INTELLI 2012

The First International Conference on Intelligent Systems and Applications

ISBN: 978-1-61208-224-0

April 29 - May 4, 2012

Chamonix / Mont Blanc, France

INTELLI 2012 Editors

Pascal Lorenz, University of Haute Alsace, France

Petre Dini, Concordia University, Canada / China Space Agency Center, China

INTELLI 2012

Foreword

The First International Conference on Intelligent Systems and Applications [INTELLI 2012], held between April 29th and May 4th, 2012 in Chamonix / Mont Blanc, France, was an inaugural event on advances towards fundamental, as well as practical and experimental aspects of intelligent and applications.

The information surrounding us is not only overwhelming but also subject to limitations of systems and applications, including specialized devices. The diversity of systems and the spectrum of situations make it almost impossible for an end-user to handle the complexity of the challenges. Embedding intelligence in systems and applications seems to be a reasonable way to move some complex tasks form user duty. However, this approach requires fundamental changes in designing the systems and applications, in designing their interfaces and requires using specific cognitive and collaborative mechanisms. Intelligence became a key paradigm and its specific use takes various forms according to the technology or the domain a system or an application belongs to.

We take here the opportunity to warmly thank all the members of the INTELLI 2012 Technical Program Committee, as well as the numerous reviewers. The creation of such a 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 INTELLI 2012. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

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

We hope that INTELLI 2012 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of intelligent systems and applications.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed their stay in the French Alps.

INTELLI Advisory Committee:

Pascal Lorenz, University of Haute Alsace, France
Petre Dini, Concordia University, Canada / China Space Agency Center, China

INTELLI 2012 PROGRAM COMMITTEE

INTELLI 2012 Advisory Committee

Pascal Lorenz, University of Haute Alsace, France
Petre Dini, Concordia University, Canada / China Space Agency Center, China

INTELLI 2012 Technical Program Committee

Syed Sibte Raza Abidi, Dalhousie University - Halifax, Canada
Witold Abramowicz, The Poznan University of Economics, Poland
Michael Affenzeller, HeuristicLab, Austraia
Samir Aknine, Université Lyon 1, France
Jose M. Alcaraz Calero, Hewlett-Packard Laboratories - Bristol, UK
Andreas S. Andreou, Cyprus University of Technology - Limassol, Cyprus
Joseph Andrew Giampapa, Carnegie Mellon University, USA
Ngamnij Arch-int, Khon Kaen University, Thailand
Wudhichai Assawinchaichote, Mongkut's University of Technology -Bangkok, Thailand
Pradeep Atrey, University of Winnipeg, Canada
Paul Barom Jeon, Samsung Electronics, Korea
Daniela Barreiro Claro, Federal University of Bahia, Brazil
Rémi Bastide, Université Champollion, France
Bernhard Bauer, University of Augsburg, Germany
Barnabas Bede, DigiPen Institute of Technology - Redmond, USA
Noureddine Belkhatir, University of Grenoble, France
Orlando Belo, University of Minho, Portugal
Petr Berka, University of Economics, Prague, Czech Republic
Félix Biscarri, University of Seville, Spain
Magnus Boman, SICS and KTH/ICT/SCS - Kista, Sweden
Luis Borges Gouveia, University Fernando Pessoa, Portugal
Abdenour Bouzouane, Université du Québec à Chicoutimi, Canada
José Braga de Vasconcelos, University Fernando Pessoa - Porto, Portugal
Stefano Bromuri, University of Applied Sciences Western Switzerland, Switzerland
Rui Camacho, Universidade do Porto, Portugal
Luis M. Camarinha-Matos, New University of Lisbon, Portugal
Longbing Cao, University of Technology - Sydney, Australia
Jose Jesus Castro Sanchez, Universidad de Castilla-La Mancha - Ciudad Real, Spain
Kit Yan Chan, Curtin University - Western Australia, Australia
Chin-Chen Chang, Feng Chia University, Taiwan, R. O. C.
Maiga Chang, Athabasca University, Canada
Yue-Shan Chang, National Taipei University, Taiwan
Naoufel Cheikhrouhou, Ecole Polytechnique Fédérale de Lausanne, Switzerland
Rung-Ching Chen, Chaoyang University of Technology, Taiwan
Li Cheng, BII/A*STAR, Singapore
Been-Chian Chien, National University of Tainan, Taiwan
Sunil Choenni, Ministry of Security and Justice, The Netherlands
Byung-Jae Choi, Daegu University, Korea

Filippo Neri, University of Naples "Federico II", Italy
Mario Arrigoni Neri, University of Bergamo, Italy
Hongbo Ni, Northwestern Polytechnical University, China
Yoosoo Oh, Gwangju Institute of Science and Technology, South Korea
Hichem Omrani, CEPS/INSTEAD Research Institute, Luxembourg
Frank Ortmeier, Otto-von-Guericke Universitaet Magdeburg, Germany
Jeng-Shyang Pan, Harbin Institute of Technology, Taiwan
Endre Pap, University Novi Sad, Serbia
Marcin Paprzycki, Systems Research Institute / Polish Academy of Sciences - Warsaw, Poland
Dana Petcu, West University of Timisoara, Romania
Leif Peterson, Methodist Hospital Research Institute / Weill Medical College, Cornell University, USA
Alain Pirott, Université de Louvain - Louvain-la-Neuve, Belgium
Agostino Poggi, Università degli Studi di Parma, Italy
Anca Ralescu, University of Cincinnati, USA
Thurasamy A/L Ramayah, Universiti Sains Malaysia - Penang, Malaysia
Fano Ramparany, Orange Labs Networks and Carrier (OLNC) - Grenoble, France
Zbigniew W. Ras, University of North Carolina - Charlotte & Warsaw University of Technology, Poland
José Raúl Romero, University of Córdoba, Spain
Danda B. Rawat, Eastern Kentucky University, USA
David Riaño, Universitat Rovira i Virgili, Spain
Daniel Rodríguez, University of Alcalá - Madrid, Spain
Oscar Mario Rodriguez-Elias, Instituto Tecnologico de Hermosillo, Mexico
Agos Rosa, Technical University of Lisbon, Portugal
Gunter Saake, University of Magdeburg, Germany
Ozgur Koray Sahingoz, Turkish Air Force Academy, Turkey
Daniel Schang, Groupe Signal Image et Instrumentation - ESEO, France
Amal El Fallah Seghrouchni, University of Pierre and Marie Curie (Paris 6) - Paris, France
Hirosato Seki, Osaka Institute of Technology, Japan
Timothy K. Shi, National Central University, Taiwan
Kuei-Ping Shih, Tamkang University - Taipei, Taiwan
Choonsung Shin, Carnegie Mellon University, USA
Elena Simperl, Karlsruhe Institute of Technology, Germany
Peter Sincák, Technical University of Kosice, Slovakia
Spiros Sirmakessis, Technological Educational Institute of Messolonghi, Greece
Alexander Smirnov, St. Petersburg Institute for Informatics and Automation of Russian Academy of Sciences (SPIIRAS), Russia
Paolo Spagnolo, Italian National Research Council, Italy
Adel Taweel, King's College London, UK
Abdel-Rahman Tawil, University of East London, UK
Jilei Tian, Nokia Research Center Beijing, China
Federico Tombari, University of Bologna, Italy
Anand Tripathi, University of Minnesota Minneapolis, USA
Juan Carlos Trujillo Mondéjar, University of Alicante, Spain
Theodoros Tzouramanis, University of the Aegean, Greece
Eiji Uchino, Yamaguchi University, Japan
Gancho Vachkov, Yamaguchi University - Yamaguchi City, Japan
Jan Vascak, Technical University of KoSice, Slovakia
Mario Vento, Università di Salerno - Fisciano, Italy

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

Swedish Speech Recognition Using Linear Time Normalization and Feature Selection Optimization

Mattias Wahde
Chalmers University of Technology
Department of Applied Mechanics
Göteborg, Sweden
Email: mattias.wahde@chalmers.se

***Abstract*—A system for Swedish isolated word recognition has been developed, intended for use in an intelligent news reader agent for elderly care. The system uses linear time normalization of feature time series, as well as optimization (with a genetic algorithm) of both feature selection and feature weighting. The optimization of feature selection results in a potentially important decrease in the time needed to recognize a spoken word, while maintaining speech recognition performance.**

Keywords-Speech recognition; optimization; linear time normalization

I. INTRODUCTION

The increase in the fraction of elderly people, relative to the working population, is a strong demographical trend at present: One recent estimate expects the fraction of elderly people (65+ years old) in the European Union to increase from around 17% in 2010 to 30% in 2060 [1]. Furthermore, the fraction of people above the age of 80 is expected to increase from around 5% to 12%, over the same time span.

In the near future, it is therefore likely that a variety of technological tools, such as assistive or partner robots [2], as well as intelligent homes [3], will come to play an important role in elderly care. Indeed, some prototypes already exist, such as the therapeutic seal robot Paro [4] and the assistive robot Kompai [5]. An increased role of such tools is probable for several reasons. For example, the (relative) decrease in the size of the working population, combined with the increase in the size of the elderly population, is likely to cause staff shortages in elderly care. Furthermore, as a quality-of-life issue, many elderly people prefer to live in their own home (rather than in, say, a nursing home) for as long as possible [2], something that can be facilitated using technological tools that, for example, can monitor medicine intakes, alert relatives or healthcare workers in case of injuries etc.

In order for such technological tools to be applicable in meaningful interactions with elderly people, they will need to be equipped with intuitive user interfaces, as one cannot expect their users to be familiar with computers, let alone robots. In addition, robots and agents must, of course, be able to interact in their users' language, which might not be one of the major languages, such as English, Japanese, or Chinese. For these reasons, the Adaptive systems research group at Chalmers University of Technology, in Göteborg, Sweden, has recently started a project (that will be described in detail elsewhere) to develop both intelligent software agents and a partner robot intended for use in elderly care in Sweden.

This paper will consider one particular aspect of that project, namely speech recognition (in Swedish) for an intelligent news reader agent that can access online news based on the user's commands. Speech recognition (henceforth: SR) is less developed (than in English) in languages spoken either by rather few people [6] or in emerging countries; see, for example, [7]. In fact, much of the SR-related robotics research is focused on English, since it is spoken by almost all researchers, regardless of nationality, but perhaps not by all people in the general population, and certainly not all elderly people. Hence, improvement of SR systems is an important step towards the development of technological tools for elderly care in countries where English (or any other major language) is not the first language.

Of course, the problems and issues encountered when developing an SR system are similar, regardless of the language considered. In general, two main forms of SR can be distinguished, namely (i) isolated word recognition (IWR), and (ii) continuous speech recognition (CSR) [8]. Over the years, SR has been approached using many different techniques, in particular dynamic time warping (DTW) (see, e.g., [8], Chapter 4) which involves a classification method based on comparison of time series of different length, and hidden Markov models (HMMs) (see, e.g., [8], Chapter 6) that, effectively, constitute a stochastic (probabilistic) extension of DTW. The HMM approach currently dominates SR research, and most of the state-of-the-art SR systems employ this method. However, deterministic approaches (such as DTW) are also useful, particularly in command-style SR systems focusing on IWR with a rather limited vocabulary [9]. DTW attempts to find the optimal alignment between time series (containing, for example, features extracted from sound data) while, simultaneously, computing a distance measure between the two series. Thus, comparing the feature time series extracted from a given utterance to feature time series stored for template sounds, one can, based on the minimum distance found, determine which word was spoken. Even though DTW is frequently applied in deterministic SR systems, some recent work, further discussed in Section IV, has indicated that DTW does not, in fact, necessarily improve performance when matching time series of different length.

 ISBN: 978-1-61208-224-0

Thus, in this paper, a simpler approach (for Swedish SR) will be evaluated, in which sound features are first computed from partially time-overlapping frames extracted from a spoken word. The time series thus obtained are normalized to unit length, and are then resampled at equidistant (relative) times, making direct comparisons between different time series (obtained from different utterances) possible, without DTW. Several time series are generated for each sound, corresponding to different sound features, and the weighted Euclidean distance between the features from the reference sounds (recorded *a priori*) and the currently spoken word is then used as a measure of dissimilarity. Furthermore, the *selection* and weighting of the features used for SR have been optimized using a genetic algorithm (henceforth: GA).

The structure of the paper is as follows: The method is described in Section II. The results are given in Section III and are discussed in Section IV. The conclusions are presented in Section V.

II. METHOD

The SR method used here, which has been implemented in C# .NET, consists of four main parts: First, sounds are preprocessed and divided into short sound frames, using a fairly standard procedure. Next, features are extracted from the sound frames. Then, in SR, the features obtained are compared to features extracted from template words, in order to determine which word was spoken. The final part involves optimization of feature selection and feature weighting.

A. Sound preprocessing

The first preprocessing step is to subtract the mean from the sound samples (which, for 16-bit sounds, range from -32768 to 32767), i.e., removing any static (DC) components from the sound. The next step is to extract the word, assuming that a single word was spoken. This is done by first moving forward along the sound samples, starting from the μ^{th} sample, and forming a moving average involving (the modulus of) μ sound samples. Once this moving average exceeds a threshold t_p, the corresponding sample, with index k_s, is taken as the start of the word. The procedure is then repeated, starting with sample $\nu-\mu+1$, where ν is the number of recorded samples, forming the moving average as just described, and then moving backward, towards lower indices. When a sample (with index k_e) is found for which the moving average exceeds t_p, the end point has been found. The sound containing the $k_\text{e}-k_\text{s}+1 \equiv m$ samples is then extracted and is henceforth referred to as a *word*. The word is then pre-emphasized. In the time domain, the pre-emphasis filter takes the form

$$s_k \leftarrow s_k - cs_{k-1}, \tag{1}$$

where s_k denotes the k^{th} sample and c is a parameter with a typical value slightly below 1. As is evident from this equation, low frequencies (for which s_k is not very different from s_{k-1}) are de-emphasized, whereas high frequencies are emphasized, improving the signal-to-noise ratio.

The next step is to divide the word into short snippets, a procedure referred to as *frame blocking*. Here, snippets of duration τ are extracted, with consecutive snippets shifted by $\delta\tau$. Note that $\delta\tau$ is typically smaller than τ, so that adjacent frames partially overlap. Once the frames have been generated, each frame is subjected to *windowing*, a procedure aimed at reducing discontinuities at the beginning and end of each frame. Thus, for each frame, the n samples are modified as

$$s_k \leftarrow s_k v_k, \tag{2}$$

where the (Hamming) windowing function takes the form

$$v_k = (1-\alpha) - \alpha\cos\frac{2\pi k}{n}, \tag{3}$$

where α is yet another parameter.

B. Feature extraction

Once the word has been preprocessed as described above, resulting in a set of frames, sound features are computed for each frame. Here, the sound features used have been (i) the autocorrelation coefficients, (ii) the linear predictive coding (LPC) coefficients, (iii) the cepstral coefficients, and (iv) the relative number of zero crossings. The autocorrelation coefficients are defined as

$$a_i = \sum_{k=1}^{n-i} \frac{(s_k-\overline{s})(s_{k+i}-\overline{s})}{\sigma^2}, \tag{4}$$

where $\overline{s}$ is the average of the samples and σ^2 is their variance. The number of extracted autocorrelation coefficients (i.e., the number of values of i used, starting from $i=1$) is referred to as the autocorrelation order.

The LPC and cepstral coefficients [10], which both are representations of the spectral envelope of a sound frame, have been used frequently in speech recognition [8]. The LPC coefficients l_i provide the best possible linear approximation of the sound, i.e., an approximation ($\hat{s}_k$) of sample s_k

$$\hat{s}_k = \sum_{i=1}^{p} l_i s_{k-i}, \tag{5}$$

for which the error $s_k - \hat{s}_k$ is minimal in the least square sense, where p is the LPC order, i.e., the number of extracted LPC coefficients.

Provided that the sound frame is quasi-stationary, which is often (almost) the case if the frame duration is set to a suitable value, the LPC coefficients provide an accurate compressed representation of the sound frame. The LPC coefficients can be derived efficiently from the autocorrelation coefficients, a procedure that will not be detailed here (see, for example, [8]). Once the LPC coefficients have been obtained, one can also compute the cepstral coefficients. These coefficients can be derived from the LPC coefficients using (non-linear) recursion. The detailed procedure can be found in [11]. The number of cepstral coefficients extracted is referred to as the cepstral order. Finally, the number of zero crossings n_zc is computed as the number of samples such that either the product $s_k s_{k-1} < 0$ or $s_k s_{k-2} < 0$ (if $s_{k-1} = 0$). Then, the *relative number of*

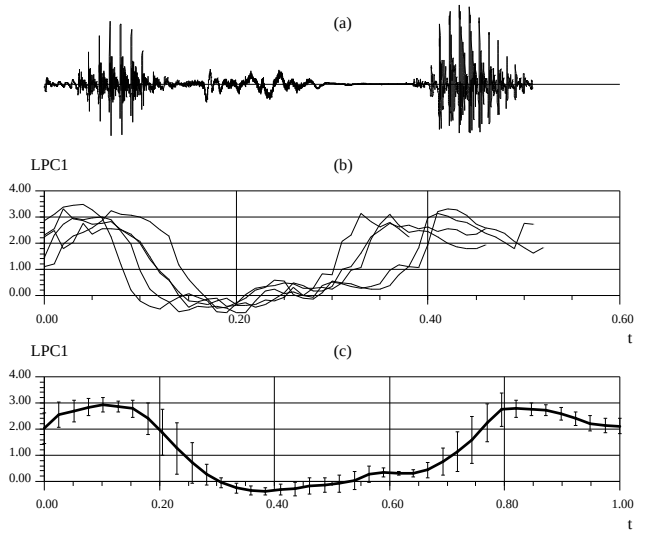


Fig. 1. An illustration of the averaging procedure (for a single feature, namely the first LPC coefficient, denoted LPC1) used when generating the template time series for a given word. Panel (a) shows one instance of the Swedish word *nästa* (meaning next). Panel (b) shows the original LPC1 time series obtained from five instances of the word *nästa*. As can be seen, the duration of the words, and therefore the number of feature points, varies a bit between instances. In panel (c), the time axes (for 10 instances of the word *nästa*) have been normalized, and the resulting series have been resampled, generating (in this case) 40 samples for each series, from which the average LPC1 time series, shown in panel (c), can be generated. This panel also shows the standard deviation (over the ten instances) for each sample.

zero crossings is formed by dividing $n_{\rm zc}$ by the number of samples (n) in the frame. With the procedure just described, several feature time series (with one set of features for each frame) are obtained for each word.

C. Speech recognition

Here, the SR (or, rather, word recognition) is based on a direct comparison between (a) feature time series stored for template words and (b) feature time series obtained from the spoken word. Thus, when generating the speech recognizer, a set of recorded template words is used. For each of the $n_{\rm w}$ words that the recognizer is supposed to cope with, $n_{\rm i}$ instances are recorded. The word instances are then preprocessed and feature time series are extracted, as described above.

Now, in order to generate a template time series for a given word, an average should be formed over the time series obtained for the $n_{\rm i}$ instances of the word in question. However, with a constant frame duration (τ) and a constant frame shift ($\Delta\tau$), the number of frames and, therefore, the number of elements in the time series, will depend on the duration of the spoken word instance. Thus, before forming the average, rather than using DTW to align the series, the time axes of all feature time series (for each instance) are simply linearly normalized to the range $[0, 1]$ and the resulting series are then re-sampled at equidistant (relative) times, resulting in an equal number of feature values for each feature time series and for each instance. Next, averages are formed over the $n_{\rm i}$ instances, which is straightforward since, after time normalization and re-sampling, all feature time series contain the same number of equidistant points. The average time series (one for each feature) are then stored in the speech recognizer. The process is repeated until all $n_{\rm w}$ words have been processed. The procedure is illustrated in Figure 1. For clarity, only five instances have been visualized in the middle panel, but the bottom panel is based on ten instances.

During speech recognition, the word to be recognized is subjected to the three steps of preprocessing, feature extraction, and resampling described above. Let F_{ijk} denote the $k^{\rm th}$ point of the $j^{\rm th}$ feature of stored word i (in the speech recognizer), and let φ_{jk} denote the $k^{\rm th}$ point of the $j^{\rm th}$ feature of the word to be recognized. The distance measure d_i is then computed as

$$d_i = \frac{1}{n_{\rm u} n_{\rm s}} \sum_{j=1}^{n_{\rm f}} w_j \sum_{k=1}^{n_{\rm s}} \left[\kappa_{jk} \left(F_{ijk} - \varphi_{jk}\right)\right]^2 . \qquad (6)$$

where $w_j \geq 0$ are feature weights. The inner sum (k) covers the number of time series points, or samples, (denoted $n_{\rm s}$) for each feature. The outer sum runs over the number of features ($n_{\rm f}$). $n_{\rm u}$ is the number of features for which w_j is different from 0. Thus, if $w_j > 0\ \forall j$, $n_{\rm u}$ is equal to $n_{\rm f}$. However, as shown below, faster SR performance can be obtained if some weights are set to 0. The κ_{jk} are scale factors (see below) that, in the basic distance measure (BDM), all take the value 1.

This distance measure is formed for each of the $n_{\rm w}$ stored words, and the index $i_{\rm r}$ of the word suggested by the speech recognizer is taken as

$$i_{\rm r} = {\rm argmin}_i d_i, \qquad (7)$$

if $d_{i_{\rm r}}$ is smaller than a threshold T. If not, the speech recognizer does not suggest any word. In order to simplify the comparison of results obtained in different runs, the threshold T was set to 1 for all runs. Note that, since the weights w_j are allowed to vary freely, in a wide range, it implies no restriction to set the threshold to a fixed value.

If the κ_{jk} are all equal to 1, all points along *a given feature* are weighted equally. However, as is evident from the bottom panel in Figure 1, the standard deviation (over the 10 different instances used when forming the average feature values) varies along the feature time series. For example, in that figure, one can see that feature points near the (normalized) time coordinate of around 0.20 have rather large standard deviation, whereas features points at time 0.60 have small standard deviation. One may thus argue that the latter points would perhaps be more useful in detecting the uttered word than feature points with larger standard deviation. Thus, a modified distance measure can be defined, henceforth referred to as the *standard deviation scaling* (SDS) distance measure[1], in which the κ_{jk} depend on the standard deviation for each feature

[1]Note that the approach introduced here is different from the *mean-variance normalization* (MVN) method [12] used in some SR systems. In MVN, the feature values are normalized using their estimated mean and variance over a sliding window. By contrast, in the approach considered here, the variance values over several stored instances are used for determining the κ_{jk} parameters.

 ISBN: 978-1-61208-224-0

point, from the stored words. For the SDS distance measure, the factors κ_{jk} are computed by first determining the values c_{jk} as

$$c_{jk} = \frac{1}{1 + \frac{\sigma_{jk}}{|\Phi_j|}} \tag{8}$$

where σ_{jk} is the standard deviation of feature point k along the time series for feature j, and $\overline{|\Phi_j|}$ is a normalization factor computed as the average modulus of the feature values along the series in question. The normalization factor is needed since different features have very different ranges. The modulus is introduced since many features have an average value near zero. Once the c_{jk} have been found, the κ_{jk} are computed as

$$\kappa_{jk} = n_{\rm s} \frac{c_{jk}}{\sum_{k=1}^{n_{\rm s}} c_{jk}}. \tag{9}$$

This slightly cumbersome procedure guarantees that the sum (over k) of κ_{jk} equals $n_{\rm s}$, making comparisons possible between runs that use the BDM (all κ_{jk} equal to 1) and runs that use the SDS distance measure.

D. optimization

Clearly, the distance measure depends on the weights w_j assigned for the different features. The simple choice of setting all weights to 1 is by no means optimal, since some features are more discriminative than others. Thus, an optimization procedure has been applied, during which the feature weights were optimized (for maximum SR performance). Note that the optimization did not involve the preprocessing, feature extraction, and resampling steps that, for a given word database, could thus be carried out once and for all.

For the purpose of optimization, a fairly standard GA [13] has been used, in which the feature weights are encoded in chromosomes (strings of floating-point numbers), and where the formation of new chromosomes takes place using standard tournament selection, single-point crossover, and creep mutations.

Three data sets were used in the GA runs: A training set, a validation set, and a test set. The results obtained over the training set were used as feedback to the GA, whereas the results obtained for the validation set, which were not provided to the GA, were used for determining when to stop the run in order to avoid overfitting. Once a run had been completed, the results obtained over the previously unused test set were taken as the true performance of the speech recognizer.

During the GA runs, the fitness Γ of an individual, i.e., a speech recognizer (with weights decoded from a chromosome), was computed as

$$\Gamma = \sum_{j=1}^{n_{\rm tr}} \gamma_j, \tag{10}$$

where the sum runs over all $n_{\rm tr}$ words in the training data set and γ_j is defined as

$$\gamma_j = \begin{cases} 1 + a(T-d) & \text{for correct identification,} \\ -b - a(T-d) & \text{for incorrect identification,} \\ a(T-d) & \text{if no word was recognized,} \end{cases} \tag{11}$$

TABLE I
THE PARAMETERS USED IN THE SPEECH RECOGNIZER.

Parameter	Value
Sound extraction threshold (t_p)	300
Sound extraction moving average length (μ)	10
Pre-emphasis parameter (c)	0.9373
Frame duration (τ)	0.030 (s)
Frame shift ($\Delta\tau$)	0.010 (s)
Hamming window parameter (α)	0.46
Autocorrelation order	8
LPC order	8
Cepstral order	12
Number of samples per feature ($n_{\rm s}$)	40

where d is the distance obtained for the word suggested by the speech recognizer, i.e., the distance d_i corresponding to index $i_{\rm r}$ in Equation (7). a and b are parameters, here set to 0.05 (a) and 0.50 (b). Thus, if $d \leq T$, a contribution (γ_j) slightly larger than 1 is given if the word was correctly identified, whereas a negative contribution is given if the wrong word was identified. Finally, if no word was recognized ($d > T$), a small negative contribution is given. With this fitness measure, the GA will attempt to find weights (see Equation (6)) that maximize the fraction of correctly identified words, as the prime objective, and also maximize the quantity $T - d$ for those words, as the second objective. Ideally, of course, $T - d$ should be as large as possible (for correctly identified words) since the risk of misidentification is then reduced for other instances of the word in question.

In some runs, described in Section III below, an effort was made to minimize the number of features used, by setting some weights to zero. In that case, the fitness Γ was multiplied by a penalty factor p defined as

$$p = 1 - \epsilon \frac{n_u}{n_f}, \tag{12}$$

where $\epsilon \ll 1$ is a positive constant. With this modification, the optimization procedure will favor speech recognizers using as few weights (and, therefore, features) as possible.

III. RESULTS

The three data sets (training, validation, and test) each contained 10 instances of 10 different words, i.e., a total of 100 sounds in each set. As the intended application is an interactive system for accessing online news (which will then be read by the agent or robot, typically as an aid to a visually impaired elderly person), the (minimal) vocabulary consisted of the ten Swedish words *ja* (yes) *nej* (no), *läs* (read), *åter* (return), *nästa* (next), *avsluta* (cancel or finish), *inrikes* (domestic (news)), *utrikes* (international), *ekonomi* (economy), and *sport* (sport, same as in English, but with different pronunciation). All sounds were sampled at 16 kHz.

