakage and easily evacuate to a safe place in the safe direction if s/he would easily know the map (Fig. 8) and ensure his/her safety by E-Safe when s/he would worry his/her own situation.

V. CONCLUSION

We described the importance of learning disaster contexts by referring the case of The Great East Japan Earthquake in 2011, a sequence of great disasters and serious accidents in Fukushima nuclear leakage, etc. We proposed eight layered disaster context library "d-Library" with cross-field agent for citizens in order to quickly evacuate to a safe place and ensure their own safety after the evacuation. We also discussed how to evacuate avoiding psychological difficulties as catastrophic occurrence forgetting. "Cross-field agent" gives citizens a chance to find precursors of serious damages and find additional solutions for the disaster emergency situations by collaborations in the Internet.

We also discuss a warning method with pictures that enabled a citizen to avoid the psychological difficulties and reminds important problems as the safety check.

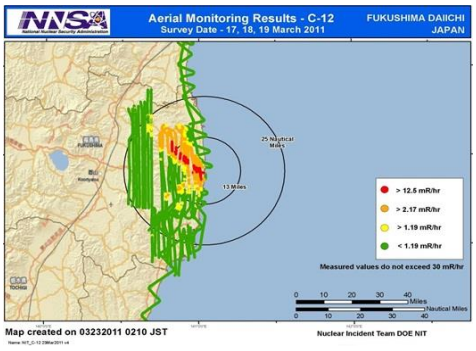


Figure 8. The map of radioactive the contaminants sent by US embassy [12]

REFERENCES

[1] F. Imamura, et al., "Irian Jaya earthquake and tsunami cause serious damage," Eos Trans. AGU, 78(19), 1997, pp. 197-204, doi:10.1029/97EO00128.

[2] S. Nomura, et al., "Mortality risk amongst nursing home residents evacuated after the Fukushima nuclear accident," "http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0060192," February 26, 2013, [retrieved: January, 2014].

[3] NHK, "Great tsunami, How people moved at the great tsunami," NHK special, 2nd, Oct. 2011, [retrieved: January, 2014].

[4] MIC Japan, "Information Sharing for Disaster Prevention," "http://www.soumu.go.jp/main_sosiki/joho_tsusin/top/local_support/ict/jirei/thema10.html," [retrieved: January, 2014].

[5] T. Miyachi, G. Buribayeva, S. Iga, and T. Furuhata, "A Study of Disaster Library System with a Field Agent to Learn a Sequence of Great Disasters," IWIN2013, Sept. 2013, pp. 91-97

[6] Shinshu radioactivity lab BLOG, "http://imeasure.cocolog-nifty.com/isotope/2013/11/post-a012.html," [retrieved: Febrary, 2014].

[7] M. McCloskey, and N. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem." in G. H. Bower (ed.) The Psychology of Learning and Motivation: NY: Academic Press, vol. 24, 1989, pp. 109-164.

[8] T. Katada, and M. Kanai, "Implementation of Tsunami Disaster Education for Children and Their Parents at Elementary School," Solutions to Coastal Disasters 2008: 2008, pp. 39-48 [9] N. Mikami, "Challenges in the Practical Stage of Participatory Technology Assessment: From the Experience of GM Consensus Conference in Hokkaido," Japanese Journal of Science Communication, No.1, 2007, pp. 84-95.

[10] T. Suzuki, "My wife was waving her hands・・・," Talk about 3.11, KHB East Japan Broadcasting, vol. 1, 2012.

[11] New Jersey Personal Defense Academy, "Normalcy Bias: Cause, Effect, Cure," "http://njpersonaldefense. com/normalcy-bias-cause-effect-cure/," [retrieved: January, 2014].

[12] U.S. Department of Energy, "Aerial Monitoring Results Cumulative," "http://energy.gov/downloads/radiation-monitoring-data-fukushima-area-32211," 25, March, 2011, [retrieved: Febrary, 2014].

Coefficient-Based Exact Approach for Frequent Itemset Hiding

Engin Leloglu
Dept. of Computer Engineering
Izmir Institute of Technology
Izmir, TURKEY
enginleloglu@ieee.org

Tolga Ayav
Dept. of Computer Engineering
Izmir Institute of Technology
Izmir, TURKEY
tolgaayav@iyte.edu.tr

Belgin Ergenc
Dept. of Computer Engineering
Izmir Institute of Technology
Izmir, TURKEY
belginergenc@iyte.edu.tr

Abstract—**Concealing sensitive relationships before sharing a database is of utmost importance in many circumstances. This implies to hide the frequent itemsets corresponding to sensitive association rules by removing some items of the database. Research efforts generally aim at finding out more effective methods in terms of convenience, execution time and side-effect. This paper presents a practical approach for hiding sensitive patterns while allowing as much nonsensitive patterns as possible in the sanitized database. We model the itemset hiding problem as integer programming whereas the objective coefficients allow finding out a solution with minimum loss of nonsensitive itemsets. We evaluate our method using three real datasets and compared the results with a previous work. The results show that information loss is dramatically minimized without sacrificing the accuracy.**

Keywords—frequent itemset hiding; exact approach; information loss

I. INTRODUCTION

Progresses in the technology give an opportunity to establish transactional databases that can reserve large volumes of data. Analyzing data and extracting meaningful information from these huge piles of data come up as a result of these advances. Data mining field has efficient techniques for this knowledge discovery process. However, improper use of these techniques caused a rise of privacy concerns. Unauthorized access to not only sensitive personal information that is stored or inferred from the data, but also commercial information that provides remarkable benefit over rivals induces privacy issues. That is why comprehensive sanitization on databases is required when the information or data from these databases is shared or published.

Sharing databases allows researchers and policy-makers to examine the data and gain significant information benefiting the society as a whole, such as the strength of a medicine or treatment, social-economic inferences that can be the guide on the road to efficient public policies, and the factors that cause vital diseases. In other words, publishing databases eventuates in utility gain for the society as a whole [13]. However, due to privacy concerns, a privacy preserving method is needed to be applied on the databases. These methods make data imprecise and/or distorted so that no sensitive knowledge is disclosed. But, this distortion causes unwanted information loss and losses in potential utility gain.

Frequent itemset hiding is one of the important and widely used methods of privacy preserving data mining field. There are several frequent itemset hiding algorithms of which methodology can be classified, such as heuristic [6], border-based [16, 17] and exact [8, 10, 11, 14]. They aim to impose small deviation in the original database to expose no sensitive itemsets. This deviation is tried to be minimized by various techniques with different quality metrics as a common feature of these algorithms. One determines the relative frequency of remaining itemsets [16] as a quality parameter while another approach uses the term of accuracy [14] that shows the impact of sanitization on transactions of the database. In addition to this, the information loss which is to conceal nonsensitive itemsets on the original database while hiding sensitive knowledge is another critical point of the hiding process [16]. Studies generally concentrate on achieving the result database which has no sensitive knowledge with small deviation and minimum information loss.

Exact approaches produce more accurate solution than other types of approaches in frequent itemset hiding. However, they are impractical when the number of itemsets and length of itemsets increase. In addition to this, they mostly focus on minimizing deviation in terms of accuracy or distance. To our knowledge there is no practical solution providing frequent itemset hiding with the objective of minimum information loss and accuracy. In this paper, we propose an exact approach for frequent itemset hiding where all sensitive patterns are concealed. Our approach is based on the combination of integer programming and heuristic sanitization. While it prevents revealing sensitive information on published database, minimum information loss and maximum accuracy are also provided.

Our approach proposes the use of coefficients in the objective function of integer programming to minimize information loss. These coefficients reminding the approaches used by utility-based mining algorithms [19] are pre-computed such that they give a measure of information loss. Integer programming allows finding the optimum solution deciding about the transactions to be sanitized. Then, heuristic sanitization algorithm is executed to remove the sensitive itemsets. The experiments with real datasets demonstrate the efficacy of our approach and give useful insight into the efforts of minimizing nonsensitive information loss.

 ISBN: 978-1-61208-329-2

TABLE I. EXAMPLE DATABASE $\mathscr{D}$

Id	Items
T_1	1 2 3 7 8 10
T_2	3 9 10
T_3	4 5 6
T_4	1 2 3 6 7 8 9
T_5	1 2 3 6 7
T_6	10
T_7	4
T_8	3 6 7 8 9
T_9	3 8 9
T_{10}	5 6 7

TABLE II. FREQUENT (NONSINGLETON) ITEMSETS FOR $\mathscr{D}$ AT $\sigma_{min} = 2$

Itemsets	σ_j	Itemsets	σ_j	Itemsets	σ_j	Itemsets	σ_j
5, 6	2	9, 6	2	2, 8, 7	2	1, 2, 8, 7	2
1, 2	3	9, 7	2	2, 8, 3	2	1, 2, 8, 3	2
1, 8	2	9, 3	4	2, 6, 7	2	1, 2, 6, 7	2
1, 6	2	r_3 →**6, 7**	**4**	2, 6, 3	2	1, 2, 6, 3	2
1, 7	3	6, 3	3	2, 7, 3	3	1, 2, 7, 3	3
1, 3	3	7, 3	4	8, 9, 6	2	1, 8, 7, 3	2
2, 8	2	1, 2, 8	2	8, 9, 7	2	1, 6, 7, 3	2
2, 6	2	1, 2, 6	2	8, 9, 3	3	2, 8, 7, 3	2
2, 7	3	1, 2, 7	3	8, 6, 7	2	2, 6, 7, 3	2
2, 3	3	r_4 →**1, 2, 3**	**3**	8, 6, 3	2	8, 9, 6, 7	2
10, 3	2	1, 8, 7	2	8, 7, 3	3	8, 9, 6, 3	2
r_1 →**8, 9**	**3**	1, 8, 3	2	9, 6, 7	2	8, 9, 7, 3	2
8, 6	2	1, 6, 7	2	9, 6, 3	2	8, 6, 7, 3	2
8, 7	3	1, 6, 3	2	9, 7, 3	2	9, 6, 7, 3	2
r_2 →**8, 3**	**4**	1, 7, 3	3	6, 7, 3	3	1, 2, 8, 7, 3	2
						1, 2, 6, 7, 3	2
						8, 9, 6, 7, 3	2

The following sections are organized as follows: Section 2 gives the background of the problem with terms, concepts and considerations. Section 3 presents our approach in detail. Section 4 is an overview of the leading studies about privacy preserving data mining. Section 5 shows the results and evaluations of the experiments to prove the effectiveness of the technique we propose. Finally, we conclude in Section 6.

II. BACKGROUND

Let $\mathscr{I}$ be a set of items. An itemset is a subset of $\mathscr{I}$ and any transaction defined over $\mathscr{I}$ is tuple $< k, \mathscr{I}_k >$, where k is the transaction id and $\mathscr{I}_k$ is the itemset. A transaction $< k, \mathscr{I}_k >$ is said to contain an itemset X iff $\mathscr{I}_k \supseteq X$. A database $\mathscr{D}$ is a set of transactions. Given a database $\mathscr{D}$, the support of an itemset $\mathscr{I}_k$ in the database $\mathscr{D}$ is denoted as *the support* $\sigma(\mathscr{I}_k, \mathscr{D})$. $\sigma(\mathscr{I}_k, \mathscr{D})$ can be represented simply as σ_k for notational convenience. For a given threshold σ_{min}, $\mathscr{I}_k$ is said to be *frequent* if $\sigma(\mathscr{I}_k, \mathscr{D}) \geq \sigma_{min}$. The set of frequent itemsets $\mathscr{F}(\sigma_{min})$ at minimum support level σ_{min} is the set of all itemsets with a minimum support of σ_{min}.

$\mathscr{F}^R(\sigma_{min}) \subseteq \mathscr{F}(\sigma_{min})$ is a group of restrictive patterns that the owner of the data would like to conceal while publishing. A transaction that supports any of these patterns is said to be *sanitized* if any alteration is made on it in such a way that it no longer supports any itemset in $\mathscr{F}^R(\sigma_{min})$. This sanitization implies reducing the support for every $j \in \mathscr{F}^R(\sigma_{min})$ below σ_{min} and concealing itemset j.

In the process of transforming a database $\mathscr{D}$ to a sanitized $\mathscr{D}'$, we have the following considerations:

1) Suppose that $\mathscr{F}'(\sigma_{min})$ be the set of frequent itemsets in the sanitized $\mathscr{D}'$. Any $j \in \mathscr{F}^R(\sigma_{min})$ in $\mathscr{D}$ should not be in $\mathscr{F}'(\sigma_{min})$. In other words, it is aimed that no sensitive knowledge is involved in the sanitized database.

2) The accuracy, which is the ratio of the number of transactions that are not sanitized and the total number of transactions in the database $\mathscr{D}$, should be maximized by keeping the number of sanitized transactions at minimum.

3) Suppose that $\mathscr{F}^n(\sigma_{min})$ be the set of non-sensitive frequent itemsets determined by $\mathscr{F}(\sigma_{min})/\mathscr{F}^R(\sigma_{min})$ in database $\mathscr{D}$. $|\mathscr{F}^n(\sigma_{min}) - \mathscr{F}'(\sigma_{min})|$ should be minimized to avoid overconcealing nonsensitive frequent itemsets and keeping the information loss at minimum.

Table 1 represents a database $\mathscr{D}$ which includes 10 transactions and 10 items. Nonsingleton frequent itemsets with the support values bigger than or equal to 2 are listed in Table 2. For example, we assume that the sensitive patterns are {8, 9}, {8, 3}, {6, 7}, {1, 2, 3} that are bold and represented with r_1, r_2, r_3 and r_4. Although, it is possible to define different support thresholds for each sensitive pattern, we assume that the support threshold $\sigma_{min} = 2$ for all patterns, which is practical and common in many circumstances. At the end of the process, we expect that 28 nonsensitive frequent itemsets that are supersets of the sensitive ones would also get concealed as the process of hiding the sensitive itemsets. The question is how to transform $\mathscr{D}$ into the sanitized database $\mathscr{D}'$ in an effective way such that aforementioned considerations 1, 2 and 3 are maintained.

Depending on the consideration 1, support values of our sensitive patterns in the database $\mathscr{D}$ should be dropped below 2, that is minimum support value. For example, the support value of r_2 is 4 and to satisfy the consideration 1, transactions that include r_2 should be found and at least one of two items in r_2 should be deleted from as many transactions as needed. The proper selection of transactions to be sanitized and the items to be removed is of paramount importance, since the number of sanitized transactions and/or the number of items to be removed should be kept at minimum. Moreover, nonsensitive itemsets that contain one of items, 8 or 3, are in danger of being concealed while the support value of r_2 is decreased.

 ISBN: 978-1-61208-329-2

According to consideration 3, the number of nonsensitive itemsets that are concealed should be at minimum.

III. COEFFICIENT-BASED ITEMSET HIDING

In this section, we introduce a novel method for the itemset hiding problem. We first define the problem using integer programming and then simply augment the objective function with coefficients in order to reduce the information loss. Thus, the method consists of three essential parts: Coefficient Computation, Integer Programming Solution and Heuristic Sanitization.

Modeling the itemset hiding problem with integer programming can be done in several ways. We follow the way of Menon et al. [14] such that the objective achieves the maximum accuracy. Our method alters the objective function such that the binary variables indicating whether a transaction is chosen or not are multiplied by some pre-computed coefficients that reflect the amount of information loss. We compute the coefficients of transactions that support sensitive patterns only.

We first start with creating the constraint matrix by eliminating transactions, which do not support sensitive itemsets from consideration, as shown below:

$$\begin{pmatrix} & t_1 & t_2 & t_3 & t_4 & t_5 & t_6 \\ r_1 & 0 & 1 & 0 & 1 & 1 & 0 \\ r_2 & 1 & 1 & 0 & 1 & 1 & 0 \\ r_3 & 0 & 1 & 1 & 1 & 0 & 1 \\ r_4 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad (1)$$

6 columns represents respectively: $t_1 \rightarrow T_1, t_2 \rightarrow T_4, t_3 \rightarrow T_5, t_4 \rightarrow T_8, t_5 \rightarrow T_9, t_6 \rightarrow T_{10}$. While t_1 supports sensitive itemsets $r_2$and r_4, t_6 contains only one sensitive itemset that is r_3.

A. Coefficient Computation

To minimize the impact on nonsensitive frequent itemsets, coefficient computation is made for each transaction which supports sensitive patterns as the first step. A coefficient of the transaction gives the information about a risk of overconcealing nonsensitive frequent itemsets on this transaction. If the value of coefficient is high, it means that the number of concealed nonsensitive itemsets included in the transaction would be high after the sanitization.

This coefficient computation is typically organized by taking into account that initial utility worth of each nonsensitive itemsets on studied database is the same. Calculating a risk of overconcealing nonsensitive frequent itemsets is made based on this assumption. If some information on the database has different worth of utility based on the area where the shared database is utilized, relevant coefficients can be computed independently of Coefficient Computation Algorithm thereby paying regard to requirements of the area.

1: **for** transactions $i \in \mathscr{D}$ such that i is to be sanitized
2: **identify** all sensitive frequent item sets $\mathscr{F}_i^R \in \mathscr{F}^R(\sigma_{min})$ supported by i
3: **identify** all nonsensitive frequent item sets $\mathscr{F}_i^n \in \mathscr{F}^n(\sigma_{min})$ supported by i
4: **while** $\mathscr{F}_i^R \neq \emptyset$
5: **calculate** $f_j = |\{k \in \mathscr{F}_i^R \mid j \in k\}|$, $\forall$ items j in i
6: **calculate** $j^* = argmax_j\{f_j\}$
7: **calculate** $g_j = |\{k \in \mathscr{F}_i^n \mid j^* \in k\}|$, $\forall$ items j in i
8: **update** coefficient $c_i = c_i + g_j$
9: **update** $\mathscr{F}_i^R = \mathscr{F}_i^R \backslash \{k \in \mathscr{F}_i^R | j^* \in k\}$
10: **end while**
11: **end for**

Fig. 1. Coefficient Computation Algorithm.

A coefficient, which is represented by the Coefficient Computation Algorithm in Figure 1 as "c_i", is calculated for each transaction that is included in the constraint matrix. For example, we may choose transaction T_1 to explain this calculation (on line 1). First, sensitive frequent itemsets and nonsensitive frequent itemsets are identified for transaction T_1 (on line 2 and 3). The item appearing in the most number of sensitive patterns supported by that transaction is selected from all items in $r_2 \cup r_4$ (Items 1, 2, 3, 8) (on line 5 and 6). The item "3" is selected. Record the number of appearances of the item "3" in the non-sensitive frequent itemsets supported by the transaction (on line 7). For our example, there are 6 appearances of item "3" in the nonsensitive frequent itemsets such as {1, 3}, {2, 3}, {10, 3}, {7, 3}, {1, 7, 3}, {2, 7, 3}. Remove all sensitive itemsets supported by the transaction contain the selected item "3" (on line 9). If sensitive itemsets remain supported by T_1, repeat the procedure (on line 4) and sum the appearances of new selected item in the nonsensitive frequent itemsets supported by the transaction with recorded value (on line 8). There is no sensitive itemset left in our example. Hence, the total summation for transaction T_1 is 6. After, all transactions which are included in the constraint matrix are taken in consideration based on this procedure, a coefficient for each transaction is found such as $T_1 \rightarrow 6$, $T_4 \rightarrow 29, T_5 \rightarrow 14, T_8 \rightarrow 6, T_9 \rightarrow 0, T_{10} \rightarrow 1$.

B. Integer Programming Solution

In this section, we describe the integer programming formulation to solve Coefficient-Based Itemset Hiding problem. Initally, give a_{ij} a binary value. Be 1 if transaction $i \in \mathscr{D}$supports itemset $j \in \mathscr{F}^R(\sigma_{min})$. Otherwise, the value of a_{ij} is 0. For the variable x_i, it will be set to 1 if transaction $i \in \mathscr{D}$ is sanitized. Otherwise, the value of x_i is 0. σ_j represents the current support for itemset $j \in \mathscr{F}(\sigma_{min})$. Recall that c_i is the coefficient that is calculated in Coefficient Computation for transaction $i \in \mathscr{D}$, which contains at least one sensitive itemset.

In the light of this information, the formulation is generated as below:

 ISBN: 978-1-61208-329-2

$$min \sum_{i \in \mathscr{D}} c_i x_i, \quad (2)$$

$$s.t. \sum_{i \in \mathscr{D}} a_{ij} x_i \geq \sigma_j - \sigma_{min} + 1 \qquad \forall j \in \mathscr{F}^R(\sigma_{min}), \quad (3)$$

$$x_i \in \{0,1\} \ \forall i \in \mathscr{D}. \quad (4)$$

Equation (2) represents the objective function that minimizes the number of transactions sanitized. Equation (3) includes the constraint that more than $(\sigma_j - \sigma_{min})$ transactions supporting each sensitive itemset have to be sanitized, that's why this line is generated for each sensitive itemset. Equation (4) imposes that x_i has only binary value. The integer programming formulation is reorganized based on the constraint matrix (1) and coefficients that are gained by Coefficient Computation in the previous section as:

$$min\ 6x_1 + 29x_2 + 14x_3 + 6x_4 + 0x_5 + 1x_6, \quad (5)$$

$$s.t. \quad x_2 + x_4 + x_5 \geq 2, \quad (6)$$

$$x_1 + x_2 + x_4 + x_5 \geq 3, \quad (7)$$

$$x_2 + x_3 + x_4 + x_6 \geq 3, \quad (8)$$

$$x_1 + x_2 + x_3 \geq 2, \quad (9)$$

$$x_1, x_2, x_3, x_4, x_5, x_6 \in \{0,1\}. \quad (10)$$

Solving this integer program results in optimal solution $x_1 = x_3 = x_4 = x_5 = x_6 = 1$ with the other variables being 0. The accuracy of the resulting sanitized database is 0.50.

C. Heuristic Sanitization

Sanitization is a kind of process that includes removing items from a transaction; thereby the sanitized version of the transaction supports no itemset in $\mathscr{F}^R(\sigma_{min})$. There are various sanitization approaches in the privacy preserving data mining literature. For instance, Verykios et al. offered two sanitization techniques in their study [18]. These are generally based on hiding itemsets that are already sorted with respect to their size and support, in a different fashion such as one-by-one and round-robin. Amiri [1] presented the Aggregate Algorithm based on removing the most sensitive and the least nonsensitive itemsets in selected transaction. The process is repeated until all the sensitive itemsets are hidden. Furthermore, three item restriction-based algorithms [15] that are known as Minimum Frequency Item Algorithm (MinFIA), Maximum Frequency Item Algorithm (MaxFIA) and Item Grouping Algorithm (IGA) selectively remove items from transactions that support the sensitive itemsets. Intelligent sanitization in the paper [14] is the variant of their IGA.

We do not focus on the development of the sanitization techniques. Since we compare our new method with the study of Menon et al. [14], we prefer to use one of heuristics in their study. One is blanket sanitization where only one item is retained from the original transaction. The sanitization occurs by eliminating support for every nonsingleton itemset supported by the transaction. The other is intelligent sanitization, where an attempt is made to remove the fewest number of items from the transaction that would result in eliminating the support for every itemset in $\mathscr{F}^R(\sigma_{min})$.

1: **for** transactions $i \in \mathscr{D}$ such that i is to be sanitized
2: **identify** all sensitive frequent item sets $\mathscr{F}_i^R \in \mathscr{F}^R(\sigma_{min})$ supported by i
3: **while** $\mathscr{F}_i^R \neq \emptyset$
4: **calculate** $f_j = |\{k \in \mathscr{F}_i^R \mid j \in k\}|$, ∀ items j in i
5: **remove** item$j^* = argmax_j\{f_j\} \in m$
6: **update** $\mathscr{F}_i^R = \mathscr{F}_i^R \backslash \{k \in \mathscr{F}_i^R | j^* \in k\}$
7: **end while**
8: **end for**

Fig. 2. Intelligent Sanitization Algorithm [14].

It is shown that the intelligent sanitization produces less distortion on nonsensitive itemsets thereby removing less number of items when this is compared with the blanket sanitization. So, we prefer to use the intelligent sanitization to hide itemsets of transactions that are identified by the method described in the previous section. In order to be self-contained, we give intelligent sanitization algorithm in Figure 2.

Let us explain this with an example; we choose transaction T_1 that is represented in the constraint matrix by t_1 (on line 1). T_1 supports two sensitive itemsets - r_2 and r_4 (on line 2). First, the item appearing in the most number of sensitive patterns supported by that transaction is selected from all items in $r_2 \cup r_4$ (Items 1, 2, 3, 8). For the example, "3" is selected, because it appears twice while each one of the others appears once. Delete "3" (on line 4 and 5) and remove all sensitive itemsets that contain selected item (on line 6). This action eliminates r_2 and r_4 at the same time. If sensitive itemsets remain supported by T_1, repeat the procedure (on line 3). For our example, there is no sensitive itemset left. The sanitized transaction is {1, 2, 7, 8, 10}. When all transaction are put in process, the sanitized database is generated with the number of modifications on the database $\mathscr{D}$ is 7. After the entire process, 17 itemsets that were previously frequent are still frequent whereas 13 itemsets that were frequent before the sanitization are no longer frequent.

IV. Performance Evaluation

We performed Coefficient-Based Itemset Hiding and the study of Menon et al. [14] on real datasets using different parameters such as number of sensitive itemsets and minimum support value. Our code was implemented in Java on a Windows 7 - PC with Intel Core i5, 2.67 GHz processor. We performed exact parts of the experiments by using GNU GLPK [12]. In this section, features of datasets, selected parameters and results are explained in detail.

A. The Datasets

All datasets we use in our experiments are available through Frequent Itemset Mining Implementations Repository - FIMI.

 ISBN: 978-1-61208-329-2

TABLE III. CHARACTERISTICS OF THE REAL DATASETS

Database name	Number of transactions	Number of items	Avg. trans. length	Number of nonsingleton frequent itemsets (support level used in the experiments)
kosarak	990,002	41,270	8.10	1,462 (0.5%)
retail	88,162	16,470	10.30	5,472 (0.1%)
				15,316 (0.05%)
mushroom	8,124	119	23.00	53,540 (20%)

The kosarak dataset, which is provided by Ferenc Bodon [4], is a very large dataset containing 990,002 sequences of click-stream data from a Hungarian on-line news portal. It has medium-level of sparsity and medium-level of density. The retail is a sparse dataset and was reported in Brijs et al. 1999 [5]. It includes the retail market basket data from an anonymous Belgian retail store. The mushroom, which was generated by Roberto Bayardo from the UCI datasets and PUMSB [3], has high-level of density. These datasets have different characteristics such as the number of transactions, varieties of items and level of sparsity - density. These variations contribute to our experiment and give a chance to measure the efficacy of our study. Table 3 includes summary information about these datasets.

B. Evaluation Methodology

We compare our approach with the approach of Menon et al. [14] in terms of the number of nonsensitive itemsets that are lost (information loss) and the ratio of the number of transactions that are not sanitized and the total number of transactions in the database (accuracy).

Execution times for coefficient computation (C), integer programming (IP) and heuristic sanitization (H) are separately recorded to maintain the total time for Coefficient-Based Itemset Hiding. Because the complexity is one of main problems for the privacy preserving data mining, the time is illustrated in detail in our experiments.

With our original sensitive itemsets, supersets of them are become hidden in the databases, since any itemsets that contains sensitive itemsets should also be hidden. Original sensitive itemsets are specified with various lengths, such as 10, 20 and 50. In addition, we use two different minimum support thresholds for the retail dataset to evaluate the impact of threshold.

C. Experimental Results

In Table 4, accuracy, time and information loss performances of Menon et al. [14] and Coefficient-Based Itemset Hiding are given. Table 5 summarizes the differences between two approaches. Negative values represent the sacrifices of our approach while positive values show outclass performance of our new method over the approach of Menon et al.

When the tables are carefully examined, it is deduced that our new method makes progress with different fluctuations based on the characteristics of databases used in experiments. Firstly, it can be noticed that the proposed method decreased the number of lost nonsensitive itemsets successfully for all kinds of databases in the experiments. However, it is obviously seen that our approach works more powerfully for sparse databases such as Retail when we compare it with the other databases in Table 4 and 5. Support level used in

TABLE IV. RESULTS FROM THE REAL DATA

DB name (σ_{min})	Sensitive itemsets (with supersets)	Approach of Menon et al.					Coefficient-Based Approach					
		(%)	Time (sec)			Itemsets(#)	(%)	Time (sec)				Itemsets(#)
		Accuracy	IP	H	Total	Info. Loss	Accuracy	C	IP	H	Total	Info. Loss
kosarak (4,950)	10 (18)	99.59	11.5	129	140.5	98	99.27	71	11.3	124	206.3	19
	20 (31)	99.23	27.7	126	153.7	182	98.5	141	67.7	116	324.7	57
	50 (65)	98.95	35,976.1	115	36,091	310	97	360	6,543	120	7,023	58
retail (88)	10 (10)	99.83	0	7	7	10	99.8	1	0.1	8	9.1	2
	20 (20)	99.6	0.1	8	8.1	61	99.42	1	0.1	8	9.1	10
	50 (65)	99.05	0.1	8	8.1	98	98.43	3	0.3	8	11.3	26
retail (44)	10 (15)	99.37	0.1	7	7.1	85	99.34	2	0.1	7	9.1	43
	20 (32)	98.77	0.1	8	8.1	335	98.54	4	0.1	8	12.1	196
	50 (97)	97.46	0.2	8	8.2	664	96.62	8	0.3	8	16.3	364
mushroom (1,625)	10 (2336)	93.4	0.1	1	1.1	19,984	93.4	89	0.1	1	90.1	19,584
	20 (2395)	93.16	0.2	1	1.2	33,049	84.94	114	0.5	1	115.5	26,791
	50 (5341)	92.32	0.3	1	1.3	35,831	79.47	141	0.8	1	142.8	31,149

 ISBN: 978-1-61208-329-2

TABLE V. DIFFERENCE OF TWO APPROACHES

DB name (σ_{min})	Sensitive itemsets (with supersets)	Perc. (%)		
		Accuracy	Total Time	Info. Loss
kosarak (4,950)	10 (18)	-0,32%	-46,83%	80,61%
	20 (31)	-0,74%	-111,26%	68,68%
	50 (65)	-1,97%	80,54%	81,29%
retail (88)	10 (10)	-0,03%	-30,00%	80,00%
	20 (20)	-0,18%	-12,35%	83,61%
	50 (65)	-0,63%	-39,51%	73,47%
retail (44)	10 (15)	-0,03%	-28,17%	49,41%
	20 (32)	-0,23%	-49,38%	41,49%
	50 (97)	-0,86%	-98,78%	45,18%
mushroom (1,625)	10 (2336)	0,00%	-8090,91%	2,00%
	20 (2395)	-8,82%	-9525,00%	18,94%
	50 (5341)	-13,92%	-10884,62%	13,07%

experiments may be another critical point. Because decreasing the support value of itemsets below low support level needs sacrificing utility gain, low support level reduces the coefficient benefit for information loss, as Retail performance result at the support level – 44 (0.05%). However, the best performance in all experiments is attained for Retail database at the level of support - 88 (0.10%).

The performance result of Kosarak database shows that new approach works well with databases which have medium sparsity and density. It has up to 80% gain on information loss while there is not above 2% accuracy loss. On other hand, since Kosarak is a very large database, total time cost is a general problem for integer programming solutions. Despite this, it is also remarkable that in case of 50 sensitive itemsets in Kosarak, our method is approximately six times better in execution time. This is quite reasonable since coefficients help branch and cut algorithms of integer programming [7] by allowing more cuts.

When the result of Mushroom database in Table 5 is examined, it is deduced that although better performance than the approach of Menon et al. [14] has is gained for the information loss, we meet undesirable accuracy loss and time cost. This result shows that in some situation such having a need of use very dense database like Mushroom, using the methods of Menon et al. [14] or different exact methods is more useful and produces better solutions.

V. RELATED WORK

One of the earlier studies, which presented the principles of privacy preserving data mining, belongs to Atallah et al. [2]. Their study proves that "association rule hiding" is NP-hard problem, due to the existence of large databases. After this research, there has been remarkable growth on the number of research on this issue recently. They are generally classified based on their proposed approach as heuristic, border-based and exact. In addition to these, as another branch of privacy preservation, the utility that involves the term of information loss has examined in detail in utility-based privacy preserving data mining.

Dasseni et al. [6] generalize the hiding problem in the sense that they consider the hiding of both sensitive frequent itemsets and sensitive association rules. The authors propose three single rule heuristic hiding algorithms that are based on the reduction of either the support or the confidence of the sensitive rules, but not both. In all three approaches, the goal is to hide the sensitive rules while minimally affecting the support of the nonsensitive itemsets. In order to achieve this, transactions are modified by removing some items, or inserting new items depending on the hiding strategy. Verykios et al. [18] extend the previous work of Dasseni et al. [6] by improving and evaluating the association rule hiding algorithms of [6] for their performance under different sizes of input datasets and different sets of sensitive rules. Oliveira and Zaïane [15] contribute to this area with a variety of heuristics. Particularly, The Item Grouping Algorithm is based on grouping sensitive association rules sharing the same itemsets. The minimum impact on the disclosed database is provided by deleting the shared items. The intelligent sanitization we use as the sanitization technique in our study is the variant of this algorithm. Amiri [1] presented three effective, multiple association rule hiding heuristics that outperform the previous heuristics studies by offering higher data utility and lower distortion, at the expense of increased computational speed. Although the algorithms by Amiri are similar in philosophy to the previous approaches, the three proposed methodologies do a better job in modeling the overall objective of a rule hiding algorithm.

The paper by Sun and Yu [16, 17] is a pioneer of border-based researches which use the border theory to hide frequent itemsets. It aims at maintaining the frequency of nonsensitive itemsets to minimize the side-effects and evaluate the impact on the result database. Gkoulalas-Divanis and Verykios used this border concept in their works [8, 10, 11] to minimize overconcealing nonsensitive itemsets. They capture the itemsets hiding process as a border revision operation and they presented a set of algorithms which enable the computation of the revised borders that pertain to an exact hiding solution.

The paper written by Menon et al. [14] includes an interesting approach to the problem of privacy preserving data mining. They were the first to present an integer programming optimization method that consisted of an exact and a heuristic part to hide frequent itemsets. The exact part of the method uses the database to formulate an integer program trying to obtain the minimum number of transactions that have to be sanitized. The researches of Gkoulalas-Divanis and Verykios [8, 9, 10, 11] are based on this exact methodology. However, they organize the integer program formulation in a way of identifying itemsets to hide, instead of transactions.

 ISBN: 978-1-61208-329-2

The information loss, which is considered a loss of utility for data mining purposes, has been examined in detail in another research area, utility-based privacy preservation. This broad approach can preserve considerable utility of the data set without violating privacy. Li and Li [13] mention the importance of utility gained by publishing database for the society as a whole and claim that it is inappropriate to directly compare privacy with utility, because of several reasons, including both technical and philosophical ones. Furthermore, they propose an integrated framework for considering privacy-utility tradeoff, borrowing concepts from the Modern Portfolio Theory for financial investment.

Although, heuristic approaches seem scalable and practical, their results are less reliable about being exact solution and providing minimum side-effect. This is not acceptable in many situations. Border-based methods give better solution on the sensitive itemset hiding and side-effect problems. However, the evaluation brings high complexity. The research of Menon et al. [14] makes progress in a way of getting exact solution. Using integer programming and heuristic together gives evaluated impact on the result data than border-based approaches give. However, it causes failing to notice side-effect of information loss. Verykios et al. present exact approaches which find a way to decrease loss of nonsensitive itemsets. But, the complexity of these approaches and not being scalable are the reasons of researching for better solution. Furthermore, Li and Li [13] and Yeh and Hsu [19] inspired us that the utility that is directly about information loss is essential for frequent itemset hiding. We realized that solutions in the literature to the utility loss problem cannot always satisfy the need of databases at different level of utility. Approaches should be flexible to be specialized in terms of utility where necessary.

VI. CONCLUSION AND FUTURE WORK

In this paper, we presented an efficient approach to minimize side-effects of accuracy and information loss in itemset hiding problem. The degree of side-effect is represented with coefficients that are placed into the objective function of integer programming. Experiments with real datasets show that our approach minimizes the number of concealed nonsensitive association rules efficiently.

Coefficient Computation Algorithm can be specialized based on the area where the published database is utilized. In this sense, coefficients in the objective function of the integer programming may be used in a more efficient way. Moreover, different optimization techniques can be achieved by exploiting the inherent characteristics of the constraints and objective function that are involved in the CSP, in a more advanced way.

REFERENCES

[1] A. Amiri, “Dare to Share: Protecting Sensitive Knowledge with Data Sanitization“, Decision Support Systems, vol. 43, iss. 1, 2007, pp. 181-191.

[2] M. Atallah, A. Elmagarmid, M. Ibrahim, E. Bertino, and V. Verykios, “Disclosure Limitation of Sensitive Rules”, KDEX ’99: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange, 1999, pp. 45-52.

[3] R. Bayardo, “Efficiently Mining Long Patterns from Databases”, Proceedings of the ACM SIGMOD, 1998, pp. 85-93.

[4] F. Bodon, “A fast APRIORI implementation”, Proceedings of Workshop Frequent Itemset Mining Implementations (FIMI’03), vol. 90, CEURWS.org, CEUR Workshop Proceedings, 2003, pp. 56-65.

[5] T. Brijs, G. Swinnen, K. Vanhoof, and G. Wets, “Using Association Rules for Product Assortment Decisions: A Case Study”, Proceeding of the 5th ACM SIGKDD Internat. Conf. Knowledge Discovery Data, Mining. ACM Press, 1999, pp. 254-260.

[6] E. Dasseni, V. S. Verykios, A. K. Elmagarmid, and E. Bertino, “Hiding Association Rules by Using Confidence and Support”, Proceedings of the 4th International Workshop on Information Hiding, 2001, pp. 369–383.

[7] M. Jünger et al., “50 Years of Integer Programming 1958-2008 From the Early Years to the State-of-the-Art”, Springer, 2010.

[8] A. Gkoulalas-Divanis and V. S. Verykios, “An Integer Programming Approach for Frequent Itemset Hiding.”, Proceedings of the ACM Conference on Information and Knowledge Management (CIKM ’06), November 2006, pp. 748-757.

[9] A. Gkoulalas-Divanis and V. S. Verykios, “A Parallelization Framework for Exact Knowledge Hiding in Transactional Databases”, Proceedings of The IFIP TC-11 23rd International Information Security Conference, IFIP 20th World Computer Congress, IFIP SEC 2008, vol. 278, September 2008, pp. 349-363, Springer.

[10] A. Gkoulalas-Divanis and V. S. Verykios, “Exact Knowledge Hiding through Database Extension”, IEEE Transactions on Knowledge and Data Engineering, vol. 21, iss. 5, May 2009, pp. 699-713.

[11] A. Gkoulalas-Divanis and V. S. Verykios, “Hiding Sensitive Knowledge without Side Effects”, Knowledge and Information Systems, vol. 20, iss. 3, August 2009, pp. 263-299.

[12] GLPK. GNU GLPK 4.32 User’s Manual. Free Software Foundation Inc., Boston, MA, 2008. Available at <http://www.gnu.org/software/glpk/glpk.html> 27.10.2013.

[13] T. Li and N. Li, “On the Tradeoff between Privacy and Utility in Data Publishing”, Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2009, pp. 517-525.

[14] S. Menon, S. Sarkar, and S. Mukherjee, “Maximizing Accuracy of Shared Databases when Concealing Sensitive Patterns”, Information Systems Research, vol. 16, no. 3, September 2005, pp. 256–270.

[15] S. R. M. Oliveira and O. R. Zaïane, “Privacy Preserving Frequent Itemset Mining”, Proceedings of the IEEE ICDM Workshop Privacy, Security Data Mining, 2002, pp. 43–54, Australian Computer Society.

[16] X. Sun and P. S. Yu, “A Border-Based Approach for Hiding Sensitive Frequent Itemsets”, Proceedings of the Fifth IEEE International Conference on Data Mining (ICDM’05), 2005, pp. 426–433.

[17] X. Sun and P. S. Yu, “Hiding Sensitive Frequent Itemsets by a Border–Based Approach”, Computing Science and Engineering, vol. 1, iss. 1, September 2007, pp. 74–94.

[18] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E. Dasseni, “Association Rule Hiding”, IEEE Transactions on Knowledge and Data Engineering, vol. 16, iss. 4, April 2004, pp. 434-447.

[19] J. S. Yeh and P. C. Hsu, “HHUIF and MSICF: Novel Algorithms for Privacy Preserving Utility Mining”, Expert Systems with Applications, vol. 37, iss. 7, 2010, pp. 4779-4786.

 ISBN: 978-1-61208-329-2

Types of Knowledge Exchange During Team Interactions: A Software Engineering Study

Pierre N. Robillard and Sébastien Cherry
Computer and Software Engineering
Polytechnique Montréal
Montréal, Canada
(Pierre.Robillard; Sebastien.Cherry) @ polymtl.ca

Abstract—**A field study performed in a professional software development environment shows that the interactions between collocated teammates have various purposes. This paper presents a comprehensive study of ad hoc communications on collocated team based on video recording of professional developers within a large organization. It is found that there are four purposes for ad hoc face-to-face communications; collaboration, cooperation, coordination and socialisation. To be able to use collective tools in distributed or virtual team environments we must be able to support at some extent the purposes of ad hoc communications that occur naturally in social presence. The main finding of this field study is that collective tools need to satisfy two different purposes. A cooperative system is needed to share the know-how needed to build the product and a collaborative system is needed to share the knowledge needed to understand the functionalities to be implemented.**

Keywords - face-to-face interactions; teamwork; field study; cooperation; collaboration

I. INTRODUCTION

Cooperative system is a general term to describe tools design to support collective activity. To better understand the needs for collective systems we studied team activity in a collocated environment. Although teams are common structures in today's workplace, the use of collective systems designed to support team activities is still an ongoing challenge [1].

Physical collocation mostly refers to a team room. The team rooms support social presence by enabling interactive continuous communications. Face-to-Face communication (FtF) is the major feature of collocated teams [2]. To be able to use collective tools in distributed or virtual teams, we must be able to support at some extent the purposes of ad hoc communications that occur naturally in social presence. The ultimate purposes of ad hoc interactions within co-located team are to exchange information and share or synchronize knowledge and mental model.

FtF communications can occur in two different ways within collocated teams. One way is planned FtF communications, which occur during scheduled meetings [3] [4], another way is ad hoc FtF communications, which occur spontaneously when teammates are working on their tasks within the collocated team environment. There is a large variety of cooperative systems to be used in scheduled or planned activities. This paper reports on a field study designed to identify the type of collective systems most likely needed to support collocated ad hoc interactions occurring during unscheduled activities. For example, when software developers are working solo on their assigned tasks.

FtF communications outside the meeting rooms occurred spontaneously without scheduling and with unknown duration. They are usually very brief but sometimes they can last very long. These FtF interactions are ad hoc and opportunistic because they are triggered on a just in time basis and their content is unpredictable. The initiator takes the opportunity to interact with a targeted teammate in hoping that this communication will provide him with relevant information to help him pursue his task or else. Spontaneous ad hoc FtF communication is one of the unique features of collocated team.

These ad hoc FtF communications are initiated during "quiet time", which is when collocated teammates are working by themselves while sitting at their workplace. Ad hoc FtF communications require immediate attention and constitute for the receiver an interruption of work activity. Many studies have shown the importance of ad hoc communications in collocated team. This paper presents the analysis of ad hoc communications that occur during this quiet time and which constitute interruptions for the responders. One of the purposes is to find the kind of knowledge that is required by the responders.

In order to achieve a better understanding of *ad hoc* FtF communication within a software development team, we go further than previous studies by analyzing the inherent patterns and content of ad hoc communications. Such an understanding will provide clues to improving the environment for collocated software teams. Moreover, while studies have shown that distance raises barriers to informal communications, resulting in a number of coordination problems [5], we believe it is reasonable to expect that a better understanding of these informal communications will pave the way to further improvements of collective tools,

 ISBN: 978-1-61208-329-2

which are likely to be more appropriate to the needs of the users [6].

This paper reports on a field study based on video recording and performed in a professional software development environment, which last for few months. The purpose is to understand these natural phenomena found in collocated team. The results of this study are useful to the participants to help them understand the purposes of this activity and improve its used. Researchers interested in knowledge sharing activities are likely to find some of the reasons and the content of such ad hoc communications, which may help them to propose adapted and optimized practices to support the ad hoc communication purposes.

Section 2 presents the methodology used to capture the information from teammate interactions. Section 3 presents the physical characteristics of the interactions. Section 4 presents the dynamic of ad hoc FtF communications. Section 5 is a discussion on the validity of this data and the usefulness for the participant, and the managers.

II. METHODOLOGY

The research was conducted in the form of a field study in an professional software engineering setting, and relied mostly on participant observations, as described by Jorgensen [7] and Babbie [8]. The goal of this field study is to observe a collocated team, where developers are free to interact with one another. Our purpose is to measure *ad hoc* FtF communications occurring in a real professional environment.

The observed team was composed of 1 project manager and 3 software developers within a team of 12 developers, with varying levels of schooling (from Bachelor to Ph.D. degrees in computer sciences and engineering), and individual experience ranging from 9 months to 5 years of service in the company.

The total observation period lasted two months. For this study, we selected 12 regular half-day sessions from the 23 recorded. These sessions are distributed over the two months of the recording time and account for 35 hours of video recording, resulting in 404 vocal communications.

A half-day session lasts 2 to 3 consecutive hours. A regular session is defined as a session where all teammates are present and where there are no special events, such as meetings, visitors, etc., which could disturb the usual *ad hoc* FtF communications occurring during normal working activities. The researchers received human-subject approval from the University and the participating Company.

In this study, participants spent almost 30% of their working time in *ad hoc FtF* communications. Most FtF *ad hoc* communications (84%) involved only two participants, which indicates that an *ad hoc* communication is directed to a particular teammate by its initiator.

III. PHYSICAL CHARACTERISTICS OF AD HOC FTF COMMUNICATIONS.

This section describes how the ad hoc communications are distributed among the various regular working sessions with respect to their purposes. After many analysis iterations, we characterize the ad hoc FtF communications according to four purposes, which are socialization, coordination, cooperation and collaboration.

FtF socialization interaction supports the process by which individuals acquire the knowledge, social skills and value to conform to the norms and roles required for integration into the group.

FtF coordination interaction, which is defined as managing *'dependencies between activities', is* done mostly in *s*cheduled meetings where participants have common objectives on which the exchanges are based [9]. Coordination is characterized by formal relationships and understanding of compatible missions.

FtF cooperation interaction occurs when individuals reach some mutual agreement, but their works together do not progress beyond this level. In cooperation, activities are mutually agreeable but not necessarily for mutual benefit. Cooperation is characterized by informal relationships that exist without a commonly defined mission, structure or effort. Information is shared as needed. Typical cooperation activities are for example, giving a help with a debugging task.

FtF collaboration interaction is a recursive process where two or more people work together in an intersection of common goals that is creative in nature—by sharing knowledge, learning and building consensus. Collaboration is usually on demand from at least two team members that want to work together on a specific task. Examples of collaboration are some forms of pair-programming. All the collaborators have a genuine interest in the same activity [10].

Figure 1 shows the cumulative duration in minutes for each ad hoc FtF communication purpose and within each session. For example, Session 1 (Column 1) sums up to more than 40 minutes of ad hoc collaboration, 80 minutes of ad hoc cooperation, few minutes of ad hoc coordination and almost 20 minutes of ad hoc socialization. We observed that the four purposes occurred in all of the observed sessions but Session 6 which had few ad hoc FtF communications. Figure 1 shows also that the total time spends in each ad hoc FtF communication is largely variable amongst the various sessions. It ranges from 20 minutes for Session 11 to 2 hours and haft for Session 8. Cooperation and collaborations are the main purposes of FtF ad hoc communications. This figure shows that coordination and socialisation are not the major reasons for ad hoc communications during working sessions.

Figure 2 shows that almost a three-quarter of these interactions are for collaboration and cooperation purposes, each with a frequency occurrence of 37% and the remaining are for coordination and socialization. This pie

chart illustrates that from all the ad hoc FtF communications observed over the recording period only 10% are for coordination purposes. We recall that this is a kind of micro-coordination that occurred during ad hoc interactions only. This type of coordination does not take into account the formal coordination meetings that occurred regularly during the life-cycle of a project.

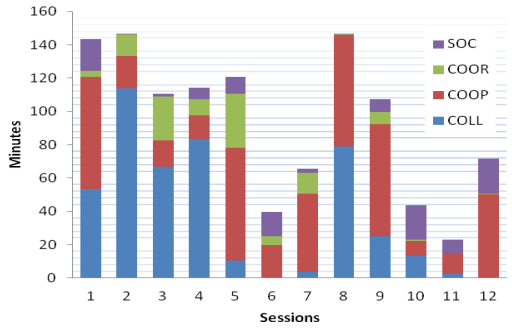


Figure 1. Cumulative duration in minutes for each of the FtF communications purposes within a recorded working session.

This first part of the analysis tells us that ad hoc FtF communications occurred mostly to satisfy two needs where one need is related to the tasks (collaboration and cooperation) and where the other need is related to the ancillary activities of the task, coordination and socialization, which are less frequent and much shorter. The first finding is that ad hoc FtF communications are mostly initiated to help participants in accomplishing their tasks through collaboration or cooperation.

According to this study, the coordination activities that occur on an ad hoc basis during the quiet time are mostly micro-coordination activities, which are related to team awareness. For example, a team member will state that his module is now ready for release or that he will test another module in the afternoon. There is little need for collective systems that will support micro-coordination activities, since it concerns less than 10% of the ad hoc interactions.

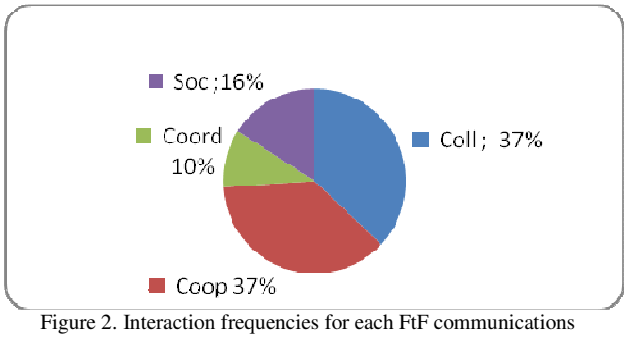


Figure 2. Interaction frequencies for each FtF communications purposes.

However, it might be useful to have a dedicated and easy to use tool, which may keep track of all these micro-coordination activities, to be able to trace them back in case of problems or to justify delays in the development plan.

IV. INITIATORS OF FTF AD HOC COMMUNICATIONS

Who are the initiators of ad hoc FtF communications? Is it everyone occasionally or is it few individuals that need to interact more often than others?

Since ad hoc collaboration and cooperation are the major cause for ad hoc FtF communications we make a closer analysis of their initiators. Figure 3 shows the relative frequencies of ad hoc collaboration initiators. Almost everyone on the team (9 out of 12 people) initiate, at some times during our recording period, an ad hoc collaboration. However, one individual (MS3) initiates more that the third of all the ad hoc collaboration. Only two other participants initiate more than 10 % of the ad hoc collaboration. It has been found that MS3 was involved in the modification of a module that has been developed some times ago by other team members and they all want this shared module to be well-maintained, which resulted in close collaboration on this task. There are six (6) other participants that share 39% of the collaboration initiatives.

In the light of this data, we can speculate that for certain tasks ad hoc FtF collaboration can be supported by collaborative tools that are likely to facilitate knowledge transfer and reduce interruptions.

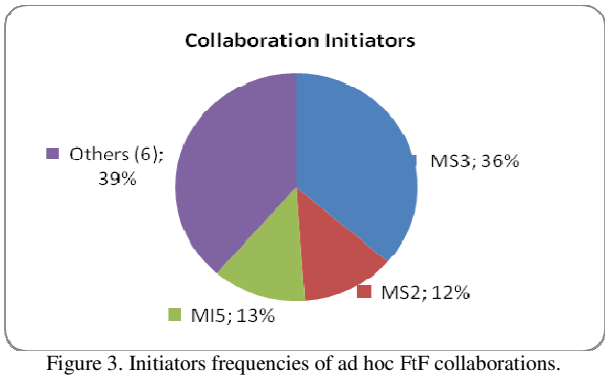


Figure 3. Initiators frequencies of ad hoc FtF collaborations.

Figure 4 shows the relative frequency of ad hoc FtF cooperation. It is observed that everyone (14 people) initiates ad hoc cooperation during the observation period. Actually, we had 2 people initiating ad hoc cooperation that was coming from outside this team of twelve (12). The leading initiator is MS1 with almost a quarter of the ad hoc cooperation initiations. MS1 is the last recruit on the team and this data shows clearly his needs for just-in-time help form others to efficiently do his task. Ad hoc cooperation from others (eleven individuals) account for 46% of the number of cooperation initiations.

 ISBN: 978-1-61208-329-2

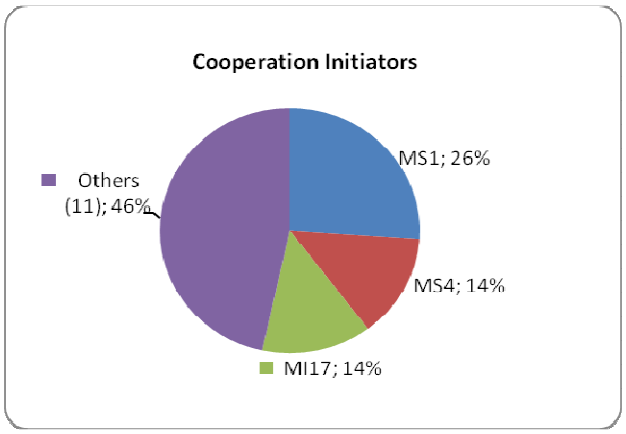


Figure 4. Initiators frequencies of Ftf ad hoc cooperation.

Closer analysis of ad hoc cooperation data suggests ways to reduce cooperation interruption. One way is to provide a cooperative system tools that will provide answers and record questions from the teammates. A dedicated FAQ (Frequently Asked Question) system could be a good example of such a tool.

V. DISCUSSION

This field study analysis shed light on the various needs for collective systems. Ad hoc FtF communications involving coordination, cooperation and collaboration are key components to maintain high level of awareness within a collocated team [11][12]. Dourish and Bly [13] define awareness to be "an understanding of the activities of others, which provides a context for your own activity." Maintaining awareness is the process by which individuals working with others transmit and acquire information, consciously or unconsciously, about their work efforts and how these efforts fit in with the on-going work of others. Our data shows that Ad hoc FtF communications contribute to the team awareness since they are not limited to few individuals but involve almost everyone within the team.

Ad hoc FtF collaboration can occur at any time during the working session, and developers address this need at the opportune time, that is, when an answer is required to complete the task with someone else sharing interest and benefit in the task. Ad hoc FtF cooperation presents the same intrinsic motivation. It occurs when questions come up about what to do or about how to do it. Developers seek the help of colleagues to obtain missing information, to ask for advice, or for guidance to pursue their tasks. Collective systems can help increase the efficiency of these interactions by providing appropriate coordination, cooperative or collaborative system tools.

Ad hoc FtF socialization is the less important purpose for ad hoc communications during normal working hours. This need for socialisation is probably well satisfied before work started, during breaks or lunch times. These periods have not been recorded during our study. To facilitate socialisation, some organisation may use collective games like ESP Game [14] applied to their own products.

Ad hoc FtF coordination occurs when there is a need to synchronize the activities of teammates on the tasks to tackle, and to plan for further activities. Coordination is not the predominant type of *ad hoc* communication in terms of frequency and time spent. Nevertheless, coordination activities are vital for maintaining synchronization of team activities, as well as for avoiding deadlocks or confusion in task organization. Coordination interactions are probably the easiest activities to support through electronic means for distributed teams. This activity is usually under the responsibility of the team leader and the agenda is well defined. Coordination is efficiently achieved today with shared calendar. However, micro-coordination in distributed teams may need some kind of support tools to maintain the team awareness on task progress.

Based on the patterns described above, *ad hoc* FtF communication seem to originate noticeably from a natural, unconscious, and *ad hoc* opportunistic process that takes place during software development activities. As discussed by Robillard [15], an *opportunistic* process, is defined by an incremental process in which knowledge is gathered as opportunities present themselves, and depends on the cognitive availability of the necessary information. On the opposite, a *systematic* process occurs when all the knowledge required to complete a task is available so that a well-structured plan can be followed.

An opportunistic process has been observed, in the software development settings studied, in the form of the *ad hoc* FtF communication of teammates to gather, in a *just-in-time* manner, the information they need to accomplish the task at hand. Collective tools may help this process by recording the information that is shared and more importantly by providing a repository where the most needed information can be found. Such collective tools are likely to reduce the interruptions initiated by FtF interactions.

A. *Task-Centric Content of Ad hoc Communications*

Studies on the topic (what) of *ad hoc* FtF communications have uncovered two contrasting content topics. On the one hand, development environment was the topic in more than 40% of the total *ad hoc* FtF communications, which relate to general *know-how* matters. Artifacts prescribed by software processes are based on knowledge related to the software to be built. These artefacts do not address *know-how* needed to build these software components. We believe that identifying and recording the recurring ad hoc FtF communications regarding these *know-how* interactions constitute a unique opportunity for optimizing cooperation interactions. For example, setting up a web based FAQ for development environment questions.

On the other hand, the results reveal that more than a third (35%) of the FtF communications concerned product

 ISBN: 978-1-61208-329-2

related topics, which are related to the software to be built. Studies on the information needs of software developers, identified the search for a design and code rationale as an important source of *ad hoc* interruptions [10]. As pointed out by previous authors, the degree to which diverse contents of the software to be built are discussed, raises questions about the adequacy and accuracy of the artifacts prescribed by software processes [16]. A Wiki structure based on concerns may make an efficient collaboration tool. Developers, who are adding or retrieving information, are sharing the same objective, which is collaborating to improve the shared understanding of the product.

Cooperation could also be efficiently computer mediated for well-defined tasks. A typical example of cooperation is when an expert helps a novice to accomplish a specific task.

Collaboration, which is based on shared goals and trust, is more difficult to computer mediate because it required some level of awareness and socialization. Efficient collaboration is based on trusted relationships, mutual respect, and on the expected capacity of the collaborators to contribute at reaching the goal. Ad hoc FtF communications are, still today, very difficult to support in distributed teams [17].

B. Ethical Issues

We considered the ethical implications of this research early on, at the planning stage, and ensured that all subjects and the hosting organization understood their rights and responsibilities before they agreed to participate [18]. All the individuals involved in our study were duly informed that their work sessions would be recorded, as well as of the nature of the study. They all signed the letter of agreement required for certification. Ethical issues were handled according to the established Canadian policies for research involving human subjects [19].

C. Scientific Value

The scientific value of this research has two components: the non-invasive study of the natural human communication in software development and the validity of the field study results. The social and collocated team aspects of software engineering constitute the major issues of this study. We believe that understanding these aspects is crucial to understanding how practices could be computer supported by appropriate tools. Global software development involves a de-located team, where verbal and FtF interactions via electronic means continue to occur. This study shows some of the parameters involved in the usual *ad hoc FtF* communications among collocated team members. These results can serve as a basis for a more in-depth study of the impacts of *ad hoc* interactions via electronic channels.

The validity of the observed results mostly relies on the fact that they faithfully and reliably represent reality. This study was not an experiment, where the various parameters could be controlled. This paper reports observations performed in a real professional environment in the course of carrying on day-to-day business. The salient outcomes of these observations, regardless of the specific setup of the organization, are a better understanding of the *raison d'être* of *ad hoc* FtF communications and how they take place and how they can be computer supported.

D. Study Limitations

One limitation of this study is inherent to most field study. The conclusions ensuing from this specific research cannot be generalized to all software development settings. However, owing to the characteristics that the featured settings have in common with the software development environments that can be encountered in the industry, we can assume that the outcomes of this study can be applied to a broader set of organizations.

Moreover, the *ad hoc* FtF communications observed in the framework of this research have been inducted in a maintenance context. Since maintenance contexts, where existing software is enhanced, predominate in software engineering settings, it is reasonable to assume that the results of this research are applicable to a broad set of contexts.

Finally, the method of video recording chosen for collecting the data poses a number of challenges, such as the background noise emerging from an open work space hosting hundreds of developers. It has several advantages, however: it can be reviewed as often as needed, and it is considered by participants to be less intrusive than having an observer take notes on their activities.

E. Validity Concerns

A coding scheme has been defined for the four purposes of *ad hoc* communication. Three coding agreement phases (inter-, intra- and extra-coder) have been applied to validate the data from the subjective coding. The first phase involved an intra-coder agreement, where a number of encoded data sequences were re-encoded a month later by the same coder. The second phase involved an inter-coder agreement, where the same coding operation was performed, by another coder who was able to understand the context and the jargon employed by both the participants and the primary coder. Finally, the third phase involved an extra-coder, were the same operation was performed by an experienced coder who was not familiar with the team work. An index proposed by Perreault and Leigh [20] was used to measure the subjectivity of the coding.

The indices obtained through the intra coder agreement were very high (.98). Inter-coder agreement show an agreement between the two coders with a value of 0.89, which is above the 0.7 limit, enabling us to deduce a strong agreement. The indices obtained with the exta-coder (.72) shows that the validity is still acceptable event when the coding is done by someone who is not familiar with the collocated team environment and its work.

 ISBN: 978-1-61208-329-2

VI. CONCLUSION AND FUTURE WORK

Through this research, it has been possible to shed some light on the main aspects of *ad hoc* FtF communications in terms of the communication mechanisms in which they are conducted, but more importantly their purposes and the content exchanged during those activities. This better understanding has revealed the opportunistic nature of ad *hoc* FtF communication, which constitutes the cornerstone for further theories and research about the phenomenon. It also paves the way for the introduction of improved practices based on collective systems to better support the various purposes of ad hoc communications, in collocated as well as de-located contexts. Further research will also be required to test the efficiency of such supporting tools in collocated as well as in de-located environments.

The main finding of this field study is that collective tools need to satisfy to different purposes.

- A cooperative system is needed for the sharing of the know-how needed to build the product
- A collaborative system is needed for the sharing of knowledge needed to understand the product functionalities.

The structures of these two systems are different. A cooperative system could be structured like a FAQ where the experts fill out the answers. A collaborative system could be structured like a Wiki where collaborating teammates are working together to fill out the information required by the shared tasks.

ACKNOWLEDGMENTS

This research would not have been possible without the agreement of the company in which it was conducted, and without the generous participation and patience of the software development team members from whom the data were collected. To all these people, we extend our grateful thanks.

REFERENCES

[1] R. F. Easley, S. Devaraj, and M. Crant, “Relating collaborative technology use to teamwork quality and performance: An empirical analysis,” J. Mng. Inf. Syst., vol. 19, no. 4, 2003, pp. 247–268.

[2] J. S. Olson, S. Teasley, L. Covi, and G. Olson, “The (currently) unique advantages of collocated work,” in Distributed Work, P. Hinds and S. Kiesler, Eds. Cambridge, MA: MIT Press, 2002, pp. 113–135.

[3] P. d'Astous and P. N. Robillard, “Empirical study of exchange patterns during software peer review meetings,” Information and Software Technology, vol. 44, no. 11, Aug 2002, pp. 639-48.

[4] P. d'Astous, F. Detienne, W. Visser, and P. N. Robillard, “Changing our view on design evaluation meetings methodology: a study of software technical review meetings,” Design Studies, vol. 25, no. 6, Nov 2004, pp. 625-655.

[5] J. D. Herbsleb and R. E. Grinter, , “Splitting the organization and integrating the code: Conway's law revisited,” Proc. 21st Intl Conf. on Software Engineering (ICSE '99), May 1999, pp. 85-95.

[6] J. D. Herbsleb and D. Moitra, “Global software development Software,” IEEE Software, vol. 18, 2, Apr 2001, pp. 16-20.

[7] D. L. Jorgensen, Participant observation : a methodology for human studies, Newbury Park, Calif.; London: Sage Publications, 1989.

[8] E. R. Babbie, The practice of social research, Belmont, CA : Wadsworth Thomson Learning, 2001.

[9] T. W. Malone and K. Crowston, “The Interdisciplinary Study of Coordination,” ACM Computing Surveys, Vol. 26, no. 1, March 1994, pp. 87-119, doi>10.1145/174666.174668.

[10] S. M. Hord, “A synthesis of research on Organizational colaboration,” Educational Leadership, Feb 1986, pp. 22-26.

[11] P. Dourish and V. Bellotti, “Awareness and coordination in shared workspace,” Proc. Computer Supported Cooperative Work, (CSCW 92), Nov. 1992, pp. 107-114.

[12] K. Schmidt, “The problem with 'awareness',” J. Comp. Supported Cooperative Work, Vol 11, Issue 3, 2002, pp. 285-298.

[13] P. Dourish and S. Bly, “Portholes: Supporting awareness in distributed work groups,” Proc, ACM Conf on Human Factors in Computing Systems (CHI 92), 1992, pp. 107-114.

[14] L. Von Ahn, "Games with a Purpose," Computer, Vol 39 (6), 2006, pp. 92–94, doi:10.1109/MC.2006.196

[15] P. N. Robillard, “Opportunistic problem solving in software engineering,” Software, IEEE, vol. 22, no. 6, 2005, pp. 60-67.

[16] T. D. LaToza, G. Venolia, and R. DeLine, “Maintaining mental models: A study of developer work habits,” Proc. 28th international conference on Software engineering, May 2006, pp. 492-501.

[17] S.E. Poltrock and G. Engelbeck, “Requirements for a virtual collocation environment,” Information and Software Technology, vol. 41, no. 6, 1999, pp. 331-339.