After extensive testing, involving (short) GA runs for each parameter setting, the parameters used for preprocessing, feature extraction, and resampling were chosen as in Table I. As can be seen from the table, the total number of feature time series (for each word) was equal to $n_{\rm f} = 8 + 8 + 12 + 1 = 29$, including also the relative number of zero crossings.

 ISBN: 978-1-61208-224-0

TABLE II
OPTIMIZATION RESULTS: COLUMN 1 INDICATES THE RECOGNIZER TYPE, WHICH IS DETERMINED BY THE PARAMETERS SHOWN IN COLUMNS 2 AND 3. DM = DISTANCE MEASURE, BDM = BASIC DISTANCE MEASURE, SDS = STANDARD DEVIATION SCALING. COLUMNS 4 TO 6 SHOW THE FRACTION OF CORRECTLY RECOGNIZED WORDS FOR THE TRAINING, VALIDATION, AND TEST SETS, RESPECTIVELY. COLUMN 7 SHOWS THE AVERAGE DISTANCE VALUE (SEE EQUATION (6)) FOR THE CORRECTLY CLASSIFIED WORDS IN THE TEST SET.

Type	n_{u}	DM	Training	Validation	Test	$\overline{d}_{\min}$
1	29 (all)	BDM	1.00	0.99	0.99	0.3209
2	5	BDM	1.00	1.00	0.98	0.1466
3	29 (all)	SDS	0.98	0.91	0.91	—
4	8	SDS	0.98	0.93	0.93	—

Next, several long GA runs were carried out. The population size was set to 30, the tournament selection parameter to 0.75, the crossover probability to 0.80, and the mutation probability to $1/n_{\mathrm{p}}$, where n_p is the number of optimizable parameters (i.e., the weights w_j). In runs using the fitness measure from Equation (10), the weights w_j took values in the range $[0, 5]$. In runs using the fitness measure from Equation (12), the weight range was the same but, in addition, weights could be set to zero (exactly) with a probability of around 0.50.

The results of the best run (i.e., the run with highest validation fitness) for each of four different speech recognizer types are given in Table II. In runs with types 1 and 3, the standard fitness measure (without weight penalty) was used whereas for runs with types 2 and 4, the weight penalty was included in the fitness measure; see Equation (12). In types 1 and 2 the BDM was used, whereas types 3 and 4 used the SDS distance measure. Rather than the fitness values, the table shows the (more relevant) fraction of correctly recognized words for the training, validation, and test sets, respectively, for the best recognizer found of each type. In the rightmost column, the average values of $d_{\min} \equiv d_{i_r}$ (i.e., the distance measure for the recognized word) are shown for the correctly classified words for types 1 and 2, but not for types 3 and 4, since their test performance was too low for the average $d_{\min}$ values to be meaningful.

In the speech recognizer with highest validation fitness (type 2 in Table II), which reached a perfect result on the training and validation sets, and nearly perfect performance on the test set, the five features used were only cepstral coefficients, namely coefficients number 3,7,8,11, and 12. Even though the type 1 recognizer reached a slightly better result on the test set, one can argue that the type 2 recognizer is better, since it has a smaller average $d_{\min}$ value.

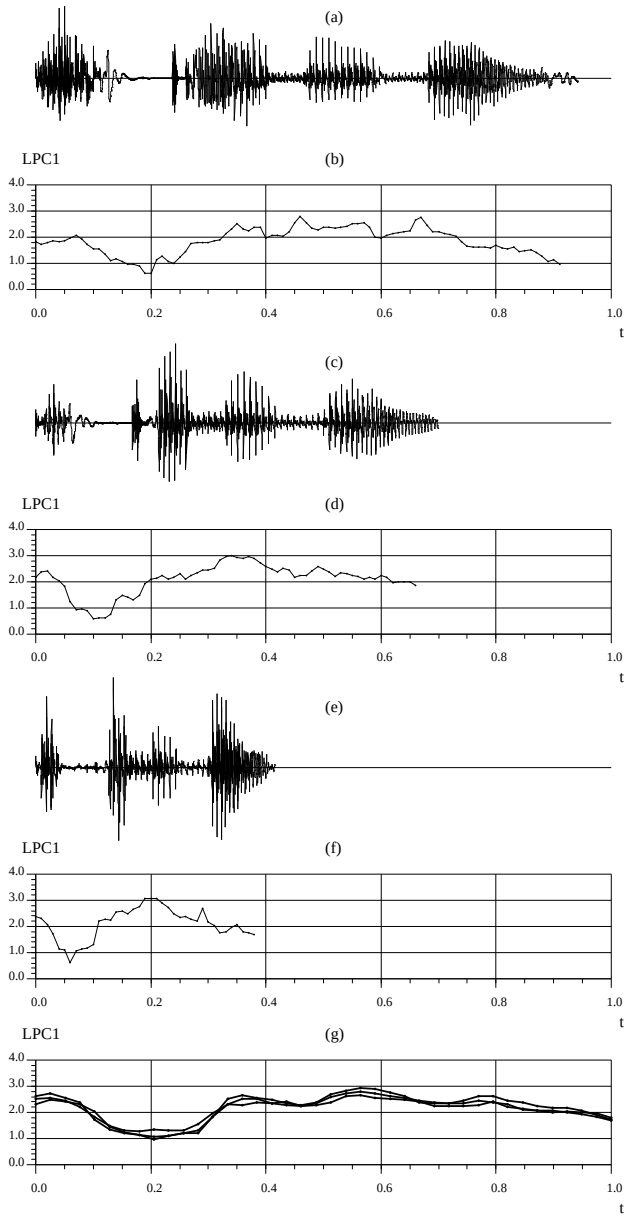


Fig. 2. An illustration of the fact that linear time normalization generates similar (average) feature time series, shown in Panel (g), regardless of the speaking speed. Panels (a)-(f) show the sound samples and the first LPC coefficient for three utterances of the word *ekonomi*, using slow, standard, and fast speaking, respectively. See the main text for a full description of the figure.

IV. DISCUSSION

The results in Table II show that the proposed method, with the basic distance measure (all κ_{jk} equal to 1), works well, and that some weights can be set to zero, without significant loss in performance. This is important, since the time needed to recognize a word may be crucial if larger vocabularies are used. The time needed for recognition consists of a part (the feature extraction) that, for a given set of features, depends only on the number of samples in the recognized word and a part (the computation of d_i) that is linear in the number of stored (recognizable) words (n_{w}). The first part is dominated by the time needed to compute autocorrelations, but can be somewhat reduced if only some LPC and cepstral coefficients are needed. In the setup used here (running on a 2.67 GHz core i7 processor) using a typical word size of around 10,000 samples, the constant part of the recognition time is around 0.0672 s, and the linear part is equal to $1.11 \times 10^{-4} n_{\mathrm{w}}$ s, if all weights are non-zero (type 1 in Table II) and $0.224 \times 10^{-4} n_{\mathrm{w}}$ s

 ISBN: 978-1-61208-224-0

for the best recognizer found, i.e., type 2 in Table II. Thus, if a maximum recognition time of, say, 0.20 s is allowed for real-time performance, the number of words that can be stored increases from around 1,200 (for type 1) to around 5,900 (for type 2).

The results also show that SDS had a detrimental effect on SR performance. A likely reason for the reduced performance is the fact that the standard deviation estimates are not very accurate since they are based on only 10 instances. Of course, the number of instances could be increased, but that would make the process of generating the speech recognizer quite time-consuming, at least for larger vocabularies than in this example. On the other hand, using SDS is not really needed, since the BDM reaches near-perfect performance.

As mentioned in Section I, direct comparison of time series in SR is usually carried out using DTW rather than (linear) time normalization followed by resampling as used here. However, the motivation for using the DTW approach is not very strong. First of all, Ratanamahatana and Keogh [14] have noted that time series of different length can simply be renormalized to equal length without loss in recognition performance. Furthermore, Boulgouris *et al.* [15] concluded that linear time normalization in fact *outperformed* DTW in the context of gait identification.

Our results confirm these findings, and a clear illustration of this fact is given in Figure 2. Here, the word *ekonomi* (economy) was uttered ten times slowly (average duration: 0.907 s), ten times with normal speed (average duration: 0.634 s), and ten times quickly (average duration: 0.433 s). The top six panels of the figure show one instance of the slow, normal, and fast utterances, along with the corresponding original time series for the first LPC coefficient (LPC1). The bottom panel shows the *average* feature time series over the ten instances of slow, normal, and fast readings, respectively, after linear time normalization and resampling. Despite the very different reading speeds, the feature values are very similar. In fact, the differences between the three curves are of the same order of magnitude as the standard deviations (not shown) over the ten instances for each curve separately. In all three cases (slow, normal, and fast), the best speech recognizers, namely types 1 and 2 in Table II, correctly identified all ten instances.

The work presented here represents the initial development stage of an intelligent agent with Swedish speech recognition, and there are some obvious limitations, namely that (i) a rather limited vocabulary was used, (ii) the sounds were recorded by a single speaker, and (iii) the system was trained using only the words in the vocabulary, i.e., no false positives were included. However, it should be noted that even with the present training setup, the speech recognizer can avoid incorrect detection of unknown words, namely in cases where the uttered word is sufficiently different from the stored words so that the resulting minimum distance exceeds the threshold T. As for speaker independence, even though the speech recognizer was trained using a single voice, some initial tests with other speakers showed promising results, which, however, will be followed up as described below.

V. Conclusion and further work

To conclude, it has been shown that linear time normalization followed by feature matching provides robust speech recognition, and that, with optimization, the recognition speed can be strongly improved (especially important for large vocabularies), by using only a subset of the available features.

Even though the size of the vocabulary used here is sufficient for the application at hand (i.e., an intelligent news reader agent; see Section I), and the intended use is as a companion to a single elderly person, the next step will be to increase the size of the vocabulary and, in doing so, use words spoken by different people. In addition, the issue of training with false positives present will be investigated. Another obvious topic for further work would be to extend the method in order to handle not only IWR but CSR as well.

References

[1] Eurostat news release 80/2011, "EU27 population is expected to peak by around 2040". Accessed Nov. 29, 2011. [Online]. Available: http://epp.eurostat.ec.europa.eu/cache/ITY_PUBLIC/3-08062011-BP/EN/3-08062011-BP-EN.PDF

[2] J. Broekens, M. Heerink, and H. Rosendal, "Assistive social robots in elderly care: a review," *Gerontechnology*, vol. 8, no. 2, pp. 94–103, 2009.

[3] M. Chan, D. Estève, C. Escriba, and E. Campo, "A review of smart homes - present state and future challenges," *Computer Methods and Programs in Biomedicine*, vol. 91, pp. 55–81, 2008.

[4] K. Wada and T. Shibata, "Living with seal robots - its sociopsychological and physiological influences on the elderly at a care house," *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 972–980, 2007.

[5] C. Granata, M. Chetouani, A. Tapus, P. Bidaud, and Dupourqué, "Voice and graphical-based interfaces for interaction with a robot dedicated to elderly and people with cognitive disorders," in *Proceedings of the 19th IEEE International Symposium on Robot and Human Interactive Communication*, 2010, pp. 785–790.

[6] A. Lipeika and J. Lipeikienė, "On the use of the formant features in the dynamic time warping based recognition of isolated words," *Informatica*, vol. 19, no. 2, pp. 213–226, 2008.

[7] F. Rosdi and R. Ainon, "Isolated malay speech recognition using hidden markov models," in *Proceedings of the International Conference on Computer and Communication Engineering*, 2008, pp. 721–725.

[8] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*. Prentice Hall, 1993.

[9] P. Wong, O. Au, J. Wong, and W. Lau, "Reducing computational complexity of dynamic time warping based isolated word recognition with time scale modification," in *Proceedings of the Fourth International Conference on Signal Processing*, 1998, pp. 722–725.

[10] J. Markel and A. Gray Jr., *Linear prediction of speech*. Springer Verlag, 1976.

[11] G. Antoniol, V. Rollo, and G. Venturi, "Linear predictive coding and cepstrum coefficients for mining time invariant information from software repositories," in *ACM SIGSOFT Software Engineering Notes*, vol. 30, no. 4, 2005, pp. 1–5.

[12] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, pp. 133–147, 1998.

[13] J. H. Holland, *Adaptation in natural and artificial systems*. University of Michigan Press, 1975.

[14] C. Ratanamahatana and E. Keogh, "Three myths about dynamics time warping data mining," in *Proceedings of SIAM International Conference on Data Mining*, 2005, pp. 506–510.

[15] N. Boulgouris, K. Plataniotis, and D. Hatzinakos, "Gait recognition using linear time normalization," *Pattern Recognition*, vol. 39, pp. 969–979, 2006.

 ISBN: 978-1-61208-224-0

Binding Data Mining to Final Business Users of Business Intelligence Systems

Ana Azevedo
CEISE/STI
ISCAP/IPP
Porto, Portugal
aazevedo@iscap.ipp.pt

Manuel Filipe Santos
Centro Algoritmi
University of Minho
Guimarães, Portugal
mfs@dsi.uminho.pt

***Abstract*— Since Lunh first used the term Business Intelligence (BI) in 1958, major transformations happened in the field of information systems and technologies, especially in the area of decision support systems. Nowadays, BI systems are widely used in organizations and their strategic importance is clearly recognized. The dissemination of data mining (DM) tools is increasing in the BI field, as well as the acknowledgement of the relevance of its usage in enterprise BI systems. One of the problems noted in the use of DM in the field of BI is related to the fact that DM models are, generally, too complex in order to be directly manipulated by business users; as opposite to other BI tools. The main contribution of this paper is a new DM language for BI conceived and implemented in the context of an Inductive Data Warehouse. The novelty is that this language is, by nature, user-friendly, iterative and interactive; it presents the same characteristics as the usual BI tools allowing business users to directly manipulate DM models and, allowing through this, the access to the potential value of these models with all the advantages that may arise.**

Keywords – Data mining; DM language; Business Intelligence; BI system; Inductive database; Inductive data warehouse; Business user.

I. INTRODUCTION

Organizations compete in environments whose complexity increases in a daily basis. Consequently there are many demands that organizations must answer in time and adequately in order to survive and gain competitive advantage in those complex environments. In this context, computerized Decision Support Systems (DSS), in particular Business Intelligence (BI) systems, play an important role in order to improve decision making and thus conducting organizations' actions. BI systems are gaining momentum each day in organizations and have a fundamental role in these issues [1][2].

The usage of Data Mining (DM) tools in BI is increasing. BI and DM, despite having roughly the same age, have different roots and as a consequence have significantly different characteristics [3][4]. DM came up from scientific environments, thus it is not business oriented. DM tools still demand heavy work in order to obtain the intended results, hence needing the knowledge of DM specialists to explore its full potential value. On the contrary, BI is rooted in industry and business, thus it is business oriented. As a result, BI tools are user-friendly and can easily be accessed and manipulated by business users.

From the literature review, it is evident that the majority of BI tools are directly manipulated by business users, allowing them to explore their potential value in a more effective way. The reason for this is related with the fact that BI tools are user-friendly, iterative, interactive, business oriented, and oriented to business activities. DM is an exception [5][6]. Despite its usage in BI systems is increasing day by day, DM models are not directly manipulated by business users who depend on reports from DM specialists. This way, business users could be unable to extract the potential business value contained in DM models. The complexity of DM models, as opposite to other BI tools, has been identified as the key factor for this.

The importance of allowing final business users to access and manipulate DM models comes up from the need of allowing business users to be more autonomous, without the permanent necessity to depend on the presence of a DM specialist. Moreover, considering that DM specialists do not usually have a complete knowledge of the business issues, making DM directly available to business users is the key element that allows obtaining all the potential business value that could be hidden in DM models. Hereby, the authors state that this can be done by means of a DM language developed, above all, to accomplish the necessities of final business users of BI systems. Consequently, it is considered in the research hereby presented, the importance of developing DM languages for BI, which are oriented to business users and, moreover, to BI activities.

Realizing the importance of the aspects mentioned above, the recognition of this reality establishes the foundations for this research. Accordingly, and based in the literature review, the research problem has been identified as: Final business users do not directly access and manipulate DM models and consequently their full potential business value could be not completely explored. The presented problem arises from the business needs existing in environments where BI systems include DM usage. Binding DM to final business users of BI systems thus inducting them into data mining models is considered a pertinent contribution.

From the literature review, it is given evidence of the necessity to develop tools for DM that present the same characteristics of BI tools, namely being user-friendly, interactive, iterative, oriented to business users, and oriented to BI activities, and thus could be directly manipulated by business users. This is also aligned with the roots of DM and Knowledge Discovery in Databases (KDD) as stated in [7]

 ISBN: 978-1-61208-224-0

where KDD is presented as an iterative and interactive process, with many decisions being made by the user.

The main contribution of this paper is a new DM language conceived and implemented in the context of an Inductive Data Warehouse (IDW). This new DM language will bind DM to final business users of BI systems, thus allowing them of being able to extract the potential business value hidden in DM models.

This paper presents the developed research and is organized as follows. It starts by presenting background and related work, in Section II. It follows with the research methodology and obtained results, in Section III. Next, in Section IV, limitations and future research design are brought in. The paper concludes in Section V.

II. BACKGROUND AND RELATED WORK

The term knowledge discovery in databases or KDD, for short, was coined in 1989 to refer to the broad process of finding knowledge in data, and to emphasize the "high-level" application of particular DM methods [8]. Fayyad considers DM as one of the phases of the KDD process. The DM phase concerns, mainly, the means by which the patterns are extracted and enumerated from data. As of the foundations of KDD and DM, several applications were developed in many diversified fields. The growth of the attention paid to the area emerged from the rising of big databases in an increasing and differentiated number of organizations. Nevertheless, there is the risk of wasting all the value and wealthy of information contained in these databases, unless the adequate techniques are used to extract useful knowledge [9][10][11]. The application of DM techniques with success can be found in a wide and diversified range of applications. One important application is in BI systems.

BI is the top level of a complex system. On its foundations lay several databases, usually based in the relational model for databases [12], that can be accessed and manipulated using specific database (DB) languages, such as SQL and Query-By-Example (QBE). On the next level, data warehouses (DW) can be manipulated using exactly the same sort of languages. Applying DM to data stored on both databases (DB) and data warehouses (DW), knowledge bases (KB) arise on the next level. KB store DM models and, traditionally, are not based on the relational model, unlike DB and DW. Nevertheless, using the framework of inductive databases (IDB), DM models can be stored in databases in the same way as data, thus DM models can be accessed and manipulated at the same level than data [13][14][15]. "Inductive databases tightly integrate databases with data mining. The key ideas are that data and patterns (or models) are handled in the same way, and that an inductive query language allows the user to query and manipulate the patterns (or models) of interest" [15, pp 69].

Using the framework of inductive databases, DM models can be obtained and manipulated through the use of DM languages, such as MineRule [16], DMQL [17], or MSQL [18]. Table I presents a comparison of the syntax of these SQL-based DM languages. The three languages are SQL extensions. The extensions are made through the implementation of a new operator that allows obtaining the DM models, namely "find classification rules" operator for DMQL, "MINE RULE" operator for Mine Rule, and "Get rules ... into ..." operator for MSQL.

TABLE I. COMPARISON OF SQL-BASED DM LANGUAGES SYNTAX [19]

Schema: student(id,gender,age,nenroll,grant,grade) Classification Rules for grade in consequent Having grade<10; support>0.1; confidence>0.2	
DMQL	use database school find classification rules as Classification Rules according to grade Related to gender, age, nenroll, grant From student Where student.grade<10 With support threshold > 0.1 With confidence threshold > 0.2
MineRule	MINE RULE ClassificationRules AS SELECT DISTINCT gender, age, nenroll, grant AS BODY, grade AS HEAD FROM student WHERE grade<10 EXTRACTING RULES WITH SUPPORT: 0.1, CONFIDENCE: 0.2
MSQL	GetRules (student) Into ClassificationRules Where consequent is {(grade<10)} and body in {(gender=*), (age=*), (nenroll=*), (grant=*)} and confidence > 0.2 and support > 0.1

These languages are very important. But, just like SQL, they are not business oriented, are not oriented to business users and are not oriented to BI activities. This is a crucial issue in organizations that is gaining momentum each day.

Codd's relational model for databases has been adopted long ago in organizations. Initially, two formal languages were defined for relational databases: relational algebra and relational calculus [20][12]. Since that time, several languages were developed in order that business users could access data stored in databases. Query-By-Example (QBE) languages [21] were developed with success. The use of QBE languages by business users in order to directly obtain answers to ad-hoc business questions is a usual practice in organizations nowadays. QBE languages are declarative, also called nonprocedural or very high level, languages. By using this type of languages the user defines "what s/he wants to do" instead of defining "how to do it", which is typical of imperative languages. According to Zloof, Query-by-Example is: "a high-level database management language that provides a convenient and unified style to query, update, define, and control a relational database. The philosophy of Query-by-Example is to require the user to know very little in order to get started and to minimize the number of concepts that s/he subsequently has to learn in order to understand and use the whole language." [22, pp 324]. QBE languages are business oriented; moreover they are oriented to business users and to BI activities.

 ISBN: 978-1-61208-224-0

III. RESEARCH METHODOLOGY AND OBTAINED RESULTS

In this research a BI system including DM was conceived and implemented. An architecture that allows an effective usage of DM with BI by business users in order to conduct to DM integration with BI, was envisaged. This architecture should bring DM into the front line business users, be iterative, visual, and understandable by front line business users, and work directly on data. Following these guidelines, an architecture for integration of DM with BI is presented in Figure 1. This architecture intends to conduct to an effective usage of DM in BI.

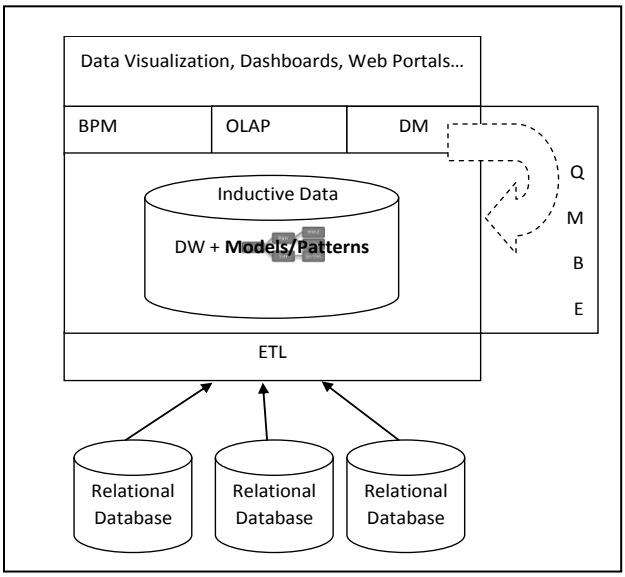


Figure 1. Architecture for integration of Data Mining with Business Intelligence.

The DM module extracts data from the DW, generates the DM models, and feeds the database with DM models. There is the possibility to include as many models as needed by the user, and new models can be included just by adding a new table.

This architecture implements the concept of Inductive Data Warehouse (IDW), which is a data warehouse storing data and data mining models at the same level, that is to say, both data and DM models are stored in data warehouse tables and can be accessed and manipulated in the same way.

An important aspect is the inductive language. Thus a new language, named as QMBE (Query Models By Example), was developed and implemented as an extension of a QBE language. Using QMBE the user is, thus, able to interact directly with the models, and to construct queries including different criteria. Table II presents several business questions commonly posed by business users involving DM models. All the business questions can be converted into queries to the system, defined in the QMBE language.

Since QMBE is an extension of QBE language, by nature it has two important characteristics, which are interactivity, and iterativeness. These characteristics are inherited from QBE languages upon which QMBE is extended. The novelty of the QMBE language is that it is oriented to business users and to BI processes. This kind of approach allows business users to directly access and manipulate data and models. This will bring DM to the front line business users, alike other BI tools, thus allowing DM integration with BI.

TABLE II. BUSINESS QUESTIONS INVOLVING DM MODELS

Queries on models	Queries on models and data
What are the characteristics of "good" students?	Select the actual students that can be "good" students.
What are the characteristics of "bad" students?	Select the actual students that can be "bad" students.
What are the characteristics of the students that do not conclude the grades according to initial schedule?	Select the actual students that cannot conclude the grades according to initial schedule.
Are there different types of students in the school?	…
…	…

Following, the concept of IDW, and QMBE language are presented.

A. Inductive Data Warehouse

In the context of BI there can be said that an IDB contains both the DW and the KB, that is to say, the DM models. This way, we can refer to this database as an Inductive Data Warehouse (IDW). Thus, an IDW is a DW which includes data and DM models, both stored in tables of the DW. This is an important concept in the realm of this research, since it focuses on making DM available to business users. In an IDW data and DM models can be accessed by business users in the same way as data. The DM models are stored in the DW in specific tables: the model tables. It is possible to include several model tables, one for each generated model.

In this research, the generated DM model corresponds to rules, since these were considered adequate for the problem under study. A rule is an IF-THEN expression of the form *IF antecedent THEN consequent*, written as:

$$antecedent => consequent$$

where antecedent and consequent are propositions of the form

$$V_1\ cond_1\ C_1\ \ AND\ \ldots\ AND\ V_N\ cond_N\ C_N$$

where $V_1, \ldots, V_N$ are variables; $C_1, \ldots, C_N$ are constants; and $cond_1, \ldots, cond_N$ stands for < or > or = or <= or >=.

In the case of classifications rules, the consequent is of the form:

$$V_i\ cond_i\ C_i$$

where V_i is the target variable; C_i is a constant; and $cond_i$ stands for < or > or = or <= or >=.

Usually BI systems are supported by special databases, namely DW. For the sake of generality, consider a DW with one fact table named FACT_TABLE, and N dimension tables named DIMENSION_1, DIMENSION_2, DIMENSION_3, ..., DIMENSION_N. The fact table has one ID column, and N columns, Dimension1, Dimension2, Dimension3, ..., DimensionN, each corresponding to one dimension table, and a column Fact. Each of the dimension

tables has got several columns, each one corresponding to a variable that can be selected for DM. Consider for instance that DIMENSION_J has M_j variables, namely, IDJ, VarJ1, VarJ2, .., VarJI, ..., $VarJM_J$.

In an IDW, DM models are stored in the database in one, or more, specific table, or tables. Without losing generality, hereby only one table will be considered and named MODEL_TABLE. The first column of the model table, ID, corresponds to the rule identifier. The next two columns, confidence and support, stand respectively for the rule confidence and for the rule support. The following column corresponds to the selected DM target variable that corresponds to one of the columns of one of the dimension tables. The L variables selected for data mining, each one corresponding to a column of one of the dimension tables included in the DW, form the rest of the table columns, namely, DMVar1, DMVar2, ..., DMVarL. Keep in mind that DMVar1, DMVar2, … DMVarL of MODEL_TABLE are selected from all the columns of tables DIMENSION_1, or DIMENSION_2, …, or DIMENSION_N. Thus, all the columns of the MODEL_TABLE are the same as some column of the dimension tables. In this manner MODEL TABLE is connected to the DW tables.The IDW general schema is presented in Figure 2.