[18] J. Singer and N. G. Vinson, “Ethical issues in empirical studies of software engineering,” IEEE Trans. on Soft. Eng., vol. 28, 2002, pp. 1171-1180.

[19] NSERC, Natural Sciences and Engineering Research Council of Canada. Ethical Conduct for Research Involving Humans. In. Interagency Advisory Panel on Research Ethics. 2005.

[20] W. D. Perreault Jr. and L. E. Leigh, “Reliability of nominal data based on qualitative judgements,” Journal of Marketing Research, vol. 26, 1989, pp. 135-148.

 ISBN: 978-1-61208-329-2

Integrating Topic, Sentiment and Syntax for Modeling Online Review

Rui Xie, Chunping Li
School of Software
Tsinghua University
Beijing, China
{harryxse, cli}@tsinghua.edu.cn

Qiang Ding, Li Li
Shannon Lab
Huawei Technologies, LTD
Beijing, China
{q.ding, jollylili.li}@huawei.com

***Abstract*— The problem of analyzing online product reviews has drawn much interest of researchers. In this paper, we propose a novel probabilistic modeling framework based on *Latent Dirichlet Allocation* (LDA), which can reveal the latent aspect and sentiment of the review simultaneously. Unlike other topic models which only consider the text itself of online review, we firstly combine the *Part-of-Speech* (POS) tag into the model. We further propose three *Tag Sentiment Aspect Models* (TSA) to integrate the syntax information into modeling. The experiments show that our models are able to achieve a promising result not only on sentiment classification but on extraction of aspects of different sentiments.**

Keywords-topic model; sentiment analysis; tag sentiment aspect model; online review analysis.

I. INTRODUCTION

Nowadays, the development of Web 2.0 [3] makes convenient for customers to express their experience with products. Websites like Amazon.com and Epinions.com offer a platform for people to praise or criticize the target product. As a result, large amount of product reviews can be easily got from the Internet. These reviews are useful resource to help the customer to make decisions whether or not to buy a specific product. By browsing these reviews, people learn the good or the bad aspects of specific product. On the other hand, not only the customers, but the product designers also pay more and more attentions to these reviews. However, facing the overwhelming amount of reviews of products, no one can read the reviews piece by piece. There is an urgent need for the approach to obtaining useful and hidden information from larger review corpus. From the perspective of designers, given a product and its reviews, the aspects these reviews talk about and what the customers' overall attitude towards these aspects are the most important issues.

There are two problems during this process. First is topic identification, for identifying the aspects of the product the reviews talk about. The other one is sentiment analysis, to determine the sentiment label (positive, negative) of opinion toward specific aspect. They are challenging, not only for the large volumes of reviews, but for unstructured property of the plain text.

Topic identification, also known as aspect discovery, has been studied for a long time. In the past, two ways were usually used to extract aspect: filtering way and expanding way. The filtering way is to firstly extract a set of frequent Noun Phrases (NP) as candidate aspects, and then filter out the candidates which are less possible to be an aspect [1][2]. The expanding way is to firstly give an aspect list as basic knowledge and then expand the original list by using various expanding methods [3]. Sentiment analysis, also known as opinion mining, aims at using automated techniques to identify semantic orientation in texts. There are many works dedicated to classify the whole document or review into positive, neutral, and negative one [4][5]. Nevertheless, the sentiment of specific aspects is usually more useful than the overall rate. Other works focus on sentiment classification on the word/phrase level [6][7], but the word's sentiment polarity is dependent on topic or domain. Modeling the sentiment along with aspect/topic is required to make the result more informative.

The proposed models in this paper tackle this problem. The *Tag Sentiment Aspect Models*, extended from *LDA* [8], can model the aspects and sentiment of online reviews simultaneously. To our best knowledge, not much work can do this except [9][10][11][12]. Our models have several differences and improvements compared to existing works: (1) *TSA* models are to incorporate the syntax information into the hierarchical Bayesian model. (2) *TSA* models are fully unsupervised while some existing works need labeled data to train. (3) *TSA* models are domain independent. By integrating different domain prior information, *TSA* models can be applied to different domains.

The way *TSA* models integrating syntax information is based on the assumption that different words in sentences play different roles. A word can appear in a sentence for several reasons. It can play a role of syntactic function, and it can play a role of semantic content [13].

The rest of the paper is organized as follows. In Section II, we discuss the related works. Section III describes our proposed three *TSA* models and corresponding inference. In Section IV we show the experimental setup and give the evaluation of the model and discussion of the results in Section V. We have the conclusion and the future work in Section VI.

II. RELATED WORK

There are two major directions to discover the hidden aspect and sentiment in reviews. One direction is to apply

 ISBN: 978-1-61208-329-2

traditional natural language processing techniques to do text mining for reviews. For aspect discovery, the Noun Phrase (NP) detection is a widely used technique. Hu and Liu [2] used POS tagging to find noun phrase and selected frequent nouns as aspect candidates. A filtering method is applied to these candidates to generate real aspects. In [1], the similar approach is used to discover hidden aspects but besides NP, text fragments in the sentence level are as well used for generating aspects. Besides *POS* tagging, linguistic rules are also used to identify product feature/aspect. Turney [6] manually designed several linguistic rules to identify feature, like 'JJ + NN + (feature)' and 'RB + VB + (feature)', etc.

Unsupervised methods often need a lexicon to decide the word's orientation. The lexicon is built by expanding from a seed list. Using lexicon, the scheme for scoring the overall sentiment of sentence or review is well designed. Lu and Zhai [14] proposed a context-aware method for constructing the lexicon to adapt for different domains. They utilized general-purpose sentiment lexicon, thesaurus, the corpus' sentiment rating information and linguistic heuristics to reassign sentiment score to the vocabulary. The reassigned vocabulary comprises the new domain dependent lexicon. There existed some works on building general-purpose sentiment lexicon as well. The famous lexicon is *SentiWordnet* [20], which is an extension of *Wordnet* [21]. The *SentiWordnet* is organized by synsets, the same way as *Wordnet* does. *SentiWordnet* assigns to each synset of *Wordnet* three sentiment scores: positivity, negativity, and objectivity.

Another direction to discover the hidden aspect and sentiment in reviews is to apply probabilistic approach to model the whole corpus. Griffiths and Steyvers [15] applied *LDA* to extract the hidden topics. They proposed a Markov chain Monte Carlo algorithm for inference of the model. Some other works extended basic *LDA* for improving the results. Brody and Elhadad [16] proposed a more sophisticated *LDA* model to discover the aspect hidden in reviews. A connectivity matrix is used to calculate the score, which decides the best hidden aspect number and iteration times. Then the scoring schema is used by selecting the representative words for each aspect. Titov and McDonald [17] distinguished general aspects and find-grained aspects. The model can capture ratable and global aspects which make the result more meaningful. Zhao and Jiang [18] introduced a background model and also treated general and specific aspects differently.

Multi-Aspect Sentiment Model (MAS) [12] is an extension of the previous work – MG-LDA [17], which only extract topics hidden in reviews regardless of sentiments. MAS model works in a supervised way because it requires every aspect to be rated by user. *Topic Sentiment Mixture Model(TSM)* [11] is extended from pLSA. TSM has the deficit of pLSA with inference of new documents and suffers from overfitting of the data. On the other hand, TSM does not consider the association between topic and sentiment. The words are drawn from either topic distribution or sentiment distribution. The words are samples of a mixture of sentiment and topic, but not a combination. This makes TSM lack the ability to exact the aspect-specific opinion words. *Joint Sentiment/Topic (JST) model* [10] is a fully unsupervised model based on LDA. It can capture topic and sentiment at the same time. *Aspect Sentiment Unification model (ASUM)* [9] is a model based on JST. It is a small adaption of the JST. But it introduced the assumption that a sentence in reviews can only be referred to some an aspect and sentiment.

III. MODELS

We propose three Tag Sentiment Aspect Models (TSA) to extend the basic LDA to incorporate syntax information in different ways. In TSA1 and TSA2 models, two kinds of hidden variable are: *z*, aspect index, and *l*, sentiment label. In TSA3 model, an additional hidden variable *x* is introduced as an indicator besides aspect index and sentiment label. We use Gibbs sampling [15] to estimate the hidden variable. The Gibbs sampling method is a simple way to implement the inference in topic modeling, with good performance comparable with other methods and tolerant to local optimization. All the notations used here are illustrated in Table I.

TABLE I NOTATION USED IN TSA MODEL

D	the number of reviews
A	the number of aspects
S	the number of sentiments
V	the number of distinct words
N_d	the number of words in review d
T	the number of distinct tags
Θ	Multinomial distribution over aspects
Π	Multinomial distribution over sentiments
Ψ	Multinomial distribution over words
Ω	Multinomial distribution over tags
Σ	Multinomial distribution over indicators
Δ	Bernoulli distribution
λ_l	Dirichlet prior vector for σ
A	Dirichlet prior vector for θ
β_l	Dirichlet prior vector for ψ for sentiment l
Γ	Dirichlet prior vector for π
μ_l	Dirichlet prior vector for ω for sentiment l

A. Tag Sentiment Aspect Model 1(TSA1)

As the POS tag of words in reviews can be got by POS tagger, it is natural to take the POS tag of words as observed

TABLE II. TOP 10 WORDS FOR SENTI-ASPECT FOR LAPTOP DATASET

Topic / Model	System and Software			Hardware and Performance			Appearance and Experience		
	Positive	Neutral	Negative	Positive	Neutral	Negative	Positive	Neutral	Negative
TSA	develop	gpu	contact	nvidia	i3	manufacture	generous	pack	desk
	email	homework	promies	surface	cell	language	bigger	release	sound
	noise	guess	bad	cheap	nvidia	city	everyday	discharge	button
	couple	easy	unacceptable	issu	i5	repute	people	lithium	bio
	software	family	stand	gpu	chip	yesterday	gamer	iron	volume
	annoy	year	experience	button	cpu	mine	wow	cycle	noise
	pro	pro	refuse	day	core	laptop	part	charge	reason
	con	os	offer	trackpad	intel	mother	fact	capac	control
	window	pc	fix	graphic	graphic	board	spec	life	fan
	laptop	mac	pay	asu	model	problem	thing	battery	compute

data in the model. The graphical presentation of TSA1 model is shown in Fig. 1(a). Tag *t* is generated conditioned on aspect index *z* and sentiment label *l*, along with the word *w*. The tag is considered as the stamp of the word. This is inspired by Wang and McCallum [19], in which the published time of the document is treated as the timestamp of words in the document.

The generative process of TSA1 model is as follows.

1. For eech aspect and sentiment pair *(z,l)*, draw a discrete distribution over words $\psi_{z,l} \sim$ Dir (β_l), and a discrete distribution over tags $\omega_{z,l} \sim$ Dir (μ_l).
2. For any a review *d*,
   a) Draw the review's sentiment distribution $\pi_d \sim$ Dir(γ).
   b) For each sentiment label *l*, draw an aspect distribution $\theta_{dl} \sim$ Dir (α).
   c) For each word w_i and tag t_i in the review,
      i. Choose a sentiment label $j \sim$ Mul (π_d)
      ii. Choose an aspect $k \sim$ Mul (θ_{dj})
      iii. Choose word w_i ~Mul($\psi_{k,j}$) and tag t_i ~Mul($\omega_{k,j}$).

The hyper-parameters α, β, γ and μ are the pseudo-counts. It carries the prior observation of the corpus. Notice that for different sentiment label *l*, there are corresponding priors β_l and μ_l.That is because we use asymetric β and μ. The asymetric priors can exploit prior sentiment information in the corpus. For instance, elements of β correponding to positive sentiment words should have small value for negative sentiment label, and vice versa; Elements of μ correponding to noun tag should have large value for natural sentiment label, because the nouns often express not opinion but aspect. In TSA1 model, θ, π, ψ and ω are all the latent variables to infer. By using Gibbs sampling, we need to calculate the full conditional probabilities $P(z_i,l_i|\mathbf{z}_{-i},\mathbf{l}_{-i},\mathbf{w},\mathbf{t},\alpha,\beta,\gamma,\mu)$, where z_i denotes the aspect assignment for w_i, l_i denotes the sentiment assignment for w_i, $\mathbf{z}_{-i}$ denotes the aspect assignment for all word tokens except w_i, $\mathbf{l}_{-i}$ denotes the sentiment assignment for all word tokens except for w_i, and **w**, **t** are the word vector and tag vector for the whole corpus. During Gibbs sampling, we draw aspect and sentiment iteratively for the word w_i according to the following probability distribution:

$$P(z_i,l_i|w,t,z_{-i},l_{-i},\alpha,\beta,\gamma,\mu) \propto \frac{\{N_{k,d}\}_{-i}+\gamma}{\{N_d\}_{-i}+S\gamma}\frac{\{N_{j,k,d}\}_{-i}+\alpha}{\{N_{k,d}\}_{-i}+A\alpha}$$

$$\frac{\{N_{w,j,k}\}_{-i}+\beta_{w_i}}{\sum_{v=1}^{V}\{N_{v,j,k}\}_{-i}+\beta_v}\frac{\{N_{t_i,j,k}\}_{-i}+\mu_{t_i}}{\sum_{t=1}^{T}\{N_{t,j,k}\}_{-i}+\mu_t} \quad (1)$$

where $N_{k,d}$ is the number of words assigned to sentiment label *k* in review *d*, N_d is the number of words, $N_{j,k,d}$ is the number words assigned to aspect *j* and sentiment *k*. $N_{wi,j,k}$ is the number the word w_i assigned to aspect *j* and sentiment *k*, and $N_{ti,j,k}$ the number the tag t_i assigned to aspect *j* and sentiment *k*. $-i$ denotes the number that excludes the i^{th} position.

Having the conditional probability, the approximate probability of θ, π, ψ and ω is estimated as follows.

$$\omega_{j,k,t} = \frac{N_{t,j,k}+\mu_t}{\sum_{i=1}^{T}\{N_{i,j,k}\}_{-t}+\mu_i} \quad (2)$$

$$\psi_{j,k,w} = \frac{N_{w,j,k}+\beta_w}{\sum_{v=1}^{V}N_{v,j,k}+\beta_v} \quad (3)$$

$$\theta_{j,k,d} = \frac{N_{j,k,d}+\alpha}{N_{k,d}+A\alpha} \quad (4)$$

$$\pi_{k,d} = \frac{N_{k,d}+\gamma}{N_d+S\gamma} \quad (5)$$

B. Tag Sentiment Aspect Model 2(TSA2)

Considering that the number of unique tag is much smaller than size of vocabulary, treating tag as stamp of words may not be proper. Additionally, as shown in Fig. 1(b), the tag is dependent on the aspect *z* and sentiment *l*, but in real situation the dependency is reverse: the aspect and sentiment are dependent on the tag of the word. When user is writing a

 ISBN: 978-1-61208-329-2

review, he first decides the sentiment he would like to express. If he wants to express neutral sentiment that means he just wants to give a description, he will decide to use nouns. If he wants to express an opinion that means he wants to praise or criticize something, he will decide to use adjective or adverb. Therefore, we adapt TSA1 to TSA2 as shown in Fig. 1(b). We extend the aspect distribution of document from each sentiment to each sentiment and tag pair. This simple change not only incorporates the POS tag information, but also makes the model simpler.

The generative process of TSA2 is as follows:

1. For each aspect and sentiment pair *(z,l)*, draw a discrete distribution over words $\psi_{z,l} \sim \text{Dir}(\beta_l)$
2. For any a review *d*,
   a) Draw the review's sentiment distribution $\pi_d \sim \text{Dir}(\gamma)$.
   b) For each sentiment label *l* and tag *t*, draw an aspect distribution $\theta_{dlt} \sim \text{Dir}(\alpha)$.
   c) For each word w_i in the review,
      i. Choose a sentiment label $j \sim \text{Mul}(\pi_d)$
      ii. Choose an aspect $k \sim \text{Mul}(\theta_{djt})$, according to the word's tag *t*.
      iii. Choose word $w_i \sim \text{Mul}(\psi_{k,j})$.

Like TSA1, the Gibbs sampling processing is the same. The full conditional probability is as follows.

$$P(z_i, l_i | w, t, z_{-i}, l_{-i}, \alpha, \beta, \gamma, \mu) \propto \frac{\{N_{k,d}\}_{-i} + \gamma}{\{N_d\}_{-i} + S\gamma} \frac{\{N_{j,kt,d}\}_{-i} + \alpha}{\sum_{v=1}^{V}\{N_{kt,d}\}_{-i} + A\alpha} \frac{\{N_{w_i,j,k}\}_{-i} + \beta_{w_i}}{\sum_{v=1}^{V}\{N_{v,j,k}\}_{-i} + \beta_v} \quad (6)$$

where the major difference with TSA1 is that the 4th part of the TSA1's conditional probability is disappear and the 2nd part is different on the subscript. In TSA2, the times of words in review *d* assigned to aspect *j* and sentiment *k* is counted on every type of tag. The approximate probability of θ, π, ψ and ω is able to esimate as follows.

$$\psi_{j,k,w} = \frac{N_{w,j,k} + \beta_w}{\sum_{v=1}^{V} N_{v,j,k} + \beta_v} \quad (7)$$

$$\theta_{j,kt,d} = \frac{N_{j,kt,d} + \alpha}{N_{kt,d} + A\alpha} \quad (8)$$

$$\pi_{k,d} = \frac{N_{k,d} + \gamma}{N_d + S\gamma} \quad (9)$$

C. Tag Sentiment Aspect Model 3(TSA3)

There is a deficit in above TSA models. The aspect distribution θ is extended by T types of tag. This is based that the tag of words indicates whether the word is about aspect or opinion. We draw a different θ exactly according to the type of the tag. This is not proper because it implies

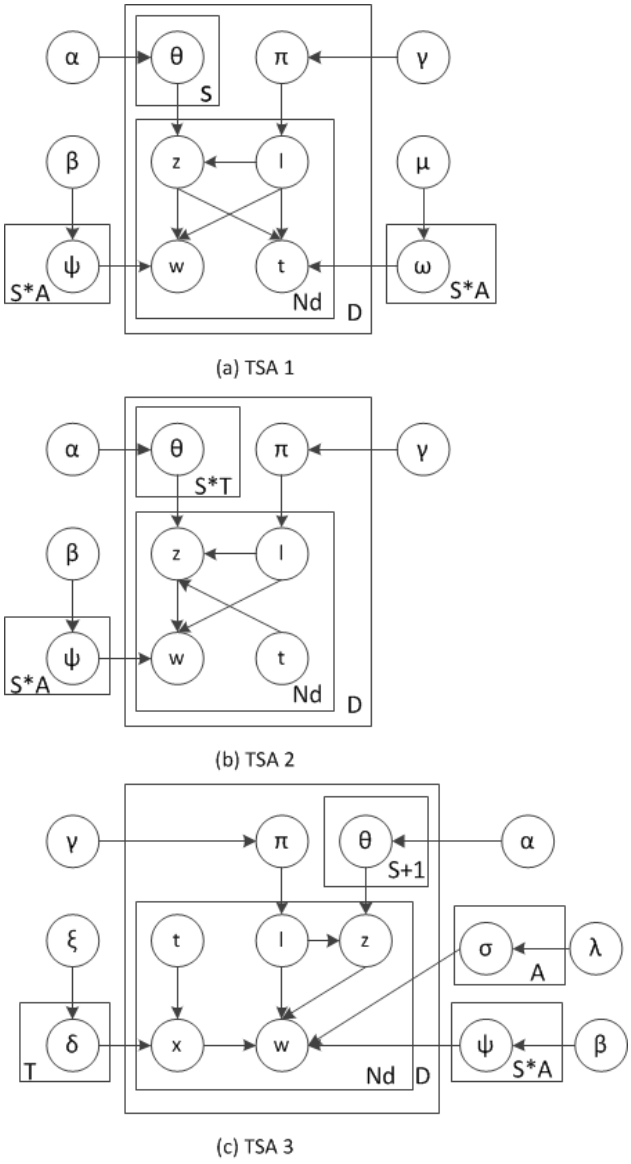


Figure 1.Tag Sentiment Aspect Models

that we applied a strict rule that a type of tag always serve as the same function. In fact, there often be exceptions. For example, the noun '*problem*' usually serves as an opinion word. To overcome this deficit, in TSA 3, we introduce a indicator variable *x* to indicate whether the word is serving as an aspect word or an opinon word. If *x* equals 0, the word serves as an aspect word and it will be drawn by distribution σ, which is a distribution set related to A aspects. If *x* equals 1, the word serves as an opinion word and it will be drawn by distribution ψ, a distribution set related to A aspects and S sentiments. The variable *x* is drawn from distribution δ. Different type of tag has an different δ. TSA3 is shown in Fig. 1(c). Notice that there is (S+1) θ for each review because the extra one θ is for the aspect word.

The generative process of TSA3 is as follows.

1. For each aspect and sentiment pair *(z,l)*, draw a discrete distribution over words $\psi_{z,l} \sim \text{Dir}(\beta_l)$, for

every aspect z, draw discrete distribution over words $\sigma_z \sim$ Dir (λ). For every tag t, draw Bernoulli distribution $\delta \sim$ Dir (ξ)

2. For any a review *d*,
   a) Draw the review's sentiment distribution $\pi_d \sim$ Dir(γ).
   b) For each sentiment label *l*, draw an aspect distribution $\theta_{dl} \sim$ Dir (α), and an extra aspect distribution $\theta_d \sim$ Dir (α).
   c) For each word w_i in the review,
      i. Choose indicator $x \sim$ Ber (δ_t), according to its tag t
      ii. If x is 0, choose an aspect $k \sim$ Mul (θ_d) and choose word $w_i \sim$ Mul (σ_k)
      iii. If *x* is 1, choose a sentiment label $j \sim$ Mul (π_d), choose an aspect $k \sim$ Mul (θ_{dj}) and choose word $w_i \sim$ Mul ($\psi_{k,j}$).

The full conditional probability is as follows.

$$P(z_t = j, l_t = k | w, t, z_{-t}, l_{-t}, \alpha, \beta, \gamma, \mu) \propto \frac{\{N_{x_l,t}\}_{-i} + \xi_{x_l}}{\{N_d\}_{-i} + \xi_m}$$

x=1

$$\frac{\{N_{w_i,j,k}\}_{-i} + \beta_{w_i}}{\sum_{v=1}^{V}\{N_{v,j,k}\}_{-i} + \beta_v} \frac{\{N_{k,d}\}_{-1} + \gamma}{\sum_{l=1}^{S}\{N_{l,d}\}_{-i} + \gamma} \frac{\{N_{j,k,d}\}_{-i} + \alpha}{\{N_{k,d}\}_{-i} + A\alpha} \quad (10)$$

x=0

$$\frac{\{N_{w_i,j}\}_{-i} + \lambda_{w_i}}{\sum_{v=1}^{V}\{N_{v,j}\}_{-i} + \lambda_v} \frac{\{N_{j,d}\}_{-i} + \alpha}{\sum_{z=1}^{A}\{N_{z,d}\}_{-i} + \alpha} \quad (11)$$

In TSA3, ξ_t is asymmetric and is incorporated with prior information. The approximate probability of δ, θ, π, ψ and σ is able to esimate as follows.

$$\delta_{x,t} = \frac{\{N_{x,t}\}_{-i} + \xi_x}{\sum_{m=0}^{1}\{N_{m,t}\}_{-i} + \xi_m} \quad (12)$$

$$\theta_{j,k,d} = \frac{N_{j,k,d} + \alpha}{N_{k,d} + A\alpha} \quad (13)$$

$$\theta_{j,d} = \frac{N_{j,d} + \alpha}{N_d + A\alpha} \quad (14)$$

$$\psi_{j,k,w} = \frac{N_{w,j,k} + \beta_w}{\sum_{v=1}^{V} N_{v,j,k} + \beta_v} \quad (15)$$

$$\sigma_{j,w} = \frac{N_{w,j} + \lambda_w}{\sum_{v=1}^{V} N_{v,j} + \lambda_v} \quad (16)$$

$$\pi_{k,d} = \frac{N_{k,d} + \gamma}{N_d + S\gamma} \quad (17)$$

IV. EXPERIMENTAL SETUP

We use the dataset of electronic device reviews from Jo and Oh [9]. We investigate the power of our models and select the Laptop and DigitalSLR categories to form our experimental dataset. We compare our TSA models by the power of sentiment classification of the review, and the power of discovering latent aspect and extracting the aspect-specific sentiment words. We also give a comparison between our models and existing models JST [10] and ASUM [9].

A. Preprocessing

For each dataset, we choose 1000 reviews including 500 positive and 500 negative. The original reviews are rated by 5-star system. We discard the 3 star reviews, and treats 1 or 2 star reviews as negative ones, 4 or 5 star reviews as positive ones. Then preprocessing is performed on DigitalSLR and Laptop dataset. NLTK (Natural Language Toolkit) [22], a software package implemented by PYTHON is used for preprocessing. First of all, the sentence segmentation algorithm is performed to obtain the sentences of each review. A POS tagger is then tagging every sentence. Afterwards, we remove the punctuation, numbers, and non-alphabet tokens. We filter out the tokens using a stop-word list [23]. For integrating syntax information, we consider 'NN', 'JJ', 'VB', 'RB', the 4 types of tag and ignore the others for the simplicity. After preprocessing, every corpus contains 1000 reviews. The DigitalSLR dataset has 83931 words with 5657 distinct words and the Laptop dataset has 81648 words with 5318 distinct words.

B. Prior Information

There are two key elements for incorporating prior information in TSA models. One key is carefully tuned hyper-parameters, the other is the initialization of Gibbs sampling. In the experiment, we use asymmetric hyper-parameters β to exploit the sentiment bias. We use a positive word list and a negative word list. If a word is in the positive list, the corresponding value of the β will be large for positive sentiment aspect and small negative sentiment, and vice versa. The exactly value to set is according to the actual experiment, which will be described in the following section. Considering the sentiment word list, we use the paradigm word list used in [9]. The sentiment words are applied in the initilization of Gibbs sampling. The word token in the sentiment word list is assigned to the corresponding sentiment label. During the iterative sampling process, the initilization effect becomes weak, so the iteration times should not be too large. We empirically set the iteration times to be 1000. The POS tag information is incorperated in the same way as the sentiment words list. In the initilization, the sentiment label of the word with 'NN' tag is drawn with distribution whose probablity is large on the neutral label, and the label of the word with 'JJ, VB, RB' tag is drawn with distribution whose probablity is large on the positive/negative label. Asymmetric prior μ_l is used for different sentiment label in TSA 1, and asymmetric prior ξ_t is used for diffferent type of tag resperctively in TSA 3.

 ISBN: 978-1-61208-329-2

C. Tasks

Two abilities of our models are investigated. The first is the ability to discover latent aspect word and aspect specific opinion word. The result of TSA models is analyzed to evaluate this ability. We compare our TSA models with existing approaches. The second ability is the power of sentiment classification. The sentiment distribution π, which indicates the proportion of each sentiment in the review, can achieve the task of sentiment classification. If the positive sentiment has a higher probability than the negative sentiment, the review is classifed as positive review, and vice versa. We compare our models with ASUM, JST and SVM with different features.

V. EXPERIMENTAL RESULTS

We first examine the automatically discovered aspects and senti-aspects from reviews by our TSA models. Then we examine the sentiment classification results.

A. Aspect and Senti-Aspect Discovery

To train our TSA models, we first set the parameters used in the model. The number of aspect is set to be 50, the prior α is set to be 1.0, according to previous works which show α should be set to 50 for topics. A symmetric prior γ is used that we asume no prior knowledge of the sentiment distribution. The value is set to 1, which means all sentiment probilities are equal. As mentioned above, prior β should be tuned carefully for its key effect to incoporate the priors. In TSA1 and TSA2, for positive aspect, we set elements of β vector to be 0 for negative words and other elements to be 0.01. For negative aspect, we set elements of β vector to be 0 for positive words and other elements to be 0.01. For neutral aspect, we use symmetric β set to be 0.01. In TSA1, the prior μ is set in the same way: 5 for elements corresponding to 'NN' and 1 for others for neutral apsect and 5 for elements corresponding to 'JJ', 'RB', 'VB' and 1 for others for positive/negtive aspect. In TSA3, asymmetric ξ is used by setting 5 for 'NN' and 1 for 'JJ', 'RB', 'VB' for x is 0, and 5 for 'JJ', 'RB', 'VB' and 1 for 'NN' for x is 1. Notice that, there are 3 sentiment labels in TSA1 and TSA2, and 2 sentiment labels in TSA3. So the β prior of TSA3 is setting as the β for positive and negative set is TSA1 or TSA2, and the new λ prior is set as the neutral one in TSA1, a symmetric prior. All these setting are used when there is prior of sentiment word list. In an random initalization context, we ignore these asymmetric priors. Instead, we use symmetric priors.

As the output of the model is the word distriutions, which is also called language model. One language model presents how frenquent the word will occur under certain aspect or aspect/sentiment pair. For TSA1 and TSA2, 3 (sentiment label) * 50 (apsect) word distribution will be obtained. For convenience of analysis, we place the extra 50 distributions with a virtual neutral label The 50 distribuions in TSA3 denotes only aspect, and neutral distribution in TSA1 and TSA2 denotes aspect and opinion with neutral sentiment label.We show the results in Table II. We select 3 aspects out of 50 for each model, and for every distribution we examine the top 10 words.

In Table II, we list three aspects drawn from the TSA models, and the labels of aspects are annotated manually. For 'system and software', the top words 'xp', 'vista', 'mac', which are different names of operating systems, are contained. The word 'program', 'windows', 'os' ,which are the concept of the 'system', are also contained. For another aspect, 'Hardware and performance', we get the words 'i3', 'i5', 'i7', which indicate a speical architecture of CPU, and we also get the words 'cpu', 'intel', 'chip' which explicitly indicate the 'hardware' aspect. The third aspect is 'appearance and experience', the appearance covers the style of laptop, screen, color, and weight, etc. The experience is about the joyment of customer to use this laptop. The second criteria requires little overlap between different aspects. It is obviously shown in Table II.

TSA models discover senti-aspect as well. Under each aspect, two distributions corresponding to postive and negative sentiment are also inferenced. For 'system and software', 'happy', 'best', 'nice', 'easy', and 'safe' are top words when talking about the positive side of the aspect, 'bad'. 'refuse', 'slow', 'hard', 'noise', and 'lose' are top words when talking about the negative side. For 'hardware and performance' aspect, when describing positive side, 'worth', 'high', 'better', and 'latest' are often used, and when describing negative side, 'hot', 'problem', 'noise', 'drop', 'bad' are often used in reviews. For 'appearance and experience', 'enjoy', 'bigger', and 'sonystyle' are used to express positive attitude and 'claim', 'provide', 'fail' and 'cancel' are used to express negative attitude. It can be noticed that the overlap of senti-aspect is high because there are two types of sentiment words. One is common sentiment words like 'good', 'bad', 'hate', and 'love', etc. The other is specific to a certain aspect, like 'hot' is positive for appearance, but negative for performance.

The power of three TSA models is different. For TSA1, the tags of words are treated as observed data. An extra distribution ω is introduced to reflect the tag informations. An intuition thought is that the word distribution ψ affected little by tag information, for this information almostly is coded in distribution ω. Therefore, the top words produced by ψ may not have high correlation with tags of words. This is shown in Table II. The top words of senti-aspect by TSA1 are the mixture of different tag types. For 'sytem and software', the positive aspect has words with noun tag like 'develop', 'email' beside words with adjective tag like 'couple', 'pro'. For 'appearance and experience' aspect, it is similar. The negtive aspect has words with adjective tag, like 'noise','bio', besides words with noun tag, 'reason', 'volumn'. For TSA2, the tag information is integrated by the document-aspect distribution, θ. The intuition is that the top words from distribution ψ will have high correlation with tags. But another problem is that as we use different θ according to the tag to draw a word, which means we use

 ISBN: 978-1-61208-329-2

the strict rule of that, words with different tag are always play different roles. So, the TSA2 may suffer with the loss of ability with exploiting nouns used in positive or negative aspect. The intuition is verified again as shown in Table III. For both positive and negative aspect, words with noun tag are hardly seen. For 'system and software', the top words in the positive aspect are all adjective except for the word '3d', and the top words in the negative aspect are all adjective except for the verb 'consist' and the noun 'drive'. Similar situation is also observed in 'hardware and performance' aspect and 'appearance and expericence' aspect. TSA3 is designed to integrate tag information to ψ distribution but not to loss flexibity and its ability to explore noun words in sentimental aspects. In TSA3, the indicator variable indicates that a word is the aspect word or the sentiment word under an aspect. The variable x is drawn from Bernoulli distribution conditioned on tag of the word. For adjective, verb and adverb, x is highly probable to be 1, indicating a sentiment word under an aspect, while noun is highly probable to be zero, indicating an pure aspect word. By introducing the Bernoulli distributions, the rule is ralaxed with the nouns as sentiment words in the low probability. In Table III, the nouns such as 'damage' and 'waste' are presented as top words in 'hardware and performance' negative aspect, and the nouns such as 'virus', 'problem' and 'lose' are presented as top words in 'system and software' negative aspect. For positive aspects, the nouns such as 'power', 'monitor', 'design' and 'budget' are also appeared as sentiment words.

The same situation is also observed when applying TSA models to the DigitalSLR dataset, we list some aspects and sentiment words in Table III.

TABLE III. ASPECT WORDS AND CORRSPONDING SENTIMENT WORDS FOR DIGITALSLR DATASET LIST

Aspect Words	Sentiment Words
compact depth picture color olympus eye iso electron	small long light higher manual fun weather love support
	heavier stain dust waste disappoint shallow horrible worry
camera focus len shoot process zoom angle digit	long smooth easy original high wide motor happier
	short expensive problem loss heavier difficult change claim
strobe battery d700 handbrake aim camcorder neon viewfinder	profession bright long light adjust detail friend improve
	amateur spend refund defect stain useless recharge lost

B. Sentiment Classification

We use distribution π for sentiment classification task on the review level. As mentioned above, we discard neutral reviews and do the evaluation only on positive and negative ones. The distribution $\pi_{k,d}$ presents the probability of sentiment k in review d, we compare the probability of sentiment positive and negative, and assign larger label to the review. We compare TSA models with existing approaches. The result is shown in Table IV. We also exploit the effect of priors, and the result shows the prior enhance the performance largely.

TABLE IV. SENTIMENT CLASSIFICATION ON DIGITALSLR AND LAPTOP DATASETS WITH Laptop TSA MODEL AND PREVIOUS APPROACHES, AND THE PLUS MEANS WITH PRIOR INFORMATION INCORPORATED

	Laptop Dataset			DigitalSLR Dataset		
	pos	neg	overall	pos	neg	overall
TSA1	58.8%	53.2%	56%	60.8%	56%	58.4%
TSA2	50.8%	54.6%	53.7%	59%	53.6%	56.3%
TSA3	59%	53.4%	56.2%	61%	58.8%	59.9%
TSA1+	86.6%	82.4%	85%	83.4%	86.2%	83.8%
TSA2+	80.8%	85.4%	83.1%	82.8%	82.4%	82.6%
TSA3+	84.8%	87.4%	86.2%	86.8%	84.4%	85.6%
ASUM	58%	52.2%	55.1%	54.8%	56.2%	55.5%
ASUM+	85.8%	84.6%	85.2%	86.4%	82.8%	84.6%
JST	54%	52.8%	53.4%	50.6%	55.2%	52.9%
JST+	78.6%	84.4%	81.5%	82.8%	76.6%	79.7%

In Table IV, first observation is that the prior information could enhance the accuracy of sentiment classification. The average accuracy is 55% and the prior improves that to about 80%. TSA2 performs worse than TSA1 and TSA3. That is because in TSA2, the strict rule that words with different tag plays the different roles in introducing the noise. TSA1 and TSA3 almost have the same power on sentiment classification, although the power of TSA3 on aspect discovery is better than TSA1. That is because the distribution π incorporates the infomation of ω but the ψ distribution does not. We also compare our models with JST [10] and ASUM [9]. JST model is worse than our TSA models. ASUM is sometimes better than TSA1 and TSA2, but not better than TSA3.

VI CONCLUSION AND FUTURE WORK

The "bag of words" assumption is suitable for traditional text classification but when comes to opinion mining, it is not good one. Opinion is expressed in the more complicated way. We need to explore more information hidden in the natural language. The tag of word is a good attempt. Our TSA models are to incorporate this kind of information. The results of our approach have better effectiveness as shown in Table IV.

The future works have two directions: one is to exploit other language information into the model. For example, the dependency relation can be used. A synonym thesaurus can used to explore the relations between aspect words. The other direction is to design the prior information better and adjust the TSA model to fit the corpus. We also can add an additional background language model to capture the

 ISBN: 978-1-61208-329-2

common across the language. In the TSA models, the prior probabilities of different tags are still fixed manually. We could further use the unsupervised machine learning method to train a model to determine these probabilities.

REFERENCES

[1] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. A. Reis, and J. Reynar, "Building a Sentiment Summarizer for Local Service Reviews," In WWW Workshop on NLP in the Information Explosion Era (NLPIX), USA, 2008.

[2] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," In Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, USA, 2004, pages 168 – 177.

[3] A. M. Popescu and O. Etzioni, "Extracting Product Features and Opinions from Reviews," In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, USA, 2005, pages 339 – 346.

[4] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs Up?: Sentiment Classification Using Machine Learning Techniques", In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, USA, 2002, pages 79 - 86.

[5] B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization based on Minimum Cuts," In Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, USA, 2004.

[6] P. D. Turney, "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews," In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, USA, 2001, pages 417–424.

[7] P. D. Turney and M. L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus," CoRR, cs.LG/0212012, 2002.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol.3, 2003, pages 993-1022.

[9] Y. Jo and A. Oh, "Aspect and Sentiment Unification Model for Online Review Analysis," In Proceedings of WSDM, 2011

[10] C. Lin and Y. He, "Joint Sentiment/Topic Model for Sentiment Analysis," In Proceeding of the 18th ACM Conference on Information and Knowledge Management, USA, 2009, pages 375 – 384.

[11] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic Sentiment Mixture: Modeling Facets and Opinions in Weblogs. In Proceedings of the 16th International Conference on World Wide Web, USA, 2007, pages 171-180.

[12] I. Titov and R. Mcdonald, "A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In Proceedings of International Conference on ACL, 2008, pages 308-316.

[13] T. Griffiths, M. Steyvers, D. Blei, and J. Tenenbaum, "Integrating Topics and Syntax," In Advances in Neural Information Processing Systems 17, 2004.

[14] Y.Lu, M.Castellanos, U.Dayal and C.Zhai, "Automatic Construction of a Context-Aware Sentiment Lexicon: An Optimization Approach," In Proceedings of WWW, 2011.

[15] T. Griffiths and M. Steyvers, "Finding Scientific Topics," In Proceedings of the National Academy of Sciences (101), 2004, pages 5228–5235.

[16] S. Brody and N. Elhadad, "An Unsupervised Aspect-Sentiment model for Online Reviews," In Proceedings of the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, USA, 2010, pages 804 - 812.

[17] I. Titov and R. McDonald, "Modeling Online Reviews with Multi-grain Topic Models," In the Proceeding of the 17th International Conference on World Wide Web, USA, 2008, pages 111–120.

[18] X. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly Modeling Aspects and Opinions with a MaxEnt-LDA Hybrid," In Proceedings of the International Conference on Empirical Methods in Natural Language Processing, USA, 2010, pages 56 – 65.

[19] X. Wang, A. McCallum, "Topics over Time: a Non-Markov Continuous-Time Model of Topical Trends," In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Dada Mining, USA, 2006.

[20] SentiWordNet, http://sentiwordnet.isti.cnr.it [retrieved: Aug. 2013]

[21] WordNet, http://wordnet.princeton.edu [retrieved: May 2013]

[22] NLTK, Natural Language Toolkit, http://www.nltk.org [retrieved: Aug. 2013]

[23] http://www.lextek.com/manuals/onix/stopwords1.html [retrieved: May 2013]

 ISBN: 978-1-61208-329-2

The Critical Dimension Problem: No Compromise Feature Selection

Divya Suryakumar, Andrew H. Sung
Department of Computer Science and Engineering
New Mexico Institute of Mining and Technology
Socorro, New Mexico 87801, USA
{divya, sung}@cs.nmt.edu

Qingzhong Liu
Department of Computer Science
Sam Houston State University
Huntsville, Texas 77341, USA
liu@shsu.edu

Abstract— **The important feature selection problem has been studied extensively and a variety of algorithms has been proposed for data analysis and mining tasks in diverse applications. As the era of "big data" arrives, the development of effective techniques for identifying important features or attributes in very large datasets will be highly valuable in dealing with many of the challenges that come with it. This paper describes work in progress regarding a related general problem: for a given dataset, is there a "Critical Dimension" or minimum number of features that are necessary for achieving good results? In other words, for a dataset with many features, how many are truly relevant and important to be included in, say machine learning and/or data mining tasks to ensure that acceptable performance is achieved? Moreover, if a Critical Dimension indeed exists, how to identify the features that need to be included? The problem is first analyzed formally and shown to be intractable. An ad hoc method is then designed for obtaining approximate solution; next experiments are performed on a selection of datasets of varying sizes to demonstrate that for many datasets there indeed exist a Critical Dimension. The significance of the existence or lack thereof in datasets is explained.**

Keywords-machine learning; ranking; feature reduction; Critical Dimension; large datasets.

I. INTRODUCTION

One of the challenges of "big data" is how to reduce the size of data without losing information contained therein. In that regard, effective feature ranking and selection algorithms [1] can guide us in significantly reducing the size of the dataset by eliminating features that are insignificant, irrelevant, or useless. In some bio- or medical- informatics datasets, for example, the number of features can reach tens of thousands. This is partly due to the reason that many datasets constructed today for intended data mining purposes, without prior knowledge about what is to be specifically explored or derived from the data, likely have included measurable attributes that are actually insignificant or irrelevant, and inevitably resulting in large numbers of useless attributes (or features) that can be deleted to greatly reduce the size of datasets without any negative consequences in data analytics or data mining [6].

We investigate in this paper the general question: Given a dataset with n features, is there a Critical Dimension, or the smallest number of features that are necessary for a particular data mining application to ensure a minimal performance requirement? The term performance in this context means the overall accuracy of the training model. That is, any machine learning, statistical analysis, or data mining, etc. tasks performed on the dataset must include at least a number of features no less than the Critical Dimension – or it would not be possible to obtain acceptable results. This is a useful question to investigate since feature selection methods generally provide no guidance on the number of features to retain for a particular task; moreover, for many complex problems to which big data brings hope of breakthrough there is very little or no prior knowledge which may be otherwise relied upon in determining this number.

The question is analyzed in the next section and shown to be intractable. In Section 3, an ad hoc method is proposed as a first attempt to approximately solve the problem. In Section 4, experimental results on selected datasets are presented to demonstrate the existence of the Critical Dimension for most of them. Section 5 provides conclusions and discussions.

II. CRITICAL DIMENSION

The intuitive concept of the Critical Dimension of a dataset with n features is that there may exist, with respect to a specific "machine" M and a fixed performance threshold T, a unique number $\mu \leq n$ such that the performance of M exceeds T when a suitable set of μ features is selected and used (and the rest $n - \mu$ features discarded); further, the performance of M is always below T when any feature set with less than μ features is used. Thus, μ is the critical number of features that are necessary to ensure that the performance of M meets the given threshold.

A. Formal Definition of Critical Dimension

Formally, for dataset D_n with n features, machine M (a learning machine, an algorithm, etc.) and performance threshold T (the accuracy of training, etc.), we call μ the T-Critical Dimension of (D_n, M) if the following two conditions hold:

1. There exists a μ-dimensional projection of D_n, which lets M to achieve a performance of at least T, i.e., **($\exists D_{\mu}$**

 ISBN: 978-1-61208-329-2

$\propto D_n$) [$P_M(D_*) \geq T$], where $P_M(D_*)$ denotes the performance of M on input dataset D_*.

2. For all $j < \mu$, a j-dimensional projection of D_n fails to let M achieve performance of at least T, i.e., ("$D_j \propto D_n$) [$j < \mu \Rightarrow P_M(D_j) < T$]

To determine whether a Critical Dimension exists for a D_n and M combination is a very difficult problem. Specifically, the problem of deciding, given D_n, T, k ($k \leq n$), and a fixed M, whether k is the T-Critical Dimension of (D_n, M) actually belongs to the class $D^P = \{L_1 \cap L_2 \mid L_1 \in NP, L_2 \in coNP\}$ (C.H. Papadimitriou et al, 1982), where it is assumed that the fixed machine M runs in polynomial time in n, the dimension of the dataset. In fact, it is shown in the next subsection that the problem is D^P-hard [4].

Since NP and coNP are subclasses of D^P (note that D^P is not the same as NP ∩ coNP), the D^P-hardness of the Critical Dimension problem indicates that it is both NP-hard and coNP-hard, and likely to be intractable.

B. Proof That CDP is D^P-Hard

The Critical Dimension Problem (CDP) is stated formally as follows: **Given D_n, T, k ($k \leq n$), and a fixed P_M (the performance of M). Is k the T-Critical Dimension of (D_n, M)?** The problem to decide if k is the T-Critical Dimension of the given dataset belongs to the class D^P under the assumption that, given any $D_i \propto D_n$, whether $P_M(D_i) \geq T$ can be decided in polynomial time of n, i.e., the machine M can be trained and tested with D_i in polynomial time. Otherwise, the problem belongs to some larger complexity class, e.g., Δ^P_2. Note here that (**NP ∪ coNP**) $\subseteq$ $D^P \subseteq \Delta^P_2$.

To prove that the CDP is a D^P-hard problem, we take a known D^P-complete problem and transform it into the CDP. Let us consider the maximal independent set problem as an example. In graph theory, a Maximal Independent Set (MIS) is an independent set that is not a subset of any other independent set. That is, it is a set S such that every edge of the graph has at least one endpoint not in S and every vertex not in S has at least one neighbor in S. A MIS is also a dominating set in the graph, and every dominating set that is independent must be maximal independent, so it is also called independent dominating sets. A graph may have many MIS's of widely varying sizes; a largest maximal independent set is called a MIS.

EXACT-MIS problem – Given a graph with n nodes, and $k \leq n$, decide if there is a maximal independent set of size exactly k in the graph is a problem is D^P-complete as proved by Papadimitriou and Yannakakis, 1982. Now we will transform this D^P-complete to an instance of CDP. To construct the instance of the *CDP*, let: dataset D_n be a representation of the given graph with n nodes (e.g., D_n can be made to contain n data points, with n features, representing the adjacency matrix of the graph), T be the value 'True' from the binary range {T, F}, $\mu = k$ be the value in the given problem and M be an algorithm that decides if the dataset represents a maximal independent set of size μ, if yes P_M = True otherwise P_M = False, then a given instance of the D^P-complete problem is transformed into an instance of the *CDP*. Three examples are shown below and explained. If the threshold is T (True) from the binary range {T,F}, then the problem is considered to be the NP-complete EXACT-MIS problem, and F (False) if it is not a NP-complete EXACT-MIS problem. Figure 1 is a graph with 5 nodes, containing an EXACT-MIS of size 3.

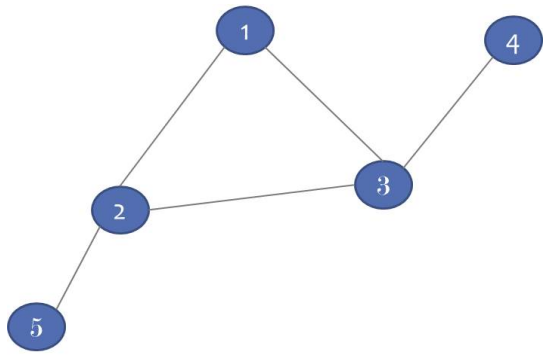


Figure 1. A Graph with 5 nodes showing exactly one MIS with 3 nodes {1,4,5}

D_5=

1	1	1	0	0
1	1	1	0	1
1	1	1	1	0
0	0	1	1	0
0	1	0	0	1

Figure 2. The adjecency matrix of graph with 5 nodes

Example 1: k=3
Threshold T = 'True' from the binary range {T, F}. μ = 3 exist; i.e., an EXACT-MIS of size 3 exists in D_5 and is as highlighted in the adjacency table shown in Figure 2. So, M is an algorithm that decides if the input dataset represents a maximal independent set of size μ, or M "verifies" that some $D\mu$ corresponds to a maximal independent set; i.e., $P_M(D_*)$ = 'True' if $D\mu$ allows M to construct a maximal independent set of G of size μ, where $D_* \propto D_n$ and D_n represents the adjacency matrix of G. Since the solution to the EXACT-MIS problem is True, solution to an instance of the CPD transformed from this is YES.

Example 2: k=4
Threshold T = 'True' from the binary range {T, F}. μ = 4 exists but is not an EXACT-MIS. From D_5 table it can be seen

that there does not exists any independent sets of size 4, so no EXACT-MIS of size 4 exists. Let *M* be an algorithm that decides if the input dataset represents a graph containing a maximal independent set of size μ, or *M* "verifies" that some $D\mu$ corresponds to a maximal independent set; i.e., $P_M(D_{\mu}) =$ 'True' if $D\mu$ allows *M* to construct a maximal independent set of *G* of size μ, where $D_{\mu} \propto D_n$ and D_n represents the adjacency matrix of *G*. In this example since no independent set of size 4 exists the solution to the instance of constructed *CDP* is NO, so $P_M(D_4) =$ 'False' for all D_4.

Example 3: k=2
Threshold *T* = 'True' from the binary range {T, F}. $\mu = 2$ exists but is not an EXACT-MIS. Again, from D_5 table it can be seen that there exist independent sets of size 2 but they not EXACT-MIS. So, algorithm *M* decides $D\mu$ \corresponds to a maximal independent set if $D\mu$ allows *M* to construct a maximal independent set of *G* of size μ, where $D_{\mu} \propto D_n$ and D_n represents the adjacency matrix of G. In this example since no independent set of size 2 exists the solution to the EXACT-MIS is 'False' so a solution to an instance of constructed *CDP* is NO, $P_M(D_2) =$ 'False' for all D_2.

The D^P-hardness of the Critical Dimension problem indicates that it is both NP-hard and coNP-hard; therefore, it's most likely to be intractable, that is, unless P = NP.

III. METHOD TO FIND THE CRITICAL DIMENSION

We can see from the CDP Problem analyzed above that even deciding if a given number is a Critical Dimension is intractable, to find that number is certainly even more difficult. So, a heuristic method is proposed in the following. The heuristic method represents a feasible and practical approach in attempting to find the Critical Dimension of a given dataset and a given performance threshold with respect to a fixed machine. Though the heuristic method actually corresponds to a different definition of the Critical Dimension, it serves to validate the concept that μ exists for datasets, though maybe not for all of them; and we will see that for most of the datasets with which experiments were conducted a Critical Dimension indeed exists. Finally, the μ determined by this heuristic method is hopefully close to the theoretical Critical Dimension as defined in the formal definition.

In the heuristic method, the Critical Dimension of a dataset is defined as that number (of features) where the performance of a specific learning machine would begin to drop below the performance threshold significantly, and would not rise again when smaller number of features is used. To make the method feasible, the features are initially sorted in descending order of significance (according to some feature ranking algorithm) and the feature set is reduced by deleting the least significant feature after each iteration of the experiment when performance of the machine is observed. For cross validation purposes, therefore, multiple experiments can be conducted when attempting to determine the Critical Dimension of a dataset: the same machine is used in conjunction with different feature ranking algorithms; also, the same feature ranking algorithm is used in conjunction with different machines; then we analyze if the different experiments resulted in similar numbers for the Critical Dimension.

A. Heuristic Method to find the Critical Dimension

The term Critical Dimension of a dataset has been described as the minimum number of features required for a learning machine to perform prediction or classification with high accuracy. Empirically, the Critical Dimension of a dataset can be defined as that number (of features) where the performance of a specific learning machine would begin to drop significantly, and would not rise again when smaller number of features is used.

In other words, it is postulated that, for a dataset there possibly exists a Critical Dimension (μ), which is a unique number for a specific machine learning and feature ranking combination and which can be determined experimentally. Specifically, let $A = \{a_1, a_2, \ldots, a_n\}$ be the feature set where a_1, a_2, …, a_n are listed in order of decreasing importance as determined by some feature ranking algorithm. Let $A_m \subseteq A$ contains the m most important features, i.e., $A_m = \{a_1, a_2, \ldots, a_m\}$ where $m \leq n$. For a learning machine *M* and a feature ranking method *R*, we call μ ($\mu \leq n$) is the Critical Dimension of $[D_n, M, T]$, if the following conditions satisfy; If *T* is a given performance threshold that is considered acceptable, when *M* uses feature set $A\mu$ the performance of *M* is $\geq T$, and whenever *M* uses less than μ features its performance drops below T_{μ}. As an illustration, the Hypothyroid disease dataset was classified using SMO (Sequential Minimal Optimization) classifier. This dataset was ranked using Chi-squared ranking algorithm. Figure 3 shows the Critical Dimension and was found to be 18; and it can be observed that this point satisfies the heuristic methods definition of a Critical Dimension.

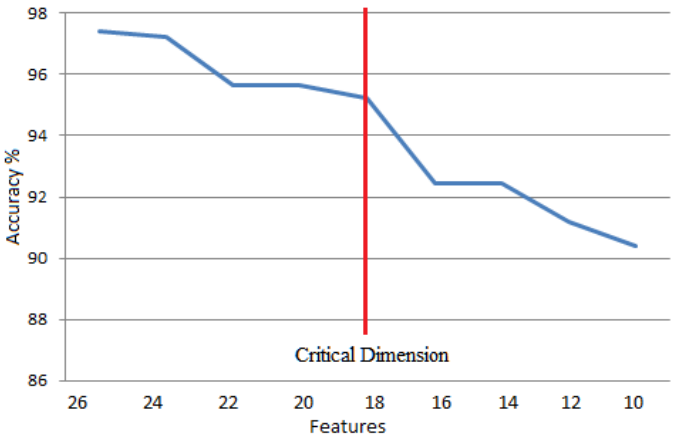


Figure 3. The Critical Dimension of Hypothyroid disease dataset

IV. FINDING THE CRITICAL DIMENSION USING FEATURE RANKING METHODS

To find approximate solutions to the Critical Dimension problem, a heuristic method based on feature ranking algorithms is applied. In this method, the performance

threshold *T* will not be specified beforehand but will be defined during the iterative process where a learning machine classifier's performance is observed as the number of features is decreased. The Figure 4 below shows the method to find the Critical Dimension.

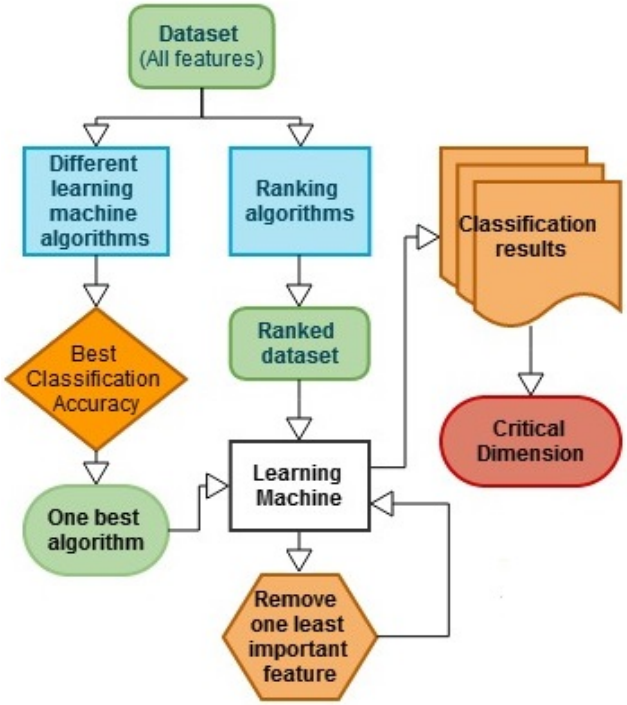


Figure 4. Method to find the Critical Dimension

A. Choosing the best classification algorithm

The dataset is first classified by building a model based on six different algorithms, namely, Bayes net, function, rule based, meta, lazy and decision tree learning machine algorithm [12][8][9]. The machine or model with the best prediction accuracy is chosen as the classifier to find the Critical Dimension for that dataset. Table 1 shows the accuracy results of the learning machine model built based on one of the six algorithms discussed above. Figure 5 shows the method in which the best classifier is choosen.

TABLE I. CHOOSING THE CLASSIFICATION ALGORITHM FOR THROMBIN DATASET

Accuracy %	Method		
	Algorithm	***Learning Machine***	***Best***
42.82	Bayes	Naive bayes	**C4.5 with 69.33%**
18	Functions	SMO	
63.59	Lazy	kStar	
66.18	Meta	Ada Boost	
68.39	Rule	Decision table	
69.33	Tree	C4.5	

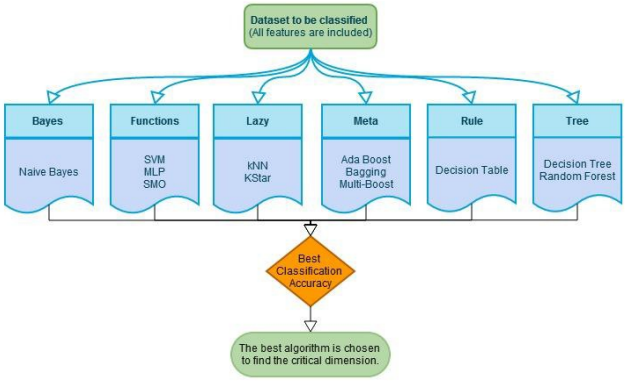


Figure 5. Choosing the best classification model

B. Ranking algorithm

The Chi-square ranking method [3] evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. There are several ways in which a chi-squared statistics is used; one such is using a contingency table. To rank features, we look at the chi-square distribution table against its degree of freedom value to find the corresponding probability level α; search method ranker, ranks these based on higher probability.

V. EXPERIMENTAL RESULTS

There are three large datasets used in this experiment. The datasets are explained and the results are discussed. The dataset for the experiments are divided into 60% for training and 40% for testing. The model is retrained by changing the parameters to decrease the error rate. Six different models are built and retrained to get the best accuracy. The model that gives the best training accuracy is used to find the Critical Dimension.

A. Amazon 10,000 dataset results

The Amazon commerce review dataset [2] is a writeprint dataset. Internet users share attractive information with openness and anonymity to the online community to freely express their opinions. People with vested interests may take the opportunity to post biased information in anonymous ways, significantly harming the purpose of the open review. Therefore, authorship identification of online texts such as verifying the authorship of emails and messages on the cyber community, plagiarism detection and personal blogs is becoming important. Similar to biological fingerprint, the unique writing-style hidden in texts is vividly described as writeprint. Online writeprint identification is the task of predicting the most likely authorship of anonymous texts by using stylistic information in language. This can be seen as a single-label multiclass text categorization problem where the candidate authors represent different classes. The key task of writeprint identification is to extract fine-grained features from texts for quantifying the style of an author. Character n-grams have been proved to be very effective for capturing

 ISBN: 978-1-61208-329-2

complicated stylistic information hidden in the texts. For example, the most frequent character 4-grams of an experimental text indicate lexical (|_the|, |_to_|, |that|), word-class (|_was|, |ing_|), and punctuation usage (|,_wh|, |,_s|). This dataset are derived from the customers' reviews in Amazon Commerce Website for authorship identification. This dataset was originally created to examine the robustness of classification algorithms. Studies conducted the identification experiments for fifty authors in the online context reviews. These are the most active users (represented by a unique ID and username) who frequently posted reviews in the Amazon newsgroups. The number of reviews collected for each author is 30. This is a classification dataset and contains 50 authors x 30 reviews each = 1500 instances. There are 10,000 attributes and they include authors' linguistic style, such as usage of digit, punctuation, words and sentences' length and usage frequency of words and so on. This is a multiclass classification problem with 50 classes. The dataset contains numerical values for all features.

The classification results and the Critical Dimension are shown below. It can be seen from Figure 6 that the Amazon 10,000 dataset shows a Critical Dimension at feature size 2486. The graph below shows the results of the Amazon dataset and the plot shows the Critical Dimension.

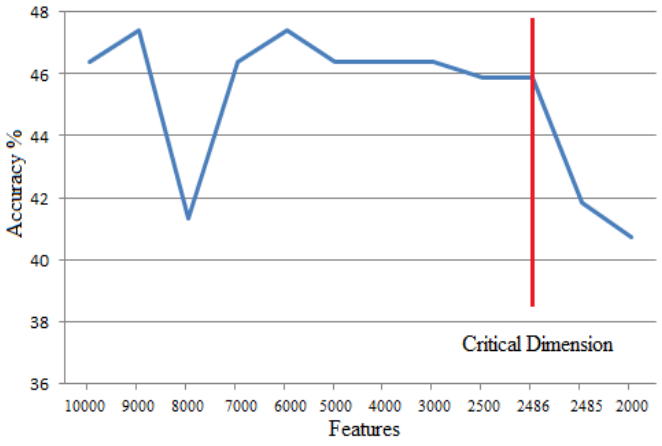


Figure 6. Critical Dimension of the Amazon 10,000 dataset

B. *Amazon ad. or non ad. dataset results*

The Amazon commerce reviews Internet advertisement dataset represents a set of possible advertisements on Internet web pages. The features encode the geometry of the image (if available) as well as phrases occurring in the URL, the image's URL and alt text, the anchor text, and words occurring near the anchor text. The task is to predict whether an image is an advertisement ("ad") or not advertisement ("nonad"). The dataset contains 459 advertisements and 2820 non ad. images, hence a total of 3279 instances. The attributes in this dataset contains 3 continuous and others binary; one or more of the three continuous features are missing in 28% of the instances. There are 1558 features in the Internet advertisement dataset.

The classification results and the Critical Dimension of the ad. and non ad. dataset is shown below. It can be seen from Figure 7 that the Ad dataset shows a Critical Dimension at feature size 383.

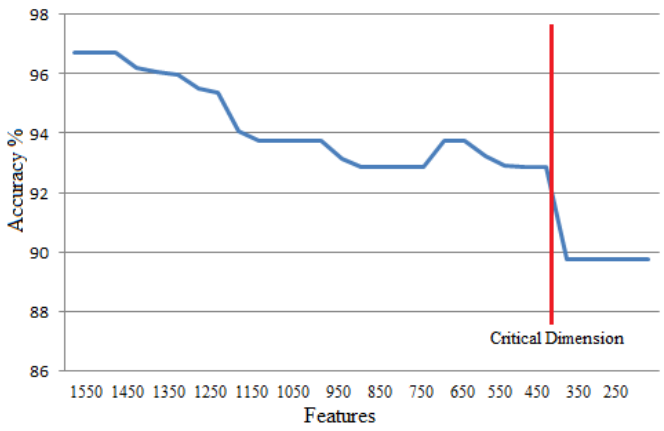


Figure 7. Critical Dimension of the Amazon ad. or non-ad. dataset

C. *Thrombin dataset results*

The present training data set consists of 1909 compounds tested for their ability to bind to a target site on thrombin, a key receptor in blood clotting. The chemical structures of these compounds are not necessary for our analysis and are not included. Of these compounds, 42 are active (bind well) and the others are inactive. Each compound is described by a single feature vector comprised of a class value (A for active, I for inactive) and 139,351 binary features, which describe three-dimensional properties of the molecule. The definitions of the individual bits are not included - we don't know what each individual bit means, only that they are generated in an internally consistent manner for all 1909 compounds. Biological activity in general and receptor binding affinity in particular, correlate with various structural and physical properties of small organic molecules. The task is to determine which of these properties are critical in this case and to learn to accurately predict the class value. Drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in the discovery of a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing many small molecules for their ability to bind to the target site. This leaves researchers with the task of determining what separates the active (binding) compounds from the inactive (non-binding) ones. Such a determination can then be used in the design of new compounds that not only bind, but also have all the other properties required for a drug (solubility, oral absorption, lack of side effects, appropriate duration of action, toxicity, etc.).

The classification results and the Critical Dimension of the thrombin dataset are shown below. It can be seen that the thrombin dataset shows a Critical Dimension at feature size 8486. Figure 8 below shows the results of the thrombin dataset and the graph plot shows the Critical Dimension.