MODEL_TABLE (ID, confidence, support, DMTarget, DMVar1, DMVar2, …, DMVarL)
FACT_TABLE (ID, Dimension1, Dimension2, ..., DimensionN, Fact)
DIMENSION_1 (ID1, Var11, Var12, …, Var1I, …, $Var1M_1$)
DIMENSION_2 (ID2, Var21, Var22, …, Var2I, …, $Var2M_2$)
DIMENSION_3 (ID3, Var31, Var32, …, Var3I, …, $Var3M_3$)
…
DIMENSION_J (IDJ, VarJ1, VarJ2, …, VarJI, …, $VarJM_J$)
…
DIMENSION_N (IDN, VarN1, VarN2, …, VarNI, …, $VarNM_N$)

Figure 2. IDW General Schema.

Each rule is introduced in the MODEL_TABLE as a line of the table. Data is introduced in a cell of the table whenever there is a constraint in the rule for the correspondent variable, and is left blank (NULL) elsewhere. Consider, for instance, a general rule:

Rule I:

DMVar1 cond1 Value1 AND ... AND DMVarK condK ValueK AND ... => DMTarget condT ValueT; where cond1, ..., condK, condT stands for < or > or = or <= or >=.

Then the line (tuple) that corresponds to that rule is:

(I, valueC, ValueS, condT ValueT, cond1 Value1, ..., condK ValueK, ...).

New models can easily be added to the IDW by the simple introduction of model tables in the IDW, one for each model.

B. QMBE Language

In the research described in this paper, a new language, named Query Models by Example (QMBE) was developed as an extension of QBE languages existing in some Relational Database Management Systems (RDBMS). Similarly to QBE languages, upon which QMBE is based on, QMBE is a declarative, also called nonprocedural or very high level, language. By using this type of languages the user defines "what s/he wants to do" instead of defining "what to do", which is typical of imperative languages.

Business users are able to interact directly with the models, and to construct queries as a way to obtain answers to ad-hoc business questions. Business questions can be converted into queries to the system, defined in the QMBE language. Like in RDBMS QBE languages, the user will be able to define different criteria, considered significant to business. Business questions can be converted into queries in the QMBE language. To construct the query, the user will have to fill in a skeleton table (Figure 3).

Table →				
Column →				
Criteria →				

Figure 3. Skeleton table for the QMBE language.

The user will have to identify which are the tables, in the first line of the skeleton table; the corresponding columns that have the necessary data to answer the intended business question will have to be identified in the second line of the skeleton table. Specific criteria can be defined for each selected column, in the next lines of the skeleton table. More than one line can be considered for criteria. If criteria are defined in the same line, they are linked with AND. If criteria are defined in different lines, they are linked with OR. There can be considered three types of QMBE queries, namely:

- queries on data, corresponding to traditional QBE languages;
- queries on models, corresponding to QMBE extensions; and
- queries on models and data, corresponding also to QMBE extensions.

In all these three cases, examples of business questions will be presented based on the IDW schema from Figure 2. There will also be presented the correspondent queries in QMBE, as well as the relational calculus sentences that correspond to each one of those QMBE queries. Relational calculus is based in a branch of mathematical logic called predicate calculus [20]. QBE languages are connected with relational calculus and so is QMBE. Just like for traditional QBE queries, all QMBE queries can be written as relational calculus queries.

 ISBN: 978-1-61208-224-0

TRADITIONAL QBE: QUERIES ON DATA

Generally speaking, queries on data involve columns from any of the tables of the IDW, except the MODEL_TABLE, for instance DIMENSION_J and FACT_TABLE. Similarly, criteria can be defined for any column.

Business Question I

What are the data from Dimension J table, which corresponds to Fact (of FACT TABLE) equal to a certain value (value)?

QMBE query I

The query is presented in Figure 4.

Table →	FACT_TABLE	DIMENSION_J	…	DIMENSION_J
Column →	Fact	Var11	…	Var1N
Criteria →	value			

Figure 4. QMBE query I.

Relational Calculus Query 1

Q1 = {f.Fact, d | FACT_TABLE(f) AND DIMENSION.J(d) AND f.Fact = value}

QMBE EXTENSIONS: QUERIES ON MODELS

Generally speaking, queries on models may involve any of the columns of the MODEL_TABLE and criteria can be defined for any column.

Business Question J

What are the rules of MODEL TABLE which correspond to DM Target equal to a certain value (value)?

QMBE query J

The query is presented in Figure 5.

Table →	MODEL_TABLE	MODEL_TABLE	…	MODEL_TABLE
Column →	DMTarget	DMVar1	…	DMVarL
Criteria →	value			

Figure 5. QMBE query J.

Relational Calculus Query J:

QJ = {m | MODEL TABLE(m) AND m.DMTarget=value}

QMBE EXTENSIONS: QUERIES ON MODELS AND DATA

Queries on models and data may involve columns from all the tables of the IDW and criteria can be defined for any column.

Business Question K

What are the data from DIMENSION J which corresponds to a pre-selected rule from MODEL TABLE, for instance, rule I?

QMBE query K:

The query is presented in Figure 6.

Table →	DIMENSION_J	…	DIMENSION_J	…	DIMENSION_J	...
Column →	VarJ1		$VarJI_1$		$VarJI_K$	
Criteria →			cond1 Value1		condK ValueK	

Figure 6. QMBE query J.

Relational calculus Query K:

QK = {dJ | DIMENSION_J(dJ) AND VarJI1 cond1 value1 AND … AND VarJIK condK ValueK AND …}

IV. DISCUSSION, LIMITATIONS, AND FUTURE RESEARCH DIRECTIONS

As a consequence of being an extension of a QBE language, this new DM language is iterative and interactive in nature. It allows business users to answer to ad-hoc business questions through queries on data or/and on DM models. QMBE allows business users to directly access and manipulate DM models. The novelty of the QMBE language is that it is oriented to business users and to BI activities. This kind of approach allows business users to directly access and manipulate data and models. This will bring DM to the front line business users, like other BI tools, allowing them to completely exploring DM potential value.

The presented architecture is being implemented as a prototype. One limitation is that, at the moment, the system is not completely automated. Another limitation is that only rules are addressed at the moment. Nevertheless, rules will be followed by clustering. This is due to the application domain, which focuses on these two DM tasks.

User interface is also a concern.

The architecture, including IDW and QMBE language, was implemented as a prototype and used in different and controlled situations, proving that the concepts are viable and can be applied in the considered environments, of BI systems using DM. Only this conceptual evaluation has been made at the moment, but a questionnaire is planned in order to obtain business users opinions.

It is expected that when tests are finished the system will be integrated in a real situation. The authors hope that the implementation on a real situation could help to bring new useful insights.

V. CONCLUSION

The goal of the research presented in this paper is to allow business users to manipulate directly DM models, thus being able to explore completely their potential business value. This is achieved by means of the use of the IDB framework in the area of BI, presenting the concept of IDW. Also, a new data mining language for BI, named as QMBE,

 ISBN: 978-1-61208-224-0

which is oriented to BI activities as well as oriented to business users, was developed. QMBE is presented as an extension of traditional QBE languages, which are included in most of the RDBMS nowadays.

The authors have introduced a BI systems' architecture that allows final business users to directly access and manipulate DM models and consequently being able of extracting their full potential business value. Consequently, the business value contained in DM models could be completely explored in BI systems that incorporate DM tools. This was achieved through means of two new important concepts: the concept of Inductive Data Warehouse and a new DM language, QMBE, which is iterative, and interactive in nature. By using this language, business questions can be converted into queries in the QMBE language, thus it is oriented to BI activities and to BI business users. This will allow business users to directly manipulate DM models, as well as data, thus bringing DM into the front line business personnel, allowing to increase DM potential to attaining BI's high potential business value.

This new DM language is extensible and flexible, since several DM models could be added just by adding new model tables to the IDW, and it is context independent, since it can be applied to any DW.

The main contribution of this paper is to verify the viability of allowing business users of BI systems to directly manipulate DM models and thus providing the possibility of exploring the potential value of applying DM in the context of BI.

REFERENCES

[1] E. Turban, R. Sharda, J. E. Aroson, D. King, Business Intelligence: A Managerial Approach, Pearson Prentice Hall, Upper Sadle River, NJ, 2008.

[2] E. Turban, J. E. Aroson, T. Liang, R. Sharda, Decision Support and Business Intelligence Systems, Pearson Prentice Hall, Upper Sadle River, NJ, 2007.

[3] G. Piatetsky-Shapiro, Data Mining and Knowledge Discovery 1996 to 2005: Overcoming the Hype and Moving from "university" to "business" and "analytics", Data Mining and Knowledge Discovery 15(1), 2007, pp: 99-105.

[4] H. Kriegel, K. M. Borgwardt, P. Kröger, A. Pryakhin, M. Schubert, A. Zimek, Future Trends in Data Mining, Data Mining and Knowledge Discovery 15(1), 2007, pp: 87-97.

[5] W. McKnight, Bringing Data Mining to the Front Line, Part 2, Information Management magazine November(2002), 2003, pp: Retrieved on July, 16th 2009 at http://www.information-management.com/issues/20021101/5980-1.html.

[6] W. McKnight, Bringing Data Mining to the Front Line, Part 1, Information Management magazine November(2002), 2002, pp: Retrieved on July, 16th 2009 at http://www.information-management.com/issues/20021101/5980-1.html.

[7] R. J. Brachman, T. Anand, The Process of Knowledge Discovery in databases, In U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining, AAAI Press/The MIT Press, Menlo Park, CA, 1996, pp: 37-57.

[8] U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth, From data mining to knowledge discovery: an overview, In U. M. Fayyad, G. Piatetski-Shapiro, P. Smyth, R. Uthurusamy (Eds.), Advances in knowledge discovery and data mining, AAAI Press/The MIT Press, Menlo Park, California, 1996, pp: 1-34.

[9] E. Simoudis, Reality Check for Data Mining, IEEE Expert 11(5), 1996, pp: 26-33.

[10] U. M. Fayyad, Data Mining and Knowledge Discovery: Making Sense Out of Data, IEEE Expert 11(5), 1996, pp: 20-25.

[11] M. Chen, J. Han, P. S. Yu, Data Mining: An Overview from a Database Perspective, IEEE transactions on Knowledge and Data Engineering 8(6), 1996, pp: 866-883.

[12] E. F. Codd, A Relational Model of Data for Large Shared Data Banks, Communications of the ACM 13(6), 1970, pp: 377-387.

[13] T. Imielinski, H. Mannila, A Database Perspective on Knowledge Discovery, Communications of the ACM 39(11), 1996, pp: 58-64.

[14] S. Dzeroski, Towards a General Framework for Data Mining, in S. Dzeroski, J. Struyf (Eds.), Knowledge Discovery in Inductive Databases - 5th International Workshop, KDID 2006, Lecture Notes in Computer Science: Vol. 4747, Springer-Verlag, Berlin, Heidelberg, 2007, pp: 259-300.

[15] L. De Raedt, A perspective on Inductive Databases, SIGKDD Explorations 4(2), 2003, pp: 69-77.

[16] R. Meo, G. Psaila, S. Ceri, An Extension to SQL for Mining Association Rules, Data Mining and Knowledge Discovery 2(2), 1998, pp: 195-224.

[17] J. Han, Y. Fu, W. Wang, K. Koperski, O. Zaiane, DMQL: A Data Mining Query Language for Relational Databases, (), 1996, pp: 27-34.

[18] T. Imielinski, A. Virmani, MSQL: A Query Language for Database Mining, Data Mining and Knowledge Discovery 3(4), 1999, pp: 373-408.

[19] A. Azevedo, M. F. Santos, A Perspective on Data Mining Integration with Business Intelligence, In A. Kumar (Ed.), Knowledge Discovery Practices and Emerging Applications of Data Mining: Trends and New Domains, IGI Publishing, Hershey, NY, 2011, pp: 109-129.

[20] E. F. Codd, A Data Base Sublanguage Founded on the Relational Calculus, (), 1971, pp: 35-68.

[21] M. M. Zloof, Query-by-Example: a data base language, IBM Systems Journal 16(4), 1977, pp: 324-343.

[22] M. M. Zloof, S. P. de Jong, The System for Business Automation (SBA): Programming Language, Communications of the ACM 20(6), 1977, pp: 385-396.

 ISBN: 978-1-61208-224-0

Group Recommendation System for User-Centric Support in Virtual Logistic Hub

Architecture and Major Components

Alexander Smirnov, Nikolay Shilov
St. Petersburg Institute for Informatics and Automation
of the Russian Academy of Sciences
St. Petersburg, Russia
{smir; nick}@iias.spb.su

***Abstract*—Currently, the society becomes more and more knowledge-intensive when a level of collaboration of different groups of people and institutes increases dramatically. One of the possible ways to assist to a group of users is collaborative recommendation systems. These systems have to recommend some solutions (related to products, technologies, tools, material and business models) based on user group requirements, preferences and willingness to compromise and to be pro-active. The paper proposes an approach to developing a group recommendation system for virtual logistic hub based on such technologies as user and group profiling, context management, decision mining. The system allows accumulation of knowledge about user actions and decisions and compromising between group and individual preferences. Proposed approach enables formulation of recommendations for users of the same group anticipating their possible further actions and decisions.**

Keywords-collaborative recommendation system; group profiling; context management; decision mining.

I. INTRODUCTION

Small and Medium businesses (SMEs) and personal travel via cars, buses and trains is usually (and reasonably) done within the radius of 450-500 kilometers. The distance between St. Petersburg, Russia and Helsinki, Finland together with nearby cities (Imatra, Lappeenranta, Kotka, Vyborg) falls into this radius. Taking into account available airports in Helsinki, Lappeenranta, and St. Petersburg as well as ferries in Helsinki, Kotka, and St. Petersburg, this region constitutes a universal hub for travelling all around the world.

In order for this hub to function, an efficient transportation system within the region has to be formed. However, today the travelling in the region is complicated due to a number of reasons, e.g., unpredictable situation at border crossing, unknown traffic condition on the roads, isolation of train, bus, and airplane schedules. The proposed approach is aimed at support of dynamic configuration of virtual multimodal logistics networks based on user requirements and preferences. The main idea is to develop models and methods that would enable ad-hoc configuration of resources for multimodal logistics. They are planned to be based on dynamic optimization of the route and transportation means as well as to take into account user preferences together with unexpected and unexpressed needs (on the basis of the profiling technology).

The small business and personal travelling is characterized by the following features: non-regular, not expensive, and safe. As a result, the proposed approach assumes developing a group recommendation system for ad hoc generation of travel plans for the region (the South of Finland and St. Petersburg region) taking into account the current situation on the roads and border crossings, fuel management aspects, travel time and distance. The increase of travelling will be a significant step towards development of the integrated economic zone in the Region.

Until recently, the most recommendation systems operated in the 2-dimentional space “user-product”. They did not take into account the context information, which, in most applications can be critical. As a result, there was a need in development of group recommendation systems based not only on previously made decisions but also on the contexts of situations in which the decisions were made. This gave a rise to development of context-driven collaborative algorithms of recommendation generation since their usage would significantly increase the quality and speed of decision making.

Besides, the proposed general framework will be a channel for collecting user’s feedback, preferences and demands for new services that users cannot find in the Region or quality of which shall be improved. What is important is that not only the problem is identified, but in most cases immediate hints/suggestions can be provided regarding what shall be done to better serve users’ needs.

The framework will also significantly benefit to the ecological situation in the region via reducing not necessary transportation and waiting time for border crossing. In accordance with Global GHG Abatement Cost Curve v 2.0 [1] in the travelling sector the carbon emission can be significantly decreased via more efficient route planning, driving less, switching from car to rail, bus, cycle, etc. As a result, evolving of flexible energy and eco-efficient logistics systems can be considered as one of the significant steps towards the knowledge-based low carbon economy.

The paper is structured as follows. The next section introduces the virtual logistic hub. It is followed by the description of the approach. Then, the group recommendation system architecture is proposed. The knowledge representation formalism used in the developed

 ISBN: 978-1-61208-224-0

approach is presented in sec. V. Sec. VI presents the user clustering algorithm, followed by the description of how the common preferences/interests are identified (sec. VII). The main results are summarized in the Conclusion.

II. VIRTUAL LOGISTIC HUB

The idea of virtual logistic hub has already been mentioned in the literature (though it could have a different name, e.g., "e-Hub" [2]), but it is still devoted very little attention in the research community. For example, [3] and [4] consider the virtual logistic hub from organizational and political points of view. Generally, virtual logistic hub represents a virtual collaboration space for two types of members: (i) transportation providers (who actually moves the passengers or cargo), and (ii) service providers (who provides additional services, e.g., sea port, border crossing authorities, etc.). These providers can potentially collaborate in order to increase the efficiency of the logistic network (solid lines in Figure 1), however, it is not always the case. The major idea of the virtual logistic hub is to arrange transportation based on the available schedules and capabilities of transportation and service providers, current and foreseen availability and occupancy of the transportation means and services ("dash-dot" lines in Figure 1). In this case, even though the schedules and actions of different members are not coordinated, the virtual logistic hub will be able to find the most feasible transportation schedule depending on the current situation and its likely future development. For the end-user (travelers or cargo owners), all this is hidden "under the hood", and only the final transportation schedule is seen (solid lines in Figure 1).

III. APPROACH

Figure 2 represents the generic scheme of the approach. The main idea of the approach is to represent the logistics system members by sets of services provided by them. This makes it possible to replace the configuration of the logistics system with that of distributed services. For the purpose of semantic interoperability, the services are represented by Web-services using the common notation described by a common ontology. The agreement between the resources and the ontology is expressed through alignment of the descriptions of the services modeling the resource functionalities and the ontology. As a result of the alignment operation the services get provided with semantics. The operation of the alignment is supported by a tool that identifies semantically similar words in the Web-service descriptions and the ontology. In the proposed approach the formalism of Object-Oriented Constraint Networks (OOCN) is used (its detailed description can be found in [20]) for knowledge representation in the ontology (see sec. V).

Depending on the problem considered, the relevant part of the ontology is selected forming an abstract context. The abstract context is an ontology-based model embedding the specification of problems to be solved. It is created by core services incorporated in the environment. When the abstract context is filled with values from the sources, an operational context (formalized description of the current situation) is built. The operational context is an instantiated abstract context and the real-time picture of the current situation. Producing the operational context is one of the purposes of resource configuration. Since the resources are represented by sets of services, the configuration of the resources is replaced with that between the appropriate services. Besides the operational context producing, the services are purposed

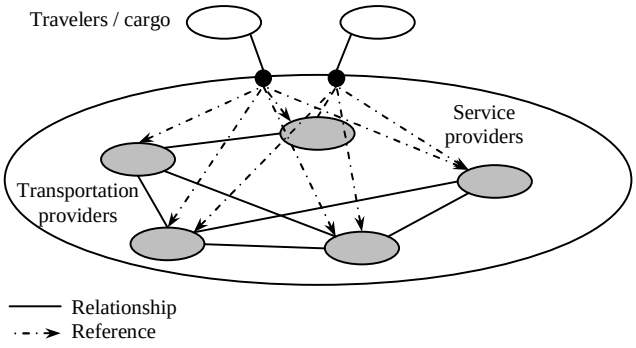


Figure 1. Generic scheme of the virtual logistic hub

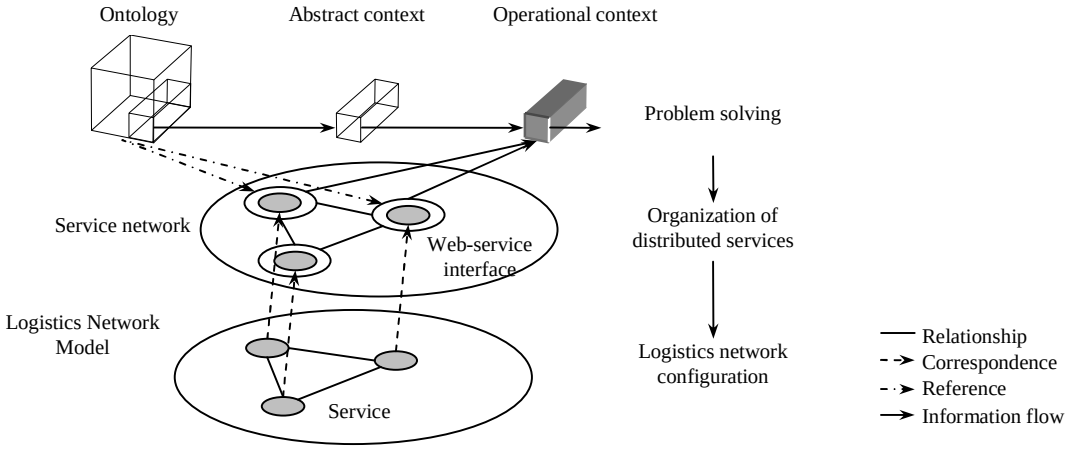


Figure 2. Generic scheme of the approach

 ISBN: 978-1-61208-224-0

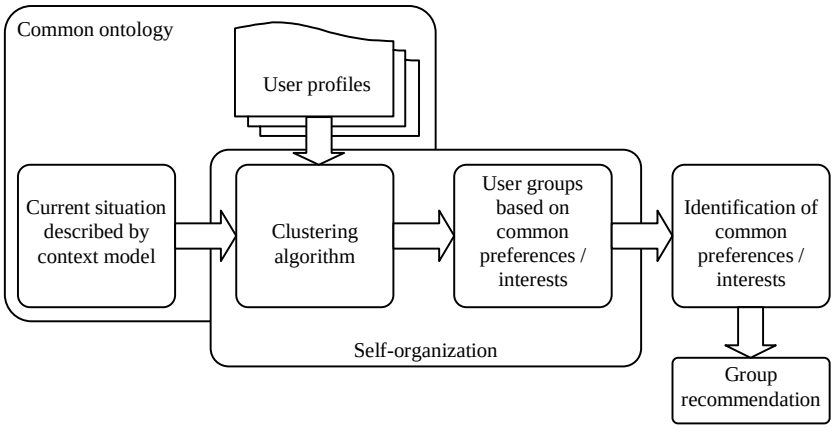


Figure 3. Group recommendation system architecture

to solve problems specified in the abstract context and to get the resources to take part in logistics plan. Due to the usage of the OOCN formalism the operational context represents the constraint satisfaction problem that is used during organisation of services for a particular task.

It can be guessed that for each particular situation there can be a large amount of feasible solutions for the users to choose from (e.g., the fastest transportation, the least amount of transfers, sightseeing routes, etc.). As a result, the paper proposes to build such a system as a group recommendation system that learns user preferences and recommends solutions, which better meet those preferences.

IV. GROUP RECOMMENDATION SYSTEM ARCHITECURE

Generation of feasible transportation plans taking account explicit and tacit preferences requires strong IT-based support of decision making so that the preferences from multiple users could be taken into account satisfying both the individual and the group [5]. Group recommendation systems are aimed to solve this problem.

Recommendation / recommending / recommender systems have been widely used in the Internet for suggesting products, activities, etc. for a single user considering his/her interests and tastes [6], in various business applications (e.g., [7], [8]) as well as in product development (e.g., [9], [10]). Group recommendation is complicated by the necessity to take into account not only personal interests but to compromise between the group interests and interests of the individuals of this group.

There are two major types of recommending systems: (i) content-based (recommendations are based on previous user choices), and (ii) collaborative filtering (recommendations are based on previous choices of users with similar interests). The second type is preferable for the domains with larger amounts of users and smaller activity histories of each user, which is the case for the logistics hub.

In literature (e.g., [11], [12]) the architecture of the collaborative filtering recommending system is proposed based on three components: (i) profile feature extraction from individual profiles, (ii) classification engine for user clustering based on their preferences (e.g., [13]), and (iii) final recommendation based on the generated groups. The core of such system is a clustering algorithms capable to continuously improve group structure based on incoming information enables for self-organization of groups [14].

The proposed group recommendation system architecture for logistics hub is presented in Figure 3. It is centralized around the user clustering algorithm [15] originating from the decision mining area [16]-[18]. The proposed clustering algorithm is based on the information from user profiles. The user profiles contain information about users including their preferences, interests and activity history (a detailed description of the profile can be found in [19]). Besides, in order for the clustering algorithm to be more precise, this information is supplied in the context of the current situation (including current user task, time pressure and other parameters). The semantic interoperability between the profile and the context is supported by the common ontology.

The user profiles are considered to be dynamic and, hence, the updated information is supplied to the algorithm from time to time. As a result, the algorithm can run as updated information is received and update user groups. Hence, it can be said that the groups self-organize in accordance with the changes in the user profiles and context information.

When groups are generated the common preferences / interests (e.g., the fastest transportation, the least amount of transfers, sightseeing routes, etc.) of the groups are identified based on the results of the clustering algorithm. These preferences are then generalized and analyzed in order to produce group recommendations.

Usage of an appropriate knowledge representation formalism is one of the keys to development of an efficient clustering algorithm.

V. KNOWLEDGE REPRESENTATION FORMALISM

Since the user profiles and the current situation context are analyzed jointly, it is reasonable to use the same formalism and terminology for their representation. In the proposed approach the formalism of Object-Oriented Constraint Networks (OOCN) is used (its detailed

 ISBN: 978-1-61208-224-0

description can be found in [20]) for knowledge representation in the ontology. It provides primitives for modelling classes, class hierarchies and other class structures, class attributes, attribute inheritance, attribute ranges, and functional dependencies.

According to this formalism the ontology A is represented by sets of classes, class attributes, attribute domains, and constraints:

$A = <O, Q, D, C>$, where

O – a set of *object classes* ("classes")

Q – a set of class *attributes* ("attributes");

D – a set of attribute *domains* ("domains");

C – a set of *constraints* used to model relationships.

The set of constraints includes six types of constraints for modelling different relationships:

C_1 – (class, attribute, domain) relation used to model triple of classes, attributes pertinent to them, and restrictions on the attribute value ranges;

C_2 – taxonomical ("is-a") and hierarchical ("part-of") relations used to model class taxonomy and class hierarchy respectively;

C_3 – classes compatibility used to model condition if two or more instances can be parts of the same class;

C_4 – associative relationships used to model any relations and axioms of external ontologies neglected by the internal formalism;

C_5 – class cardinality restriction used to define how many subclasses the class can have;

C_6 – functional relations used to model functions and equations.

Such representation of knowledge can be interpreted as a constraint satisfaction task and used by a constraint satisfaction / propagation engines for reasoning and optimization.

Below, some example constraints are given:

- an attribute *costs* (q_1) belongs to a class *ride* (o_1): $c^{I}_1 = (o_1, q_1)$;
- the attribute *costs* (q_1) belonging to the class *ride* (o_1) is a real number: $c^{II}_1 = (o_1, q_1, \mathrm{R})$;
- a class *cargo* (o_2) is compatible with a class *truck* (o_3): $c^{III}_1 = (\{o_2, o_3\}, True)$;
- an instance of the class *ride* (o_1) can be a part of an instance of a class *travel* (o_4): $c^{IV}_1 = <o_1, o_4, 1>$;
- the *truck* (o_3) is a *resource* (o_5): $c^{\mathrm{IV}}_1 = <o_3, o_5, 0>$;
- an instance of the class *cargo* (o_2) can be connected to an instance of the class *truck* (o_3): $c^{V}_1 = (o_2, o_3)$;
- the value of the attribute *cost* (q_1) of an instance of the class *travel* (o_4) depends on the values of the attribute *cost* (q_1) of instances of the class *ride* (o_1) connected to that instance of the class *travel* and on the number of such instances: $c^{VI}_1 = f(\{o_1\}, \{(o_4, q_1), (o_1, q_1)\})$.

VI. USER CLUSTERING ALGORITHM

Due to the specific of the tasks in the considered domain the implemented algorithm (adapted from [15]) of user clustering is based on analysing user preferences and solutions selected by users and has the following steps:

1. Preliminary linguistic analysis of preferences (tokenisation, spelling and stemming).
2. Extract words/phrases from the preferences and solutions (text processing).
3. Find ontology elements occurring in the extracted words and phrases.
4. Construct weighted graph consisting of ontology classes and attributes, and users. Weights of arcs are calculated on the basis of (i) similarity metrics (i.e. they are different for different user solutions) and (ii) taxonomic relations in the ontology.
5. Construct weighted graph consisting of users (when classes and attributes are removed, arcs' weights are recalculated).
6. Cluster users graph.

Finding ontology elements occurring in the extracted words and phrases is done in two ways: (i) via syntactic similarity, and (ii) via semantic similarity.

The syntactic similarity is calculated via the algorithm of fuzzy string comparison similar to the well-known Jaccard index [21]. It calculates occurrence of substrings of one string in the other string. For example, string "motor" has 5 different substrings (m, o, t, r, mo) contained in the string "mortar". The total number of different substrings in "motor" is 13 (m, o, t, r; mo, ot, to, or; mot, oto, tor; moto, otor). The resulting similarity of the string "motor" to the string "mortar" is 5/13 or 38%.

The semantic similarity (or distance) is based on the machine-readable dictionary Wiktionary [22]. The ontology is represented as a semantic network where names of classes and properties constitute nodes of the network. The nodes corresponding to the ontology elements are linked to nodes representing their synonyms and associated words as this is given in the machine-readable dictionary. The links between the nodes are labelled by the weights of relations specified between the concepts represented by these nodes in the machine-readable dictionary. Weight w of a relation specified between two concepts t_i and t_j is assigned as 0,5 if t_i and t_j are synonyms; 0,3 if t_i and t_j are associated words; and $\propto$ if t_i and t_j are the same words. The nodes representing extracted words and phrases are checked for their similarity to nodes representing ontology elements. As a measure of similarity semantic distance *Dist* is used:

$$Dist(t_j, t_j) = 1 / \left(\sum_{S} \prod_{k} w_k \right),$$

where S is a set of paths from t_i to t_j, formed by any number of links that connect t_i and t_j passing through any number of nodes (k).

For example, let us suppose that the set of words came of parsing the profile comprises two words: *trip* and *lorry*. An illustrative piece of the semantic network built based on this table and is represented in Figure 4. The Figure illustrates three names for classes and attributes in the ontology corresponding to the extracted words: *Trip*, *Ship*, and *Truck*. The semantic distances are as follows:

$Dist(trip, trip) = 1 / \propto = 0$

$Dist(lorry, ship) = 1 / (0.5*0.3 + 0.3*0.5) = 3{,}33$

$Dist(lorry, truck) = 1 / (0.5*0.3 + 0.3*0.3*0.3 + 0{,}5) = 1.48$

 ISBN: 978-1-61208-224-0

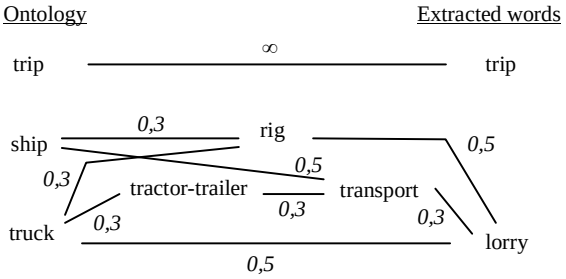


Figure 4. A piece of semantic network relevant to WSDL-attribute "Accident point".

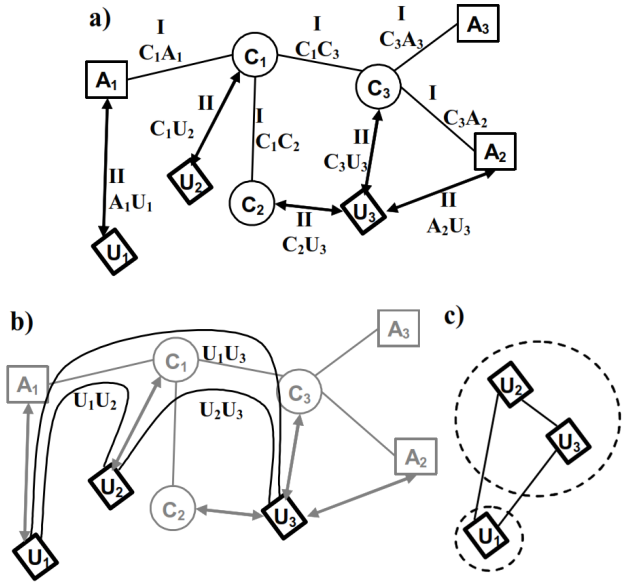


Figure 5. Weighted user – ontology graph and user clustering procedure

It can be seen that for the distance between the concepts *lorry* and *truck* is much shorter than between the concepts *lorry* and *ship*. So, the class *truck* is aligned to the concept *lorry*.

For the clustering procedure, a weighted user – ontology graph is considered. It contains three types of nodes: C – classes from the ontology, A – their attributes, and U – users.

The graph consists of two types of arcs. The first type of arcs I (CA, CC) is defined by the taxonomy of classes and attributes in the ontology. The second type of arcs II (CU, AU) is defined by relations between user solutions and classes/attributes (Figure 5a).

Weights of arc between nodes corresponding to classes and users CU_{weight} and corresponding to attributes and users AU_{weight} are defined via the similarity CU_{sim} and AU_{sim} of the class or attribute (calculated via the fuzzy string comparison algorithm described above). The similarity is a property of relations between class – user/solution or attribute – user/solution. Weights of arcs are defined as follows: $CU_{weignt} = 1 - CU_{sim}$; $AU_{weight} = 1 – AU_{sim}$.

Arcs CA and CC tying together classes and attributes via taxonomic relations (defined by ontology relations class-class, class-attribute) have CA_{weight}, $CC_{weight} \in (\varepsilon, 1)$ defined empirically. CC_{weight} means arcs' weight of linked classes in the ontology. CA_{weight} – arcs' weight of linked attributes and classes.

Since users are represented by their solutions, based on this graph the solutions and weight consequently users are clustered on the basis of the lowest weights of connecting arcs. This is performed in the following sequence. First, the shortest routes between users are calculated (Figure 5b). E.g., weight of the arc U_1U_2 will be calculated as follows: $U_1U_2{}_{weight} = A_1U_1{}_{weight} + C_1A_1{}_{weight} + C_1U_2{}_{weight}$; weight of the arc U_2U_3 can be calculated in 3 ways, it is considered in Figure 2b that $U_2U_3{}_{weight} = C_1U_2{}_{weight} + C_1C_3{}_{weight} + C_3U_3{}_{weight}$ is the shortest one; etc. Based on the calculated weights a new graph consisting of the users only is built (cf. Figure 2c). The value of the parameter D_{max} is set empirically. Assuming that $U_1U_2{}_{weight} > D_{max}$, $U_1U_3{}_{weight} > D_{max}$, and $U_2U_3{}_{weight} < D_{max}$, two clusters can be identified: the first cluster includes users U_2 and U_3, and the second one includes customer U_1 (dashed circles in Figure 5c).

The algorithm can run as updated information is received and update user groups thus providing for self-organizations of user groups in accordance with the changes in the user profiles and context information.

The developed ontology-based clustering algorithm has the following advantages compared to other clustering techniques: (i) *domain-specific knowledge filter* using the ontology; (ii) *natural language processing*; (iii) *term extraction*, such as ontology classes and attributes, units of measures (e.g., "km" and "hrs") can be extracted from the user preferences.

VII. IDENTIFICATION OF COMMON PREFERENCES/INTERESTS AND GROUP RECOMMENDATIONS

User preferences consist of attributes (properties) and/or their values, classes (problem types), relationships (problem structure) and/or optimization criteria that are usually preferred or avoided by the user. The preference revealing can be interpreted as identification of *patterns of the solution selection* (decision) by a user from a generated set of solutions. The ability to automatically identify patterns of the solution selection allows to sort the set of solutions, so that the most relevant (to user needs) solutions would be in the top of the list of solutions presented to the user.

Currently, three major tasks of identification of user preferences can be selected:

1. Identification of *user preferences based on solutions generated for the same context*. In this case, the problem structure is always the same, however its parameters may differ.
2. Identification of user *preferences based on solutions generated for different contexts*. This task will be more complex then the first one since structures of the problem will be different.
3. Identification of *user preferences in terms of optimization parameters*. This task will try to identify if a user tends to select solutions with

 ISBN: 978-1-61208-224-0

minimal or maximal values of certain parameters (e.g., time minimization) or their aggregation.

Based on the clusters built, the user preferences can be identified as common preferences of the users grouped into the clusters.

VIII. CONCLUSION

The paper presents an approach to development of group recommendation system for virtual logistic hub. Virtual logistic hub performs ad-hoc transportation scheduling based on the available schedules, current and foreseen availability and occupancy of the transportation means and services even though they do not cooperate with each other. The approach is based on application of such technologies as user and group profiling, context management, decision mining. It enables for self-organization of user groups in accordance with changing user profiles and the current situation context.

Presented research is at an early development stage. The future work is aimed at implementation of the proposed system in a limited domain for validation of its applicability and efficiency.

ACKNOWLEDGMENT

Some parts of the research were carried out under projects funded by grants # 11-07-00045-a, # 12-07-00298-a, # 12-07-00302-a of the Russian Foundation for Basic Research, project 2.2 of the Nano- & Information Technologies Branch of the Russian Academy of Sciences, and project of the research program “Information, control, and intelligent technologies & systems” of the Russian Academy of Sciences.

REFERENCES

[1] Global GHG Abatement Cost Curve v 2.0, 2009, https://solutions.mckinsey.com/ClimateDesk/default.aspx. Retrieved: June, 2011.

[2] E. Chang, T. Dillon, W. Gardner, A. Talevski, R. Rajugan and T. Kapnoullas, “A Virtual Logistics Network and an E-hub as a Competitive Approach for Small to Medium Size Companies,” Web and Communication Technologies and Internet-Related Social Issues – HIS 2003, Lecture Notes in Computer Science, vol. 2713, 2003, pp. 167-168.

[3] Working Group on Logistics, “Developing Singapore into a Global Integrated Logistic Hub,” Report, 2002. http://app.mti.gov.sg/data/pages/507/doc/ERC_SVS_LOG_MainReport.pdf. Retrieved: February, 2012.

[4] E. Sweeney, “Supply Chain Management in Ireland:the Future,” Logistics Solutions, the Journal of the National Institute for Transport and Logistics, Vol. 5, No. 3, pp. 14-16, June 2002.

[5] K. McCarthy, M. Salamo, L. Coyole, L. McGinty, B. Smyth, P. Nixon, “Group Recommender Systems: A Critiquing Based Approach,” IUI '06: Proceedings of the 11th international conference on Intelligent user interfaces, 2006, pp. 267-269.

[6] I. Garcia, L. Sebastia, E. Onaindia, C. Guzman, “Group Recommender System for Tourist Activities,” EC-Web 2009: Proceedings of E-Commerce and Web Technologies, The 10th International Conference (2009), Springer, LNCS 5692, 2009, pp. 26-37.

[7] T. Hornung, A. Koschmider, A. Oberweis, A. “A Recommender System for Business Process Models,” In WITS’09: Proceedings of the 17th Annual Workshop on Information Technologies & Systems, 2009. Electronic resource. http://ssrn.com/abstract=1328244. Retrieved: February, 2012.

[8] L. Zhena, G. Q. Huangb, Z. Jiang, “Recommender system based on workflow,” Decision Support Systems, Elsevier, vol. 48, no. 1, 2009, pp. 237-245.

[9] S. K. Moon, T. W. Simpson, S. R. T. Kumara, “An agent-based recommender system for developing customized families of products,” Journal of Intelligent Manufacturing, Springer, vol. 20, no. 6, 2009, pp. 649-659.

[10] Y.-J. Chen, Y.-M. Chen, M.-S. Wu, “'An expert recommendation system for product empirical knowledge consultation,” ICCSIT2010: The 3rd IEEE International Conference on Computer Science and Information Technology, IEEE, 2010, pp. 23-27.

[11] E.-A. Baatarjav, S. Phithakkitnukoon, R. Dantu, R. “Group Recommendation System for Facebook,” OTM 2008: Proceedings of On the Move to Meaningful Internet Systems Workshop (2008), Springer, LNCS 5333, 2008, pp. 211-219.

[12] S. E. Middleton, D. De Roure, N. R. Shadbolt, “Ontology-Based Recommender Systems,” Handbook on Ontologies (Ed. By S. Staab, R. Studer), Springer, 2003, pp. 477-498.

[13] H. C. Romesburg, “Cluster Analysis for Researchers,” Lulu Press, California, 2004, 344 p.

[14] G. W. Flake, S. Lawrence, C. L. Giles, F. Coetzee, “Self-Organization and identification of Web Communities,” IEEE Computer, vol. 35, no. 3, 2002, pp. 66-71.

[15] A. Smirnov, M. Pashkin, N. Chilov, “Personalized Customer Service Management for Networked Enterprises,” ICE 2005: Proceedings of the 11th International Conference on Concurrent Enterprising, 2005, pp. 295-302.

[16] A. Smirnov, M. Pashkin, T. Levashova, A. Kashevnik, N. Shilov, “Context-Driven Decision Mining,” Encyclopedia of Data Warehousing and Mining. Hershey (Ed. by J. Wang), New York, Information Science Preference, Second Edition. vol. 1., 2008, pp. 320 – 327.

[17] A. Rozinat, W. M. P. van der Aalst, “Decision Mining in Business Processes,” BPM Center Report no. BPM-06-10, 2006.

[18] P. Petrusel, D. Mican, “Mining Decision Activity Logs,” BIS2010: Proceedings of Business Information Systems Workshops, Springer, LNBIP 57, 2010, pp. 67–79.

[19] A. Smirnov, T. Levashova, A. Kashevnik, N. Shilov, “Profile-based self-organization for PLM: approach and technological framework,” PLM 2009: Proceedings of the 6th International Conferecne on Product Lifecycle Management, Electronic proceedings, 2009.

[20] A. Smirnov, T. Levashova, N. Shilov, “Semantic-oriented support of interoperability between production information systems,” International Journal of Product Development, Inderscience Enterprises Ltd., vol. 4, no. 3/4, 2007, pp. 225-240.

[21] P.-N. Tan, M. Steinbach, V. Kumar, “Introduction to Data Mining,” Addison Wesley, 2005, 769p.

[22] Wiktionary, 2012. http://www.wiktionary.org. Retrieved: February, 2012.

 ISBN: 978-1-61208-224-0

Smart Implementation of Text Recognition (OCR) for Smart Mobile Devices

Ondrej Krejcar
Department of Information Technologies,
Faculty of Informatics and Management,
University of Hradec Kralove,
Hradec Kralove, Czech Republic
ondrej.krejcar@remoteworld.net

***Abstract* –The paper deal with a development of a mobile application for capturing digital photography and its subsequent processing by OCR (Optical Character Recognition) technologies.The developped solution adds to existing Smart Device a capability of a virtual keyboard to which it is possible to transfer recognized text for further work in SMS or text editor. For example, based on the limitation of mobile devices it is mainly targeting at short text sections (internet references, complex adresses, etc.). The accent is targeted on the simple, fast and intuitive working with a mobile device. Practical realization is verified at several Smart Devices with Windows Mobile OS.**

Keywords – OCR; Smart Device; Windows Mobile; Image Processing; Virtual Keyboard

I. INTRODUCTION

The Smart Phones, such as cell phones and PDA (Personal Digital Assistant), especially MDA (Mobile Digital Assistant) are the phenomenon nowadays. The number of cell phone users over 16 years old in the Czech Republic for the year 2009 climbed up to 91%. For the population in the age group from 16 to 54 years the number is equal to 98% [1]. A great boom in the field of cell phones and their performance was caused by using the OS (Operating System), such as Symbian, Android or Windows Mobile. Many of these devices use large colorful displays with touch screen and fast 32bit CPU. Moreover, the GSM module is usually integrated within the standard PDA together with WiFi module. The result is the incorporation of cell phones and PDA as Smart Phones. Based on the usage of efficient 32bit CPUs it is possible to develop power applications for computation.

The primary input system of these devices is the keyboard in a classic "physical" design or in the form of virtual keyboard on a display in the case of touch screen. These types of keyboards provide a comfortable method of information inscription. Nevertheless, the typing is approx. 4x slower than in the case of computer keyboards [2]. However, this typing speed may be insufficient if we would like to use a Smart Device as a tool for fast information recording (e.g. business card copying or copying parts of text). Most commonly integrated CCD (In many PDAs, more precisely in cell phones the cheaper CMOS sensors are used) chips enables the photographing or recording of a video-sequence. Therefore, it is a convenient and instant way of capturing information. Moreover, if this information is time-limited (e.g. it must return within certain time limit or it is only displayed for short time period) then it is the only method.

Nevertheless, sometimes there is a need to further process this captured text. The text retyping from these images is lengthy. Furthermore, if it is necessary to retype using the PDA it should be accounted for switching often between an application with displayed image and the text editor.

In these cases the usage of OCR (Optical Character Recognition) technology is the best solution. The first mobile application OCR was released to the market already in 2002 [3]. Certain factors complicate the usage of OCR in PDA which mostly originated from the low quality of copies acquired by CCM (Compact Camera Module, module with integrated CCD (Charge Coupled Device) or CMOS (Complementary Metal Oxide Semiconductor) sensor, simple optics and electronics). Finally, it is necessary to mention that the common source for OCR application is a scanner.

A PDA which is supplied by OCR has many options in a way of utilization. If the user notices an URL address in some printed document, he can look at it by taking a picture which consequently opens the link in a browser. After this picture the business card with user's data is saved into contacts, etc.

The problem we would like to deal with in this paper is based on a development of mobile OCR application for current Smart Phones at Windows Mobile platform. Such application is necessary for solving of problems mentioned before with the goal in development of virtual keyboard with embedded OCR engine.

Firstly an evaluation of existing solutions will be made in (Section II).

II. EXISTING OCR ENGINES FOR MOBILE DEVICES

The accuracy of OCR depends mainly on the quality of recognizable under layer. The most common usage of OCR on scanned documents achieves quite satisfactory results. Using of OCR in PDA with CCM as a data source recognizer carries number of problems [4], especially:

 ISBN: 978-1-61208-224-0

- Relatively low computational performance (Usually 1/10 of PC performance)
- Low quality of images for OCR (Generally meant as low resolution, blurring, background noise, anomalies caused by compression, etc.)
- Tilt (perspective deformation), skew and rotation
- Incoherent lighting and shadows

Mainly due to these complications is OCR in PDA limited to just small parts of text. Therefore, the insufficient quality of acquired images is compensated by the size proportion of symbols in the overall resolution. The existing applications may be good examples, because they are usually specialized on business card scanning.

A. Existing mobile applications

1) Nokia Multiscanner

Nokia Multiscanner [5] is a freeware application designed for cell phones with Symbian OS. The application supports picture taking and consequently sending it through MMS, Bluetooth or via infrared. It is possible to transfer the image into a text and save it and at the same time the selection of certain area can be made by dragging. Another possibility is to send the image for business card recognition. This option automatically recognizes contact details on the business card and fills in the details for adding a new contact. The OCR engine supports post-processing on the basis of language dictionaries (Technology for replacement of recognized words by words from a dictionary according to their relevance), including the Czech language.

However this solution do not support real virtual keyboard nor clipboard Copy/Past features. Also only Symbian OS is supported.

2) CameraDictionary OCR for Moto

An application [6] for cell phones with Android, Symbian and Windows Mobile systems. It operates on the basis of recorded text recognition and its immediate translation to another language. Even though, this recorded language is available in Chinese or English, the translation is extended by couple of other languages. Furthermore, it enables the text recording with consequent signing of the translated text or so called "Video" regime during which the cursor appears on the screen. The text below the cursor is immediately translated.

However, the main disadvantages are the price and the necessity of internet connection when used.

3) CamCard - Business Card Reader

CamCard [7] is an application specialized on reading business cards. It is targeted at cell phones which run on OS Android, iOS (OS of iPhone cell phones), or Windows Mobile and BlackBerry phones. Furthermore, the CamCard is an extensively automated business card reader with detection of a rotation and a language. The whole recording takes just couple of presses.

The main disadvantage is the narrow specialization on business cards and its price.

4) Babel Reader-LE

Babel Reader-LE [8] is a particular version of Babel Reader for Windows Mobile distributed as a freeware. It enables capturing of an image and subsequent storing of this image in a form of text. Babel Reader-LE is a very simple application. Moreover, it is possible to adjust the captured image before the actual recognition e.g. by background noise removal.

As in the case of Nokia solution a clipboard and keybard option is not possible.

B. Problems of Existing Mobile OCR Solutions

Nokia Multiscanner is the closest application to the one we needed. However, it is designed only for OS Symbian. CameraDictionary OCR and CamCard are commercial applications which are very specialized and not free. Finally, the last mentioned application called Babel Reader was only invented for text recognition. The selection of these applications with OCR for cell phones is significantly limited and the broader application with OCR which would work as an alternative for a virtual keyboard is still missing.

These reasons lead us to develop a new application which is described in this article. We expect to develop a solution which fills a space on current market.

C. Selection of OCR engine

Due to the extent of this application, it is planned to use the existing OCR engine. Following types of engines were chosen as the most suitable:

- **Tesseract OCR** [9] – OCR Engine developed by HP Company in since 1985 until 1995. Nowadays, it is being improved by Google. It is offered in C/C++ language.
- **Ocrad** [10] – another open-source OCR engine. One of his main advantages is mainly an automatic transformation of an input image. It does not accomplish post-processing on the basis of language dictionaries. It is written in C/C++ language.
- **Puma.NET** [11] – an engine for implementation in C# projects with .NET framework.
- **ABBYY Mobile OCR Engine** [12] – a commercial engine used here just for comparison of results. Not available for end users, tested by ABBYY FineReader Online service**.**

A script in PHP language was created in order to accomplish an objective comparison of recognition accuracy. This script is not included within the topic of this article and therefore will not be described in the text. The accuracy of the match is calculated by following formula.

The Greek letter ω is going to represent the number of symbols in a reference text and ω_{err} is the number of errors (substituted, missing symbols or additional symbols). Then the accuracy of match γ_{acc} is defined as:

$$\gamma_{acc} = \frac{\omega - \omega_{err}}{\omega} 100 \quad [\%] \qquad (1)$$

 ISBN: 978-1-61208-224-0

In order to identify the accuracy of recognition, the reference text [Fig. 1] was used.

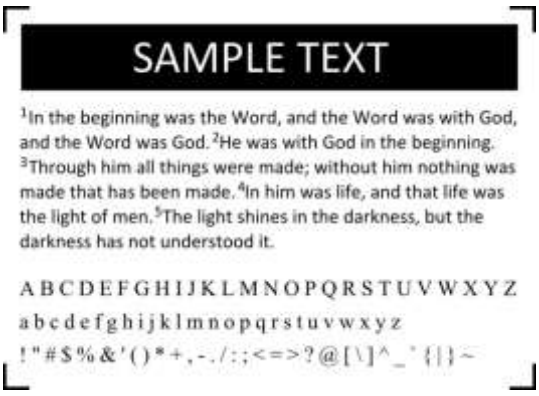


Figure 1. Reference image.

Furthermore, this sample was photographed by Canon PowerShot S3 IS and MDA HTC Touch 2. Consequently, this sample was transferred back to the text form using the above mentioned OCR and compared to the reference text. The accuracy of the match is expressed in percentages in [

]. Column "original" mean the reference text in form of an image.

TABLE I. COMPARISON OF OCR ENGINE ACCURACY

OCR	Original	Canon	HTC
Tesseract	89,78 %	94,72 %	85,25 %
Ocrad	93,30 %	92,71 %	74,36 %
Puma.NET	92,05 %	90,41 %	25,55 %
ABBYY	95,96 %	94,41 %	87,77 %

The comparison shows that the most exact engine is ABBYY Mobile OCR Engine. At the same time it may be noticed that the decreasing quality of sample results in a gradual increase in number of errors. The most significant is the rapid increase of errors when using the Puma.NET engine. In this case, the application was able to correctly recognize approximately one quarter of the text from an image taken by HTC Touch 2. On the other hand, the least sensitive engine considering the quality is Tesseract OCR. Even though, the Ocrad does not realize post-processing on the basis of language dictionaries, it was proven to be very precise.

The most suitable from the point of the evolving application would be to use the OCR engine Puma.NET. This engine is designed for .NET framework environment, contains an analysis of a document structure, writing styles and filter (It accomplishes filtering out of dot-matrix anomalies and contains a regime for processing documents which are sent through fax) [11]. Nevertheless, due to the insufficient recognition accuracy of low quality images, its use has to be denied. On the other hand, it is necessary to choose an engine with good results even for bad resolution photographs. Therefore, the Tesseract OCR engine will be used for this purpose.

III. IMPLEMENTATION

The programmating language C# with a connection to the developing environment Microsoft Visual Studio 2008 was designed for the development of the described application. Microsoft Windows Mobile 5.0 runs as the end platform. Therefore, the application should function well on a PDA with this OS or a higher type of OS.

The application is composed of couple components (by component we mean an incased object (UML expression), not the Visual Studio component), whose relationship is presented on a components'diagram in UML [13] on [Fig. 2].