 ISBN: 978-1-61208-329-2

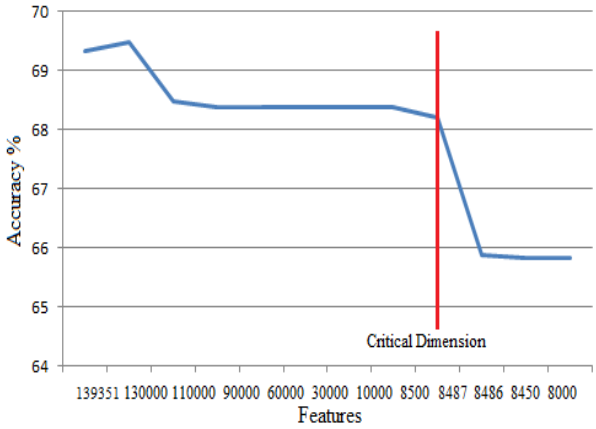


Figure 8. Critical Dimension of the thrombin dataset

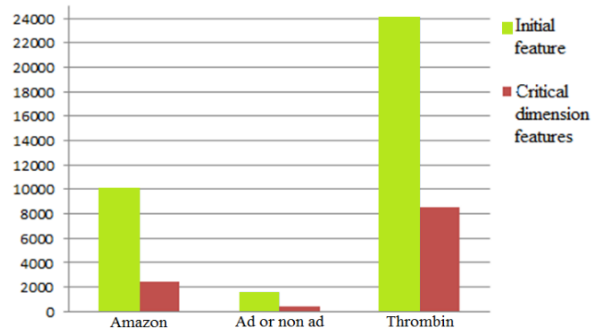


Figure 9. Reduction in the feature size of three large datasets at the Critical Dimension

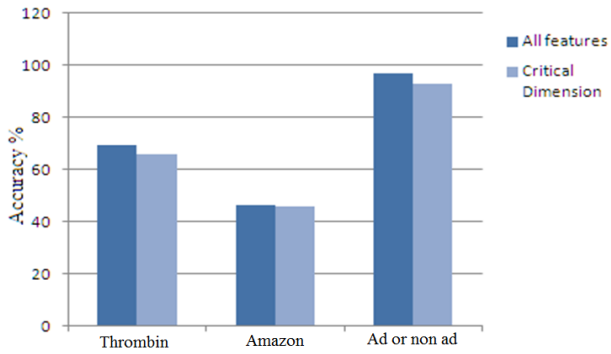


Figure 10. Prediciton accuracy of three large datasets at the Critical Dimension and intial condition (all features included)

Three large featured sized datasets namely the Amazon 10,000, the ad. or non ad. and the thrombin datasets were studied in this experiment. All three datasets shows an obvious existence of Critical Dimension. Figure 9 shows that the feature size has largely decreased at Critical Dimension and the performance is of each of these datasets are maintained 'high'. Figure 10 shows difference in accuracies at initial condition and at Critical Dimension. The initial Amazon 10,000 dataset contains 10,000 features and the accuracy with which the bagging classifier predicted the 50 classes was 46.39%. However, at the Critical Dimension the number of features was reduced to 2486 features and the accuracy to predict 50 classes using the bagging classifier was 45.88%. Similarly, for the binary class classification ad. or non ad. dataset, the initial number of features was 1559 and at Critical Dimension was 383 features. The random forest classifier the accuracy of classifying the initial dataset into the two classes was 96.71% and at Critical Dimension was 92.74%. The largest dataset namely the thrombin dataset contained 139351 features initially and using a bagging classifier, the classification accuracy to predict the two classes was 69.33%. The Critical Dimension for this dataset was then found and the number of feature at this Critical Dimension was found to be 8487 which is an enormous decrease in the feature size. At this Critical Dimension the classification accuracy was 65.87% using the bagging classifier. The results of this paper show us that a Critical Dimension is not only found in smaller datasets but also in much larger datasets. Results of 16 different datasets that were studied earlier are shown below [5]. The chart in Figure 11 shows the accuracies of all datasets. It can be seen that the accuracies at initial condition and at Critical Dimension are not very different; infact for some datasets like the Parkinson's disease and some text mining datasets the performance of the model to correctly classify has increased at Critical Dimension when compared to the performance accuracy measured at the initial condition. While the performance is maintained 'high', the feature size has decreased largely. This is shown in Figure 12.

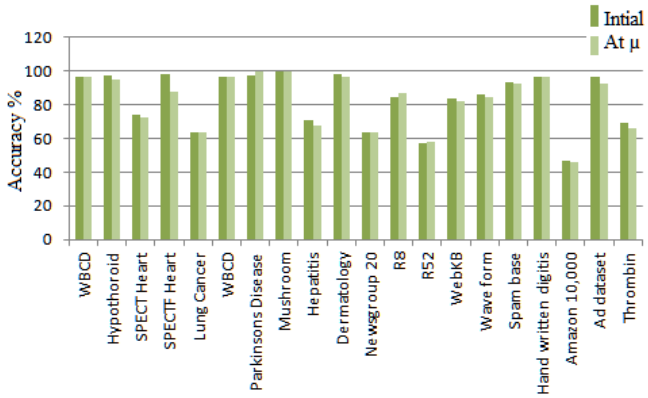


Figure 11. Accuracies of All Datasets at Initial Condition and at Critical Dimension

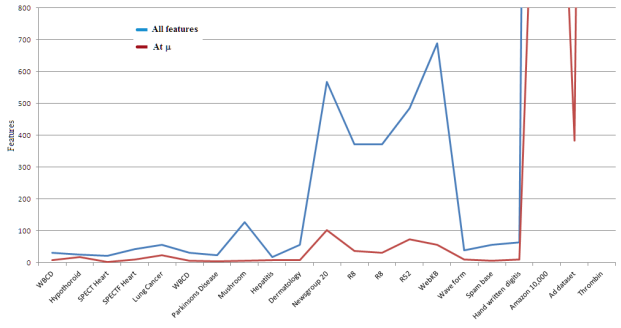


Figure 12. Reduction in the feature size of all datasets at the Critical Dimension.

VI. CONCLUSIONS AND FUTURE WORK

The concept of a Critical Dimension of datasets and a heuristic method for finding it are introduced. Firstly, we ahve shown that finding the Critical Dimension is an intractable problem, and therefore, justifies the use of heuristic methods for finding the Critical Dimension. We also presented the results showing that the heuristic method succeeded in finding the Critical Dimension of some large datasets.

Even though different feature ranking methods are used in the heuristic method, it is emphasized that this paper is not about feature tanking or selection–rather, it is about finding the Critical Dimension. The feature ranking algorithms were merely employed in the heuristic method as a means to help determine if a Critical Dimension exists for a given dataset.

However, many interesting questions are raised that deserve further investigation:

- The heuristic method may find a Critical Dimension for a given dataset and a given machine. In addition, the method identifies the features to be included. But how good the solution is (relative to the formal definition of Critical Dimension) is really unclear, since the method relies on a selected feature ranking algorithm which may well have overlooked the effect of combinations of features–though this seems inevitable for all general feature ranking algorithms that do not take into account prior knowledge about the features and/or the specific problem or application underlying the datasets.
- Using different ranking algorithms and different machines and apply the same heuristic method may lead to very different CD values, what does that mean?
- What does the existence of a CD mean for a dataset? Does it mean that the quality of data is low–since insignificant and/or useless features are included? Or does it mean that the amount of data is in fact insufficient–once the dataset is expanded with more data, might the CD disappear? Both seem to be possibilities.

More fundamentally, how do we count features? What is a feature? In many problems, features are developed and computed from the collected simple measurements (e.g., the TF-IDF feature in text classification). But this seems a different issue regarding prior knowledge. The authors are pursuing, as the next steps of this work, to develop and experiment with more sophisticated (than the linear) heuristic methods for finding the features that constitute the Critical Dimension, and apply the methods to larger datasets.

ACKNOWLEDGMENT

Support for this work received from ICASA (Institute for Complex Additive Systems Analysis) of New Mexico Tech and the National Institute of Justice, U.S. Department of Justice (Award No. 2010-DN-BX-K223) is gratefully acknowledged.

REFERENCES

[1] A. Blum and P. Langley, “Selection of relevant features and examples in machine learning”, Artificial Intelligence, vol. 97, pp. 1-2, 1997.

[2] A. Frank and A. Asuncion, UCI Machine Learning Repository, Irvine, CA: University of California, School of Information and Computer Science, http://archive.ics.uci.edu/ml, [retrived: May, 2010].

[3] A. M. A. Mesleh, “Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System”, Journal of Computer Science, ISSN 1549-3636, pp. 430-435, 2007.

[4] C. H. Papadimitriou and M. Yannakakis, “The complexity of facets (and some facets of complexity)”, Proceedings of the fourteenth annual ACM Symposium on Theory of Computing, pp. 255-260, 1982.

[5] D. Suryakumar, A. H. Sung, and Q. Liu, “Determine the Critical Dimension in data mining (experiments with bioinformatics datasets)”, In Intelligent Systems Design and Applications, 2011 11th International Conference, pp. 48-486, 2011.

[6] H. Almuallim and T. G. Dietterich, “Learning with many irrelevant features”, Ninth National Conference on Artificial Intelligence, MIT Press, pp. 547-552, 1991.

[7] H. Buhrman and J. M. Hitchcock, “NP-hard sets are exponentially dense unless coNP ⊆ NP/poly,” IEEE Conference on Computational Complexity, IEEE Computer Society Press, pp. 600-601, 2008.

[8] I. Guyon, A. Elisseeff, “An Introduction to Variable and Feature Selection”, Journal of Machine Learning Research, pp. 1157-1182, 2003.

[9] J. G. Dy and C. E. Brodley, “Feature Subset Selection and Order Identification for Unsupervised Learning”, Proceedings of the Seventeenth International Conference on Machine Learning, pp. 247-254, 2001.

[10] J. R. Quinlan, “C4.5: Programs for Machine Learning”, Morgan Kaufmann Machine Learning series, 1993.

[11] L. Breiman, “Random forests”, Journal Machine Learning, Vol 45(1) pp. 5-32, 2001.

[12] W. Buntine, “Theory refinement on Bayesian networks”, Proceedings of the Seventh, Annual Conference on Uncertainty in AI, pp. 52–60 1991.

[13] X. Geng, T. Liu, T. Qin, and H. Li, “Feature selection for ranking”, In Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 407-414, 2007.

 ISBN: 978-1-61208-329-2

Application of Business Process Quality Models in Agile Business Process Management

Michael Gebhart
Gebhart Quality Analysis
(QA) 82 GmbH
Karlsruhe, Germany
michael.gebhart@qa82.de

Marco Mevius, Peter Wiedmann
HTWG Konstanz
Constance, Germany
mmevius/pewiedma@htwg-konstanz.de

***Abstract*— The increasing complexity of business process management projects requires a methodology that supports the tight and efficient collaboration between customer and process analysts. For that purpose, the agile methodology that is well-known in software projects has been transferred to business process management. However, in these agile environments governance regarding the designed processes is necessary for ensuring their high quality. This article demonstrates how to apply quality models for business processes in agile business process management environments and its specific challenges. To illustrate the application, a business process in the context of offer management has been captured by means of this approach.**

Keywords-business process; design; quality; agile; ISO 25000.

I. INTRODUCTION

High business process flexibility is required for companies to counter current challenges. Fast and efficient adaptability to business processes becomes an increasingly important competitive factor [1]. Explicit knowledge about the structure and functionality of business processes is essential for the understanding of organizational sequences [2]. A targeted enhancement of Business Process Management (BPM) with the help of agile advantages generates new significant potential for the automation, modeling, interaction and optimization of business processes. Therefore, different (agile) approaches have been developed. The idea of agility is described as the ability to balance flexibility and structure [3] and to minimize risks for instance by conforming project changes rapidly [4]. One of these approaches is called BPM(N)$^{Easy1.2}$.

With BPM(N)$^{Easy1.2}$ an agile BPM method is introduced [5][6]. BPM(N)$^{Easy1.2}$ describes a combination of Business Process Management and Business Process Model and Notation (BPMN) with the ambition of making BPM easier. The major intention of the method is to provide aspects of agile software engineering for BPM. The approach extents and supports the interaction between every participant with focus on more coherency without confronting them with unneeded complexity. Furthermore, it follows an empirical, incremental and iterative concept to increase predictability of the process quality and to reduce project risks [5]. Hereby the efficiency and effectiveness of BPM will be enhanced.

However, within the prediction and control of the business process quality the participants have to know what constitutes a good process and how to evaluate processes [7]. But there are no general rules which define what a good process is. Aspects, such as the customer value, process standardization, and the employee well-being, can be a signal [8]. But this information is not sufficient to perform a systematic or even automatic quality analysis of business processes. Aggravating this situation, contradictory constraints and needs – for instance speed and quality – generate the need to focus on the delivering value [9].

To enable a systematic quality assurance in agile BPM, this paper introduces the application of quality models and quality gates. Quality gates define a specific point within a project to evaluate determined maturity and sustainability [10]. These quality gates ensure the synchronization and acceptance of all participants. For instance, an automated business process has to correspond with all predefined requirements and expectations. The introduced quality gates are supposed to close the gap especially at the beginning and during the business process modeling step. In the area of BPM and business process quality measuring, different approaches already exist. As quality model, these existing quality assurances are reused as well as evaluated and adapted [11] for applying and measuring them in BPM environments. Especially agile environments with short iterations and high interaction are suited for the continuous monitoring of the business process quality.

The paper is organized as follows: Section 2 analyses relevant literature regarding quality models for BPM and their application for agile BPM. In Section 3, the application of a certain business process quality model in agile business process management is illustrated by means of a scenario from offer management. Section 4 introduces the BPM(N)$^{Easy1.2}$ method and demonstrates where quality models can be applied within an agile approach. In addition, a possible tool support is shown. The last section presents a conclusion and outlook.

 ISBN: 978-1-61208-329-2

II. RELATED WORK

This section describes the fundamental terms and existing work in the context of measuring the quality of business processes. For that purpose, work that targets the quality from both a functional and a technical point of view is considered. Furthermore, this work is examined in detail regarding its applicability in agile BPM environments.

The International Organization for Standardization (ISO) and the International Electro technical Commission (IEC) have created standards regarding the quality of software products. Both ISO/IEC 9126 [13] and the successor ISO/IEC 25000 ff. [14] define relevant terms for software product quality. Furthermore, they describe quality characteristics, their subcharacteristics, and their final quality measure elements. They hereby provide a wide overview of measuring the quality of software products. In order to apply these standards on business processes, the term "business process" has to be distinguished from "business process model". As the standards refer to software products, they can only be directly applied on business process models as software artifacts. Also, in this case, only a subset of described characteristics is applicable. Heinrich et al. [16], Sánchez-González et al. [17], and YeonSeok et al. [18] show the adaptation of these standards on business process models. However, according to the introduction, we focus on the quality of business processes and their content instead of the models as software artifacts and their syntactical correctness etc. For that reason, the standards cannot be applied directly. Nevertheless, they provide good hints about characteristics that might be important for business processes as well.

Further standards regarding quality management focus on quality management systems. Examples are ISO 9000 ff. [15], or branch-specific manifestations, such as the European Norm (EN) 9100 for aerospace. There also exist standards for the quality management in projects, such as ISO 10006. Even though they consider the quality in business domains and in some cases also describe business processes, the quality of the business processes themselves is not explained in detail. This is also the case when choosing Capability Maturity Model Integration (CMMI) or IT Infrastructure Library (ITIL).

In [8], Krogstie describes criteria for so-called good processes. He introduces dimensions of value that is valid for most customer groups. Furthermore, he summarizes heuristics for good business processes. Even though no metrics are provided, these heuristics can be good starting points to derive more concrete quality aspects that again enable a systematic and automatic evaluation of business processes. In addition, this work helps to understand the purpose of business processes and why it is important to have good processes. Thus, it forms the framework for a quality model as it focuses on the motivation and strategic goals of business processes.

In order to enable a more systematic quality analysis of business processes, Kneuper created the quality model Gokyo Ri based on existing standards, such as ISO 9000, CMMI, and ITIL [19]. It refines the quality of business processes so that their quality can be determined. Even though this quality model focuses on business processes and their content, the quality model is still too abstract to be used in agile business process management environments. In agile projects the quality has to be determined in short intervals best automated based on modeled business processes. Thus, Gokyo Ri has to be further refined until at least a subset of the quality attributes can be determined automatically or with short user interaction intervals.

Similarly, Lohrmann et al. introduce quality attributes for business processes [7]. Also, in this case the quality attributes are derived from business-related quality concerns and focus on the content of the business process and not the artifact. Lohrmann et al. distinguish between the efficacy and efficiency of business processes that can be either determined on basis of business process models and running instances. Former is called business process design and implement efficacy and efficiency. Latter is described as business process enactment efficacy and efficiency. Even though Lehrmann et al. do not describe an entire quality model, they introduce quality attributes that are relevant for the business process quality as considered in this article. Nevertheless, similar to the quality model introduced by Kneuper the quality attributes are still too abstract to be applied in an agile environment. They first have to be refined so that they can be determined either based on business process models or by answering simple questions by process analysts.

Regarding a more technical point of view, Suarez et al. [20] describe best practices for modeling business processes using certain languages, such as the Business Process Model and Notation (BPMN). Even though this article also focuses on BPMN as modeling language, these best practices mostly consider syntactical correctness of created models or related issues. The content of the processes and their quality from a functional point of view is not considered. The described best practices are also not aligned with a holistic quality model. So, the impact of these best practices on abstract quality characteristics is not obvious. The best practices can increase the quality of modeled business processes. They are also applicable in agile business process management environments as they can be easily determined or can be even measured automatically by tools. Nevertheless, they do not target the kind of business process quality considered in this article.

Thus, this overview shows that there exists work considering the quality of business processes from a functional point of view as required in this article. However, the introduced quality attributes are too abstract to be measured directly and especially too heavyweight to be determined in agile environments. Other work focuses on fine-grained quality aspects, such as syntactical correctness that can be easily determined, however does not provide value for the quality of business process from a functional point of view. This article shows how to fill this gap by reusing existing work as introduced by Lohrmann et al. [7] and breaking these quality attributes down into aspects that can be either directly measured on business process models or easily answered by process analysts.

 ISBN: 978-1-61208-329-2

The methodology applied in this article has already been successfully applied for service-oriented architectures [12]. Also, in this case, existing abstract quality attributes were refined to enable a fully or partially automated quality assurance. As result, a solution was created to ensure the systematical creation of a flexible and maintainable architecture.

III. SCENARIO

A sample business process model has been selected to apply the quality model. The business process model originates from a real business process model repository of an industry partner. The model describes the business process of an "offer creation". The activities consider the aspects from setting up a new offer until sending it out to a potential customer. The business process model is modeled with BPMN$^{Easy1.2}$. BPMN$^{Easy1.2}$ is a business process modeling language which uses BPMN 2.0 [21] but reduces the complexity of the first modeling step. In the second step, the model can be enriched and used, e.g., for business process automation. Fig. 1 shows the business process model.

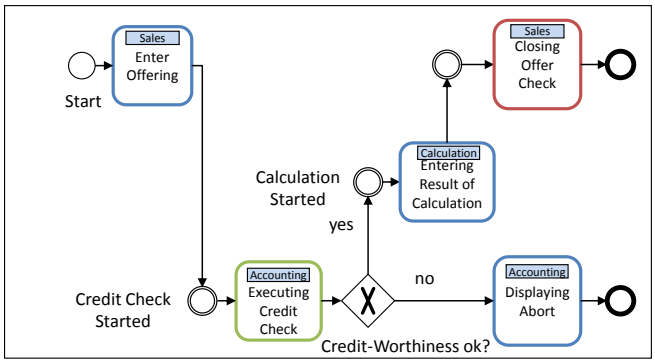


Figure 1. Offer creation business process.

The business process requires three roles: Sales, Calculation and Accounting and follows two different paths. In case of a successful credit check the Sales can finalize the offer otherwise the business process will be aborted. During the scenario the first draft of the business process model has been designed. BPMN$^{Easy1.2}$ provides three different activity types: manual (green form), semi-automated (blue form) and automated (red form). For instance, the "Enter Offering" activity is computer-aided and can be defined as a semi-automated activity. In addition, the required user stories have been described according to an agile methodology. To verify the correctness, participants interacted with each other closely. To prove the quality of the business process model it was necessary to use a specified quality model. In the following section, the developed quality model will be applied to this business process.

IV. APPLICATION OF BUSINESS PROCESS QUALITY MODELS IN AGILE BUSINESS PROCESS MANAGEMENT

In order to apply a certain business process quality model in an agile business process management project, several questions have to be answered. These constitute the structure of this section.

A. Agile BPM

There are different approaches to agile BPM e.g. [5] [22] [23]. In the following section the agile approach BPM(N)$^{Easy1.2}$ [5] is used to show when (time of application) during the methodology the quality model is expected to be applied. BPM(N)$^{Easy1.2}$ enables highly sophisticated agile Business Process Management. It covers all aspects of Business Process Management – from process design and process execution to process controlling with focus on the integration of all process participants. The following Fig. 2 provides an overview of the approach and the including quality gates. Latter are displayed as stars:

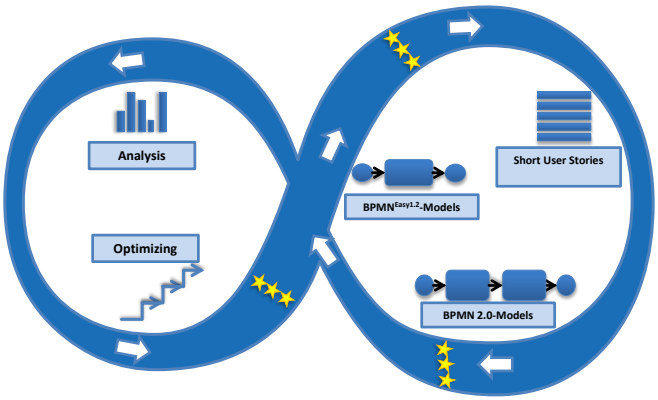


Figure 2. Illustration of the BPM(N)$^{Easy1.2}$ approach.

The approach consists of two connected cycles. One cycle is used to capture new BPMN$^{Easy1.2}$ models and short user stories. Both BPMN$^{Easy1.2}$ models and short user stories formulate the requirements of the activities within a business process. The BPMN$^{Easy1.2}$ models are used to design the flow in general and set up a first model very easily. The short user stories describe additional information, e.g., additional business rules. The formulated requirements are the basis for the modeling and implementation of an enriched BPMN 2.0 business process. For the enrichment a BPMN$^{Easy1.2}$ model and a number of user stories are selected to work on. Furthermore, the business process is modeled on the business user's point of view. In addition, in consultation with a business user, an IT expert is able to use the business process model to automate the process. Once the modeling and implementation stages are completed the resulting BPMN 2.0 models are transferred to a final control. Within this control all participants assure that the result e.g. an automated business process corresponds with the BPMN$^{Easy1.2}$ models and formulated short user stories (synchronization and acceptance).

 ISBN: 978-1-61208-329-2

Immediately after the acceptance, new requirements can be taken and transformed into a business process model or implementation. If defined key performance indicators show optimization potential (analysis and optimizing cycle), new BPMN$^{Easy1.2}$ models or short user stories will be generated. The several iteration and high collaboration between every participant allows the continuous monitoring of the business process quality [5].

However, in general there are still different weak points in agile methods. Mohammad [24] says short response times and high interaction during the agile development do not require the writing of documents which can lead to a reduced quality of documentation. Furthermore, Mohammad [24] mentions the increased collaboration time of the participants. But in fact in some circumstances there is not enough time for the required coordination or the participants are not at the same (physical) location [24]. In [25], agile methods are described as a risk of large or complex projects. The magnitude of uncertainty is increased. Therefore agile methods are mistrusted in most organizations. To counteract these disadvantages and related lack of quality it is required to introduce quality checks during the application of an agile approach. In [26], quality checks are suggested to be applied to different steps of agile approaches.

According to [26] and with the assumption that software engineering has the same goals as Business Process Management, e.g., cost reduction, collaboration enhancement, the quality gates listed in Table I are suggested for agile BPM approaches:

TABLE I. QUALITY GATES

Quality Gate	Time of Application	Comment
1	Formulation of user stories	Continuous feedback and collaboration between every participant
2	Modeling of business process	
3	Automation of business process	Test of process application
4	Acceptance testing	

Today, some of the quality gates have already been implemented to assure the determined quality. For instance, the quality gate 1 can be applied by a continuous feedback process between every participant or by means of standard assurance tests of the process application [27]. For quality gates 1, 3 and 4 methods already exists, which can be used to assess the quality e.g. real tests of a process applications.

Therefore, in this article, the quality gate number 2 that is applied during the modeling of business process is considered to improve and guarantee the expected quality.

B. Quality Model Choice and Adaptation

In the previous section, quality gates during an agile methodology have been identified. One quality gate considers the quality of modeled business process. In order to support this quality assurance, an appropriate quality model has to be prepared. For that purpose, first the most appropriate existing quality model has to be identified. Afterwards, its direct applicability has to be verified. As described in Section II, appropriate quality models are those introduced by Lohrmann et al. [7] and Kneuper [19]. However, in both cases, the introduced quality attributes have to be adapted for requirements in agile environments: As mentioned before, the quality of business processes has to be determined in short intervals, which again requires a quality analysis to be easy and lightweight. This requirement cannot be fulfilled by these existing quality models and the contained quality attributes. They are not formalized using metrics which hampers their automatic determination based on business process models. Furthermore, the informal description requires interpretation effort that can result in misunderstandings and thus wrong measures. This is a typical issue when performing quality analyses and has already been identified for other domains, such as the quality analysis of service-oriented architectures by Gebhart et al. [28][29].

Thus, after choosing a certain quality model, the quality attributes have to be refined if necessary until more fine-grained and comprehensible quality attributes are identified so that no interpretation is necessary any longer. They are called quality indicators, formalized as metric, and return a measure. It is not necessary that a quality indicator can be fully automatically measured on process models. If this is not possible as they require further knowledge, such as domain knowledge, the only condition is that it is possible to formulate unambiguous questions that can be answered by experts and do not require interpretation. Summarized, for every function and variable used within a metric, the criteria listed in Table II have to be fulfilled.

TABLE II. CRITERIA FOR FUNCTIONS AND VARIABLES IN METRICS

Criterion	Description
Technology Representation *for variables and functions*	A variable or function represents a certain aspect within the considered technologies, i.e., business process models in this case. This enables an automatic measurement.
Comprehensible Question *for variables and functions*	If Technology Reflection is not fulfilled, for example if expert knowledge is necessary, a comprehensible and unamabigious question can be formulated that can be answered by experts and does not require interpretation.
Composition *for functions*	If the previous criteria are not fulfilled, the considered function is composed of other functions using automatically measurable operators.

In this article, the quality model and its attributes introduced by Lohrmann et al. [7] are chosen. The refinement and application in an agile environment is exemplified by means of two quality attributes and their correlating quality predicates.

 ISBN: 978-1-61208-329-2

1) Controlled resource consumption in activities: According to Lohrmann et al. a business process fulfills this predicate when activities within the process are designed to avoid materials waste and capacity waste. This information is too abstract to be comprehensible on a certain business process model as it is not explained how this waste is reflected in process design. For that reason, the predicate and its quality attribute have to be refined into quality indicators.

For this purpose, best practices that could be identified in earlier projects are tested for their suitability to represent the considered predicate and its quality attribute. One best practice suitable in this case is that for every role at least two persons have to be available. This ensures that in case of a person being absent still another person can continue the work and other persons do not have to wait and to be idle, which represents a capacity waste. As the predicate refers to the business process as a whole, also the refinement has to be measured on the entire process. Thus, the indicator measures the degree to which the participating roles have more than one person assigned. This indicator can be formalized as metric (1) similar to the ones introduced by Gebhart et al. in [28]. Table III describes the used elements.

$$PAR(bp) = \frac{|\,F(R(bp),r,HSP(r))\,|}{|\,R(bp)\,|} \quad (1)$$

TABLE III. VARIABLES AND FUNCTIONS USED FOR PAR (1)

Element	Description
PAR(bp)	*Person Availability of Roles:* Degree to which roles in business process bp have more than one person assigned
R(bp)	*Role of Business Process:* roles used in business process bp
F(e, v, c)	*Filter:* filter the elements e by condition c that uses the variable v as iterator
HSP(r)	*Role Has Several Persons:* true if role r has more than one person

TABLE IV. FULFILLED CRITERIA FOR PAR (1)

Element	Fulfilled Criteria
bp	Technology Representation: The considered business process is represented by the BPMN process file
PAR(bp)	Composition: This function is composed of other functions and all operations can be automated.
R(bp)	Technology Representation: The roles are represented by the pools and lanes within the BPMN business process model
F(e, v, c)	Composition: This function is requires other functions as input and the filter operation can be automatically performed.
HSP(r)	Comprehensible Question: This aspect is not measurable on standard BPMN 2.0 artifacts. Thus, it has to be answered by an expert, but the question is easily to understand, unambigious and comprehensible: "Are more than one person assigned to role r?" As input, a boolean value is expected.

In order to prove the suitability of this quality indicator as quality indicator in an agile environment, in Table IV for every element used in the formalization the criteria introduced in Table II are checked. As mentioned before, we assume business process models using BPMN 2.0 [21].

For the sake of simplicity, we focus on this best practice as solely quality indicator for the considered predicate. If further best practices, standards, or guidelines can be identified as influencing quality indicators they can be added later and have to be weighted.

Applied on the scenario introduced in Section III, the metric returns a value less than 1 as we assume that not every role is filled by at least two persons yet, i.e., HSP(r) is not true for all roles. Table V shows how to interpret this value. In order to fulfill the predicate of controlled resource consumption in activities, the metric is expected to return 1 as desired value. Thus, the business analyst is made aware to ensure that some further persons have to be assigned to roles with only one person. Even though if this is not possible, the business analyst gets the information that this fact represents a critical point for the efficiency of the business process.

TABLE V. INTERPRETATION OF VALUES FOR PAR (1)

Value	Interpretation
0	No role within the business process is filled with at least two persons
Between 0 and 1	Some roles are filled with less than two persons
1	All roles within the business process are filled with at least two persons

2) Controlled skill employment: A business process can only be efficiently performed when skill employment is controlled. According to Lohrmann et al. [7], this quality attribute or predicate is fulfilled when all activities are documented and trained. This refinement can be used as measurement. In BPMN, these activities are represented by manual tasks or tasks that are not further specified yet.

$$CSE(bp) = \frac{DT(bp)+TT(bp)}{2} \quad (2)$$

$$DT(bp) = \frac{|\,F(MT(bp),t,D(t))\,|}{|\,MT(bp)\,|} \quad (3)$$

$$TT(bp) = \frac{|\,F(MT(bp),t,T(t))\,|}{|\,MT(bp)\,|} \quad (4)$$

TABLE VI. VARIABLES AND FUNCTIONS USED FOR CSE (2, 3, 4)

Element	Description
CSE(bp)	*Controlled Skill Employment:* Degree to which skill employment is controlled in business process bp
DT(bp)	*Documentation of Tasks:* Degree to which manual tasks in business process bp are documented.
D(t)	*Documentation:* true if task t is documented
TT(bp)	*Training of Tasks:* Degree to which manual tasks in business process bp are trained.
T(t)	*Training:* true if task t is trained

 ISBN: 978-1-61208-329-2

Also, in this case, all used functions and variables are described in Table VI. They fulfill the required criteria described in Table II. The manual tasks represent certain aspects within the technology and the other functions are either composed of others or comprehensible questions can be formulated as for D(t) and T(t).

Applied on the scenario we assume that all tasks represent manual tasks as automation has not been specified yet. When the metric is calculated, the business analyst has to answer, whether all these tasks are documented and trained. We assume that the business analyst realizes now that this is not the case. Only some tasks are documented and trained. Thus, the metric returns a value less than 1. The interpretation of this value is shown in Table VII.

TABLE VII. INTERPRETATION OF VALUES FOR CSE

Value	Interpretation
0	No manual task within the business process is documented or trained
Between 0 and 1	Some manual tasks within the business process are documented and trained
1	All manual tasks within the busienss process are documented and trained.

By applying the refined metrics, the business analyst is made aware that the documentation and training is important for the efficiency of the business process. If the metric returns a value less than 1 the analyst gets the information that further documentation and training effort is necessary.

C. Tool Support

In order to increase the efficiency of quality analyses especially in agile environments, an appropriate tool support is necessary. For that purpose the already existing QA82 Analyzer [30] (Fig. 3) can be applied as it is suited for agile environments and hybrid quality indicators identified in the previous section.

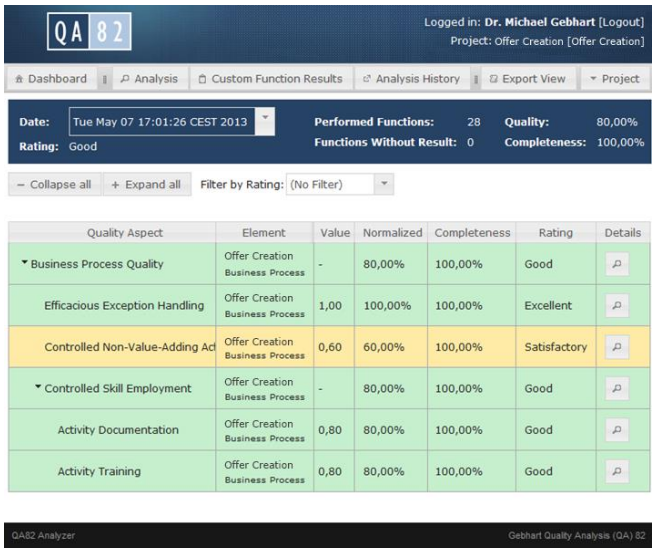


Figure 3. QA82 Analyzer to analyze business process.

First, it supports the integration of custom quality models and combines the measure of model elements with questions that can be answered by experts, i.e., process analysts in this case. Second, the QA82 Analyzer can be integrated in business process modeling tools, such as BPM(N)$^{Easy1.2}$, using web services. This enables the display of quality analysis results directly in existing environments. Finally, the QA82 Analyzer allows the provisions of advices about how to improve the quality. As result, process analysts can model business processes using their modeling tool and directly get hints about how to design the process to improve their quality based on the custom quality model.

For that purpose, the quality model based on the quality attributes of Lohrmann et al. and the derived quality indicators has to be formalized and integrated into the QA82 Analyzer. This includes the mapping of functions to technology, i.e., to BPMN 2.0 artifacts, and the formulation of appropriate questions if necessary. As result, the QA82 Analyzer can be used to apply the identified quality indicators on any BPMN 2.0 compliant business process.