The application consists of 5 parts of basic components as it may be noticed in this diagram. These are *CameraControl*, *ImageProcessing*, *MainApplication*, *OCRInterface* and *OCR*. The image data are obtained by CCM while considering the data flow. Consequently, the components *CameraControl* and *ImageProcessing* manage their accuracy and new modifications. After the start-up of the *CameraControl* application, it portrays the image data in preview regime on the device's display (*ImageProcessing* is spanned in this stage – it is in bypass regime). After taking a picture and pressing the touchscreen, the *MainApplication* component records this event and ends the preview. A static image is on the output of *CameraControl*. This image is adjusted by *ImageControl* component and portrayed. At this moment, the user is able to choose the words for OCR processing. *MainApplication* sends the coordinates of the selection to the *ImageControl* component which crops the image and forwards it further through the *OCRInterface* to the OCR engine. Finally, the recognized data may be saved inside a folder or copied to a Windows folder.

 ISBN: 978-1-61208-224-0

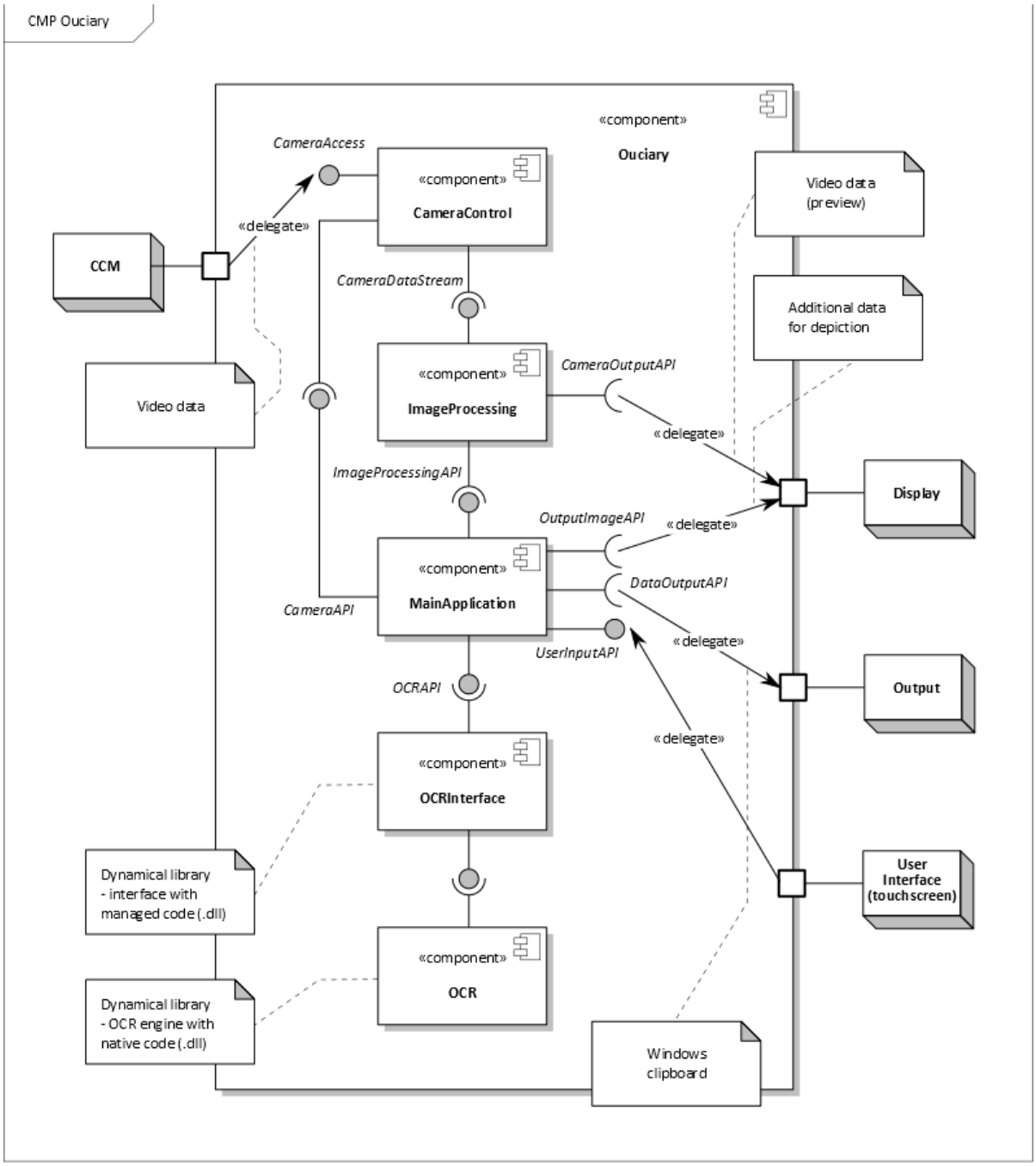


Figure 2. Components‘ diagram of application.

A. Ouciary Application

The application is practically created in form of a guide. After the initiation user is able to turn on the camera and capture the recognized image or simply choose from a file. This is displayed on the following images of the application [Fig. 3].

Consequently, the image is modified. According to the settings, the normalization, automatic rotation and saturation removal take place.

Furthermore, there is the area selection screen for recognition. Here it is possible to rotate the image manually and choose an area for recognition. Character recognition (described above) is very helpful during the text selection if this function is allowed. Moreover, this screen might be absolutely left out (In this case the image is modified according to the settings and the whole image is accounted for).

The selection happens by dragging (“rectangle drawing”). If it is necessary to cancel the whole selection it

 ISBN: 978-1-61208-224-0

is enough to press anywhere on the image. If no area is chosen, the application automatically calculates with the whole image.

Figure 3. Application after the initiation and text capturing.

After the selection of the area it is possible to establish individual recognition process. The progress of recognition is shown here. After the termination of recognition process the application automatically moves on to the form for storage.

The resulting text can be seen here in a textbox and at the same time may be saved into a file or Windows mailbox.

As it was already mentioned, there is a space here for adding the functionality in a form of automatic events in relation to recognized text, eventually to a "templates" usage for contact creation according to a business card etc.

IV. TESTING OF DEVELOPED APPLICATION

The application was tested continuously. The comparison of Tesseract with other OCR engines can be found in the section [*Section II*] and concretely in the table [*Table 1*]. Moreover, there is also a visible influence of the The testing of recognition quality is established by OCR engine which is a product of a third party and therefore is not a direct part of this work. The OCR engine was only a component of this work by a transfer (porting) to Windows Mobile (More precisely Windows Pocket PC 2003) in form of DLL library and to create suitable API. The function of recognition was not interrupted by anything; therefore the testing of quality recognition is not a direct component of this work.

Furthermore, it was established that the best results are obtained by using Tahoma font. In all of the tested samples the results were around 95% which is very sufficient. This is given mostly by the used language dictionary. In cases of graphically different fonts high results can be reached again, however Tesseract must firstly study these writings. The output is formed again by the language dictionary. It is possible to find more information about the preparation of language dictionaries on Tesseract webpage [9].

The results and quality of recognition are completely identical with Tesseract for PC.

A. *Testing of the speed of OCR on PDA*

According to the significantly smaller calculation performance of mobile devices it is advisable to undertake the measurement for the influence of an image characters number and the recognition of time of OCR process.

This measurement was accomplished using TesseractCLI application. The device which was used for testing was PDA HP iPAQ hx4700. This device has a 32 bite processor ARM which works using frequency of 624MHz and RAM memory of 64MB. Windows Mobile 6.5 was used as an operating system.

Moreover, 2 colored TIFF images were tested in resolutions of 268x240 (~64 kPx), 536x480 (~257 kPx) and 1072x960 (~1 MPx). All of these images were in variants with 81,202 and 335 characters. The tested images are shown on [Fig. 4].

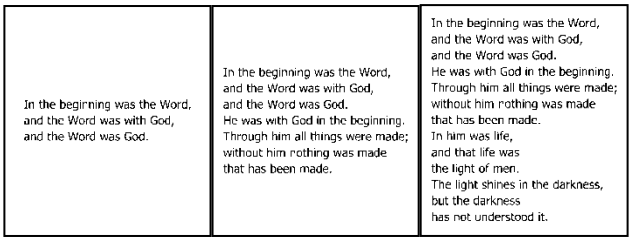


Figure 4. Tested images.

The testing was proceeding after the reset of the device. Each of the recognition measurements were established 5x and the final times were stated as an average. There were only small number of anomalies (maximum of 10% around average) among individual measurements and it was found out that the reset of the device does not have an influence on OCR operating period. The table [*Tab. 2*] gives the test results.

TABLE II. OCR OPERATING PERIOD IN RELATION TO THE NUMBER OF IMAGE CHARACTERS

Time OCR (s)		**Image resolution (px)**		
		268x240	*536x480*	*1072x960*
Number of characters	*81*	3,106	4,609	4,477
	202	9,529	8,450	10,523
	335	24,767	15,581	16,389

There was a strong dependence of the number of characters on the OCR operating period visible from the test's results. On the other hand the resolution dependence did not show to be as relevant, however the results here are still quite interesting.

The OCR time fluctuation concerning the recognition would be most likely given by the function of Tesseract. It can be expected that if the small resolution is used, there will be more faults in recognition and therefore the usage of dictionary will be increased in order to fix those faults. This may have impact on the time of transfer.

 ISBN: 978-1-61208-224-0

On the other hand, high resolutions will have effect on the on image processing time and recognition of individual characters.

According to the mechanism, the minimal time for the transfer is given by a compromise between image resolution (small resolution – dictionary usage, high resolution – long processing and image recognition). This assumption is identical to the measurement (the measurement which was intended for chosen combinations of resolution and for number of characters was repeated with practically same results.).

Generally, the processing is 10x slower than on PC, however it is bearable comparing to usage on PDA.

V. CONCLUSIONS

Main contribution of this project is development of mobile OCR application for Smart Devices with Windows Mobile OS. Used OCR technology can significantly speed up the work using text recognition where there is no requirement for manual transfer of text from an image (e.g. URL address capturing and consequent its display in the browser which is much faster than rewriting the address manually, especially in case of long and complicated addresses). Solution can be also used in the case of business cards digitalization or any other printed material which need to be rewritten. The mobile devices are always nearby [1] and therefore this method brings instant capturing of printed texts.

During the development phase some problem arose out from the real implementation, where the biggest one was the porting of selected desktop OCR Tesseract to the end platform (Windows Mobile). Nevertheless, due to its minimal demands on other components, Tesseract is one of the few free OCR engines which are possible to port using relatively small interferences.

Another inportant issue was in launching of CCM, where used DirectShow is quite complicated in the sense of a slow framerate of preview (around 2 fps).

Developped solution is very well functional, useful and positive.

ACKNOWLEDGEMENTS

This work was supported by „SMEW – Smart Environments at Workplaces“, Grant Agency of the Czech Republic, GACR P403/10/1310. We also acknowledge strong support from Ales Kurecka during development phase and testing of application.

REFERENCES

[1] J. Zelenka, “Information and Communication technologies in tourism – influence, dynamics, trends”, In *E & M EKONOMIE A MANAGEMENT*, Vol. 12, Issue 1, pp. 123-132, 2009

[2] ZY. Liu, HN. Zhou, “Segmenting Texts From Outdoor Images Taken By Mobile Phones Using Color Features”, Proceedings of SPIE, vol. 7874, Article Number: 78740B, DOI: 10.1117/12.872149, 2011

[3] J. Berclaz, N. Bhatti, SJ. Simske, and JC. Schettino,”Image-Based Mobile Service: Automatic Text Extraction and Translation”, Proceedings of SPIE, Vol. 7542, Article Number: 754204, DOI: 10.1117/12.840279, 2010

[4] AP. Pozo, AW. Haddad, M. Boutin, and EJ. Delp, “A Method for Translating Printed Documents Using a Hand-Held Device”, IEEE International Conference on Multimedia and Expo (ICME), JUL 11-15, 2011, DOI 10.1109/ICME.2011.6011940.

[5] Nokia Multiscanner - Recognize Texts With Camera. SymbianV3.Com. [Online] 6. 11 2008. http://symbianv3.com/nokia-multiscanner-recognize-texts-with-camera/.

[6] Camera-Dictionary Software (Net) for Android. [Online] 2010. http://www.hotcardtech.com/eng/OCRUserGuideen.doc.

[7] CamCard (Business Card Reader). IntSig Information. [Online] 3. 10 2010. http://www.intsig.net/home/us/android-/51-camcard-.

[8] Babel Reader-LE 1.0. WareSeeker - Search and Free Download PDA Software. [Online] 16. 10 2010. http://pda.wareseeker.com/Business/babel-reader-le-1.0.zip/15ca4f31ed.

[9] tesseract-ocr - Project Hosting on Google Code. Google Code. [Online] [Citace: 23. 10 2010.] http://code.google.com/p/tesseract-ocr/.

[10] Ocrad - The GNU OCR. Ocrad - GNU Project - Free Software Foundation (FSF). [Online] 11. 05 2010. http://www.gnu.org/software/ocrad/.

[11] Puma.NET. CodePlex - Open Source Project Hosting. [Online] 09. 01 2010. http://pumanet.codeplex.com/.

[12] ABBYY Mobile OCR Engine is a Software Development Kit (SDK). ABBYY - OCR, ICR, OMR, Data Capture and Linguistic Software. [Online] http://www.abbyy.com/mobileocr/.

[13] M. Kadavova, A. Slaby, and F. Maly, “Key factors involving the design of the system of virtual university”, 7th WSEAS International Conference on Applied Computer and Applied Computational Science, pp. 678-683, april 6.-8. 2008

[14] F. Lefley, F. Wharton, L. Hajek, J. Hynek, V. Janecek, “Manufacturing investments in the Czech Republic: An international comparison”, International Journal of Production Economics, vol. 88, Issue: 1, pp: 1-14, DOI: 10.1016/S0925.5273(03)00129-4, 2004

[15] P. Mikulecky, “Remarks on Ubiquitous Intelligent Supportive Spaces”, 15th American Conference on Applied Mathematics/International Conference on Computational and Information Science, Univ Houston, Houston, TX, pp. 523-528, 2009.

[16] I. Bridova, M. Vaculik, and P. Brida, “Impact of Background Traffic on VoIP QoS Parameters in GPON Upstream Link”, Elektronika ir Elektrotechnika, No. 8(104), pp. 113-118, 2011.

[17] V. Bures, “Conceptual Perspective of Knowledge Management”, E & M Ekonomie a Management, vol. 12, Issue: 2, pp. 84-96, 2009

[18] D. Vybiral, M. Augustynek, and M. Penhaker, “Devices for position detection”, Journal of Vibroengineering, vol. 13, Issue: 3, pp. 531-535, Sep. 2011

[19] J. Tariq, U. Nauman, M.U. Naru, “α-Soft: An English Language OCR”. Second International Conference on Computer Engineering and Applications (ICCEA 2010), IEEE Xplore, DOI 10.1109/ICCEA.2010.112, 2010

 ISBN: 978-1-61208-224-0

Distributed Control of Job-shop Systems via Edge Reversal Dynamics for Automated Guided Vehicles

Omar Lengerke, Hernán González Acuña
Universidad Autónoma de Bucaramanga, UNAB
Mechatronics and Control Research Group
Bucaramanga, Santander, Colombia
{olengerke, hgonzalez3}@unab.edu.co

Max Suell Dutra[1], Felipe França[2]
Felix Mora Camino[3]
[12]COPPE UFRJ, Brazil
[3]ENAC, MAIAA Laboratory Toulouse, France
[1]max@mecanica.coppe.ufrj.br, [2]felipe@cos.ufrj.br,
[3]felix.mora@enac.fr

***Abstract*—Flexible Manufacturing Systems (FMS), in which the use of Automatically Guided Vehicles (AGVs) is typical, are a growing trend in many industrial scenarios. A novel, distributed, algorithmic approach to the execution control of activities (work-center oriented) is introduced in this paper, as is, in an integrated way, transportation (AGV oriented) scheduling. The relationship between *jobs*, modeled as *processes*, and *work centers*, modeled as *resources*, and *sinks* defines an undirected graph *G* representing a target Job-shop system. Analogously, the transportation performed by AGVs, also modeled as processes, and their corresponding physical paths, modeled as resources, can also be seen as a dual Job-shop problem. The new approach is based on the Scheduling by Edge Reversal (SER) graph dynamics which, from an initial acyclic orientation over edges, that can be defined via traditional and/or efficient heuristics, let jobs and AGVs proceed in a deadlock-and-starvation-free fashion without the need for any central coordination.**

Keywords-Job-shop; Distributed algorithm; Flexible Manufacturing System; Graph dynamics; Scheduling by Edge Reversal.

I. INTRODUCTION

With the current interest in Flexible Manufacturing System (FMS), there is a growing need for scalable Job-shop solutions. This article presents a new approach to the distributed representation and control of Job-shop systems. The novel approach consists of mapping a Job-shop system into an undirected graph $G = (N, E)$, where $N = \{1, \dots, n\}$ is the set of activities and E is defined as follows: if R_i is the set of resources used by node i in order to operate, an edge $(i, j) \in E$ exists whenever $SR_i \cap SR_j \neq \emptyset$, that is, activities *i* and *j*, share at least one atomic resource.

Next, an initial acyclic orientation *w* is defined over *E*. As shown in the following sections, this setup can be produced via well-known heuristic criteria, such as Earliest Due Date (EDD), Shortest Processing Time (SPT) and Priority (P). The Scheduling by Edge Reversal (SER) dynamics is then applied over *G*, where activities having all of its edges oriented to themselves have the right to operate upon shared resources and then reverse all associated edges, becoming source nodes in a new acyclic orientation *w'*. This ensures that neighboring activities in the system cannot operate simultaneously upon shared resources. In this context, SER acts as a decentralized control mechanism, ensuring mutual exclusion, coordinating all planned activities, regardless of whether they are concurrent or sequential. Besides, the proposed algorithm takes into consideration transport times, integrating transport and activity schedules, and also providing scalable solutions. In addition, it produces optimal minimum make-spam solutions comparable to traditional methods, while creating a deadlock-and-starvation-free system by construction.

SER is our subject in Section II. The two sections that follow (Section III and IV) are devoted to contextualizing the Job-shop and dispatching problem into the FMS domain. Sections V and VI discuss the construction of the proposed algorithm, and show the effective use of SER for distributed control of Job-shop systems, as well as the final conclusions.

II. SCHEDULING BY EDGE REVERSAL

In order to implement a distributed scheduling algorithm for decentralized control of Job-shop systems employed throughout, we decided to use a scheduling scheme which ensures by construction a deadlock-and-starvation-free system. The adopted approach is based on the algorithm presented in [1][2][3] to ensure mutual exclusion on distributed asynchronous systems, namely Scheduling by Edge Reversal (SER). In this context, SER is a simple and powerful distributed algorithm, originally conceived to support Distributed Systems under heavy load condition, when processors are constantly demanding access to all resources that they use.

Important SER properties, and the NP-completeness of the problem of finding optimal concurrency amounts provided by the SER dynamics over a given distributed system, are established in [3]. SER works as follows: (*i*) the target distributed system is described by an undirected graph $G = (N, E)$, where $N = \{1, \dots, n\}$ is the set of processing nodes and E is defined as follows: if SR_i is the set of resources used by node i in order to operate, an edge

 ISBN: 978-1-61208-224-0

$(i,j) \in E$ exists whenever $SR_i \cap SR_j \neq \emptyset$, that is, nodes i and j share at least one atomic resource; (*ii*) an initial acyclic orientation w is defined over E; (*iii*) all, and only sink nodes in w, i.e., nodes having all of its edges oriented to themselves, have the right to operate upon shared resources and then reverse all associated edges, becoming source nodes in a new acyclic orientation w'. This ensures that neighboring nodes in the target distributed system cannot operate simultaneously upon atomic shared resources. SER is the graph dynamics defined by the endless iteration of (*iii*) over G (Figure 1).

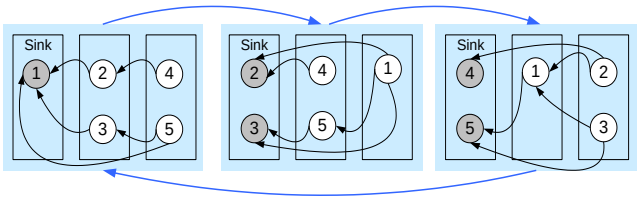


Figure 1. SER Operation

Considering G finite and, consequently, a finite number of possible acyclic orientations over G, eventually a repetition, i.e., a *period* of length l, will occur. An interesting property of SER lies in the fact that, inside any given period, each node operates, i.e., becomes a sink, the same number q of times, ensuring "fairness", in the long run operation, among all processing elements of G [3].

Many works devised a powerful family of SER-based distributed algorithms in different contexts [4]: presented how a SER dynamic can be used for sharing resources at non-uniform rates, allowing different processor priorities, breaking the symmetry rule that every processor should become a sink the same number q of times in a given period;[4] illustrates how to perform an optimal mapping of processors or machines in neighborhood-constrained systems and [5] demonstrated a novel algorithm named Scheduling by Edge Reversal with Hibernation (SERH), a distributed algorithm for scheduling of atomic shared resources in the context of dynamic load reconfiguration, where processors or nodes are able to relinquish the right of execution, allowing the reconfiguration of the whole of the distributed system.

Due to its simplicity, SER is being currently applied to different domains. Among them, we could list: (*i*) industrial plants, where process are jobs and resources are machines, Automated Guided Vehicles (AGV), consumption, etc.; (*ii*) computational grid scheduling, where processes are computing jobs, and resources are *CPUs*, data; disk space and network links are grid data movement, where applications geographically distribute every datum to be used by a distributed computation.

III. FLEXIBLE MANUFACTURING SYSTEM AND JOB-SHOP SCHEDULING

Our interest in distributed Job-shop algorithms comes from the increasing interest in Flexible Manufacturing System (FMS) [6][7][8]. Flexibility measures the ability to adapt to a wide range of possible environments. The term FMS refers to a class of highly automated systems that consist of set of computer-numerically-controlled (CNC) machine tools and supporting workstations that are connected by an automated material handling system. The resulting system is controlled by a central computer that coordinates machine tools, material handling, and parts [9]. Especially, we consider the FMS composed by several Flexible Manufacturing Modules (FMM) or Flexible Manufacturing Cells (FMC), and, at least, one Material Handling System (MHS) consisting of one or more Automatic Guided Vehicles (AGVs). FMS scheduling is significantly different from traditional Job-shops where the human being is concerned. Deadlock situations may occur in FMS due to jobs in a circular waiting of resources (robots, buffers or paths). Consequently deadlocked situations have been identified as one of the most critical problems in the scheduling and control of FMSs.

In multi-operation shops, jobs often have different routes. More specifically, in a Job-shop, each part has its own route. Such environment is known as a generalization of a flow shop (a flow shop is a Job-shop in which each and every job has the same route). The simplest Job-shop models assume that a job may be processed on a particular machine at most once on its route through the system. In others a job may visit a given machine several times on its route through the system.

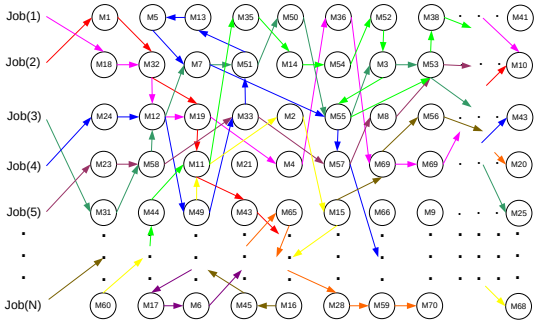


Figure 2. Job-shop Problem

These shops are said to be subject to re-circulation, which increases the complexity of the model considerably, besides the NP-completeness of the Job-shop problem [10].

In our formulation of the Job-shop problem, we assume that there are many jobs on each route. In practice, routes may correspond to various production processes, or to various types of products manufactured in a factory. In that case, the jobs may correspond to parts or lots, and there will indeed be many such jobs for each route. A generalization of the Job-shop is the flexible system with work centers that

 ISBN: 978-1-61208-224-0

have multiple machines in parallel. From a combinatorial point of view, the flexible Job-shop with re-circulation is one of the most complex machine environments. It is a very common setting in the semiconductor industry

In the Job-shop scheduling problem, a set J of n jobs $J_1, J_2, \dots, J_n$, has to be processed on a set M of m different machines $M_1, M_2, \dots, M_n$. Each job J_j consists of a sequence of m_j operations $O_{j1}, O_{j2}, \dots, Ojm_j$, that must be scheduled in this order (Figure 2). Moreover, each operation needs to be processed only on a specific machine among the m available ones. Pre-emption is not allowed and machines can handle at most one operation at a time. Operation O_{jk} has a fixed processing time p_{jk}. The objective is to find an operating sequence for each machine to minimize the make-span $C_{max} = max_{j=1,n}, C_j$, where C_j denotes the completion time of the last operation of job $J_j (j = 1, \dots, n)$ [11]. Operations of the jobs in a Job-shop have to be scheduled to minimize one or more objectives, such as the make-span C_{max} or the number of late jobs.

	Operation 1		Operation 2		Operation 3	
	Proc. time	Mach. number	Proc. time	Mach. number	Proc. time	Mach. number
J_1	3	M_A	1	M_B	5	M_C
J_2	9	M_C	10	M_B	3	M_A
J_3	10	M_B	8	M_C	6	M_A

TABLE I. SAMPLE SCHEDULING PROBLEM (Liao and You (1992))

Step t	A_t	e_k	m^*	k
1	$O_{11}O_{21}O_{31}$	0 0 0	B	O_{31}
2	$O_{11}O_{21}O_{32}$	0 0 10	C	O_{21}
3	$O_{11}O_{22}O_{32}$	0 10 10	A	O_{11}
4	$O_{12}O_{22}O_{32}$	10 10 10	B	O_{22}
5	$O_{12}O_{23}O_{32}$	20 20 10	C	O_{32}
6	$O_{12}O_{23}O_{33}$	20 20 18	A	O_{33}
7	$O_{12}O_{23}$	20 24 24	B	O_{12}
8	$O_{13}O_{23}$	21 24 -	C	O_{13}
9	O_{23}	26 26 - 27	A	O_{23}

TABLE II. Construction of Schedule for Example TABLE I

We present a procedure that will allow the generation of as many non-delay schedules as desired [9]. Basically, we construct a schedule by scheduling one operation at a time using the following algorithm: *(i) Initialization.* Let stage *t = 1*, $S_1 = 0$ (where, S_t is the partial schedule of (*t-1*) scheduled operations. A_1 contains the first operation of each ready job (where, A_t is the set of operations available to be scheduled at stage *t*, that is, all predecessor operations are in S_t). *(ii) Selection*. Find $e^* = min_k \in A_t e_k$ (where, $e_k$is the earliest time that operation $k \in A_t$ can be scheduled, that is, predecessors are completed and the needed machine is available). If several e^* exist, the algorithm chooses it arbitrarily. Let m^* be the machine needed by e^*. Choose any $k \in A_t$ that requires $m^*$and has $e_k = e^*$. *(iii) Increment.* Add the selected operation k to S_t to create S_{t+1}. Remove k from A_t and add the next operation for its job unless that job is completed; this creates A_{t+1}. Set *t = t + 1*. If *t = MJ* stop; otherwise go to (i). As an illustration, consider the following Job-shop problem (TABLE I), presented by [12]. The process continues until all 9 operations are assigned. Steps are summarized in TABLE II., and this new algorithm also produces the best known make-span of 27.