V. CONCLUSION AND OUTLOOK

In this article we demonstrated the application of business process quality models to support agile business process management and to assure a high quality of created solutions. For that purpose, we exemplarily chose the quality model introduced by Lohrmann et al. [7]. We identified the essential challenges and showed how to address them. First, the application of business quality models was aligned with an agile methodology. As essential deficit the abstraction of available quality attributes was identified. To solve this issue, we demonstrated how these quality attributes can be refined to be applicable in agile environments. Finally, we illustrated necessary tool support to increase the efficiency of quality analyses.

To illustrate our work, a scenario in the context of a real offer creation business process was chosen. The refined quality attributes enabled the systematic analysis of this process and the results helped the process analysts to revise the process and its environment in a quality-oriented manner. Even though the quality of a business process includes a lot of further aspects not covered in this article, the application of a fine-grained quality model increases the awareness of relevant aspects and supports the creation of high-quality business processes.

Thus, our approach enables companies and their process analysts to increase the quality of created business processes whilst reducing at the same time effort and costs for quality assurance. Process analysts can create business process models using their preferred modeling tool, such as BPM(N)$^{Easy1.2}$, and directly receive feedback about their quality. Finally, derived advices are shown and help them to improve the created business models with regard to quality attributes that influence business-related goals.

Next, we will consider further quality attributes and derive appropriate quality indicators to enhance the created quality model. As described in this article, we will focus on reuse of existing quality attributes.

 ISBN: 978-1-61208-329-2

We also plan to refine the tool support. In particular, the integration with existing modeling tools has to be enhanced. Finally, the approach is expected to be applied in further business process management projects to identify advantages and also weaknesses that have to be examined.

REFERENCES

[1] T.M. Bekele and Z. Weihua, "Towards collaborative business process management development current and future approaches", in Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference, 2011, pp.458-462.

[2] J. Eder and H.Pichler, "Business Process Modeling and Workflow Design", in D.W. Embley and B. Thalheim (eds), Handbook of Conceptual Modeling, Springer, 2011, pp. 259-286.

[3] J. Highsmith, "Agile Software Development Ecosystems", Pearson Education, 2002, pp. 26-34.

[4] Agile Manifesto, http://agilemanifesto.org. [accessed: May 21, 2013]

[5] M. Mevius, R. Stephan and P. Wiedmann, "Innovative Approach for Agile BPM", in The Fifth International Conference on Information, Process, and Knowledge Management eKNOW 2013, ISBN 978-1-61208-254-7.

[6] M. Mevius and P. Wiedmann, „BPM(N)$^{Easy1.2}$ –Gebrauchssprachliche Gestaltung IT-basierter Prozesse. BSOA 2013. 8. Workshop „Bewertungsaspekte service- und cloudbasierter Achitekturen" der GI Fachgruppe „Software-Messung und -Bewertung", Basel, 2013, pp. 31-46.

[7] M. Lohrmann and M. Reichert, "Understanding business process quality", in Business Process Management - Theory and Applications, Springer, 2013, pp. 41-73.

[8] J. Krogstie, "Model-Based Development and Evolution of Information Systems – A Quality Approach", Springer-Verlag, London, 2012.

[9] J. Highsmith, "Agile Project Management – Creating Innovative Products", Addison-Wesley Professional, 2009, pp. 15-16.

[10] F. Salger, M. Bennicke, G. Engels , and C. Lewerentz, "Comprehensive Architecture Evaluation and Management in Large Software-Systems", in 4th International Conference on the Quality of Software Architectures, 2008, pp.205-219.

[11] E. Mnkandla and B. Dwolatzky, "Defining Agile Software Quality Assurance", International Conference on Software Engineering Advances, IEEE 0-7695-2703-5/06, 2006.

[12] M. Gebhart, "Service identification and specification with SoaML", in Migrating Legacy Applications: Challenges in Service Oriented Architecture and Cloud Computing Environments, Vol. I, A. D. Ionita, M. Litoiu, and G. Lewis, Eds. 2012. IGI Global. ISBN 978-1-46662488-7.

[13] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), ISO/IEC Standard No. 9126: Software Engineering – Product Quality; Parts 1–4.

[14] International Organization for Standardization (ISO) and International Electrotechnical Commission (IEC), ISO/IEC Standard No. 25000: Software Engineering – Software Product Quality Requirements and Evaluation (SQuaRE) - Guide to SQuaRE.

[15] International Organization for Standardization (ISO), ISO Standard No. 9000: Quality Management Systems – Fundamentals and Vocabulary.

[16] R. Heinrich and B. Paech, "Defining the quality of business processes", in Lecture Notes in Informatics Vol. P-161, 2010, pp. 133-148.

[17] L. Sánchez-González, F. Ruiz, F. García, and M. Piattini, "Improving quality of business process models", in Evaluation of Novel Approaches to Software Engineering, 2013, pp. 130-144.

[18] L. YeonSeok, B. JungHyun, and S. Seokkoo, "Development of quality evaluation metrics for BPM (business process management) system", in Proceedings of the Fourth Annual ACIS International Conference on Computer and Information Science, 2005, pp. 424-429.

[19] R. Kneuper, „Gokyo ri: messung und bewertung der qualität von entwicklungsprozessen am beispiel des v-modell xt", in O. Linssen, M. Kuhrmann (Eds.): Qualitätsmanagement und Vorgehensmodelle, Shaker, 2012, pp. 25-34.

[20] G. N. Suarez, J. Freund, and M. Schrepfer, "Best practice guidelines for BPMN 2.0", in BPMN 2.0 Handbook, 2011, pp. 153-165.

[21] OMG, "Business process model and notation (BPMN)", Version 2.0, 2011.

[22] R. Meziani and I. Saleh, "Towards a Collaborative Business Process Management Methodology", International Conference on Multimedia Computing and Systems (ICMCS). 2011.

[23] F. Schnabel, Y. Gorronogoitia, M. Radzimski, F. Lecue, N. Mehandjiev, G. Ripa,, et al."Empowering Business Users to Model and Execute Business Processes", BPM 2010 Workshops, Springer, 2011, pp. 433-448.

[24] A. H. Mohammad, T. Alwada'n , and J. Adabneh, "Agile Software Methodologies: Strength and Weakness", in International Journal of Engineering Science and Technology, 2013, ISSN : 0975-5462, pp.455-459.

[25] J. Barlow, J. Giboney, M. Keith, D. Wilson, R. Schuetzler, P. Lowry, and A. Vance, "Overview and Guidance on Agile Development in Large Organization", Communications of the Association for Information Systems, Vol. 29, No. 2, 2011, pp. 25-44.

[26] M. Huo, J. Verner, L. Zhu , and M. Babar, "Software Quality and Agile Methods", 28th Annual International Computer Software and Applications Conference, IEEE 0730-3157/04, 2004.

[27] A. Janus, A. Schmietendorf, R. Dumke , and J. Jäger, "The 3C Approach for Agile Quality Assurance", Emerging Trends in Software Metrics (WETSoM), 2012 3rd International Workshop, 2012.

[28] M. Gebhart and S. Abeck, "Metrics for evaluating service designs based on SoaML", International Journal on Advances in Software, 4(1&2), 2011, pp. 61-75.

[29] M. Gebhart and S. Abeck, "Quality-oriented design of services", International Journal on Advances in Software, 4(1&2), 2011, pp. 144-157.

[30] Gebhart Quality Analysis (QA) 82 GmbH, QA82 Analyzer, http://www.qa82.com. [accessed: May 21, 2013]

 ISBN: 978-1-61208-329-2

A Study on Innovation Diffusion Understanding with Multi-Agent Simulation

Takao Nomakuchi
Faculty of Economics
Wakayama University
Wakayama, Japan
e-mail: tnoma@eco.wakayama-u.ac.jp

Masakazu Takahashi
Graduate School of Innovation and Technology
Management, Yamaguchi University
Ube, Japan
e-mail: masakazu@yamaguchi-u.ac.jp

***Abstract*— This paper describes the Innovation diffusions with Multi Agent Simulation. Among the Innovation diffusion theories, there are five classification types such as Innovator, Early adopter, Early majority, Late majority, and Laggard, so far. Among each classification, there are cracks in the innovation adaptation. In particular, in the high-tech industries, a big slot called Chasm, proposed by Moor, is made between Early adopter and Early majority. These adaptations are based on human homogeneous behavior in social contacts from the results of observation in the real world. This theory is even heuristics intelligence and one cannot capture the conditions for the crack made. Based on the innovation diffusion backgrounds, we made a simulator for Chasm observation. The simulation results confirmed that a Chasm crack was made in the industry with Multi-Agent Simulation. We have attempted to acquire new knowledge for the industry.**

Keywords-Innovation; Innovation Diffusion; Multi-Agent Simulation; Simulation; Chasm.

I. INTRODUCTION

According to Matsuoka [1], the products diffused generally a lose competitive power for catching up with newly emerging countries, such as China, South Korea, and Taiwan. As a result, the domestic industry is forced to severe market competition by losing their global competitiveness. For example, there are household electrical appliances and high tech products, such as cell phone and television, which the Japanese companies are good at making. Therefore, in Japan, producing an innovation that will be prosperous for the high-tech industry is advocated.

Although many policies for future industry generation have been made by the Japanese Government, they have not contributed to the recovery from the long term recession in Japan. It is one of the reasons it is not possible to identify under what kind of conditions innovation starts so far. Then, the diffusion process of an innovation is considered with the multi-agent simulation that reproduces Chasm based on the diffusion theory of the innovation by Moore [2].

According to the survey about mid- and long-term research and development of companies in our country that contributes to innovation creation by the Ministry of Economy Trade and Industry Japan [3], mid- and long-term research and development of the companies in Japan contribute to innovation creation. In this survey report, it is necessary for Japanese companies not to consider business plans focusing on short-term profits. In addition, future vision plans to create new venture business are needed for Japanese companies.

Hasegawa [4] mentioned that there is a need to emphasize the existing market for large companies. Therefore, the creation of new markets is difficult for large companies. Also, in these domains, there is a view that a venture business is suitable depending on the industry type. However, a venture business does not develop easily in our country. As it became clear from this investigation, it may be necessary to challenge to enter into a new domain. Concentrated effort and investment by industry, and administrative and academic sectors in these domains are not being fully performed by either minor or leading companies.

The advanced expansion to overseas markets that correspond to an emerging country is also included in new market creation. The reason why venture company's leading research does not develop is because of the duration it takes to disseminate the technology. This is because the body of knowledge stagnates and a gap (called Chasm) opens. Since this stagnation exists and a gap develops, destructive innovation is assumed and the entire industry of one country is damaged. Therefore, in order to be successful in venture businesses, a time reduction method for diffusion of innovation is required. Suggestions should be given to companies, including venture companies, regarding stagnation and the gap in the spread of innovation.

Suppose there are five adoptive type individuals, an "Innovator," an "Early adopter," an "Early majority," a "Late majority," and a "Laggard." These classification types are seen in the spread theory of an innovation. There are cracks among the five-adopter classifications. It is also assumed that a big gap, which is called Chasm, exists between an Early adopter and an Early majority. In this paper, five adoptive type individuals are based on the observation by Moore [2]. Therefore, differences among five adoptive type behaviors can be realized by the homogeneous behavior and the differentiation behavior with the same kind adoptive type and the diffusion preceded type. We tried to reproduce the Chasm phenomenon in a simulation, by using a Multi-Agent Simulation (MAS). We considered whether Chasm existed, at what frequency Chasm was generated, and what kind of case brings rise to Chasm generated by performing simulation experiments using MAS.

This paper consists of the following components: Section 2 presents the related work. Section 3 gives the experimental setup for the simulation. Section 4 discusses

 ISBN: 978-1-61208-329-2

the results. Section 5 concludes the paper and describes future work.

II. PREVIOUS RESEARCH

This section, firstly, reviews the previous research of the innovation. We referred to the Chasm concept of Moore [2] in order to create MAS. Therefore, this section surveys the research regarding Chasm. Next, we review the previous research that tried the development and experiment using MAS regarding the diffusion of innovations.

A. Inovatinon

Schumpeter [5] defined innovation as "Producing something that already exists by a new method, or producing a new thing." Utterback and Abernathy [6] divided innovations broadly into the product innovation that destroys the existing technical concept, and the process innovation that elaborates on the best at the lowest price.

Rosenberg [7] insisted that a long-term ripple effect does not exist without cumulative and continuous innovation, and the qualitative development does not exist, without an epoch-making innovation. Abernathy et al. [8] and Abernathy and Clark [9] described four innovation types based on both the technical and the market condition. Those types are defined as follows:

- Architectural innovation: The innovation that sets a base to the systematized techniques that destroy the existing systematized techniques, and reclaims a completely new market.
- Revolutionary innovation: The innovation which reclaims the existing market while it is in the systematized techniques which destroy the same existing systematized techniques.
- Niche creation: The innovation that reclaims a completely new market while aiming to strengthen the existing systematized techniques technically.
- Regular innovation: The innovation that strengthens the existing systematized techniques and moreover cultivates the existing market.

Christensen [10] mentioned as follows: Big companies think that the market for innovation is small and not attractive when compared with the conventional large-scale business. Since there is a risk of destroying the conventional primary business, the adoption of innovation is overdue. Therefore, the big company will lag behind the new companies. Then he named the dilemma of the innovation.

B. Chasm

It is said that there is a deep gap which checks the shift to the leading market from the initial market in the diffusion of the innovations that makes new products and new technology permeate a market in the high-tech industry.

Moore [2] advocates the Chasm concept. The strategy of overcoming the Chasm from the concept of the Chasm is called the Chasm theory.

Rogers [11] classified the customer into five adoptive types as Innovator, Early adopter, Early majority, Late majority, and Laggard, in the innovation diffusion model. In this theory, it is supposed that innovation spreads rapidly from the Innovator to the Early adopter (more than 16% of diffusion rate).

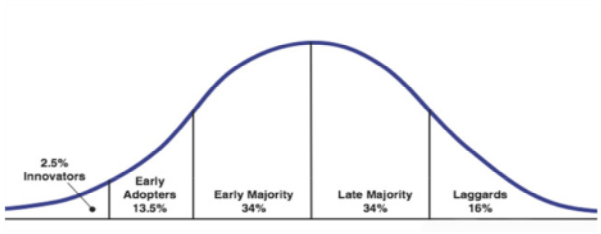


Fig. 1. Technology Life Cycle by Rogers

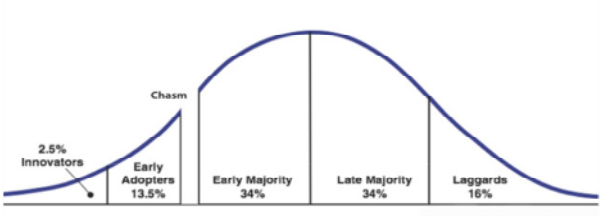


Fig. 2. Chasm by Moore [2]

Then, it is assumed that the key to new product spread is what is advertised to an innovator and an Early adopter.

Fig. 1 indicates the Rogers innovation spread model. In the high-tech product that forces a user's behavioral pattern change, Moore [2] discovered a crack among the five-adopter classifications. He named this Chasm, and it supposes that there is a deep gap between the Early adopter and the Early majority. The gap in Fig. 2 indicates the image of Chasm.

The Early adopter layer adopts new technology positively. The Early majority layer tends to think about stability and relief as important. Therefore, the uneasiness of an Early majority layer is not canceled in the place where the Early adopter layer is only a part of the adopted market. Both demands differ fundamentally, and in order to shift to a leading market from an initial market exceeding Chasm, it is necessary to change the approach of marketing according to the spread stage of an in-house product. Moore [2] observed the following rates within the five-adopter classifications:

- Innovator: (2.5%) People and companies that adopt technology aiming for differentiation from novelty.
- Early adopter: (13.5%) People and companies that adopt technology in the first stage aiming at differentiation not from technology but from an actual profit position.
- Early majority: (34%) People and companies that check a preceding person's success example and adopt by imitation.
- Late majority: (34%) Prudent people and companies that copy large majority uses.
- Laggard: (16%) People and companies that hate new things technically and practically.

An Early adopter tends to adopt new technology as a "means of change". They aim at the action of a

differentiation strategy by staying ahead of their competitor and adopting new technology.

They introduce new technology with the determination to overlook the risk, in order to obtain a competitive advantage by differentiation. They also often make excessive demands on previously trusted vendor.

On the other hand, the Early majority (utilitarian) positions the product as a "means of an operational efficiency improvement." This is the situation where a trial-and-error method with unripe technology is avoided. They also copy the example of the usage of the new technology of the other companies in the same industry. They want to take action with a strategy of homogeneous behavior. However, since an Early majority specifies the product and technology that were introduced as a company standard in many cases, technology vendors can expect a high profit ratio. Therefore, Early majority is an important customer for vendors. In the Chasm theory, there are different demands for the Early adopter and the Early majority, and in order to shift to a leading market exceeding Chasm, the marketing approach needs to be changed according to the diffusion stage of an in-house product.

The differences among these five-adopter classes are what is derived from the strategic activity principle in a management strategy theory called differentiation behavior (behavior by the snob effect), and homogeneous behavior (behavior by the bandwagon effect). Strategic behavior was mentioned by Leibenstein [12], Porter [13], Porter et al. [14], and Asaba [15]. This paper examines the conditions of generating Chasm based on two strategic behaviors such as differentiation behavior and homogeneous behavior as agents activities. Meaning of homogeneous behavior is that of action to mimic the behavior of others. The snob effect definition is as follows: People do not want the same product others bought, and want something different from the product others bought. The bandwagon effect definition is as follows: More people support certain products and services, and the effect of satisfaction and sense of security that the customer obtained by the products and services will increase.

Moreover, they mention that the technologies that could not exceed Chasm are Video conference systems, Artificial Intelligence, Pen computing system and so on. Regarding music devices, Compact Discs (CDs) and Digital Versatile Discs (DVDs) have exceeded Chasm, but Laser disc and Mini Disc (MDs) have not. Chasm is a big gap that exists before the diffusion of a high-tech product through the mainstream market. In order to exceed Chasm, the basic strategy that Moore [2] asserts is responding to the utilitarianism of the Early majority who is a customer segment of the beginning of the mainstream market. However, he suggests that the innovation vendor must not provide all early majorities with a product. The concrete method exceeding Chasm is concentrating the best in one area. It is important to complete the perfect product quickly toward a certain specific customer segment.

The greatest reason against the overall market is that the demand level of Early majority who is utilitarian wants 100% of the solution.

He insists on that the Early adopter who is constitution of the initial market expects and dreams product usefulness in the future.

This approach is explained by the lane of a bowling alley metaphor. Each customer segment is also equivalent to knocking over one pin. Knocking down one pin causes all others to also fall.

In other words, success with one customer segment is used as a springboard, and success with a new customer segments is then gained.

Eventually a "strike" is made and it can create rapid growth in whole market. The analogy of the bowling alley lane serves as reference when developing MAS.

Moreover, the approaches for exceeding Chasm are the following three steps.

1. Though it is small, a positive foothold is made somewhere in one mainstream market as soon as possible.
2. When innovation diffuses in the mainstream market, the strategy that was conscious in the overall market is promoted, and it should be remade to spread widely as a standard product.
3. Return to the approach of a client centered again and append added value to a product through mass customization. Mass customization is building the product to individual specification in large quantities.

Moreover, Markides and Geroski [16] stated: If the second runner is not called the “Fast Second,” then it cannot generate "radical innovation." This is the reason why there is this big gap called "Chasm" between soliciting some Innovators, and public acceptance in a market. This is also presupposed, because the second runner has the advantage to exceed Chasm.

The second runner who has made the market expand raises business that disturbs the existence of a customer's customs and the existing company, such as in the mobile phone and an online bookstore. It can be said that strategic behavior called homogeneous behavior and differentiation behavior show also that the second runner has taken advantage of innovation.

C. MAS of Innovation diffusion

Washida [17], Washida et al. [18], and Matsuka et al. [19] developed MAS of innovation emergence in the innovation diffusion processes. They are referring to the diffusion model of the innovations from Rogers (1986) [8], the Chasm from Moore [2], the small-world network structure from Watts [20] and the scale-free network structure from Barabsi [21]. Small-world network structure is a small world character network structure that appears in both a network natural and artificial (a nervous system and a transmission network). Moreover, small-world network structure follows "A power law Distribution." [17] [18] [19] "A "power law Distribution” is a network structure without a specific type value. They stated that the innovation is not based on the development of a supplier's technology, but the

discovery of the utility value by the consumer from the experimental results of multi-agent simulation.

They develop the multi-agent model of innovation that was generated by consumer user's conversion value phenomenon. The developed multi-agent model assumes that the case of the mobile phone which carried out conversion to e-mail and a ringtone to identify a specific caller and the case of the development of the station wagon type car from a regular sedan . However these two products were developed for a niche of the market, these became mainstream goods in high demand.

Kitanaka [22] set up four kinds of agents, namely, maker, wholesale, retail store, and consumer. Three diffusion course networks were stretched for each agent with MAS. Three spread courses were a distribution channel network, an advertising and promotional network, and a word-of-mouth network. The distribution channel network was made into a tree structure. The advertising and promotional route was made into an emanated type network structure. The word-of-mouth network was also made into the scale-free network. A distribution channel network and an advertising and promotional network spread innovation through a consumer agent according to dropping resources. In the word-of-mouth network, it was set up so that the consumer agent was recognized as a hub because of the number of links it could use to dispense innovation. By experimenting, the researchers were able to reproduce the difference that appears between the diffusion of innovations and the active degree of a word-of-mouth network (the number of hub consumer agents) by experiment.

Morioka [23] developed MAS of brand value. He set up that an agent gave with the bandwagon effect (effect which makes it take homogeneous behavior), and the snob effects (effect which makes it take differentiation action) by communicating market share information. As the result, the change of the market share is reproduced with MAS. A market share became higher, so that the threshold value of a market share when giving the bandwagon effect is higher as a result. However, it was found that the market share is balanced with a fixed value.

D. Suggestions from previous works

We considered that Moore [2] proposed a five-adopter classification for the spread of an innovation. This spread depends on how to take homogeneous behavior and differentiation behavior into behavior called strategic behavior. Therefore, after giving a definition to an agent as to how to use strategic behavior differently with all five-adopter classifications, MAS should be developed with regard to diffusion of innovations. The purpose of this paper was to obtain implications about the conditions for generating Chasm from the experiment of MAS.

III. CONFIGURATIONS OF CHASM IN MAS

In this paper, we used Artisoc3.0 (http://mas.kke.co.jp/index.php) as Multi-Agent simulator. Artisoc3.0 is a software simulator of the MAS that KOZOKEIKAKU Engineering Institute (http://www.kke.co.jp/en/) provides. We focused on the consumer market as the simulation market for a group of companies in a certain industry targeted for the diffusion of innovations. The case where an innovation spread through industry is assumed in this paper. The setup was as follows:

- Space Industry (as default setups) was added to the Universe.
- Agent High Tech1, which expresses as an innovation of the Space Industry, was added.
- As an agent showing a company as an Innovator, Early adopter, Early majority, Late majority, and Laggard were added.
- The number of agents for each company could be set from 0 to 200 in the control panel.
- The real type variable, which expresses each agent's diffusion rate in the Universe, was added. INDiffusion was added to Innovator and EADiffusion to Early adopter, EMDiffusion was added to Early majority, LMDiffusion was added to Late majority, and LADiffusion was added to Laggard.
- The output setup was the real type variable Diffusion showing the entire diffusion rate was added.
- The real type variable speed which specifies the speed that corresponds to each company agent was added.
- We added a real type variable SHIYA to specify the size of the field of view to observe the movement of intra-industry competitors by each company agent. SHIYA means a field of view company agent to look for other company agents.
- We added a real type variable NAKAMA to specify the number of others to observe as a condition of taking the homogeneous behavior by the bandwagon effect by each company agent. NAKAMA means the number of peer company to be homogenized by company agents.
- We added a real type variable KYOGO to specify the number of conflicts within the field of view as a condition by taking the behavior by differentiation, the snob effect on each company agent. KYOGO means the number of competitors that is the subject of differentiation by company agents.
- An output map of the Space Industry was added as an element for each company agent on the map.

The diffusion from HighTech1 agent to each company agent was set up as follows.

- We have defined the state of Innovation diffusion as the analogy that the company agent is facing the direction of 0 degree the same as the high-tech 1 agent.
- Agent High Tech1 acted in the direction of 0° , and it added the function that made Innovator to 0 ° direction as a function to transmit an innovation to the Innovator in the field of view within less than 15.
- Agent Innovator has the capability to make the Early adopter to 0° direction in the field of view within 3, as a function of diffusing the innovation.

 ISBN: 978-1-61208-329-2

- Agent Early adopter has the capability to make the Early majority to 0° direction in the field of view within 1, as a function of diffusing the innovation.
- Agent Early majority has the capability to make the Late majority to 0° direction in the field of view within 1, as a function of diffusing the innovation.
- Agent Late majority has the capability to make Laggard to 0° direction in the field of view within 1, as a function of diffusing the innovation.

Fig. 3 illustrates the innovation diffusion.

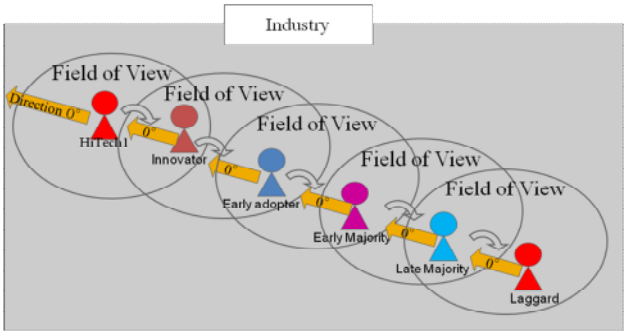


Fig. 3. Innovation Diffusion Model

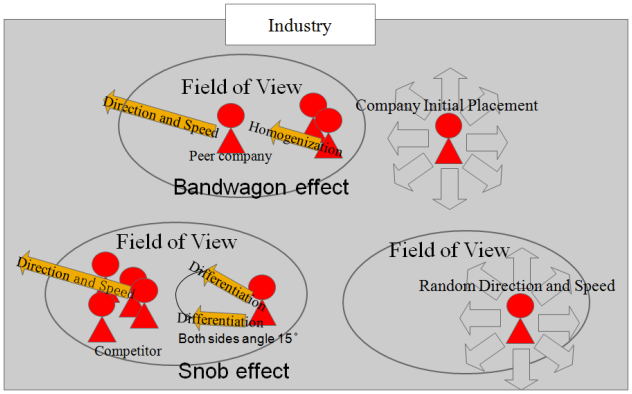


Fig. 4. Homogeneous Behavior with MAS

Each company agent shall take homogeneous behavior or differentiation depending on the following configuration.

- As analogy that takes the homogeneous behavior by the bandwagon effect, with every agent set up as follows.
- When number of company agents of the same kind within a view size was more than the NAKAMA number, it was made to progress at the same speed and the same direction as a company agent of the same kind.
- When there were many agents of the same kind who turned to and followed the same direction behavior according to the snob effect, it was set up as follows for every agent. When the number of agents of the same kind within a view size was more than the KYOGO number, it was made to progress in a different direction in a range of 15 on both sides.

Fig. 4 indicates homogeneous behavior with MAS.

The flow of the company agents contains the following configurations.

1. At first, random position, direction, and speed was used.
2. If more than the fixed number (the number is the Variable NAKAMA), of the other agents of the same kind are within the surroundings (width of a view), the company agents take the same direction and speed as the other agents of the same kind, because of the bandwagon effect. This action was defined as homogeneous behavior.
3. Unite the direction and speed of your company with the direction and speed of one company of the homogeneous partners (the number is the Variable NAKAMA).
4. If more than the number (Variable KYOGO) whose agents of the same kind are in the surroundings (width variable SHIYA of the view) , the company agents take the different direction and speed as the other agents of the same kind because of the snob effect.
5. Change the direction in the direction of another company of the differentiation partners (the number is the Variable KYOGO) to the direction of 15 on both sides. However, the present Speed is not changed.
6. There is neither a homogeneous partner nor a differentiation partner, change of direction or speed suitably.
7. If there is an affecting target agent in the view, it will turn in the direction of 0.

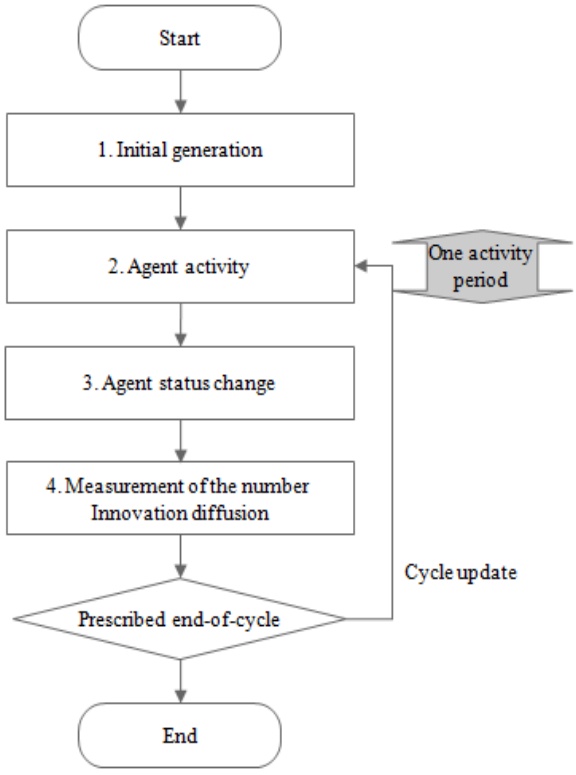


Fig. 5. Simulation Flow

 ISBN: 978-1-61208-329-2

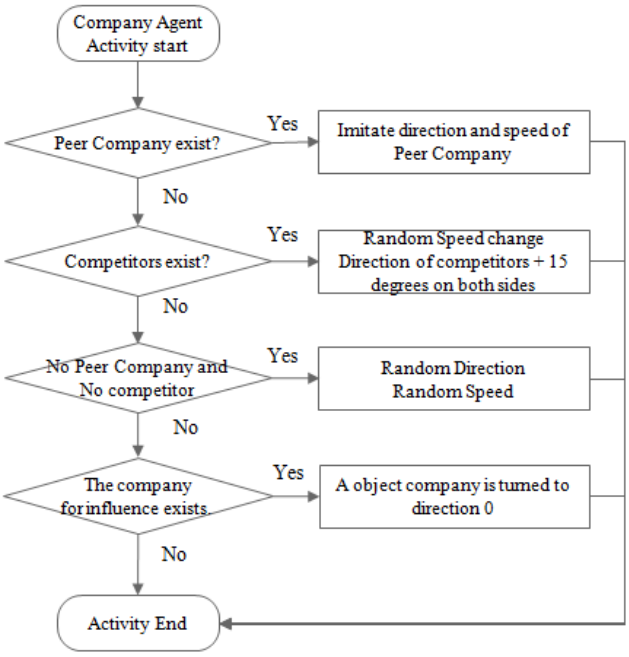


Fig. 6. Company Agent Activity Flow

The simulation flow is shown in Fig. 5. The flow of the company agent activity is shown in Fig. 6.

IV. EXPERIMENTS

We set 200 companies in the same industry with two variables of KYOGO as the competitor and of NAKAMA as another company and generated agents with the ratio that Moore [2] proposed. Each agent takes homogeneous behavior or without the judgment of agent's sight variables used in agent's decision. The experimental set-ups are shown in Table I.

Based on the above configurations, 10 times trials within each 10,000 steps were operated. Figures 7 to 16 show the experimental results.

TABLE I. EXPERIMENTAL SET-UPS

	Innovator	Early Adopter	Early Majority	Late Majority	Laggard	Sum Total
SHIYA	2	2	2	2	2	
NAKAMA	1	1	2	3	3	
KYOGO	10	10	10	10	10	
Existing ratio	0.025	0.135	0.340	0.340	0.160	1.000
Number of existence	5	27	68	68	32	200

The top line in gray color indicates the sum of diffusion of innovation in the industry. Figures 7 to 16 indicate the result of ten trials. Only one trial was not observed in the crack in Fig. 11.