IV. DISPATCHING RULES

Dispatching rules have received much attention from researchers over the past decades [13][14][15]. In general, whenever a machine is freed, a job with the highest priority in the processing queue is selected to be run on a machine or work center.

Dispatching is the job selection process from a queue, its immediate setup and processing, when a processor becomes available. Simple dispatching rules are often used in shop scheduling and a list of the more popular ones follows: (*i*) Shortest Processing Time (SPT): Highest priority is given to the waiting operation with the shortest imminent operation time. Processing time (p_{ij}) represents the time job *j* has to spend on machine *i*. Subscript *i* is omitted if the processing time of job *j* does not depend on the machine or if it only needs processing on one machine. If there are a number of identical jobs that all need a processing time p_i on one machine, then we refer to this set of jobs as items of type *j*.

The production rate of type *j* items is denoted by $Q_j = \frac{1}{p_j}$ (number of items per unit time). (*ii*) Longest Processing Time (LPT): Highest priority is given to the waiting operation with the longest imminent operation time. (*iii*) Earliest Due date (EDD). Select a job with minimum processing time. The due date d_j of job *j* represents the committed shipping or completion date (the date the job is promised to the customer).

Completion of a job after its due date is allowed, but a penalty is then incurred. When the due date absolutely must be met, it is referred to as a deadline. (*iv*) Most Work Remaining (MWKR): Highest priority is given to the waiting operation associated with the job having the most total processing time remaining to be done. (*v*) Least Work Remaining (LWKR): Highest priority is given to the waiting operation associated with the job having the least amount of total processing time remaining to be done. (*vi*) Total Work (TWORK): Highest priority is given to the job with the least total processing requirement on all operations. (*vii*) First In First Out (FIFO): Highest priority is given to the waiting operation that arrived at the queue first. (*viii*) Last In First Out (LIFO): Highest priority is given to the waiting operation that arrived at the queue last. (*ix*) RANDOM (Random): Select a job *au hazard*.

 ISBN: 978-1-61208-224-0

V. SER ON JOB-SHOP SYSTEMS FOR ROUTING PLANNING OF AGVS

AGV routing planning is an important problem in the transportation, distribution and logistics fields. Route is the customary series of stops during a trip (programming of a succession of procedures). Computing the firing sequence of transitions which will yield an optimal result and also avoid deadlocks which might be present is important to real-time control of the modeled system. If an FMS is modeled, the routes planning of AGVs using SER, an optimal firing sequence is an optimal schedule for the system. Hence a method to find an optimal firing sequence of transitions is beneficial to both SER and FMS scheduling. A perpetual deadlock can happen in FMS due to a number of works which are expected to move resources to each other. Therefore, a model that can handle such complex systems is necessary. Several works conducted these analysis types using different methods such as Petri networks [16][17][18], but most of these methods have limitations when there are several types of tasks or activities and large quantity of machinery and do not solve the problem of routing and perpetual deadlock.

A. Definition

The problem of Job-shop systems can be developed from a scheduling distributed algorithm to control this category of decentralized systems. This is possible through a mapping of the Job-shop target in a graph *G = (N,E)* where each element of *N* is one of the planned activities, with pre-established time, to be implemented in exclusive mode on a limited set of resources, which access restrictions defined the edges set *E*. It is also shown as an acyclic orientation is performed directly on *E* the basics of criteria such as traditional heuristic EDD, SPT and P. The dynamics of scheduling by edge reversal can then be applied to *G*, acting as a decentralized control mechanism of coordination of the implementation of various activities planned, whether concurrent or sequential. Implementation of SER in such systems is a new concept that provided a description of the form of sharing (AND, OR, XOR, negative, among others) to solve the problem of planning routes of the AGVs.

Binary Operators: For OR sharing operates a single resource *M (machine)* in a process *J (job)* (Figure 3). The resource is released (edge reversal) when the processing time finishes (p_{ij}) in each of the process operations (*O*). For AND sharing are illustrated in Figure 4.

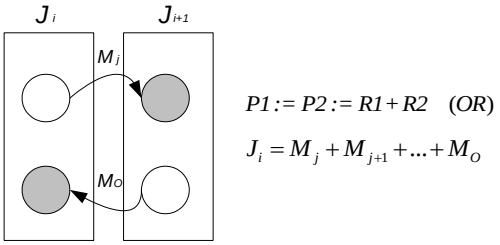


Figure 3. OR Sharing

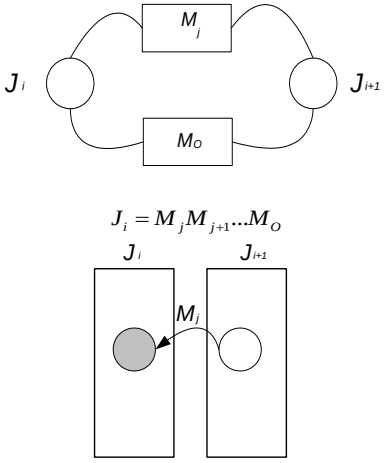


Figure 4. AND Sharing

Example: Applying the concept of SER for the example of TABLE I and represented by Figure 5, which was used the concept of algorithm, to solve problems Job-shop.

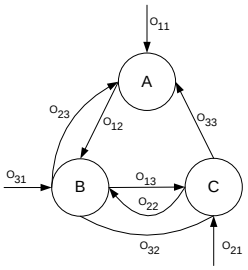


Figure 5. Schematic Diagram of the Problem - Table 1

The three jobs are represented as an expression given by the Equation (1) of the XOR sharing $(M_A \oplus M_B \oplus M_C)$.

$$J_1 = J_2 = J_3 = (M_A \oplus M_B \oplus M_C) = M_A\overline{M_B}\,\overline{M_C} + \overline{M_A}M_B\overline{M_C} + \overline{M_A}\,\overline{M_B}M_C \quad (1)$$

The dynamics of edge reversal corresponding of the system proposed by Equation (1), is shown in Figure 6. In this case, the initial acyclic orientated adopted is determined from criteria or classic dispatching rules (EDD, MWKR, Priority J_1 and Random). According to the dynamic and orientation criteria, the first set of operations to be processed

 ISBN: 978-1-61208-224-0

is $A_1 = O_{11}, O_{21}, O_{31}$ where A_i are sinks, and the processing time (p_{ij}) is given by $\min t \in A_I$, $t_p = 3$ and remainder time (t_f) of $O_{21} = 6$ and $O_{31} = 3$, while O_{11} is completed. The following edge reversal is selected operations $A_2 = O_{12}, O_{21}, O_{31}$. The next steps are summarized in Figure 6 and Figure 7, where the make-span is 27. For simplifying:

$$A = M_A\overline{M_B M_C}, B = \overline{M_A}M_B\overline{M_C} \; and \; C = \overline{M_A M_B}M_C.$$

An immediate benefit of this approach is the decentralization of the job control, which enables the distributed control to deal with any modification of the due time (asynchronous algorithm).

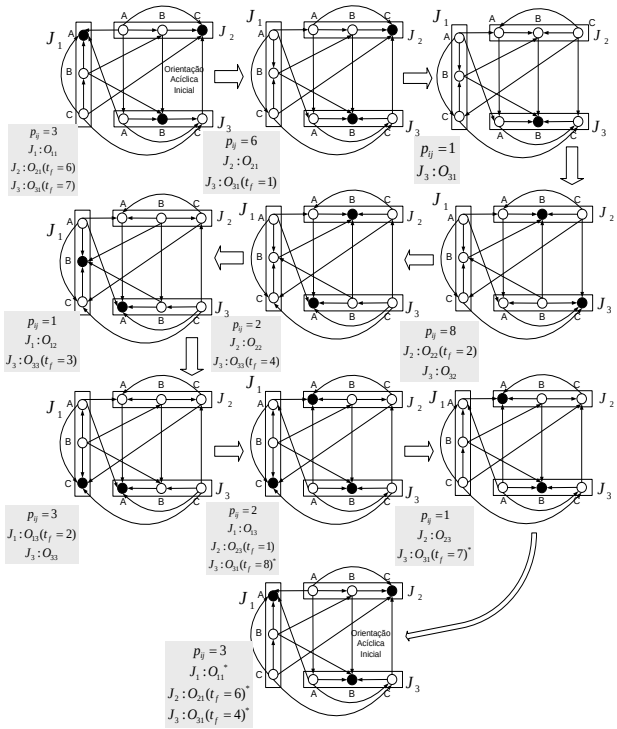


Figure 6 Example of SER

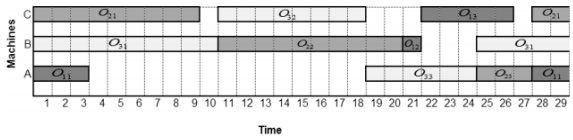


Figure 7 The Generated Schedule with make-spam of 27, with the endless operation of the SER

VI. SER ON JOB-SHOP SYSTEMS FOR PATH PLANNING OF AGVS

The scheduling by edge reversal can be also used in the creation of a mechanism for dynamic planning programming of paths, allowing traffic concurrent of AGVs by the various regions (R) that constitute the layout of a FMS. Each layout of FMS is presented in a schematic diagram in order to show the paths, roads connected and ways of vehicles traffic. Each AGV needs some regions to complete its scheduled displacement, this displacement is related to an operation time or displacement (t_0). The region number is defined as $R = R_1, R_2 \ldots, R_m$. In the example presented at Figure 9, different AGVs can compete for one or more regions (shared resources) that constitute the FMS. If there is a conflict, classic rules for dispatching (such as EDD, SPT, Priority) can be used.

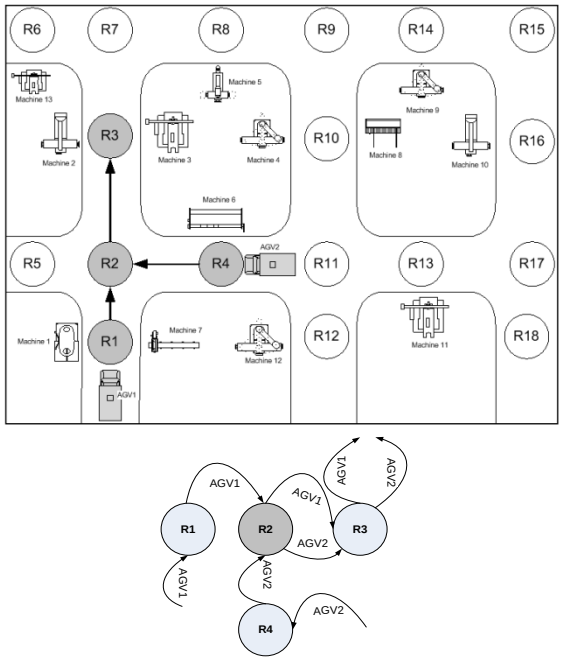


Figure 8. Planning Example for Path Planning Scheduling

Figure 8 shows a schematic example of scheduled displacement of AGVs into an FMS connected to 3 sequential regions, all of them aligned in the same direction. Vehicles can move around in accordance to the following: (*i*) AGV1 moves through the subsequent path, $R_1 \rightarrow R_2 \rightarrow R_3$ and (*ii*) AGV2 moves through $R_4 \rightarrow R_2 \rightarrow R_3$. (*iii*) we also know when (*t*) each AGV is willing to use each shared resource. The description of the processes and resources (Figure 8) is given by:

$$AGV1 = R_1(4t_0) \rightarrow R_2(3t_0) \rightarrow R_3(4t_0) \quad (2)$$
$$AGV2 = R_4(3t_0) \rightarrow R_2(4t_0) \rightarrow R_3(2t_0) \quad (3)$$

The Boolean expression that represents the dynamic is represented by:

$$AGV1 = R_1 \oplus R_2 \oplus R_3 = \underbrace{R1\overline{R2}R3}_{A} + \underbrace{\overline{R1}R2\overline{R3}}_{B} + \underbrace{\overline{R1R2R3}}_{C} \quad (4)$$

$$AGV2 = R_4 \oplus R_2 \oplus R_3 = \underbrace{R4\overline{R2}R3}_{D} + \underbrace{\overline{R4}R2\overline{R3}}_{E} + \underbrace{\overline{R4R2}R3}_{F} \quad (5)$$

In the dynamics of edge reversal for the example system, the initial acyclic orientation adopted is determined from the criteria of EDD, priority AGV1 and SPT. The first operations being processed are the displacements of AGV1 and AGV2 on $R_1 = R1, R4$ where $R_i$are sinks and operation time t_0 is given by (t_f) $\min t \in R_i, t_0 = 3$ and remainder time to finish the displacement (t_f) of R1 is 1, while R_4 is completed. In the following edge reversal, operations R_2 = R1, R2 and $t_0 = 1$ are selected. The next steps are summarized in Figure 9 and Figure 10. The previous shows that the problem of path planning can be treated as a Job-shop problem.

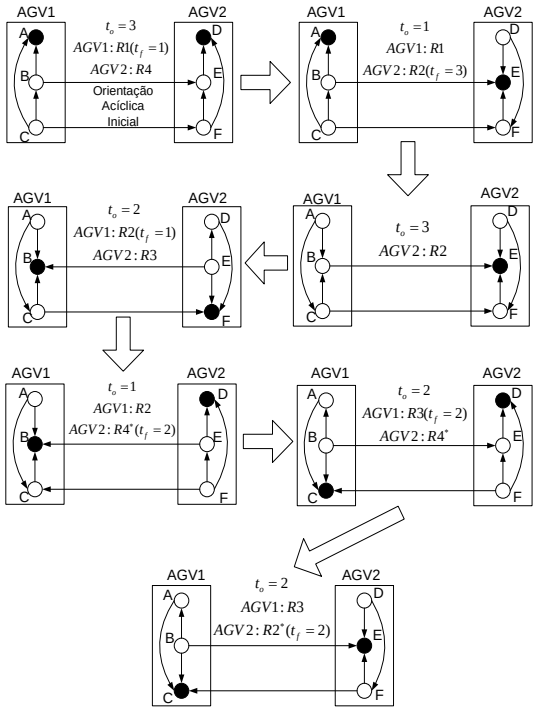


Figure 9. Example of SER on Path Planning

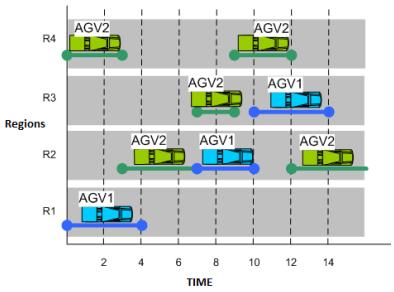


Figure 10. The Generated Schedule

CONCLUSION

With the current growing interest in Flexible Manufacturing System (FMS), there is a growing need for distributed Job-shop algorithms. This article presents an implementation of a distributed scheduling algorithm for decentralized Job-shop systems that can be used for FMS control and scheduling. This novel approach allows decentralization of the job control and enables the distributed control to deal with any modification of the due time, caused by its asynchronous nature. The next step is the use of this algorithm in two real conditions: (*i*) AGV traffic control in automated container terminal and automated large scale freight transport systems and (*ii*) computational grid scheduling and grid data movement.

REFERENCES

[1] Vieira, F.R.J., Rezende, J.F., Barbosa, V.C. and Fdida, S., "Scheduling links for heavy traffic on interfering routes in wireless mesh networks". Computer Networks, 2011.

[2] Barbosa, V.C., "An Introduction to Distributed Algorithms", MIT Press, 1996.

[3] Barbosa, V.C. and Gafni, E., "Concurrency in heavily loaded neighborhood-constrained systems", ACM Transactions on Programming Languages and Systems 11(4), 1989, pp. 562-584.

[4] França, F.M.G. and Faria, L., "Optimal mapping of neighbourhood-constrained systems", In: Proceedings, Irregular'95, Springer-Verlag, Lyon, France, Lecture Notes in Computer Science, 1995 , pp. 165-170.

[5] Carvalho, D, Protti, F., Gregorio, M.D. and França F.M.G., "A novel distributed scheduling algorithm for resource sharing under near-heavy load". Lecture Notes in Computer Science. 2005, pp. 431-442.

[6] Herrero-Perez, D. and Martinez-Barbera, H., "Modeling Distributed Transportation Systems Composed of Flexible Automated Guided Vehicles in Flexible Manufacturing Systems", IEEE Transactions on Industrial Informatics, vol. 6, 2010, pp. 166 – 180.

[7] Hartley, J., FMS at Work. IFS Publications Ltd., North-Holland Publishing Company, Division of Elsevier Science Publishers B. V, 1984.

[8] Harrison, D.K. and Petty, D.J. "Systems for Planning and Control in Manufacturing". Butterworth-Heinemannm, Elsevier Science, 2002.

[9] Askin, A.G, Standridge, C.R, "Modeling and Analysis of Manufacturing Systems". John Wiley & Sons, 1993.

[10] Vinod, V. and Sridharan, R., "Simulation modeling and analysis of due-date assignment methods and scheduling decision rules in a dynamic job shop production system". International Journal of Production Economics, 129(1), 2011, pp. 127–146.

[11] Leung, J.Y.T., "Handbook of Scheduling - Algorithms, Models, and Performance Analysis". Chapman & Hall CRC Press LLC, 2004.

[12] Liao, C.J., and You, C.T., "An improved formulation for the job-shop scheduling problem". The Journal of the Operational Research Society 43(11),1992, pp.1047- 1054.

[13] Lu, H.L., Huang, G.Q. and Yang, H.D., "Integrating order review/release and dispatching rules for assembly job shop scheduling using a simulation approach". International Journal of Production Research 49(3), 2011, pp. 647 – 669.

[14] Chan, F.T.S., Chan, H.K. and Lau, H.C.W.,"Analysis of dynamic dispatching rules for a exible manufacturing system". Journal of Materials Processing Technology 138, 2003, pp.325-331.

[15] Dominic, P.D.D., Kaliyamoorthy, S. and Kumar, M.S., "Efficient dispatching rules for dynamic job-shop scheduling". The International Journal of Advanced Manufacturing Technology 24(1-2), 2004, pp. 70-75.

[16] Meng, J, Soh, Y. and Wang, Y, "A tcpn model and deadlock avoidance for fms job shop scheduling and control system". In: IEEE International Workshop on Emerging Technologies and Factory Automation, Paris, France, 1995, pp 521- 532.

[17] Wu, N. and Zhow, M., "Modeling and deadlock control of automated guided vehicle systems", In: IEEE/ASME Transactions on Mechatronics, vol. 9, 2004, pp. 50-57.

[18] Zhang, H., Li, D., Yang, S. and Wang, W., "A new model of exible manufacturing system based on petri nets". In: International Conference on Mechatronics and Automation, ICMA 2007, 2007, pp. 3894-3899.

 ISBN: 978-1-61208-224-0

The Design of a Self-Localization Estimation Method for Indoor Mobile Robots using an Improved SURF Algorithm

Xing Xiong
College of Information and Communication
Daegu University
Gyeongsan-City, Gyeongbuk, Korea
e-mail: GaleWing@gmail.com

Byung-Jae Choi
School of Electronic Eng.
Daegu University
Gyeongsan-City, Gyeongbuk, Korea
e-mail: bjchoi@daegu.ac.kr

***Abstract*— We present an improved self-localization estimation algorithm in this paper. The algorithm uses a modified SURF method to extract the interest points, using it to extract the orientation and a descriptor of the interest point in order to lessen the computation time. A robot using this method can estimate its indoor self-localization according to matched interest points. A number of intermediate results will also be discussed. The intermediate results show that the displacement method could correctly match the interest points in two images.**

Keywords-Ceiling Key Point Extraction; SURF (Speeded-Up Robust Features); DSP (Digital Signal Processor).

I. INTRODUCTION

Mobile robot self-localization is a mandatory task for accomplishing full autonomy during navigation. Various solutions in the robotics community have been developed in order to solve the self-localization problem. The solutions can be categorized into two groups: relative localization (dead-reckoning) and absolute localization. Although very simple and fast, dead reckoning algorithms tend to accumulate errors in the system since these methods only utilize the information coming from proprioceptive sensors, such as odometer readings (e.g. incremental encoders on the robot wheels). Absolute localization methods are based on exteroceptive sensor information. This method yields a stable locating error but is more complex and costly in terms of computation time. Relative localization requires a high sampling rate in order to maintain an up-to-date pose, whereas absolute localization is applied periodically with a lower sampling rate to correct relative positioning misalignments [3].

With the furthering development of science and technology, visual positioning methods play an important role in the self-localization of autonomous mobile service robots working in indoor environments [5]. Generally, prior knowledge of an indoor environment can be used to determine the position and orientation of a mobile robot via visual positioning approaches. The features used by different approaches for mobile robot localization range from artificial markers, such as barcodes, to the placement and orientation of ceiling lights and tiles, for example. Indeed, the selected visual features have significant influence on the positioning approach performance.

The remainder of this paper is organized as follows: Section II presents some of the related studies. Section III lays out the composition of the proposed algorithm. Section IV discusses some of the intermediate results. We draw our conclusions in Section V.

II. RELATED WORK

In the field of image processing, the Speeded Up Robust Features (SURF) algorithm [6] is an efficient and high-speed algorithm, which is considered to be an improved version of the Scale-invariant Feature Transform (SIFT) algorithm [10]. The SURF algorithm mainly consists of two parts: interest points extraction and an orientation and descriptor of the interest points extraction. For the interest point extraction, the SURF method uses an integral image and box filter to replace the Gaussian filter and the DoG (Difference of Gaussian) method found in the SIFT algorithm. This allows for a greatly reduced computation time.

However, in the orientation and descriptor section the SURF method scans the neighborhood region twice. In the first scan the orientation of the interest point is extracted. The second scan, according to the orientation of the interest point, is used to extract the descriptor. In low-speed devices such as a DSP board, the two scans increase the amount of computation time. Furthermore, in the case of images that only rotate and move, a simpler method can be used to obtain the orientation and descriptor.

In our paper, we use an alternative method to obtain the orientation and descriptor. This method only scans the neighborhood region interest points once. Our self-localization estimation algorithm contains three parts: interest points extraction (using SURF), orientation and descriptor extraction (using the improved method), and the interest points matching and self-localization estimation. Because there is only one scan the necessary amount of calculation is reduced.

III. THE COMPOSITION OF THE ALGORITHM

In an indoor environment, the floor is assumed to be planar. The ceiling usually consists of a series of blocks that form a chessboard pattern parallel to the floor. In this study, a camera is mounted onto the top of the mobile robot working on the floor. The camera points to the ceiling, as shown in Figure 1.

 ISBN: 978-1-61208-224-0

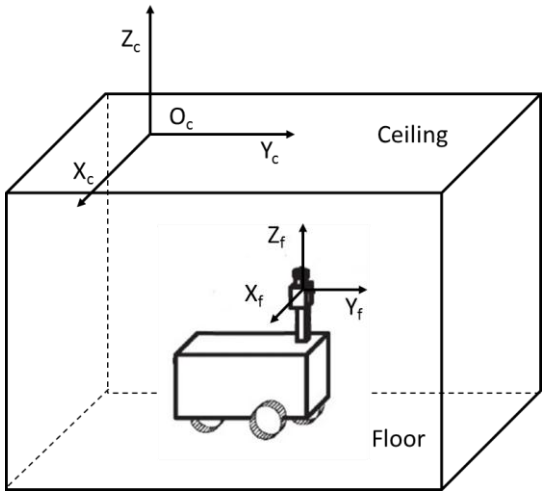


Figure 1. Ceiling based visual positioning

In our research, the SURF algorithm is used to extract the interest points. The SURF's replacement method is used to extract the orientation and the descriptors of the interest points. The self-localization of the mobile robot is then estimated according to the different positions of the same interest points in two images. The flow chart of the algorithm is shown in Figure 2. The interest points extraction section is broken down in Figure 3.

Simply put, in the rapid interest points detection method, the NMS (Non-Maximum Suppression) method used after obtaining the Fast-Hessian matrix in the conventional SURF algorithm is changed to a Non-minimum suppression method to obtain the feature points whose gray value is high. In addition, the conventional order of the box filter scale [6] is changed to 75, 51, and 99. Not only does this remove the impact of the image edge, reducing the amount of calculation, but also leads to an increase in the interest points. The interest points extraction results are shown in Figure 6.

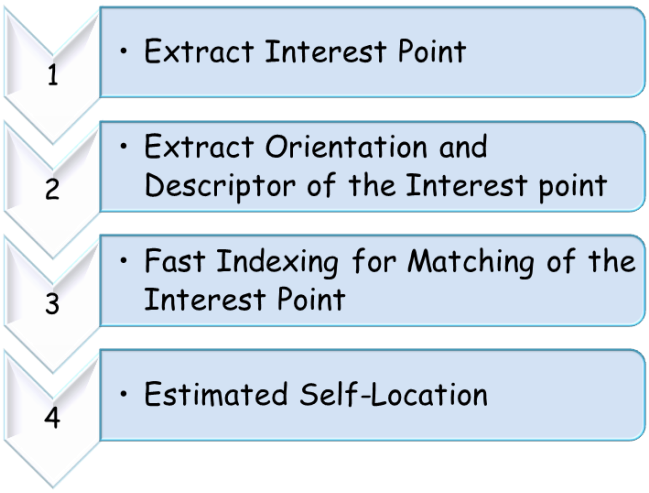


Figure 2. The SURF algorithm flow chart

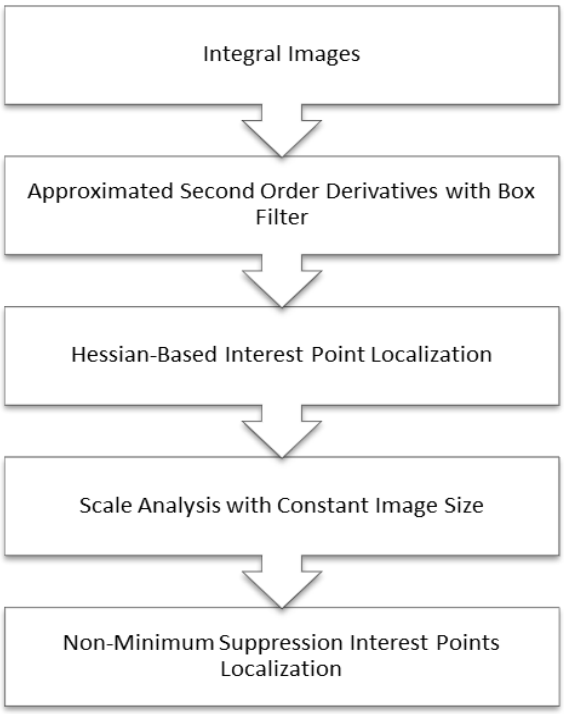


Figure 3. The interest points extration

Regarding the orientation and descriptor extraction, our descriptor method is similar to that of the SIFT algorithm. The flow chart describing the replacement algorithm is shown in Figure 4. The details regarding this algorithm can be found in [11].

To date, in a simple environment, for the case of the translation and rotation of the image obtained from the ceiling, the replacement algorithm has been verified to be feasible. The results from the replacement algorithm are shown in Figure 7.

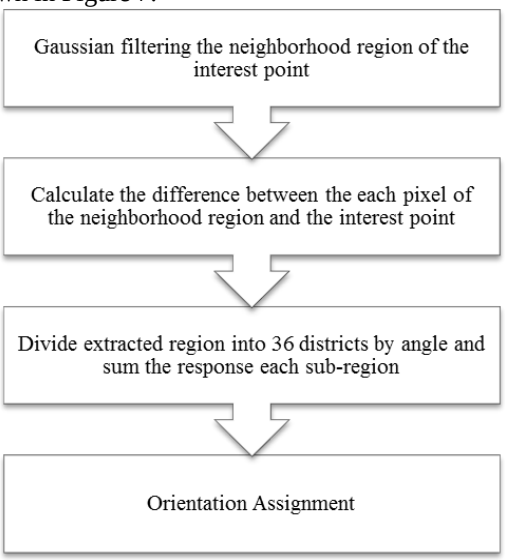
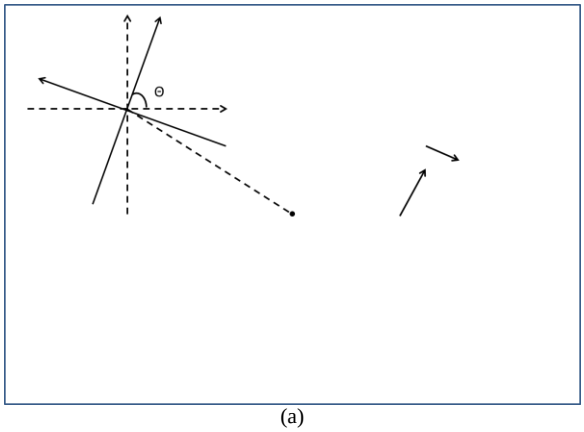


Figure 4. The modified algorithm flow chart

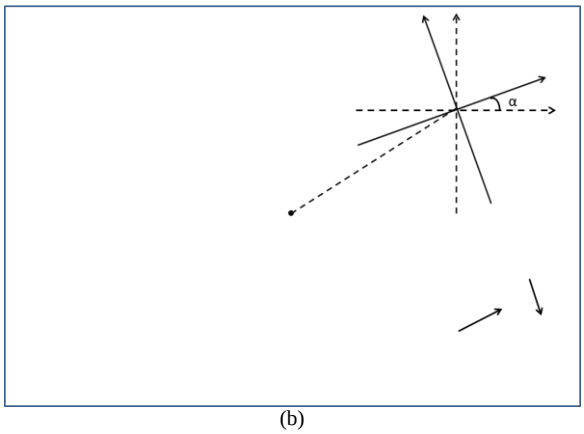
 ISBN: 978-1-61208-224-0

The self-localization estimation method is illustrated in Figure 5.

The two illustrations shown in Figures 5(a) and 5(b) show differences obtained over a small time interval. Figure 5(a) illustrates a baseline before the camera is moved. After the camera is moved (5 seconds), Figure 5(b) illustrates the position and orientation differences. We assume that there are three interest points and a center point. The center point represents the self-localization of the mobile robot. One point amongst three interest points have two coordinates (Dash Line Coordinate System and Solid Line Coordinate System).



(a)



(b)

Figure 5. The self-localization estimation method

The dashed line coordinates represent the image coordinates whereas the solid line coordinates represent the interest points. The X-axis of the solid line coordinate represents the orientation of the interest point.

The center point's coordinate *(x, y)* in the dashed line coordinate system can be changed to the solid line coordinate system $(\hat{x}, \hat{y})$ by:

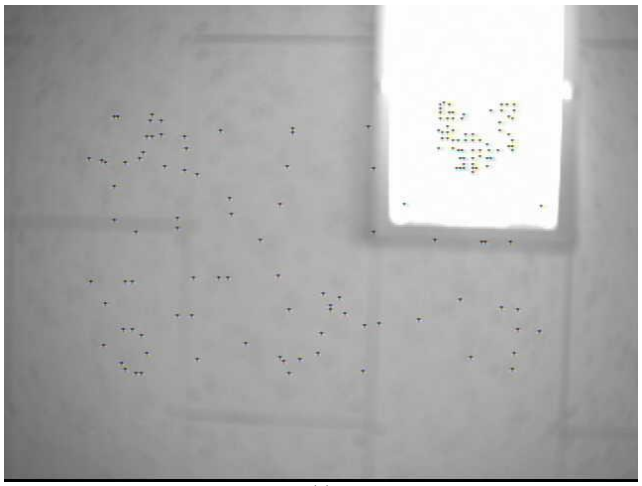
$$\begin{bmatrix} \hat{y} \\ \hat{x} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} y \\ x \end{bmatrix} \quad (1)$$

As shown in Figure 5(b), there are two new coordinates $(\hat{x}_a, \hat{y}_a)$ and $(\hat{x}_b, \hat{y}_b)$. According to the characteristics of the invariant properties [10], $(\hat{x}_a, \hat{y}_a)$ and $(\hat{x}_b, \hat{y}_b)$ are in the same coordinate system. In fact, the location change of the mobile robot is from $(\hat{x}_a, \hat{y}_a)$ to $(\hat{x}_b, \hat{y}_b)$. The relative displacement of the mobile robot can therefore be expressed simply as:

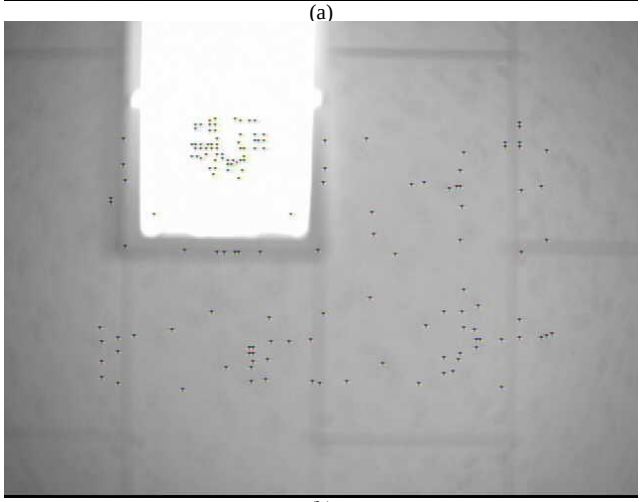
$$D = \sqrt{(\hat{x}_a - \hat{x}_b)^2 + (\hat{y}_a - \hat{y}_b)^2} \quad (2)$$

IV. THE INTERMEDIATE RESULTS

In this section we discuss some intermediate results. Figure 6 shows the extracted interest points results. The black dots represent the interest points. The captured image after camera moved is shown in Figure 6(b). Comparing the two images, most of the interest points are retained.



(a)



(b)

Figure 6. The extracted interest points simulation results: (a) before and (b) after the camera moved

 ISBN: 978-1-61208-224-0

Because the interest points being matched by the results from the DSP is not very intuitive, Figure 7 shows the results which were simulated using the same method in MatLAB.

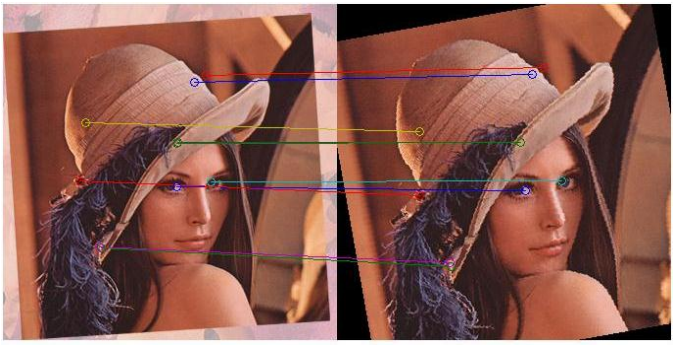


Figure 7. Descriptor extraction and matching using our improved algorithm in MatLAB

V. CONCLUSION

In this study, we used Non-minimum suppression to replace Non-maximum suppression in interest points extraction. As a result, we present a modified SURF algorithm used to extract the orientation and descriptor of the interest points. The simulation results verify the modified algorithm has good interest point matching results. In future work, we will write a program to verify the proposed self-localization estimating design in a DSP board. We will also address camera rotation in regards to the self-localization algorithm and verify it.

ACKNOWLEDGMENT

This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology under Grant 2010-0006588.

REFERENCES

[1] David C. K. Yuen and Bruce A. MacDonald: Vision-Based Localization Algorithm Based on Landmark Matching, Triangulation, Reconstruction, and Comparison, IEEE Transactions on Robotics, vol. 21, no. 2, pp. 217-226, April. 2005

[2] Andreja Kitanov, Sanjin Bišević, and Ivan Petrovíc, "Mobile robot self-localization in complex indoor environments using monocular vision and 3D model", IEEE/ASME international conference on Advanced intelligent mechatronics, pp. 1-6, 2007

[3] Alexander Koenig, Jens Kessler and Horst-Michael Gross: A Graph Matching Technique for an Appearance-based, visual SLAM-Approach using Rao-Blackwellized Particle Filters," IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1576-1581, 2008

[4] Kuan-Chieh Chen and Wen-Hsiang Tsai: "Vision-Based Autonomous Vehicle Guidance for Indoor Security Patrolling by a SIFT-Based Vehicle-Localization Technique", IEEE Transactions On Vehicular Technology, Vol. 59, No. 7, pp. 3261-3271, 2010

[5] De Xu, Liwei Han, Min Tan, and You Fu Li: "Ceiling-Based Visual Positioning for an Indoor Mobile Robot With Monocular Vision", IEEE Transactions On Industrial Electronics, Vol. 56, No. 5, pp. 1617-1628, 2009.

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool, "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding(CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

[7] Paul Viola and Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. I-511 – I-518, Kauai. HI. USA, 2001.

[8] Han Bing and B. Boyd, "Direct Replacement Algorithms of Fast Computing Integral Image in SURF", Journal of Projectiles, Rockets, Missiles and Guidance, Vol. 31, No. 3, pp. 211-15, 2011.

[9] Wang Jun-ben, LU Xuan-min and HE Zhao, "An Improved Algorithm of Image Registration Based on Fast Robust Features", Computer Engineering & Science, Vo1.33, No.2, pp. 112-118, 2011

[10] David G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints", International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004

[11] Xing Xiong and Byung-Jae Choi, "A Replacement Algorithm of Fast Computing Interest Point's Orientation and Descriptor in SURF", International Conference on Advances in Electrical and Electonicas Engineeing, pp. 336-338, Penang. Malaysia, 2012

 ISBN: 978-1-61208-224-0

Functional Roots and Manufacturing Tasks

Ingo Schwab
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
Ingo.Schwab@hs-karlsruhe.de

Norbert Link
Karlsruhe University of Applied Sciences
Karlsruhe, Germany
Norbert.Link@hs-karlsruhe.de

***Abstract*—Many production processes consist of repetitive, almost identical sub-processes. Process models are needed for state estimation and control purposes. Models are frequently formed from an analysis of input-output data relations of the overall process. For a repetitive process, the model of the repeated process is a functional root of the relation. Functional roots are introduced and symbolic approaches are presented. We propose to find functional roots via Symbolic Regression to model repetitive processes. As a first proof of principle we show the suitability of this approach with two basic and well-known problems in the scientific field of physics and nonlinear dynamics. The exact solutions of these problems are available from textbooks and can be used to assess the results of our approach. The first step in our project work therefore is to develop suitable concepts and technologies. The next steps will include analyzing real world data in cooperation with our project partners.**

Keywords- Symbolic Regression; Manufacturing; Functional Roots; Machine Learning.

I. INTRODUCTION

Many manufacturing tasks and processes are composed of a repetition of some simple process steps, since the necessary power of the repeated process must only be a fraction compared to the power needed in a single-step process. In fact, repeating manufacturing tasks represent an important group of manufacturing tasks and are of high practical relevance.

One of the problems is that manufacturing conditions restrict the observation of the material properties during the process, which therefore can often not be quantified. In such cases only the initial and final state of the material or work piece is known. The knowledge of the intermediate material qualities is mandatory for optimal process control. It is represented by a process model, which has to be established for the process under consideration.

There are several methods used to model the dynamics of nonlinear complex systems [1]. Conceptually, they can be split into two classes. The first class includes prior domain knowledge from human experts. For example numerical simulations like finite elements or phase field methods simulate the behavior of systems with domain knowledge from human experts. The second approach is to use phenomenological or general base function models which try to fit the observed behavior of the systems as good as possible. The latter approach includes many machine learning, data mining and statistical methods.

The second class can be further refined in modeling via symbolic [2] (e.g., general formula expressions) and subsymbolic (e.g., dedicated base function class, support vector machines or neural networks) representations. Symbolic learning representations can be interpreted by human domain experts and they can help to understand the process in a more formal way. Therefore this class does not only aim to model the system behavior. Sometimes the human experts are able to identify previously unknown facts of the observed process.

In contrast subsymbolic representations are black boxes. In most cases it is very difficult or impossible to interpret the behavior of the learnt representation. In our approach, we interpret mathematical formulas as one form of symbolic representation which can be used to gain additional insight into the system behavior.

The remaining part of the paper is organized as follows: In Section 2, we introduce the relation between industrial processes and functional roots. Section 3 gives a summary of the background and of related work. Additionally the proposed method is further described. Section 4 introduces the sample experiments and Section 5 the results of the method application. A summary is drawn in Section 6 with an outlook to future work.

II. INDUSTRIAL PROCESSES AND FUNCTIONAL ROOTS

One of our project tasks is to develop algorithms which are able to model the behavior of manufacturing processes. In the first steps we identified an important class of recurring problems which will be described in the following part of this subsection.

Technical processes like steel rolling or annealing are often recursive repetitions of some simple processes where the repeated application fulfills the original task. The main reason for such a recursive process is that the elementary process is much easier to handle. Figure 1 shows a schematic example of a steel mill. Stripes of metal are rolled in a sequence of up to seven identical stands where the task is to reduce their initial thickness of some centimeters down to some millimeters. That means that the resulting semi-

 ISBN: 978-1-61208-224-0

manufactured products of a subprocess are the input of the next almost identical sub-process. This continues until the target properties are reached.

In Figure 1, a block of steel with known property x_{in} is transformed by *n* stands to a stripe with the measurable property x_{out}.

The total process *F* can be modeled as a whole, but revealing a description of a single stand f_i is equivalent to computing functional roots of *F*. Intermediate values x_i are not accessible, but might be important to know for optimal process control.

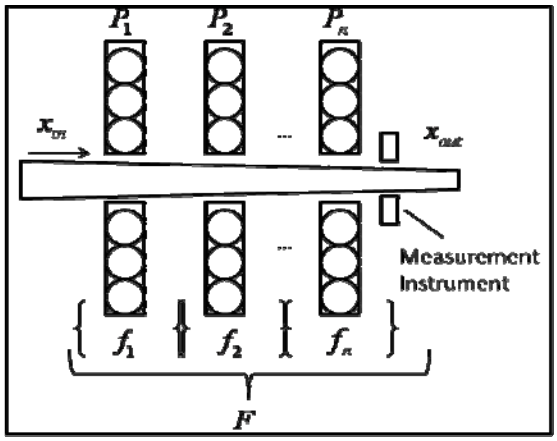


Figure 1. Model of a steel mill.

Due to technical reasons, it is impossible to measure some parameters like the profile of the stripes between the stands and the intermediate processing steps. However, this information is essential for optimal process control. Therefore, a model of a single stand can be generated from the measured values of the incoming and outgoing material and the fact that the transformation occurred in a number of identical steps. In [4], the whole process line is successfully modeled by a neural network. In [5], system identification of *F* and f_i is done with neural networks. The disadvantage of this method is that the results are subsymbolic and cannot be interpreted by a human expert.

III. BACKGROUND AND RELATED WORK

A. Functional Roots

In one sense, the concept of functional roots (aka iterated functions) is the inverse problem to the well-known compositions of a function with itself. The function *f(x)* is not known, but its composition with itself is given. For example, what is *f(x)* such that *f(f(x))=F(x)*, where *F(x)* is a given function. This question is an important part of the theory of functional equations and the areas of application appear in various fields such as computer science (e.g., recursions), dynamic systems or chaos theory. Little mathematical theory is known to find functional roots. It can be shown that functional roots of all orders exist for at least all continuous and strictly increasing real-valued functions [7]. Theoretical solutions for the problem do only exist for specific cases, such as monotonic functions. There is no formal way to find solutions for the general case.

Nevertheless, they have practical significance and few tools can solve them. Symbolic Regression is one solution method [8] and in this paper we present our first results.

Definition: Given an arbitrary function $F(x):\Re \to \Re$, the function *f(x)* with f(f(x))≡F(x) is called a functional or iterative root of F.

Higher order roots can be defined as $f^k(x) = f(f(...f(x)...)) \equiv F(x)$ and the function $f = F^{1/k}$ is a *k*-th iterative root of F.

Some simple examples are shown in Table 1.

Functional Root	Solution
$F(x) = x$	$f(x) = x$
$F(x) = x + 1$	$f(x) = x + \frac{1}{2}$
$F(x) = x^2$	$f(x) = \lvert x \rvert^{\sqrt{2}}$
$F(x) = x^4$	$f(x) = x^2$

Table 1. Functional roots.

To find a functional root to a problem seems on the first sight appealing because of its apparent simplicity and its natural idea. But, already the simple function $F(x) = x^2 - 2$ requires deep mathematical insight to be solved. In [8], it was shown that one analytical solution is

$f(x) = 2\cos(\sqrt{2}\cos^{-1}(\frac{x}{2}))$, which is not intuitive at first sight.

As a final remark, it should be mentioned that functional roots represent a universal concept and their use is not limited to the optimization of industrial processes. Applications range from data analysis to chaos theory.

B. Classical Regression Analysis and Symbolic Regression

Regression analysis [9] is one of the basic tools of scientific investigation enabling identification of functional relationship between independent and dependent variables. The general task of regression analysis is defined as identification of a functional relationship between the independent variables **x** = [x_1, x_2, … , x_n] and dependent variables **y** = [y_1, y_2, … , y_m], where *n* is a number of independent variables in each observation and *m* is a number of dependent variables.

The task is often reduced from an identification of a functional relationship *f*() to an identification of the

parameter values of a predefined (e.g., linear) function. That means that the structure of the function is predefined by a human expert and only the free parameters are adjusted. From this point of view Symbolic Regression goes much further.

Like other statistical and machine learning regression techniques Symbolic Regression also tries to fit observed and recorded experimental data. But unlike the well-known regression techniques in statistics and machine learning Symbolic Regression tries to identify an analytical mathematical description and it has more degrees of freedom in building it. A set of predefined (basic) operators is defined (e.g., add, multiply, sin, cos) and the algorithm is mostly free in concatenating them. Unlike the classical regression approaches which optimize the parameters of a predefined structure also the structure of the function is free and the algorithm both optimizes the parameters and the structure of the basis functions.

There are different ways to represent the solutions in Symbolic Regression. For example informal and formal grammars have been used in Genetic Programming to enhance the representation and the efficiency of a number of applications including Symbolic Regression [10].

Since Symbolic Regression operates on discrete representations of mathematical formulas non-standard optimization methods are needed to fit the data. The main idea of the algorithm is to focus the search on promising areas of the target space while abandoning unpromising solutions (see [3] for more details). In order to achieve this, the Symbolic Regression algorithm uses the main mechanisms of Genetic and Evolutionary Algorithms. In detail they are mutation, crossover and selection [6] and they are used to operate on an algebraic mathematical representation.

This representation is encoded in a tree [6] (see Figure 2). Both the parameters and the form of the equation are subject to search in the target space of all possible mathematical expressions of the tree.

In Symbolic Regression, many initially random symbolic equations compete to model experimental data in the most promising way. Promising are those solutions which are a good compromise between correct prediction quality of the experimental data and the length of the symbolic representation.

The operations are nodes in the tree (Figure 2 represents the formula 6x+2) and can be mathematical operations such as additions (add), multiplications (mul), abs, exp and others. The terminal values of the tree consist of the function's input variables and real numbers. The input variables are replaced by the values of the training data set.

Mutation in a symbolic expression can change the mathematical type of formula in different ways. For example a div is changed to add, the arguments of an operation changed (e.g., change 2*x to 3*x), delete an operation (e.g., change 2*x+1 to 2*x), or add an operation (e.g., change 2*x to 2*x+1).

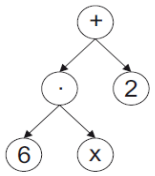


Figure 2. Tree representation of the equation 6x+2.

The fitness objective in Symbolic Regression, like in other machine learning and data mining mechanism, is to minimize the regression error on the training set. After an equation reaches a desired quality level of accuracy, the algorithm returns the best equation or a set of good solutions (the pareto front). In many cases the solution reflects the underlying principles of the observed system.

C. *Proposed Method*

In this article, we introduce an approach which uses Symbolic Regression to model the intermediate processing steps of manufacturing tasks. Mathematically, this is equivalent to the problem of computing iterative or functional roots: Given the equation $F(x)=f(f(x))$ and an arbitrary function $F(x)$ we seek a solution for $f(x)$. The major advantage of this approach is the interpretability of the identified solutions.

IV. SAMPLE EXPERIMENTS

In the following two subsections, we give a brief description of the two application scenarios of our first experiments. It should be noted that the next stage of our project is to evaluate the quality of the proposed methodologies on real-world data from industrial partners

A. *Free Fall*

The Free Fall textbook problem belongs to the elementary problems in physics and every first-year student in physics will probably be familiar with it. Nevertheless, we used it as starting point to gain a better understanding of the developed methodologies and functional roots.

In a nutshell, the free fall describes a vertical motion of an object falling a small distance close to the surface of a planet. It is a good approximation in air as long as the force of gravity on the object is much greater than the force of aerodynamic resistance, or equivalently the object's velocity is always much smaller than the stationary velocity.

B. *The Logistic Function*

A discrete map is the inverse to a functional root and is basically a sequence defined by the successive compositions of a function with itself. If, for example, we consider a function f from R to R, for each value in the domain we can define a sequence $(x, f(x), f^2(x),..., f^n(x))$, whereby

$f^k(x)$ describes the k times concatenation $f \circ f \circ f \ldots \circ f$.

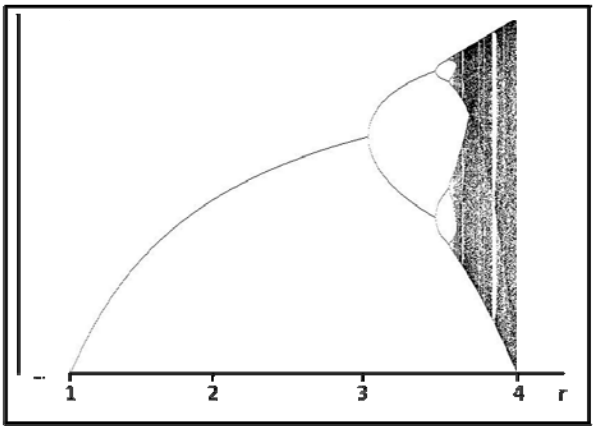


Figure 3. Bifurcation diagram.

There are many reasons why we may be interested in a sequence of this sort. For instance, the iteration of a suitable function can be successful in describing an event in the real world which is considered at discrete steps, such as the growth of a population of rabbits through its generations.
The Logistic Function is defined as

$$x_n = r x_{n-1}(1 - x_{n-1}), \text{ r} > 0. \quad (1)$$

It is easy to check that this is the equation of an upside-down parabola, which goes through the origin and the intercepts the x-axis at x = 1. This function is a good model of growing populations, but it has also peculiar mathematical properties.

The function has three different defined ranges with different behavior.
$0 \le r \le 1$: the function converges to 0.
$1 < r \le 3$: the function converges to the attractor 1 – 1/r.
$3 < r \le 4$: the function shows a periodic-doubling bifurcation. It starts with one attractor and approaches chaos via period doubling.

The logistic function is particularly interesting when $r > 2 + \sqrt{5}$. In this case, the dynamic system shows a deterministic chaotic behavior. That means that the system behavior is very sensitive to its initial conditions and infinitesimal variations for a dynamic system lead to large variations in behavior.

Figure 3 shows the Bifurcation or Feigenbaum diagram. The bifurcation parameter r is shown on the horizontal axis of the plot and the vertical axis shows the possible long-term population values of the logistic function. Only the stable solutions are shown here, there are many other unstable solutions which are not shown in this diagram. The bifurcation diagram shows the forking of the possible periods of stable orbits from 1 to 2 to 4 to 8 etc. Each of these bifurcation points is a period-doubling bifurcation.

V. EXPERIMENTS AND RESULTS

In our project, we have developed a Symbolic Regression framework. Additionally we adapted this algorithm to search for solutions for functional roots ($F(x)=f(f(x))$).
One of the main challenges posed in this paragraph is to modify algorithms to determine mathematical equations which are able to interpolate observed systems behavior. These data were measured at different points in time. In other words, we want to learn a function which is able to interpolate the dynamics of a system for nonlinear behavior.

A. Free Fall

As a starting point of our project we analyzed the well-known physical free-fall problem. The experiment setup is easy: An object is falling from attitude h_0 to h_1. On level h_0 it has the velocity v_0 and on level h_1 v_1. The starting velocity is varied and the resulting speed is measured on level h_1.
With knowledge of the necessary physical laws it is easy to find the correct answer. E.g., with knowledge of the energy theorem, attitude m and gravitation a the function is

$$\frac{1}{2}mv_0^2 + ma(h_1 - h_0) = \frac{1}{2}mv_1^2 \quad (2)$$

The task was to determine a formula which satisfies the following conditions for time

$$t_m = t_0 + \frac{\Delta t}{2}$$
$$(v_1, h_1) = g(v_m, h_m) = g(g(v_0, h_0)) = f(v_0, h_0) \quad (3)$$

Replacing g with the function:

$$v_m = v_0 + \frac{1}{2}a\Delta t$$
$$(v_m, h_m) = g(v_0, h_0) \text{ with } h_m = h_0 + \frac{1}{2}v_0\Delta t + \frac{1}{8}a\Delta t^2 \quad (4)$$

the iterated function is f and g is the iterated function of f.

In a first step we generated a training set of 40 learning examples.

Then we used the Symbolic Regression algorithm to search for the solution. The operation set contained addition, subtraction, multiplication, division, sine, cosine, exponential, logarithm function. The terminal values consisted of the function's input variables and real numbers.

The main task was to learn a functional root for this function. Several experiments showed that the developed Symbolic Regression system had no problem in finding the iterated function for this first sample experiment.

It was good starting point, but a more complex problem was needed.

B. The logistic function

The logistic function is defined as $x_n = r x_{n-1}(1 - x_{n-1})$.