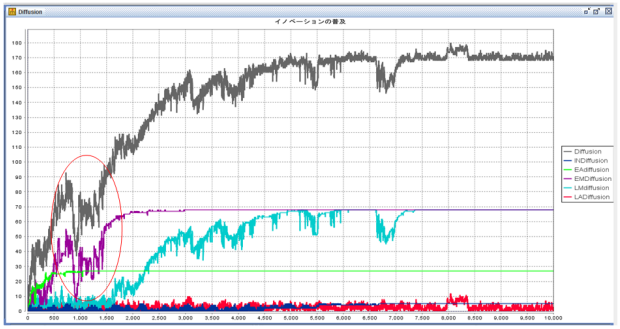


Fig. 7 Simulation Result (1st trial)

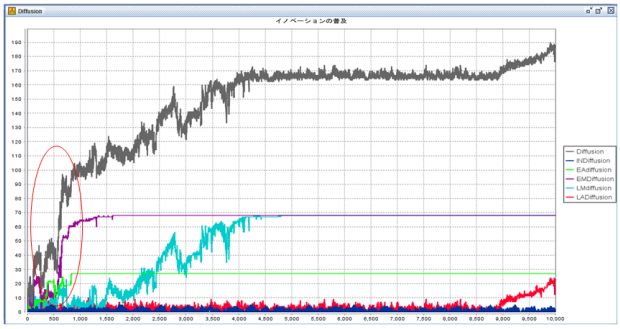


Fig. 8 Simulation Result (2nd trial)

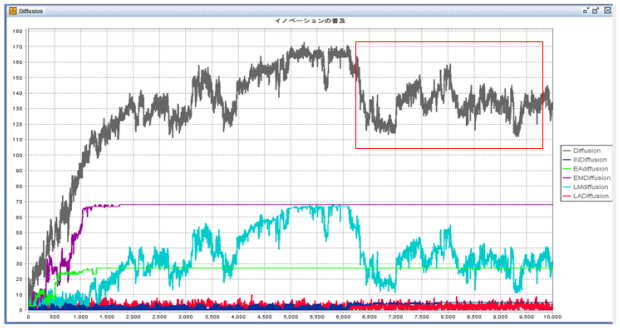


Fig. 9 Simulation Result (3rd trial)

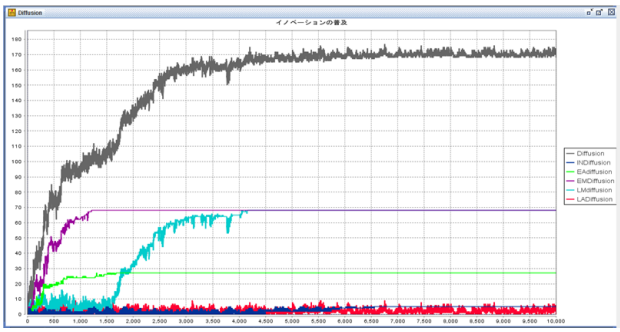


Fig. 10 Simulation Result (4th trial)

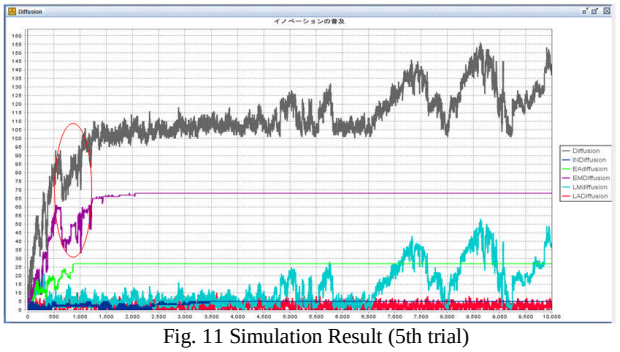


Fig. 11 Simulation Result (5th trial)

 ISBN: 978-1-61208-329-2

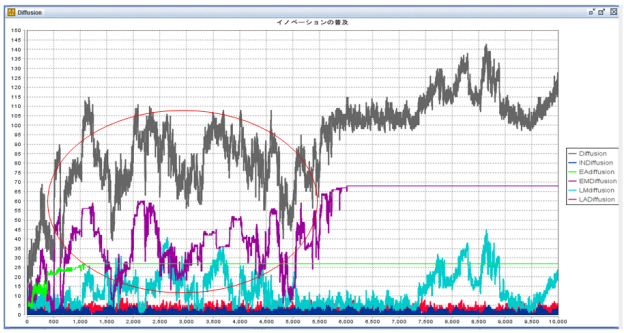


Fig. 12 Simulation Result (6th trial)

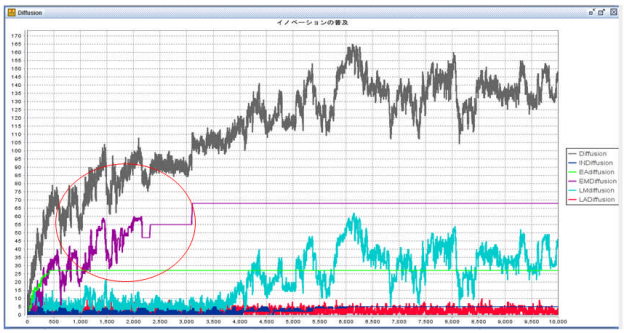


Fig. 13 Simulation Result (7th trial)

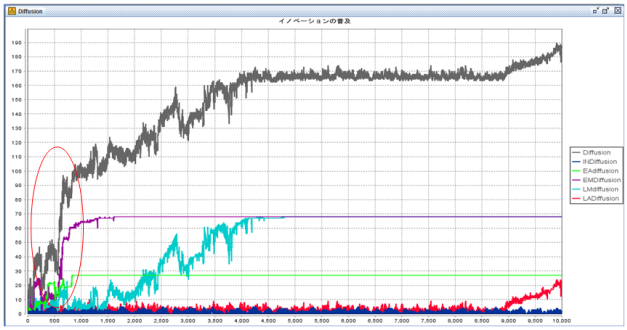


Fig. 14 Simulation Result (8th trial)

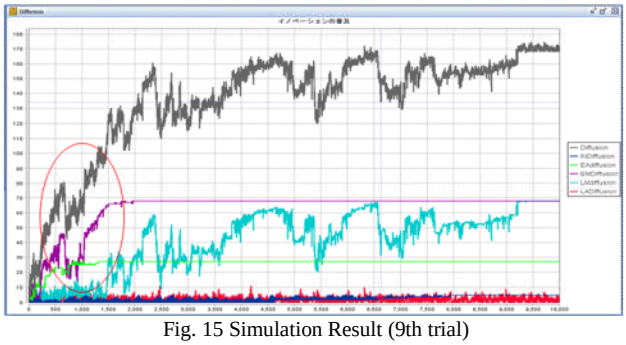


Fig. 15 Simulation Result (9th trial)

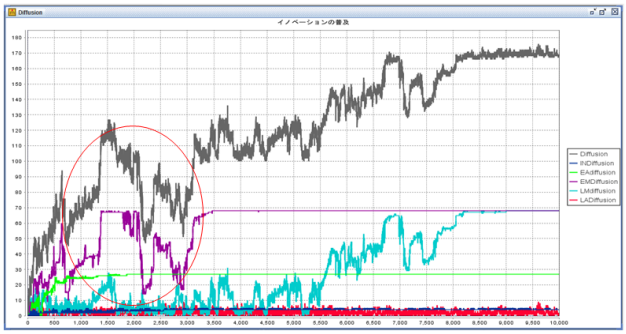


Fig. 16 Simulation Result (10th trial)

As seen in the simulation results in Figures 7 to 9 and Figures 11 to 16 (except Fig. 10), we succeeded in crack generation.

V. CONCLUDING REMARKS

This paper described the Innovation diffusion with MAS. In the Innovation diffusion theory, the market is classified in five types, namely: Innovator, Early adopter, Early majority, Late majority, and Laggard, so far. Among each classification, there are cracks in the innovation adaptation. Especially in High-tech industries, a big gap called Chasm is made between Early adopter and Early majority that were proposed by Moore (1991) [2]. Since this paper proposes the innovation diffusion model with MAS as a preliminary trial, we translated the heuristic knowledge into a computer simulation model.

At first, we recounted the related work. Based on the innovation diffusion theory, we defined parameters and made a simulator for Chasm observation. From the results of the simulation, we succeeded in crack generation. By statistical analysis and by case, we identified the conditions for generating the Chasm within the spread of innovation. In contrast, by utilizing MAS, the possibility of identifying the condition is confirmed. As an example of the Chasm in recent years, the mobile phone standard is unique to Japan, but did not spread to the global market while gaining the function of an Internet connection terminal and the like. Currently, the mobile phones of Japan's own standard are called Galapagos mobile phones. We were able to recognize that Galapagos mobile phones have fallen into a Chasm of innovation diffusion. We believe that we reproduced such phenomena as the Galapagos mobile's Chasm by MAS.

Our future work is as follows: (a) Capture the conditions for the crack generation, (b) Parameter tunings of corporate indicators such as sales, costs, assets, and capital and so on, and (c) Compose simulation for a new technology as Innovation goes into the industry.

ACKNOWLEDGMENT

We wish to express our gratitude for the cooperation and fruitful discussion from the companies we analyzed.

REFERENCES

[1] H. Matsuoka, "Japanese consumer electronics manufacturers in Asia and white goods market Efforts," General Foundation Japan Asia Pacific Ocean Institute Macroeconomic Analysis, project, March, 2012.
http://www.apir.or.jp/ja/research/files/2013/03/423_02.pdf (accessed on January 25th, 2014)

[2] G. Moore, "Crossing the Chasm: Marketing and Selling Disruptive Products to Mainstream Customers," Collins Business Essentials, Harper Business, 1991.

[3] Ministry of Economy Trade and Industry, "Survey about mid- and long-term research and development of company in our country which contributes to innovation creation," Industrial Technology Investigation Report in Heisei 23 fiscal year, 2011.
http://www.meti.go.jp/policy/economy/gijutsu_kakushin/innovation_policy/kaihatu-hyoka-22.htm (accessed on November 5th, 2013)

[4] H. Hasegawa," Introduction to Venture Management business creation," Tokyo: Nikkei Publishing Inc., 2010.

 ISBN: 978-1-61208-329-2

[5] J. A. Schumpeter, “The Teory of Economic Development:An Inquiry into Projits, Capital, Credit, Interest, and the Business Cycle,” Cambridge: Harvard University Press,1934.

[6] Utterback. J. M. and W. J. Abernathy, “A Dynamic Model of Process and Product Innovation."Omega.Vol.3, No.6, ,1975, pp. 639-656.

[7] N. Rosenberg, “Inside Black Box: Technology and Economics,” New York: Cambridge University Press,1983.

[8] W. J. Abernathy, K. B. Clark, and A. M. Kantrow, “Industrial Renaissance: Producing a Competitive Future for America,” Basic Books.1983.

[9] W. J. Abernathy and K. B. Clark, "Innovation: Mapping the Winds of Creative Destruction." Research Policy, vol. 14, no. 1, 1985, pp. 3-22.

[10] C. M. Christensen, “The Inovation's Dilemma when New Technologies Cause Great Firms to Fail”, Boston: Harvard Business School Press. 1997.

[11] E. M. Rogers, “Diffusion of Innovations,” Glencoe: Free Press., 1962.

[12] H. Leibenstein, “Bandwagon, Snob, and Veblen Effects in the Theory of Consumers' Demand,” The Quarterly Journal of Economics vol. 64, no. 2, 1950, pp. 183-207.

[13] M, E. Porter, ”What is Strategy” Harvard Business Review, November-December, 1996, pp. 61-78.

[14] M. E. Porter, H. Takeuchi, and M. Sakakibara, ”Can Japan Compete?,” Basingstoke Macmillan, 2000.

[15] S. Asaba, “Empirical Analysis of homogeneous behavior - Competition of Japanese Companies,” Tokyo: Toyo Keizai Inc., 2002.

[16] C. C. Markides and P. A. Geroski, ”Fast Second: How Smart Companies Bypass Radical Innovation to Enter and Dominate New Markets,” John Wiley & Sons, 2004.

[17] Y. Washida, ” The heterogeneity of an information propagation network and structural analysis of a value conversion phenomenon in a diffusion process : Possibility of the innovation which the demand side tows,” The University of Tokyo doctoral dissertation, 2008.

[18] Y. Washida and K. Ueda, ”Research on the demand side network propagation structure that generates innovation ideas,” Journal of Information Processing of Japan. vol. 49, pp. 1515-1526, 2008.

[19] T. Matsuka, Toshihiko, H. Honda, Y. Washida, and K. Ueda, ”Exploratory multi-agent modeling studies examining influence of social structures on development of innovative ideas,” The 27th Annual Conference of the Japanese Society for Artificial Intelligence, 2013, 4F1-4.

[20] D. J. Watts and S. H. Strogatz, “Collective Dynamics of Small-World Network,” Nature, vol. 393, pp. 440-442, 1998.

[21] A. L.Barabsi and R. Albert, ”Emergence of Scaling in Random,” Network Science, Vol. 286, p509-512 ,1999.

[22] H. Kitanaka, “Diffusion of Strategic Innovation and Social Network Structure: A Discussion Using Agent Based Approach,” Takushoku management accounting research,81, 2007, pp.27-63.

[23] K. Morioka, ”An understanding by brand value generation and a change, the explanation based on the dynamics social system theory of a market share, and a multi-agent simulation,” The Mita commercial science research, vol. 52, no. 1, 2009, pp. 87-110.

 ISBN: 978-1-61208-329-2

MLPM: A Multi-Layered Process Model Toward Complete Descriptions of People's Behaviors

Zhang Zuo, Hung-Hsuan Huang, Kyoji Kawagoe
Graduate School of Information Science and Engineering
Ritsumeikan University
Kusatsu, Shiga JAPAN
e-mail: {gr0186rk@ed, huang@fc, kawagoe@is}.ritsumei.ac.jp

***Abstract*—Despite the rapid progress in the development of sensor technologies, as well as of information management, no technology exists for recording all the activities of people in the many varieties of human societies. In this paper, we propose a novel process meta model for describing people's activities that uses a Multi-Layered Process description Model, MLPM. The meta model allows models for various kinds of human social activities, such as sport plays, medical treatment, and agent communications, to be easily described. The significance of this model is that it can be used not only for searching a subpart of people's activities given a process query, but also for fostering a young novice by presenting to him/her behavior patterns, differentiating between those of the expert and the novice. The meta model can also be used for detecting outliers in process databases. In this paper, we describe the components and structures of the MLPM. It is mentioned that the MLPM is a model suitable toward complete descriptions of people's behaviors by comparing it with other methods.**

***Keywords*—process; behavior; action, meta model; searching.**

I. INTRODUCTION

Recently, the use of various kinds of sensor devices, such as motion and location sensors, has become widespread. As a result of this rapid popularization, huge amounts of their monitored data have been obtained and analyzed for application-oriented purposes. For example, human motions can be easily detected by using motion sensor devices such as Microsoft Kinect [1] and Leap Motion [2]. Human trajectories can be traced by using position sensing devices, such as GPS and OptiTrack [3]. These advanced sensor devices are currently used primary for analyzing object movements, as well as for visualizing them.

Despite the fact that many data related to human activities, recorded by the sensors, the data exist, they cannot be managed for the purpose of processing in a unified way. In particular, motion data are stored in a proprietary format and no common formats have yet been proposed. The raw location data for human activities basically comprises a combination of human motions and positions. There are some common formats for representing human motions and positions, such as H-Anim [4] and BVH (BioVision Hierarchical data) [5]. However, these formats are used only for storage and exchange of the data, not for representations of all human activities.

In the area of business process management, some process description models exist, such as BPMN (Business Process Modeling Notation) [6], XPDL (XML Process Definition Language) [7], and BPDM (Business Process Definition Metamodel) [8]. Although these models are used for business management representation and business system development, their main purpose is not to represent people's behaviors, and thus, representations of human motions and positions are outside their scope.

In this paper, we propose a new and novel process meta model for describing a model for people's behaviors that uses a Multi-Layered Process description Model, MLPM. The model allows various kinds of human social activities, such as sport game plays, medical treatment, and agent communications, to be easily described.

For example, suppose that a skilled doctor is fostering junior doctors in order to impart to them better skills for the task of giving intravenous or subcutaneous injections. It is difficult for them to understand and to perform the task without any practical experience of it. Even if they have some experience of giving the injections, appropriate real-time comments from experienced doctors are very necessary and helpful. However, when they use a subcutaneous-injection simulator, such useful comments are not available to them. There is thus no opportunity to improve their skills in such situations. When our model can be applied to provide this medical treatment education, a system based on our process model will be able to support young doctors by presenting to them the differences between the activity as executed by a skilled doctor and by a junior doctor, with specific details. This is because all the activities in the injection process can completely be represented by our proposed model, and the difference can be detected by real-time checking of the distance between two processes.

Our proposed model is composed of seven fundamental components: Process, Task, Entity, Activity, Action, Motion, and Expression. These components are linked to many types of associations. The main characteristics of MLPM are:

- All processes related to people's behaviors can be modeled by using our MLPM. A process can be represented by a hierarchical structure, as well as by linked data among components.
- MLPM can give researchers in many fields a way of describing behaviors in a common representation format.
- Many functions and tools can be incorporated in the basic structure of MLPM. Some examples of such functions are process similarity searching, process outliers

 ISBN: 978-1-61208-329-2

detection, and process classification.

The contribution of this paper is that the proposed MLPM is the first representation meta model for describing overall people's behaviors comprehensively. Throughout this paper, it is assumed that the representation model is a model for describing people's behaviors that occurred in the past. The prediction of a future behavior from past behaviors is beyond the scope of this paper. The term "behavior" is used to define all the aspects of people's activities, including tasks, motions, positions, and interactions, because the term is defined in [10] as "behaviors refer to those activities that represent actions, operations or events as well as activity sequences conducted by human beings under a certain context and environment."

The remainder of this paper is structured as follows. Some previous works related to our paper are described in Section II. Next, in Section III, a detailed description of our model, MLPM, including examples and formal specifications, is presented. Some expected applications and discussions are then given in Section IV, followed by some concluding remarks in Section V.

II. Related Work

We describe the related work from three viewpoints: business process models, process mining, and multi-agent models.

Many studies on the business process model have been reported [7], [8], [9], [11], [12], [13], [14]. Three innovative business process models were proposed and standardized: Business Process Definition MetaModel (BPDM) [8], Business Process Modeling Notation (BPMN) [6], and XML Process Definition Language (XPDL) [7]. These proposed models were intended to represent all the processes performed in an enterprise. For example, the BPDM model is composed primary of Common behavior model, Activity model, and Interaction protocol model. The common behavior model is composed of two detailed models: Behavior model and Interactive behavior model. These models allow many types of processes used in a specific enterprise application field to be defined. In addition to these models, many similar models have been proposed. Typical examples of such models are explained in survey papers [9], [12]. Two examples are Object-Oriented Business Process Model [11] and ADONIS BPMS model [15]. Although all the business process models can be used for process representations in an enterprise, they can never be employed for representing people's behaviors including their motions and moving trajectories. The mining of business processes also has been studied recently [13], [14].

Process mining is a new area in the data mining research field. Many mining methods for acquiring useful and effective rules in terms of processes have been proposed [16], [17], [18], [19], [20]. The definition of the objective of the process mining is to discover, monitor, and improve real processes by extracting knowledge from the event logs readily available in today's (information) systems [16]. However, this definition is not appropriate in the case of descriptions of all aspects of people's behaviors. It is also difficult to perform mining from a large collections of logs of people's motions, trajectories, and actions, because of issues related to similarity definition, dimensionality reduction, and data cleansing.

Lastly, the multi-agent model is a model for simulating human behaviors by using virtual agents or humanized robots. Many studies on the multi-agent model have been proposed [21], [22], [23], [24]. The model usually includes the agents' motions as well as interactions with other agents or with humans. Examples of multi-agent models are AALAADIN [21], Swarms [23], and IPC/Q [24]. The common concepts applied in the models are based on the object-orientation concept and the procedural process description. This category of the behavior model has disadvantages in that in general it is difficult to archive all the process data, although a process can be described using the procedures provided by the model. Therefore, it cannot be used for representing people's behaviors comprehensively for the purpose of archiving.

In addition to the above related work, it should be noted that a behavior model has recently been proposed [10], [25]. The proposed behavior model is very general and can be implemented for many applications that require behavior mining. However, it cannot be used for archiving an overall process, to the best of our knowledge.

Recently, Neumuth et al. proposed a process model for surgery [26]. Although the basic idea seems similar to that of our proposed model, their model is basically a hierarchical structure built from the descriptive format viewpoint, using natural language sentences, ontological description, formal mathematical description, and actual description. In contrast, our model is a layered structure that uses the abstract-to-concrete viewpoint.

III. MLPM (Multi-Layered Process description Model)

A. Requirements

Before describing the MLPM proposed in this paper, we describe some requirements for process representation that we need to specify. The process model should meet the following requirements to specify the overall process of people's behaviors.

- People's behaviors should be described in various aspects, because they are recognized from an abstract view as well as from a concrete view.
- Usually, a business is run or a task is performed by a team of people who uses various tools. Therefore, their interactions should be represented by the process model.
- In terms of the actual people's behavior, the motions of each person can be captured using various kinds of equipment, such as cameras and GPS sensors. Therefore, the model should be able to use the data generated by these devices.

B. Basic concepts

We describe the basic concepts of our MLPM for process representation. We introduce the multi-layered structure to represent the overall process, which contains various kinds of descriptive aspects. Our proposed model is composed of three layers to meet the requirements mentioned above: the process/task layer, activity layer, and motion layer.

 ISBN: 978-1-61208-329-2

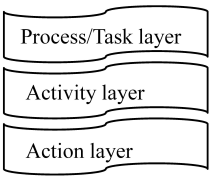


Figure 1. The layered structure.

- In the process/task layer, a sequence of tasks conducted in a process and their properties can be described. The business process model [8] or the work flow process model [7] can be introduced in this layer to describe the sequences and task properties.
- In the activity layer, activities representing each task can be described. In this layer, a task is decomposed into activities and represented by relationships with entities, such as individuals and instruments. The properties of the activity and the entity are also described here.
- In the action layer, an activity is decomposed into a sequence of actions. An action is further decomposed into a set of motions, which constitute the basic component corresponding to a human being's actual movement. Actual motions can also be described using various forms of expression, such as video and trajectories. Concurrent motions can be represented by multiple actions to which several such expressions are attached. Clearly, actions, motions, and expressions all have properties that describe them.

Figure 1 shows an outline of the layered structure of our MLPM.

C. Fundamental components

In our MLPM, seven fundamental components are used for describing an overall process. A brief explanation of these components follows.

1) The process/task layer
   a) Process: A process is an abstract unit of functions in the specific application field. People's behavior are first defined as a set of processes. Examples of a process are ordering-by-customer, injecting-drug, and ballroom-tango-dancing. The process definition in BPDM [8] or XPDL [7] can be used to describe the process.
   b) Task: A task is also an abstract unit of the sub-functions that compose a process. Each process is represented by a sequence of tasks as in BPDM or XPDL, although the name of the task is either the sub-process or the activity, representatively. The task can be defined using such a standard specification.
2) The activity layer
   a) Entity: An entity is an abstract class of objects, which can be modeled using MLPM. Entities are used to perform a process/task. They are divided into either user or instrument classes. Other classes can be introduced according to the specific application. The user classes are related to human groups, such as doctors and patients. The instrument classes are related to the machines, goods, or tools that support users' activities.
   b) Activity: An activity is a relationship between an entity and a task that is used to represent a task execution. It is possible to represent multiple activities for an entity executing one specific task. Moreover, there are multiple entities for one specific activity. One example is that a nurse inserts a syringe into a patient's vein in a blood collection task. In this example, Nurse, Injector, and Patient are described as entities, and the injection activity is an activity, which is followed by the activity of removing the inserted syringe.
3) The action layer
   a) Action: An action is an abstract movement related to one entity. One entity is related to a sequence of actions used to perform an activity associated with other entities.
   b) Motion: A motion is an actual entity's movement used to represent a specific action. The motion data are aggregated and integrated from various kinds of motion expressions.
   c) Expression: An expression is a view of one motion. The motion data can be extracted from various devices. Examples of these expressions include pictures, videos, voices, trajectories, and textual annotations.

In addition to these fundamental components, the following types of associations are introduced for each layer in our MLPM.

1) Temporal associations: A sequence is a kind of temporal association. In addition, there are other types of temporal associations between components such as those between activities and between motions. Other temporal associations can be introduced from concepts of temporal relations [27]. Examples include "done at the time" "after" "before" and "during."
2) Spatial associations: A spatial association is an association between components in terms of their locations. When two entities are located in certain region together, there is a type of spatial associations between them called "located together." Other types of spatial association can be defined and to be used to describe the overall process.
3) Link associations: There are other types of association in addition to spatial or temporal associations between components. These types of association are called link associations in our MLPM. For example, a nurse attaches a patient name label to a blood collection vessel after taking blood. There is a type of link association called "attaching" between the "Nurse," "Blood collection vessel," and "Patient name label" entities.

An example of MLPM representation for medical treatment process modeling is shown in Figure 2. In this Figure 2,

 ISBN: 978-1-61208-329-2

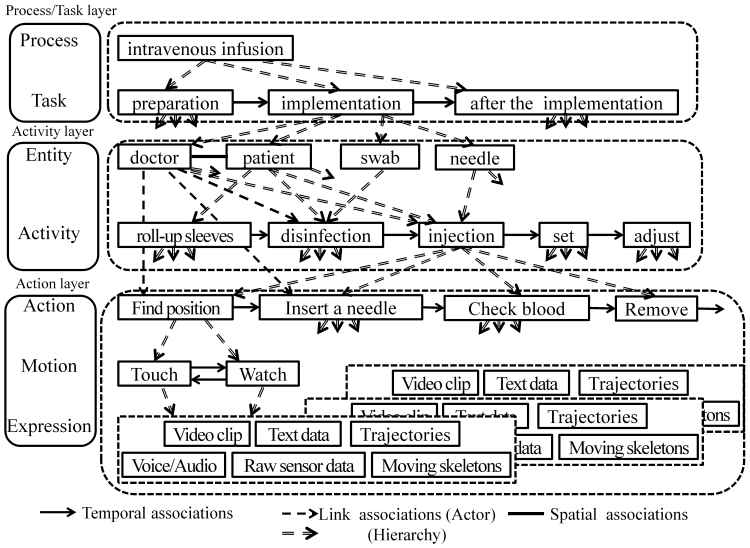


Figure 2. MLPM structure example.

not all the nodes are illustrated because of space limitations. Therefore, some associations with no destination nodes appear.

D. MPLM specifications

Basically and intuitively, an actual model, described using MLPM, is represented as a directed graph, which is composed of nodes and links. Both a node and a link have a type, which is the special attribute for identifying its MLPM component. The formal specifications of the fundamental components in MLPM are described here.

First, the fundamental component FC is defined, which is the set of fundamental component classes FC_i, as $FC = \{FC_i\}$. When instances of the fundamental components FC are denoted by $\{FI\}$, the set of fundamental component instances are described as $FI = \{FI_j\}$. As the fundamental components FC are composed of seven actual fundamental components- Process $P = \{P_j\}$, Task $T = \{T_j\}$, Entity $E = \{E_j\}$, Activity $AV = \{AV_j\}$, Action $AT = \{AT_j\}$, Motion $M = \{M_j\}$ and Expression $E = \{E_j\}$, a fundamental component class FC_i is a class of one of these component classes. In addition, a fundamental component instance set FI_j is the set of instances of one of these classes: P, T, E, AV, AT, M, and E. Moreover, the entity class set E is decomposed into of Human Entity HE and Artificial Entity AE. That is, $E = HE \cup AE$ where $HE \cap AE \neq \emptyset$.

In order to formalize the relationship between a class and its instance, we define it as follows. For each class P_j in P, the instance set is

$$PI_j = \{PI_{j,k} | PI_{j,k} \in FI, PI_{j,k}\ is\_an\_instance\_of\ P_j\}$$

In the same way as PI_j, the instance set in the other components can be defined. For each class FC_i, in FC, the class has multiple attributes, represented by $\{AC_k\}$, where AC_k denotes the domain of an attribute of FC_i. By using these $\{AC_k\}$, we can define the set of the instances of FC_i as $FC_i \subset AC_1 \times \ldots \times AC_{p_i}$.

Second, because the structure of MLPM is basically hierarchical, it is necessary to introduce the definition of the hierarchical relationships between multiple fundamental component classes. The hierarchical relationship HR is defined as the set of the hierarchical relationship classes: $HR = \{HC_j\}$. A hierarchical relationship class HC_j has the relationship instances HI_j, where $HI_j \subset C_1 \times \ldots \times C_{ph_j} \times AH_1, \times \ldots \times AH_{qh_j}$. In this definition, C_k is a set of instances belonging to one of the classses of FC, and AH_k is the domain of an attribute of HC_j. An attribute is a property of the hierarchical relationship and its value changes depending on the HC_j instance. For example, if C_1 is a set of instances belonging to one process class P_k, then $C_1 = PI_k$, where PI_k is the set of P_k instances.

Finally, there are three types of associations: Temporal, Spatial, and Link. Therefore, we introduce the following definition of these associations. A link association LA_k is described as $LA_j \subset Label_{link} \times C_1 \times \ldots \times C_{pl_j} \times AL_1, \times \ldots \times AL_{ql_j}$, where $Label_{link}$ is the set of label names for identifying the type of link associations and AL_k is the domain of an attribute of LA_j. Similarly, a temporal association TA_k is described $TA_j \subset Label_t \times C_1 \times \ldots \times C_{pt_j} \times AT_1, \times \ldots \times AT_{qt_j}$, where $Label$ is the set of label names for identifying the type of link associations and AT_k is the domain of an attribute of TA_j. The formalization of the spatial associations is defined in a similar way.

The fundamental components described above are only a part of the components in MLPM. There are other components, necessary for describing the overall processes using MLPM: Event, Role, and Environment. These additional components are all related to other types of relationship. The event is used to represent a thing that occurs at a certain time hav-

 ISBN: 978-1-61208-329-2

TABLE I. MODEL COMPARISON RESULTS.

	Main Components	Target Applications	Similarity. Definition
MLPM	- Process and Action Models - Motion model incl. movements & expressions - Temporal/Spatial assoc.	- Medical process - Sports	- (TBD)
BPDM [8]	- Business Process Model - Business Semantics - Rules and Policies	- Business	- None
Behavior model [10]	- Abstract Behavior Model - Actor/Operation/Coupling - Temporal/Inferential/Party - Behavior Aggregator - Risk and Impact	- Business - Data analysis	- Behavior Feature Matrix-based similarity
Process model [26]	- Natural Lang. level - Conceptual levels - Formal level - Implementation level	- Surgical . process	- Combination of five similarity definitions

ing relationships with other components, the role is defined as the set of relationships among component instances, and the environment can be represented as the space containing many component instances. These components contain their attributes, as do the fundamental components.

Evaluations of the specifications of MLPM are necessary and are currently in progress in a medical treatment application field. We are also developing a detailed design of all the MLPM specifications, for which we are considering the related standards.

IV. DISCUSSIONS

A. Comparison with other models

Although it is difficult to compare our MLPM with other related models, the preliminary comparison results are shown in Table I. In Table I, three features for each model are described: the main components, target applications, and similarity definition. The last feature is important for extracting processes, activities, or motions that are similar to given process data from a large collection of process data. In this table, it can be seen that the point of the MLPM is an integration of the existing process and motion models. However, it is important to define the similarity function, which will be done in future studies.

B. Functions to be realized based on MLPM

There are many functions that can be realized based on MLPM. In particular, the followings are important and basic functions, which we are developing.

- Process database: In order to realize the proposed MLPM, we need to develop a method to manage the detailed specifications of our process model, in order to realize an automatic process construction method by developing data aggregation and abstraction methods.
- Process classification: When the process database has been generated, we need to develop a method for comparing two processes and then a method for classifying processes for process management.
- Process matching and similarity search: It is important to develop a matching or search function of similar processes for improving the current process management or for supporting a user's process management.
- Process mining: Process association rules or process correlations need to be extracted from an MLPM-based process database.

C. MPLM limitations

Currently, the following types of process data cannot be represented using MLPM: 1) continuous motions and trajectories, and 2) ontological relationships among component instances. After specifying MLPM in detail and developing several applications, the model will be extended to nullify these limitations.

D. Applications

As mentioned in the previous section, suppose that an junior doctor has to learn how to do an intravenous injection. Although the doctor possesses useful how-to books for self-study of the intravenous injection process, it is impossible to practice the injection on a medical volunteer without a skilled doctor being present. The most difficult problem is how to impart the doctor the know-how and the difference between his/her execution of the injection process and that of skilled doctor, by dynamically checking the injection process. If it were possible to realize such support for a junior doctor, this would draw a better and more natural way to practice a process than looking at a how-to-book.

Moreover, it is also impossible in general to impart the know-how of skilled doctors to junior doctors, because it tends to be difficult to describe the know-how precisely. If the know-how can be extracted by comparing the junior doctor's and skilled doctor's processes and by generating a process rule from many skilled doctors' processes, the know-how can be expressed and thus transferred to junior doctors.

Finally, if the know-how can be extracted, described, and visualized based on a process model and its database, skilled doctors can identify their own skills and know-hows and offer them to aid the development of the best process. Currently, such doctors watch their past videos many times in order to develop the best process, to improve the current process or to solve a problem related to the current process. A process model-based technique, such as our MLPM, can help them to do this more effectively.

Another application of our MLPM, in addition to that for the medical field, is a process model for team sports. A team sport is one that involves people playing together to accomplish a specified goal. Although there are many team sports in the world, we consider ballroom dancing, a sport that involves pairs, as an example. In a ballroom dancing competition, the competitors are judged according to several factors using posture, timing, togetherness, and musicality. After archiving and describing the entire process of a ballroom dance using our MLPM, the dancers can check their process from the abstract to the concrete level by examining the results of the

 ISBN: 978-1-61208-329-2

differentiating function of a MLPM-based system or mining the processes using a method based on the MLPM.

The model can also be applied to improve worker's skills. After skilled worker's movements have been captured and the captured data have been annotated and aggregated, the worker's skill processes can be archived using our model. Then, when a novice worker is learning how to do a job, the archived processes can be used effectively by using functions based on the model. That is, the model enables the workers to improve their skills by checking the differences between a skilled worker's process and their own process. The model can also produce a set of rules to a worker learn the skills more easily. Clearly, many studies on the model refinement, model application experience, and MLPM-based mining method development are still necessary.

V. Conclusion and Future Work

In this paper, we proposed a meta model for representing the overall processes from the higher abstract level to the lower actual motion level. The point of the meta model is to introduce a multi-layered process meta model to represent various kinds of representations of processes in an integrated way. We described the basic concept and fundamental components for developing the process meta model.

The proposed MLPM requires further work. In particular, it is necessary to develop 1) detailed specifications of MLPM, 2) methods for matching, searching, and classifying processes using our MLPM, and 3) a new method of process mining from MLPM-based process databases. We also plan to develop a system based on the proposed MLPM after the design of the architecture has been completed.

Acknowledgment

This work was partially supported by MEXT-Supported Program for the Strategic Research Foundation at Private Universities, 2013-2017. We also thank the reviewers for providing valuable comments and suggestions.

References

[1] Kinect for Windows, http://www.microsoft.com/en-us/kinectforwindows/ , Microsoft, 2010 [retrieved: Jan. 7, 2014].

[2] The Leap Motion Controller, https://www.leapmotion.com/, Leap Motion, Inc., 2013 [retrieved: Jan. 7, 2014].

[3] OptiTrack, http://www.naturalpoint.com/optitrack/ , NaturalPoint, Inc., 2013 [retrieved: Jan. 7, 2014].

[4] H-anim: Specification for a Standard Humanoid, http://h-anim.org/ , Humanoid Animation Working Group, 2000 [retrieved: Jan. 7, 2014].

[5] M. Meredith and S. Maddock, "Motion capture file formats explained," http://www.dcs.shef.ac.uk/intranet/research/public/resmes/CS0111.pdf, 2001, [retrieved: Dec. 22, 2013].

[6] Business Process Model and Notation (BPMN), FTF Beta 1 for Ver. 2.0, Object Management Group OMG Specifications, http://www.omg.org/cgi-bin/doc?dtc/09-08-14.pdf, Sept. 2009 [retrieved: Jan. 7, 2014].

[7] Workflow Management Coalition Workflow Standard Process Definition Interface- XML Process Definition Language, The Workflow Management Coalition The Workflow Management Coalition Specification WFMC-TC-1025, Ver.2.2, http://www.xpdl.org/standards/xpdl-2.2/XPDL2.2(2012-08-30).pdf, Aug. 2012 [retrieved: Dec. 22, 2013].

[8] Business Process Definition MetaModel (BPDM), Ver. 1.0, Object Management Group OMG Specifications, http://www.omg.org/spec/BPDM/1.0/volume1/PDF/andhttp://www.omg.org/spec/BPDM/1.0/volume2/PDF/, Nov. 2008 [retrieved: Dec. 22, 2013].

[9] W. M. P. van der Aalst, A. H. M. ter Hofstede, and M. Weske, "Business process management: A survey," in BPM 2003, LNCS 2678, W. M. P. van der Aalst et al., Ed., 2003, pp. 1–12.

[10] L. Cao, "Behavior informatics and analytics: Let behavior talk," Proc. IEEE International Conference on Data Mining Workshops, 2008, pp. 87–96.

[11] P. Kueng, P. Bichler, P. Kawalek, and M. Schrefl, "How to compose an object-oriented business process model?" Proc. IFIP Method Engineering, 1996, pp. 94–110.

[12] R. S. Aguilar-Saven, "Business process modelling: Review and framework," International Journal of Production Economics, vol. 90, no. 2, 2004, pp. 129–149.

[13] W. M. P. van der Aalst, H. A. Reijers, A. J. M. M. Weijters, B. F. van Dongen, and A. K. A. de Medeiros et al, "Business process mining: An industrial application," Information Systems, vol. 32, no. 5, 2007 pp. 713–732.

[14] A. l. Rebugea and D. R. Ferreirab, "Business process analysis in healthcare environments: a methodology based on process mining," Information Systems, vol. 37, no. 2, 2012, pp. 99–116.

[15] D. Karagiannis, S. Junginger, and R. Strobl, "Introduction to business process management systems concepts," Business Process Modelling, 1996, pp. .81–106.

[16] R. S. Mans, M. H. Schonenberg, M. Song, W. M. P. van der Aalst, and P. J. M. Bakker, "Application of process mining in healthcare: A case study in a dutch hospital," Biomedical Engineering Systems and Technologies, Communications in Computer and Information Science, vol. 25, 2009, pp. 425–438.

[17] M. Castellanos, F. Casati, and U. Dayal, "A probabilistic-based approach to process model discovery," Proc. IEEE ICDE Workshop, 2011, pp. 232–237.

[18] L. Wen, J. Wang, W. M. P. van der Aalst, B. Huang, and J. Sun, "A novel approach for process mining based on event types," J. Intell. Inf. Syst., vol. 32, no. 2, 2009, pp. 163–190.

[19] W. van der Aalst, A. Adriansyah, A. K. A. de Medeiros, F. Arcieri, and T. B. et al., "Process mining manifesto," Proc. BPM 2011 Workshops, Part I, LNBIP 99, Springer, 2012, pp. 169–194.

[20] W. M. P. van der Aalst, M. Pesic, and M. Song, "Beyond process mining: From the past to present and future," Proc. CAiSE 2010, LNCS 6051, Springer, 2010, pp. 38–52.

[21] J. Ferber and O. Gutknecht, "A meta-model for the analysis and design of organizations in multi-agent systems," Multi Agent Systems, 1998, pp. 128–135.

[22] J. Dijkstra and H. Timmermans, "Towards a multi-agent model for visualizing simulated user behavior to support the assessment of design performance," Proc. ACADIA '99, Automation in Construction, vol. 11, no. 2, Elsevier, 1999, pp. 135–145.

[23] N. Minar, R. Burkhart, C. Langton, and M. Askenazi, "The swarm simulation system: A toolkit for building multi-agent simulations," Santa Fe: Santa Fe Institute, 1996.

[24] "Scenario description for multi-agent simulation," Proc. AAMAS03, 2003, pp. 369–376.

[25] E. L. Cao and P. S. Yu, Eds., Behavior Computing: Modeling, Analysis, Mining and Decision, Springer, 2012.

[26] D. Neumuth, F. Loebe, H. Herrec, and T. Neumuth, "Modeling surgical processes: A four-level translational approach," Artificial Intelligence in Medicine, vol. 2011, no. 51, 2011, pp. 147–161.

[27] J. F. Allen and G. Ferguson, "Actions and events in interval temporal logic," Journal of logic and computation, vol. 4, no. 5, 1994, pp. 531–579.

 ISBN: 978-1-61208-329-2

How can Start-up Business Firms Keep the Motivations of Employees?

Analyzing organizational management strategies through an agent based model

Tomomi Kobayashi
Waseda University
Tokyo, Japan
kbys@triton.ocn.ne.jp

Satoshi Takahashi, Masaaki Kunigami,
Atsushi Yoshikawa, Takao Terano
Tokyo Institute of Technology
Yokohama, Japan

Abstract— **This paper describes an agent based model of a start-up business firm for analyzing the conflict between the organizational performance and its employee motivation. Start-up business firms tend to change its management strategies with the growth of the firm in order to increase the productivity and business performance. However, those changes may cause negative impacts on the motivation or entrepreneurship of its employee, and they might weaken the vitality of the firm for sustainable growth. According to those considerations, we have conducted the agent based simulation and have gotten the following suggestions. 1) Building management structures increases organizational performance while decreasing employee motivation. 2) Keeping the initial informal management style by not building a management structure makes employee motivation increase, however, it makes organizational performance decline. 3) Informal networks among diversified employees can ease the negative impact of building a management structure.**

Keywords- Agent based modeling; start-up business firm; organizational life cycle.

I. INTRODUCTION

Companies tend to build their management structures with the growing size of the organization in order to keep or enhance their organizational performance and profitability. "Management structure" means, for example, building organizational hierarchy, formalizing communication, creating a system of rewards, and so on. It has some advantages of enhancing efficiency of the company's operation and establishing an orderly growth. However, it also has some disadvantages of reducing organization member's entrepreneurship and motivation, because the members role and power is restricted by formalized management system. Under those considerations, we have made following assumptions.

- Underlying conflicts between organizational performance and employee motivation exist in a start-up firm.
- The changes of organizational management strategies may effect on those conflicts.

According to the assumptions, we propose an agent based model, which consists of organization utility and individual utility functions which represent organizational performance, and employee motivation.

The first purpose of this paper is to present an agent based model for analyzing the effects of building a management structure for both organizational performance and member's motivation. The second purpose is to detect the factors for the mitigating disadvantage of building a management structure.

In organizational life cycle theory, there are many definitions of organizational growth stages [1], and they are frequently used in organization management because easy to be understood intuitively. However, they are criticized that they tend to fall into tautology, for example, organizations go into "formalization stage" because they formalize their management [2]. In order to overcome the tautology and make the discussion in the organizational life cycle to be more meaningful, it is important to focus on not only management style itself, but also its effect on organizational members' motivation, because the organizational growth stage transition should be decided considering the conflict between organizational performance and employee motivation. For that reason, we have built the model for analyzing the conflict.

The rest of the paper is organized as follows: Section 2 explains our model; Section 3 describes the simulation experiment settings; Section 4 shows the experimental results; Section 5 shows the experimental result which focuses on the informal network and diversity in employees; and Section 6 presents our findings and remarks as a conclusion

II. AGENT BASED MODEL

This section describes our agent based model for analyzing the effect of management style transition, which simplifies a real structure of an organization and the relation between an organization and individuals. We have applied the agent based modeling method [3], [4] in order to examine the bottom up changing process of organizational performance and employee motivation. In this model, hierarchical utility landscape is implemented based on the landscape theory [5], [6] that consists of two classes: individual utility and organizational utility.

Fig. 1 shows an outline of the hierarchical utility landscape in our model. The utility function of individuals means experience and values of each agent. The utility function of the organization means strategy and business model of a company. When agents choose their actions, their own utility and their contributions to organizational utility are determined. Organizational utility is distributed to agents through a reward system.

 ISBN: 978-1-61208-329-2

A. Structure of the Model

Fig. 1 shows an outline of the hierarchical utility landscape in our model.

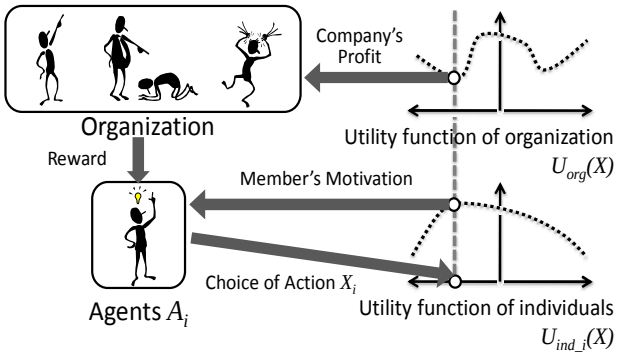


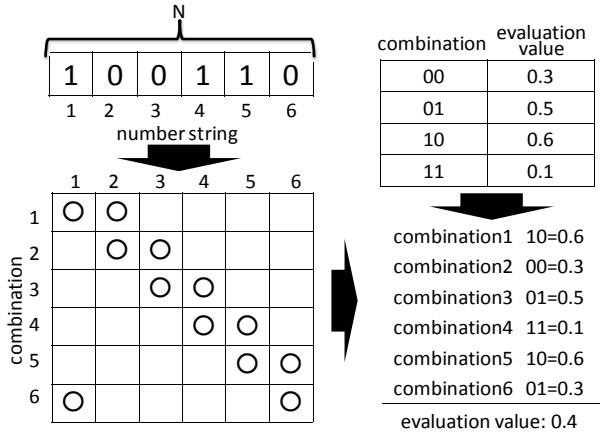
Figure 1. Structure of the Agent Based Model.

In Fig. 1, a hierarchical organizational structure which consists of two layers is brought into our model, because it is commonly seen in many companies. The utility function of individuals means the values of each agent. The utility function of the organization means the business model of a company.

In this model, agents choose their actions according to the rewards from organizations and information from another agent. As a result, their utility production amount for the organization is determined based on utility functions. Agents can recognize their own utility, however, they cannot completely recognize organizational utility.

B. Utility Function

The utility functions described in the previous section, are based on the NK fitness landscape model [7], [8]. The NK model determines the values of N integer sequences, and utility landscape is defined by the combinations of K integers. Fig. 2 shows a sample of integer combinations and their values, in case of N=6 and K=1.



combination	evaluation value
00	0.3
01	0.5
10	0.6
11	0.1

Figure 2. NK Model.

The variations of utility functions are described by number sequences and their evaluation values. The evaluation value is between 0 to 1 depending on combinations of integers. The complexity of the utility landscape depends on the number of integers and their combinations.

C. Choosing Actions of Agents

Equation (1) describes that all agents changes their action in order to increase their satisfaction. The degree of agents satisfaction increases along with the rising of their individual utilities: *Uind_i(X)*, and rewards from organization: Re_i. The index *i* means the number of agents.

$$S(U_{ind_i}(X), Re_i) = U_{ind_i}(X) + Re_i \quad (1)$$

Equation (2) describes that agents imitate the actions of other agents whose actions are similar to them and receiving more rewards from the organization. P_j means the probability that agent$_i$ imitating the action of agent$_j$. *k* means the number of agents. *Lij* means the similarity of action between agent$_i$ and agent$_j$. The agents evaluate their satisfaction after imitation, and then return to original action when their degrees of satisfaction have been declined by the imitation.

$$P_j = \frac{Re_j \times L_{ij}}{\sum_{k \neq i} Re_k \times L_{ik}} \quad (2)$$

The agents produce their own utility and contribute to organizational utility as the result of their actions. The contributions of agents are accumulated in an organization.

D. Organizational Structure

Fig. 3 shows the hierarchical tree structure is applied to our model.

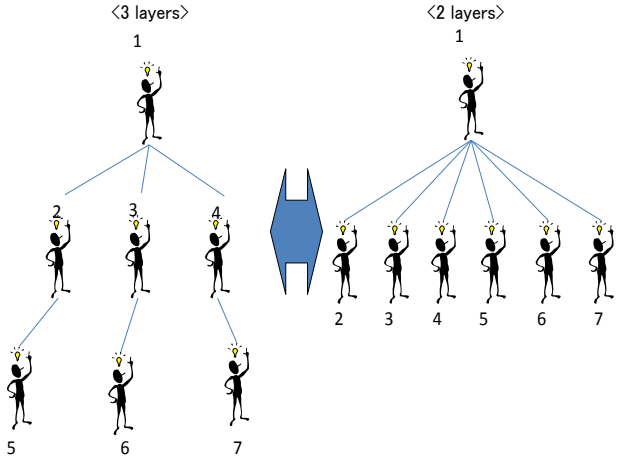


Figure 3. Changing the hierarchy, keeping the number of agents.

 ISBN: 978-1-61208-329-2

We change the number of layers by controlling the number of subordinate agents of each upper layer agent as shown in Fig.3.

III. EXPERIMENT SCENARIOS AND SETTINGS

Based on the descriptions of the model in previous section, we have developed the simulator according to agent based computational architecture [9] in Java language. This section describes the scenarios and parameter settings of the agent based simulation experiment.

In this experiment, we set the two types of management transition scenarios with three parameters based on organizational life cycle theory, as shown in Table 1. Those are 1) Building management structure with growth stage transition, 2) Keep initial management style throughout growth stages. We set each experimental condition, and analyze the difference of individual and organizational utility production amount depending on those scenarios. In scenario 1), organizational hierarchy is enlarged, ratio of informal network is lower, and the degree of result-based reward is higher with the transition of the growth stage. In scenario 2), all three parameters are maintained at initial condition throughout growth stages. The number of agents is increasing from 5 to 50, and the ratio of diverse agents is increasing from 0% to 70% with progress from stage 1 to stage 4.

TABLE I. EXPERIMENT SCENARIO AND PARAMETER SETTINGS

Growth stage		Stage 1 Conception	Stage 2 Commerciali-zation	Stage 3 Growth	Stage 4 Stability
Number of agents		5	20	40	50
Diversity of agents		0%	20%	50%	70%
Scenario 1	Organization hierarchy	3	4	4	5
Build formal management structure with growth stage transition	Ratio of informal network	100%	70%	40%	20%
	Degree of result-based reward	1.1	4	18	36
Scenario 2	Organization hierarchy	3	3	3	3
Keep initial management style throughout growth stages	Ratio of informal network	100%	100%	100%	100%
	Degree of result-based reward	1.1	1.1	1.1	1.1

In the next subsections, simulation experiments are organized according to the scenarios which are described in Table 1.

IV. THE RESULTS OF COMPUTER SIMULATION

A. Experimental Results of Organizational Utility Production

At the beginning, Fig. 4 represents the result of organizational utility production change with the growth stage transition. Agents produce more organization utility in experiment scenario 1 than scenario 2.

This result means that building a management structure is increasing the performance of the organization. On the other hand, organizational performance is decreasing by keeping the initial informal management style.

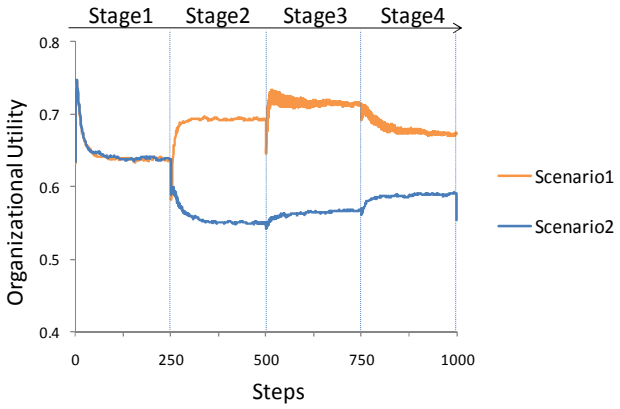


Figure 4. Difference of organizational utility production curve with growth stage transition by experiment scenarios.

B. Experimental Results of Individual Utility Production

Fig. 5 shows the result of Individual utility production change with the growth stage transition. Agents produce less individual utility in experiment scenario 1 than scenario 2.

This result means that building a management structure is decreasing the motivation and entrepreneurship of organization members. On the other hand, the motivation of organization members is maintained by keeping with the initial informal management style compared to formalization.

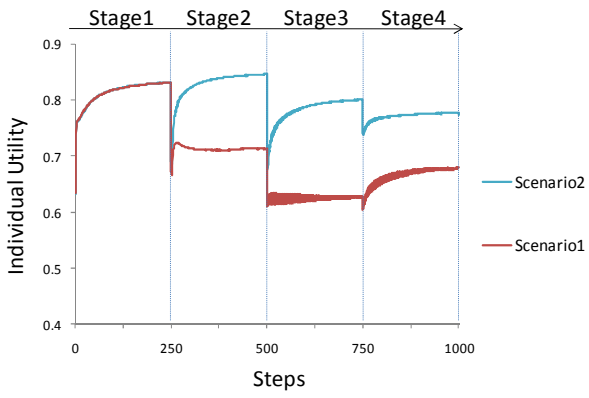


Figure 5. Difference of individual utility production curve with growth stage transition by experiment scenarios.

 ISBN: 978-1-61208-329-2

C. Conflict between Organizational and Individual Utility Production

In this subsection, the gap between organizational and individual utility production is analyzed. Fig. 6 shows the difference of gap comparing scenario 1 and scenario 2.

In scenario 1, the gap between organizational utility and individual utility production is narrowing with growth stage transition. On the other hand, it is maintained throughout the growth stages in scenario 2 compared to scenario 1.

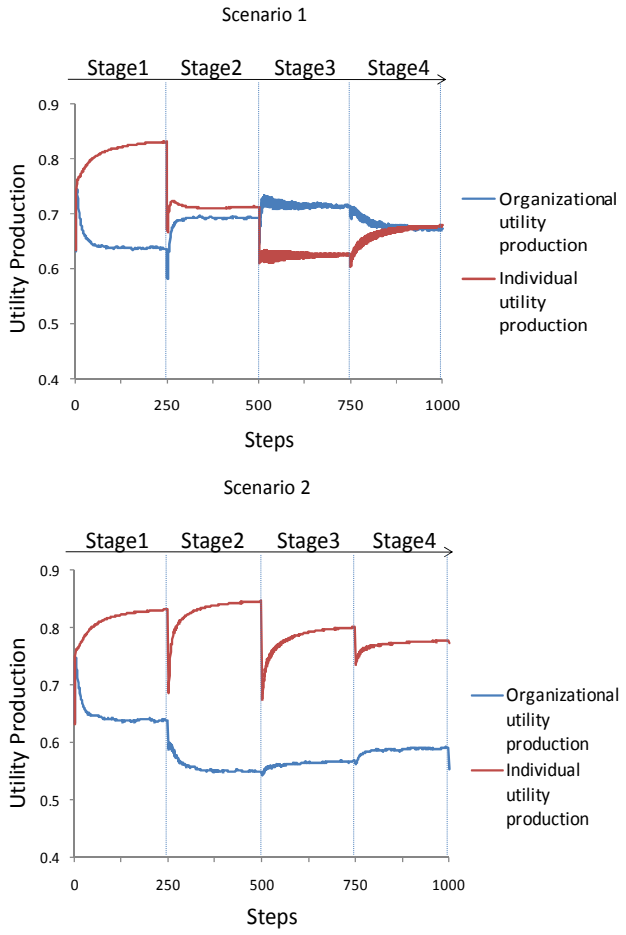


Figure 6. The comparison of the gap between organizational and individual utility production in experimental scenario 1 and 2.

The result in Fig. 6 means that building a management structure mitigates the conflict between organizational performance and individual motivation while decreasing member's motivation. On the other hand, organization members behave pursuing their motivation while neglecting their contribution to organizational profit by maintaining informal management style.

V. THE KEY FACTORS FOR EASING CONFLICT

As described in the previous subsection, building a management structure mitigates the conflict between organizational performance and individual motivation. However, it is achieved by sacrificing individual motivation and this may be a cause of preventing organization for sustainable growth. Therefore, it is necessary to achieve an appropriate balance between organizational performance and individual motivation.

A. Informal Network

The experimental results of simulation focusing on informal network are shown and discussed in this subsection. Fig. 7 presents the gap between organizational and individual utility production curve in experimental scenario 1 except for informal networking.

In this experiment, the informal network ratio has maintained 80% and 0% throughout all stages in order to fix the informal communication volume among the agents. Other conditions; organization hierarchy and degree of result-based reward, are the same as scenario 1.

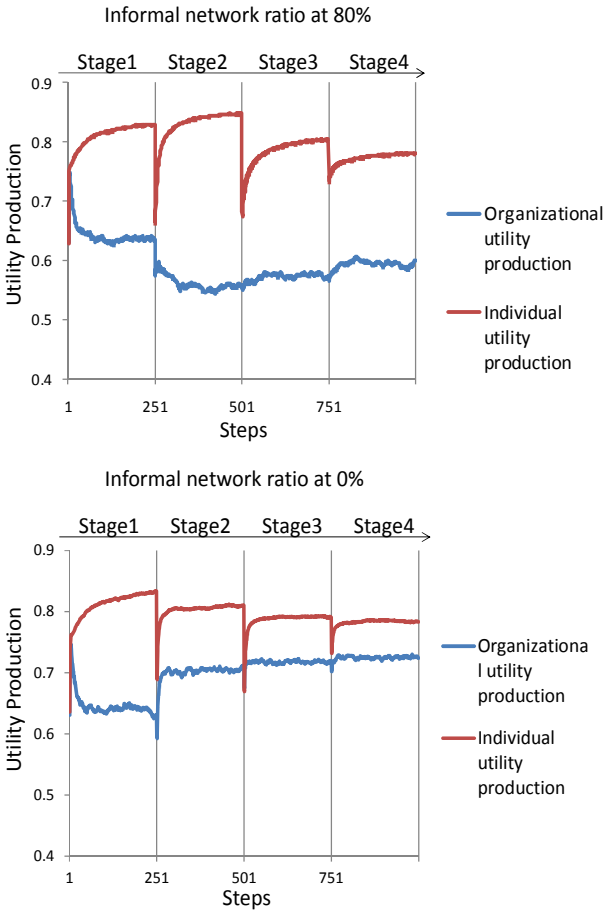


Figure 7. The comparison of utility production amount between the informal network ratio 80% and 0% throughout all stages in experimental scenario 1.

As seen in Fig. 7, the result of maintaining an informal network ratio 80% is similar to scenario 2 in Fig. 6. This result means that company employees tend to pursue their

 ISBN: 978-1-61208-329-2

motivation while neglecting their contribution to organizational profit by maintaining an informal network at high ratio even though building hierarchal organization structure and result-based reward system.

On the other hand, organizational utility production with the condition of informal network ratio at 0%, is higher than at 80% in Fig. 7. And individual utility production has been kept higher than the scenario 1 in Fig. 6. This result suggests that building a formal communication style in the early stages is more effective for balancing the organizational performance and employee motivation than building it in later stages.

Those experiments have been conducted under the consideration that a communication strategy is one of the factors in achieving the balance. Some start-up companies have overcome the stagnation by acquiring new capabilities with spontaneous collaboration [10]. And in the organizational life cycle theory, it is described that decentralization of organizational structure is necessary to maintain organizational flexibility and achieve sustainable growth [1]. The previous studies suggest the importance of communication strategies based on the informal networks [11], [12].

B. *Diversity in Organization*

Fig. 8 presents the comparison of utility production between uniform agents group and diversified agents group based on the previous study on diversity in organization [13], [14]. The experimental conditions are as same as in Fig. 7, and an informal network ratio is maintained at 80%.

As seen in Fig. 8, the organizational utility and the individual utility productions are more balanced in the diversified group than in the uniform group. Furthermore, its individual utility production amount is higher than that of scenario 1 in Fig. 6, and the organizational utility production amount is higher than that of scenario 2 in Fig. 6. In stage 1, there is no utility production in diversified group, because there are no diversified agents on stage 1 according to the condition setting.

Those results suggest that informal networks may enhance the mutual communication among organization members, and within uniform agents, they could have imitated the behavior which increases individual utility production because they have the same individual utility function; personal value or experience. As a result, they could have neglected contribution to organizational performance because they could increase their satisfaction without reward from the organization.

On the other hand, within diversified agents, they could have imitated the behavior which increases contribution to organizational utility in order to maintain or increase their satisfaction with the reward form organization. The reason is that it is difficult to increase individual utility by mutual imitation for diversified agents because their individual utility functions are different from each other. Those behaviors are caused by the choosing action and maintaining satisfaction mechanism of the agents, which is defined in (1).

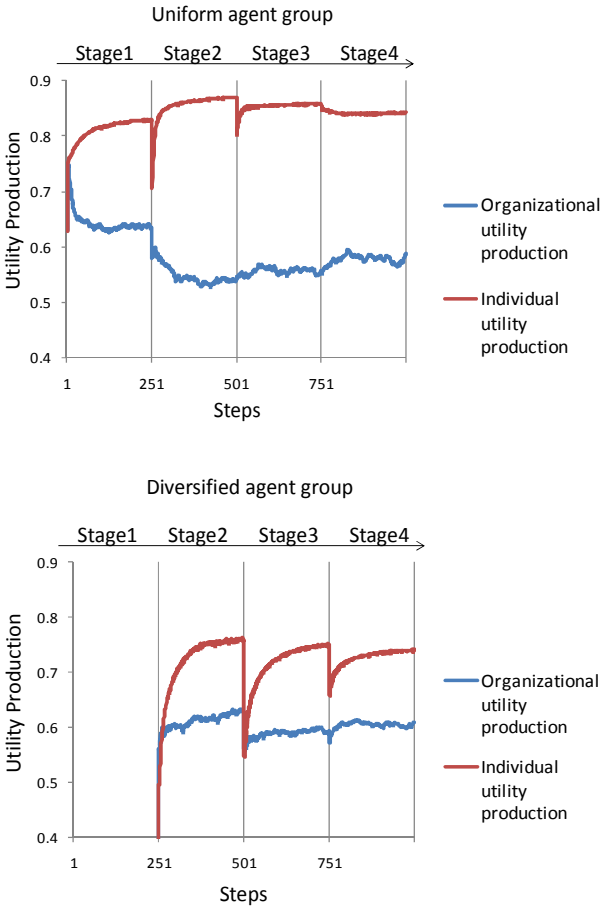


Figure 8. The comparison between uniform agents group and diversified agents group at the same condition in Fig. 7.

The results and considerations in Fig.8 suggest that enhancing diversity of organization is a key factor in balancing organizational performance and individual motivation by optimizing the effects of the informal network.

VI. CONCLUSIONS AND FUTURE WORK

This paper has presented an agent based model for analyzing the effect of building a management structure in start-up business firms. In this paper, we have intended to contribute to organizational life cycle theory by analyzing the effect of management style transition. The advantage of our model is to enable analysis of the management structure's effect to organizational performance and member's motivation in an integrated view.

Many start-up companies intend to increase their organizational performance by building management structure, but they tend to fall into stagnation by failure of keeping their growth abilities. The results of experiments in this paper show that the company employees tend to increase their contribution to organizational performance while sacrificing their individual motivation by building the management structure. This may be a cause of preventing start-up companies from sustainable growth. On the other

 ISBN: 978-1-61208-329-2

hand, company employees pursue their motivation while neglecting organizational performance when the initial informal management style is maintained.

This paper also describes the effect of informal network and diversity in employees as follows.

- When informal networks are expanded in uniform agent group, the agents tend to behave selfishly and neglect organizational performance. However, in the diversified agent group, informal networks are effective to balance organizational performance and individual motivation. This experimental result suggests that an informal network in diversified organization is a key factor for achieving sustainable growth by mitigating the conflict between organizational performance and employee motivation.
- Building a formal communication style in the early stages is more effective for balancing the organizational performance and employee motivation, compared to building it in the later stages.

In the further work, we would conduct additional experiments and analysis, and detect more key factors for sustainable growth of start-up business firms by balancing the organizational performance and employee motivation.

REFERENCES

[1] J. B. Quinn, and K. Cameron, “Organizational life cycles and shifting criteria of effectiveness Some preliminary evidence,” Management Science, vol. 29, no. 1, pp. 33-51, 1983.

[2] R. K. Kazanjian, and R. Drazin, “A stage-contingent model of design and growth for technology-based new ventures, ” Journal of Business Venturing, vol. 5, pp. 137-150 , 1990.

[3] T. Terano, “Perspective on Agent-Based Modeling, ” JSIS & JASI1-8, 2006.

[4] T. Terano, “Why Agent-Based Modeling in Social System Analysis? ” Oukan, Vol.4, No.2, pp. 56-62, 2010.

[5] R. Axelrod, “The Complexity of cooperation, ” Princeton Univ. Press, 1999.

[6] K. Kijima, “Generalized Landscape Theory: Agent-based Approach to Alliance Formations in Civil Aviation Industry, ” Journal of System Science and Complexity, vol. 14, no. 2, pp. 113-123, 2001.

[7] S. Kauffman, “The Origins of Order: Self-Organization and Selection in Evolution, ” Oxford University Press, 1993.

[8] S. Kauffman, “At Home in the Universe: The Search for Laws of Self-Organization and Complexity, ” Oxford University Press, 1995.

[9] R. L. Axtell, “Why Agents? On The Varied Motivations for Agent Computing in the Social Sciences, ” Center on Social and Economic Dynamics Working Paper No. 17, 2000.

[10] T. Kobayashi, “A research on the organizational development of start-up companies which are operating business of software - Based on the resource and capability relation framework, ” Graduate School of System Management University of Tsukuba, 2007.

[11] J. Katzenbach, and Z. Khan, “Leading outside the lines: How to Mobilize the (in) Formal Organization Energize Your Team and Get Better Results, ” Booz & Company, Inc, 2010.

[12] R. Kraut, R. Fish, R. Root, and B. Chalfonte, “Informal communication in organizations: Form, function, and technology, ” I S. Oskamp & S. Spacapan (Eds.), Sage Publications, pp. 145-199, 1990.

[13] L. Hong and S. E. Page, “Groups of diverse problem solvers can outperform groups of high-ability problem solvers, ” PNAS, vol. 101, no. 46, pp. 16385-16389, 2004.

[14] S. E. Page, “The Difference, ” Princeton University Press, 2007.

 ISBN: 978-1-61208-329-2