 ISBN: 978-1-61208-224-0

At first, we generated a training set of 70 data sets. 4 times r was varied (see Figure 4). The vertical lines show the different r. In this first example each data set consists of a multitude of triples with x_{n-1}, x_n and r.

Figure 4. Training Data.

In a first step of our experiments, we tried to learn r with given x_{n-1} and x_n.

The algorithm started with the following operation set: addition, subtraction, multiplication, division, sine, cosine, exponential, logarithm functions. The terminal values consisted of the function's input variables and real numbers. As expected the system was able to detect the correct formula:

$$r = \frac{x_n}{x_{n-1} - (x_{n-1})^2} \tag{5}$$

The next experiment was to find a formula which is able to predict x_n without a given r. To solve the problem, it is not sufficient to make x_{n-1} available to the learning algorithm. Therefore, we added x_{n-2} (predecessor of x_{n-1}) to the data set und detected the formula which describes each point x_n of the Feigenbaum-Diagram with only two given points x_{n-1}, x_{n-2}:

$$x_n = \frac{x_{n-1}^3 - x_{n-1}^2}{x_{n-2}^2 - x_{n-2}} \tag{6}$$

Remarkably, this formula is able to describe every x_n with only two given data points x_{n-1}, x_{n-2} and without given *r*.

Functional root

The final experiment for the logistic function experiment was to determine the functional root of the logistic function with given *r*. Unlike in the former experiments, our algorithm was not able to find an exact analytical solution to this problem. But, experiments with a separated validation data set showed that they are good approximation to this problem.

Again, our Symbolic Regression algorithm was searching for the solution with the operation set of addition, subtraction, multiplication, division, sine, cosine, exponential, logarithm. The terminal values consisted of the function's input variables and real numbers.

Two runs of the Symbolic Regression algorithm found the following solutions:

$$\begin{aligned} &x_n = f(f(x_{n-1}, r), r) = \\ &1.0578035 * \sin(x_{n-1}) * \sqrt{r} * \cos(3.7038517 + x_{n-1}) \end{aligned} \tag{7}$$

$$\begin{aligned} &x_n = f(f(x_{n-1}, r), r) = \\ &0.30775887 + 0.42397907 * \sqrt{r} * \\ &\cos(1.9426149 + 2.3924117 * x_{n-1}) \end{aligned} \tag{8}$$

VI. CONCLUSIONS

In this paper, we address the task to find mathematical formulas to functional roots with Symbolic Regression. A practical real-world application is the interpolation of recursive repetitions of manufacturing tasks. This problem arises in many scientific fields but few existing tools can be used to find the functional root analytically or to analyze them. Our approach is applicable to arbitrary problems, and does not require deep mathematical insight into this research field. It is especially favorable for analyzing systems in which little expert knowledge is available.

In a first step of our project, we demonstrated the feasibility of this approach by two well-known problems. Based on the results from our Symbolic Regression analyses we found a solution for the logistic function which is able to predict the next time step with arbitrary and unknown *r* and only with two previous data measurements.

Our results show that Symbolic Regression is a suitable tool for modeling the dynamics of systems and to find functional roots for iterated processes of arbitrary behavior and dynamics.

REFERENCES

[1] R. O. Duda, P. E. Hart, D. G. Stork, "Pattern Classification", 2nd ed., Wiley Interscience, 2000.

[2] I. Schwab, N. Link, “Reusable Knowledge from Symbolic Regression Classification”, Genetic and Evolutionary Computing (ICGEC 2011), 2011.

[3] S. Choenni ,”Design and Implementation of a Genetic-Based Algorithm for Data Mining,” in VLDB 2000, pp. 33-42, 2000.

[4] L. Kindermann, A. Lewandowski, and P. Protzel, “A framework for solving functional equations with neural networks,” in Proc. Eighth Int'l Conf. on Neural Information

 ISBN: 978-1-61208-224-0

Processing - ICONIP'2001, Fudan University Press, Shanghai (2001), pp. 1075-1078.

[5] L. Kindermann, and P. Protzel, “Computing iterative roots with second order training methods,” Proc. of the International Joint Conference on Neural Networks (IJCNN'2001), Washington DC 2001, pp. 424-427.

[6] J. R. Koza, “Genetic Programming: On the Programming of Computers by Means of Natural Selection”. Cambridge, MA, USA: MIT Press, 1992.

[7] M. Kuczama, and B. Choczewski , R. Ger, “Iterative Functional Equations”. Cambridge University Press, Cambridge, 1990.

[8] M. Schmidt, and H. Lipson, “Solving Iterated Functions Using Genetic Programming”, Proc. of Genetic and Evolutionary Computation Conference (GECCO'09), Montreal, Canada.

[9] D. A. Freedman, “Statistical Models: Theory and Practice,” Cambridge University Press, 2005.

[10] M. O'Neill, C. Ryan, "Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language," Kluwer Academic Publishers, Dordrecht Netherlands, 2003.

 ISBN: 978-1-61208-224-0

Optimizing the End-to-End Opportunistic Resource Sharing using Social Mobility

Constandinos X. Mavromoustakis
Department of Computer Science,
University of Nicosia
46 Makedonitissas Avenue, P.O.Box 24005
1700 Nicosia, Cyprus
mavromoustakis.c@unic.ac.cy

***Abstract*—Opportunistic resource sharing, and contacts' interaction in opportunistic networks, faces several resource challenges that need to be faced via intelligent combinatorial practices. This work proposes a scheme which takes into account the non-synchronized motion of the devices in an urban area where an intelligent opportunistic socially oriented caching scheme is presented. The concept of social centrality is being introduced and modeled, which takes into consideration interactions among users. Through the proposed model the users' interactions can be exploited through time according to the contact frequency, in order to enable in an efficient way opportunistic resource sharing among mobile peers. The collaborative opportunistic communication with the proposed combined social-oriented model and the gossip-based replication scheme is thoroughly evaluated through experimental simulation, which takes measures for the end-to-end reliability of the resource sharing scheme. The proposed scheme enables efficient resource sharing via the social model and the evaluated interactions, minimizing at the same time, the delay variations between packets and maximizing the efficiency of resource exchange between mobile peers.**

Keywords-Temporal Social Metrics; Resource Exchange Scheme; Social Interaction Metrics; Opportunistic Communication Performance; Opportunistic Optimistic Replication.

I. INTRODUCTION

Today, wireless networks are used in many real-time applications that offer specialized services 'on-the-move', where these services require reliable communication and continuous end-to-end connectivity. When dealing with resource sharing applications, low latency and high responsiveness should be supported by the device in a decentralized way, whereas the ease of interaction, and the "always-on" connectivity should be engaged with users' demands and their requested data [1]. Current devices' limitations host many problems in the end-to-end communication, and therefore, are unable to handle resource sharing in a reliable manner. The challenging problems expressed by these devices are considered very disastrous and impair significantly the high responsiveness when devices face temporary and unannounced loss of network connectivity while they move, whereas they are usually engaged in rather short connection sessions since they need to discover other hosts in an ad-hoc manner where, then, the requested resources may not be available. Therefore, a mechanism that faces the intermittent connectivity problem and enables the devices to react to frequent changes in the environment, such as change of location or the context conditions, the variability of network connectivity, will affect significantly the end-to-end reliability and will face the unavailability and the scarceness of wireless resources. This work proposes a model for combining the resource sharing characteristics with the opportunistic content sharing procedure in order to offer higher reliability and availability of the requested resources. The interaction model that is introduced, and the social centrality principle [1] allow the users to share resources among devices when a shared contact rate threshold is satisfied, and devices follow a stationary non-synchronized motion while the connectivity is maintained. Social centrality is evaluated according to the usefulness of the location of a node in public areas. Usefulness is formed within a context of the connections allowed and the bandwidth served for a certain location. Opportunistic object/resource sharing [2] takes place in order to enable efficient dissemination of files chunks [3] whereas, the designed model guarantees the end-to-end connectivity maintenance, in a mobility-enabled cluster-based communication.

In this work, the proposed socially-oriented model for storing-and-forwarding for a certain time-cycle requested objects, utilizes the network resources (capacity and temporal connectivity) and enables high resource exchange. The high resource availability is as a result of the utilization of the social centrality principle as simulation results show. This work primarily addresses the problem of resource sharing in opportunistic systems and uses a constrained social caching mechanism. Through the proposed model the ability to accommodate in an adaptive way the requested data increases, whereas it enables a specified maximum number of concurrent users to share resources in a cluster according to social-interaction parameters of the users who are interacting. The model strengthens or weakens the resource exchange scheme according to the social contacts and the replication scheme exploited by the user's interaction parameters. In addition, different types of traffic can be supported where the adaptability and the robustness is shown by the proposed scheme.

The structure of this work is as follows: Section II describes the related work done and Section III follows presenting the proposed social-enabled mechanism for opportunistic resource sharing. Section IV presents the performance evaluation of the proposed scheme through simulation followed by Section V with the conclusions and foundations as well as potential future directions.

II. RELATED WORK

As location-based social networks have already appeared (i.e., Foursquare [3]) with great acceptance from the social community, different models were extracted in order to link

 ISBN: 978-1-61208-224-0

communicational problems and connectivity maintenance during a communication among peers. Mobility in autonomic communication is considered an essential parameter, where, along with the user's demands, they pose the vision of what self-behaving flexibility should encompass in next-generation self-tuning behavior of the devices[4]. The opportunistic communication aggravates the capacity of the nodes [2][5], where the requested information is being forwarded. Obviously, the need of modeling the social contacts behavior becomes timely nowadays since smartphones are now capable to process efficiently any requested information, whereas at the same time they can gather information from any hosted application (i.e., location aware or social contacts) in order to better utilize the network resources.

From the object sharing perspective, research has extensively proposed efficient architectures [6] [7] [8], which rely on local information and local devices' views, without considering the global networking context or views, which may be very useful for optimizing load balancing, enable adaptive routing, energy management, and even some self-behaving properties like self-organization. Mavromoustakis and Karatza in [9] propose the HyMIS scheme, which extends the advantages offered by the Hybrid Mobile Infostation System architecture, where the Primary Infostation (PI) is not static but can move according to the pathway(s) of the roadmaps. However, the HyMIS does not consider the social parameters-like the history contract rate and the temporal parameters of the users.

This work's contribution is to link the file sharing scheme with the underlying social parameters in order to optimize the efficiency of the resource sharing process. Event dissemination protocols use gossip to carry out multicasts. These gossips may be even more efficient in broadcasting information, if social parameters can be hosted and evaluated in a way that they affect the end-to-end resource sharing. Different caching approaches were used in the past, for enabling the requested data content to be available and discoverable [10] [11] at any time such that content can be discovered in a peer-to-peer manner without having network partitioning problems. Additionally if requested data was at some time window back available then through the proposed scheme we can keep an adaptive track of the resources and their availability. Mavromoustakis and Karatza in [12] consider the impact of impatience on optimal content dissemination scheme and a general model to capture this impact and show that under very general assumptions, the impatience function. However the contact relation and the history of the mobile peers is not yet explored due to the complexity and the dynamic nature that these environments impose. However selective and criteria-based dissemination procedures that take into consideration the social mobility and the social interactions in order to gather an allocation index for each ranked request by each peer, based on mobile nodes' content requirements is still a relatively unexplored area.

This work proposes an efficient way to optimize the end-to-end resource sharing reliability by enhancing the replications of the high ranked requested objects by users using a social-oriented methodology. The social-oriented model introduces the social centrality, and is utilized for selectively storing -for a certain time-cycle- the requested objects, whereas it considers the motion and movement characteristic of the devices for enabling optimized reliability, reduced traffic and generated overhead. The proposed scheme combines the strengths of both selective replication in opportunistic communication systems utilizing the outsourcing concept and attempts to fill the trade-offs between user's mobility, reliable file sharing and on-demand requested file availability limitation in the end-to-end path. Examination for the effectiveness of the proposed scheme is performed through simulation taking into consideration the offered reliability by the collaborative-social caching replication scheme within the mobility context. Thorough evaluations have been performed for the throughput optimization and the variation in the grade of robustness during the file sharing process among mobile peers, as well as for the throughput response.

III. PROBABILISTIC RANDOM WALK MOTION FOR EFFICIENT END-TO-END RESOURCE SHARING USING OPPORTUNISTIC SOCIALLY ORIENTED CACHING

Assuming that a source needs to send requested packets or stream of packets (file) to a destination where the destination moves from one location to another. This implies that, in a non-static multi-hop environment, there is a need to model the motion and the requested resources in the end-to-end path such that the resources can be efficiently shared among users, whereas any redundant transmissions and retransmissions are avoided. This work proposes a clustered-based mobility configuration scenario, which is set in Figure 1. Clusters enable the connectivity between nodes and the local (within a cluster) control of a specified area. On the contrary with [12][13] in this work a different mobility scenario is modeled and hosted in the scheme, which enables a parameterized feedback provision through the modeled scheme. Unlike the predetermined Landscape in [12], in this work, the mobility scenario used is Fractional Random Walk. The random walk mobility model was derived from the Brownian motion, which is a stochastic process that models random continuous motion [14]. In this model, a mobile node moves from its current location with a randomly selected speed in a randomly selected direction as real time mobile users act. However the real time mobility that the users express, can be defined by spotting out some environmental stimulating elements (adverts, cinema, shopping mall et.c) where users' decisions may be affected. In the proposed scenario the new speed and direction are both chosen from predefined ranges, $[v_{min}, v_{max}]$ and $[0, 2\pi)$, respectively [15]. The new speed and direction are maintained for an arbitrary length of time randomly chosen from$(0, t_{max}]$. At the end of the chosen time, the node makes a memoryless decision of a new random speed and direction. Figure 1(a) shows the scenario where the associations and the potential coverage area of a node is depicted. The movements are shown as a Fractional Random Walk (FRW) on a Weighted Graph.

Taking into consideration the movement of each device and by using the graph theoretical model, a device can perform random movements according to the topological graph G = (V,E) where it comprises of a pair of sets V (or V(G)) and E (or E(G)) called vertices (or nodes) and edges (or arcs), respectively, where the edges join different pairs of vertices. This work considers a connected graph with *n* nodes labeled *{1, 2, . . . , n}* in a cluster L^n with weight $w_{ij} \geq 0$ on the edge (i, j). If edge (i, j) does not exist, we set $w_{ij} = 0$.We assume that the graph is undirected so

that $w_{ij} = w_{ji}$. A particle walks from node to node in the graph in the following random walk/movement manner. Given that the device/particle is currently at node *i*, the next node *j* is chosen from among the neighbors of *i* with probability:

$$p_{ij}^{L} = \frac{w_{ij}}{\sum_{k} w_{ik}} \quad (1)$$

where in (1) above the p_{ij} is proportional to the weight of the edge (i, j). then the sum of the weights of all edges in the cluster *L* is:

$$w^{L}{}_{ij} = \sum_{i,j:j>1} w_{ij} \quad (1.1)$$

By using the motion notation, we can express the track of the requests as a function of the location (i.e. movements and updates p_{ij}^{L}) as: $R_i(I_{ij}, p_{ij}^{L})$ where R_i is the request from node i, I_{ij} is the interaction coefficient measured as in eq. 2. This work uses the representation of the interactions by utilizing notations of weighted graphs (equation 1). An example of social network is represented in Figure 1(a) where each node represents one person. The weights associated with each edge linking two persons (two devices) of the network are used to model the strength of the interactions between individuals [16]. The assumption made lays within the context that these weights are expressed as a measure of the strength of the social relations of the linking parts. Then the degree of social interaction between two people/devices can be expressed as a value in the range [0, 1]. These social interaction coefficients have a simplistic mean that, when 0 is expressed, it indicates that there is no interaction; whereas when 1 is expressed, it indicates that there is a strong social interaction. This aspect will affect the outsourcing degree of the requests in order to be available by other users in any cluster as 3.A's section show. Therefore, the connectivity of interactions in the network of Figure 1 can be represented by the 5×5 symmetric matrix (matrix is based on the population in the network and the hosted clusters), where the names of nodes correspond to both rows and columns and are ordered based on the interaction and connectivity. Matrix I_{ij} is referred to, as the Interaction Matrix. The generic element *i,j* represents the interaction between two individuals *i* and *j* where the diagonal elements represent the relationships that an individual has with himself and are set, conventionally, to 1. In (2), the I_{ij} represents all the links associated to a weight before applying the threshold values, which will indicate the stronger association between two individuals.

$$I_{ij} = \begin{bmatrix} 1 & 0.66 & 0.13 & 0.87 & 0 \\ 0.12 & 1 & 0.99 & 1 & 0.31 \\ 0 & 0.21 & 1 & 0.54 & 0.65 \\ 0.21 & 0 & 0 & 1 & 0.84 \\ 0 & 0 & 0.95 & 0.09 & 1 \end{bmatrix} \quad (2)$$

The threshold value is estimated according to the enhancement of the relation of the individuals as follows:

$$I_{ij} = \frac{I_{ij} + \Delta I}{1 + \Delta I} \quad (2.1)$$

where I_{ij} is the enhanced or weakened (if less than $\nabla I_{ij} = I_{ij_{(\tau)}} - I_{ij_{(\tau-1)}}$) association between two individuals and ΔI is the difference according to the previous I_{ij} association between *i* and *j*. Since while sharing resources time plays a major role, this work models a time-oriented enforcement of enabling an association to fade, i.e. if two individuals are not in contact for a prolonged time period. This association increases or reduces progressively with the time using the equation:

$$\Delta I_{ij} = \frac{a}{t_{age}} + b, \forall t_{age} < T_{R_L} \quad (2.2)$$

where t_{age} is the time that has passed since last contact and is measured until the individuals abandon the clustered plane L. *a* and *b* are proper constants[1] chosen by the designer of the network (*a*= 0.08, *b*=0.005). The proposed model encompasses the impact of the mobility on the interaction elements I_{ij} as the derived matrix consisting of the elements of $w^{L}{}_{ij}$ and I_{ij} as follows:

$$M_{ij} = I_{ij} \cdot p_{ij}^{L} \quad (2.3)$$

where the element w_{ij} derived from the p_{ij}^{L} matrix of the plane area L, is the likelihood of an individual to move from *i* to a certain direction to *j*, as Figure 1 shows.

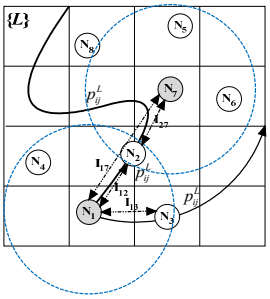


Figure 1. Inter-cluster communication and connectivity between nodes (within a cluster) with FRW model and social interactions (I).

A. Using social centrality for replicated object policy

One of the key tasks in wireless network analysis is determining the relative importance of individuals based on their positions, connectivity structure and motion through time. One concept in measuring and combining these aspects of the behavior of wireless networks is referred to as the centrality of individual devices with respect to the placement and behavior of each individual device in the cluster. Thus by using centrality approximation, a subset of the individuals in the network is sampled, and an induced subgraph consisting only of these individuals, and the links among them is produced, as a representative sample at time t. The centrality computation, then, is performed on this induced subgraph instead of the entire network, with the centrality scores of the sample being used as approximations. This work uses the social centrality as a measure

[1] Constant values for *a*= 0.08, *b*=0.005 are design parameters and were found to consist a suitable set (see [16] for more calculations on these estimations), based on the network's dimension at a certain time.

 ISBN: 978-1-61208-224-0

of the generic centrality of the system and the relative associations. Thus in the system with moving devices, a node that is directly connected with many nodes and has high I_{ij} can be considered as a high degree node. In other words, the lower degree nodes need the high degree nodes to serve as a bridge in order to connect with other lower degree nodes. According to the high degree nodes, the degree or centrality $D_c(aj)$ can be measured by: $D_c(aj) = \sum_{i=1}^{n} d(ai, aj)$,where $d(ai, aj) = \begin{cases} 1 \forall ai, aj \in D \\ 0 \forall ai, aj \notin D \end{cases}$, D denotes the direct connectivity. A maximum number of connected nodes for a certain graph is *n−1*. Therefore, the formula to calculate the centrality of the node by using the proportion of the number of adjacent nodes to the maximum number (n−1) is as follows:

$$D'_c(aj) = \frac{\sum_{i=1}^{n} d(ai, aj)}{n-1} \tag{3.1}$$

Centrality indicates the relative importance of a node in a network [17] and the relative contribution of this node to the communication process (in terms of duration and distance covered with the frequency, and parameterized in the context of avoiding communication partitioning problems). The social centrality is a relative measure of the betweenness centrality of two or more nodes. Social centrality is a type of centrality, that measures the number of times a node is chosen to host the 'best effort' parameters (in terms of storage, capacity and betweeness location) for time *t* in L, for, which requested data can be outsourced to this node. Therefore, a node with high social betweenness centrality β_{ai} can have a strong ability to interact with other nodes in the cluster *L*, and can be measured as:

$$\beta_{ai} = \frac{\sum_{1}^{j} P_{aj \to ak}}{\sum_{1}^{k} P_{ij} \forall P \in ai} \tag{3.2}$$

where $P_{aj \to ak}$ is the number of paths in the cluster via, which a requested object can be retrieved between the *aj* and *ak*, and P_{ij} is the number of paths in the cluster that include ai, $\forall P \in ai$. We introduce the social-oriented stability parameter $\sigma_C(t)$ for a time t, and is estimated as:

$$\sigma_C(t) = \left[\frac{R_{ij|t} \cdot (1 - norm(\beta_{ai})) \cdot N_{C(i \to j|t)}}{\inf(C_r) \cdot R_{C(t)}} \right] m_{ij}(t) \tag{3.3}$$

where R_{ij} is the normalized communication ping delays between *i* and *j* nodes at time t, β_{ai} is the normalized [0..1] social betweenness centrality showing the strong ability to interact with other nodes in the cluster *L*, $N_{C(i \to j|t)}$ is the successfully downloaded chunk capacity files over the total file capacity, C_r is the multi-hop channel's available capacity, $m_{ij}(t)$ is the interaction measures derived from eq. 2.3 at the time interval *t*, and $R_{C(t)}$ is the end-to-end delay in the cluster's pathway. The social-oriented stability parameter $\sigma_C(t)$ indicates the capability and transmitability of the node *i* to diffuse a certain requested object according to the ranked criteria of each requested object in *L* for a time *t*.

1) Ranking requested resources according to users' demands

In order to define which requested objects should be outsourced for being available for future requests, a ranking model has been applied as follows: To find the rank of an object $a1\ a2\ \ldots\ a_m$, one should find the number of objects preceding it. It can be found by the following function:

function rank($a_1, a_2, \ldots, a_m$ |*L*)// ranking the $a_1, a_2, ..$ in L cluster
rank← 1 ;
for *i* ← 1 **to** *m* **do**
while (*k has any neighbor with* a_i) ***do***
rank ← *rank* + *N*($a_1, a_2, \ldots, a_{i-1}$)

where the function above indicates which resources are highly demanded and are ranked according to these demands for the first *k-hop* nodes in the path.

2) Cluster merging and transfers' minimization

This work has utilized a cluster merging notation, where, if a resource is available from a nearby user in another Cluster accessible in a specified number of hops, a virtual merging mechanism has been enabled similar with the cluster merging in [18]. This means that by the utilization of the MinMax Cut algorithm proposed in [18], the cluster is partitioned in a way to make a distinction between pair-wise similarities with regards to the requested objects.

B. Optimistic intracluster outsourcing scheme using social interactions

In order to enable the proposed combination of the weights of the motion, with the interaction matrix and its elements of the matrix denoted as the interaction indicators we have modeled a mechanism for diffusing the requested high ranked resources to be outsourced. The diffusion policy used is using a resource sharing model, which reflects the impact of the mobility and the impact of the interaction indicators onto the proposed diffusion mechanism. Equation 4 shows the quantitative resource sharing approach by using the M_{ij} as a function of the contact rate and the number of users that are interacting (or not) in the cluster. The model is similar with the [12] in the sense that, the epidemiological model that is used in [12] is being replaced by the social interactions and the elements of the interaction indicators. We assume that a common lookup service is followed by all devices in the network cooperating via a shared platform. This model is providing feedback via the interactions performed M_{ij} and the number of users that are not interacting within time frame *t*. Suppose there are *u* hosts in the system, then a host is sharing a resource with *β(u−1)* other hosts per unit time. $\tilde{I}$ *(k−1)* are the number of users that are not interacting with (i,j) taking into account the threshold values, and/or are newly entered users in the cluster and have no relation/interaction with i,j. Therefore,

 ISBN: 978-1-61208-224-0

the diffusion transition rate for the resource sharing process taking into consideration the above, becomes:

$$\Phi_{ij} = M_{ij} \cdot \beta(u-1) \cdot \widetilde{I}_L \qquad (4)$$
$$\widetilde{I}_L = \frac{\bar{I}}{(u-1)}$$

where Φ_{ij} is the resource sharing from *i* to *j*, β is the contact rate, *u* are the hosts in the cluster as in [12]. It stands that all $u \in L$. The download rate can be measured as $\delta = M_{ij} \cdot \gamma \forall i, j \in L$, where γ is the supported transmission rate by the channels for downloading resources from *j* to *i* for time t. In order to avoid caching saturation [4] onto nodes, which are hosting the requested cached resources, this work uses a countermeasure to delete the requested resources from any node after time t. This is to clean up with redundant files the devices' storage units and memories. The delete or Purging Enforcement policy encompasses the Time-To-Leave duration of a node in the cluster, which is enforced by the motion weight w^L_{ij} as follows:

$$d_{TTL} = \frac{m_{ij}\tau}{C_{ij}} \cdot C_{ij}(\tau) \qquad (4.1)$$

where m_{ij} is the *i,j* element of the matrix of eq. 2.3, τ is the duration since last claim of the file from destination *j*, $C_{ij}(\tau)$ is the relative reserved capacity according to the channel's available capacity from *i* node to *j*, and C_{ij} is the total channel's capacity from node *i* to node *j* within the time duration τ .

IV. PERFORMANCE ANALYSIS THROUGH SIMULATION AND DISCUSSION

In the implementation-simulation of the proposed scenario, C/Objective programming language libraries were used as in [2]. The movement patterns generator was implemented to produce primarily traces for the ns-2 simulator [19]. All mobile nodes collaborate via a shared application that uses a distributed look-up service. Radio coverage is small compared to the area covered by all nodes, so that most nodes cannot contact each other directly. Additionally, we assume IEEE 802.11x as the underlying radio technology supporting the Cluster-based Routing Protocol (CRP). Queries regarding the resources that are available by peers for sharing, are generated dynamically and are selectively cached onto nodes according to Section III.A (eq. 3.4), where requested resources are ranked according to users' demands. The community is following the FRW on a Weighted Graph and consists of 250 users diffused in a landscape. The interaction of a particular user affects the contact rate and the interaction matrix entries of the next moment as real time users do and follows the measure of the equation 2.2. This association increases or reduces progressively with the time. Figure 2(a) shows the distribution of contacts' duration in *msec* with the Successful packet Delivery Ratio (SDR), whereas Figure 2(b) shows the SDR with the speed of each device comparing three different schemes.

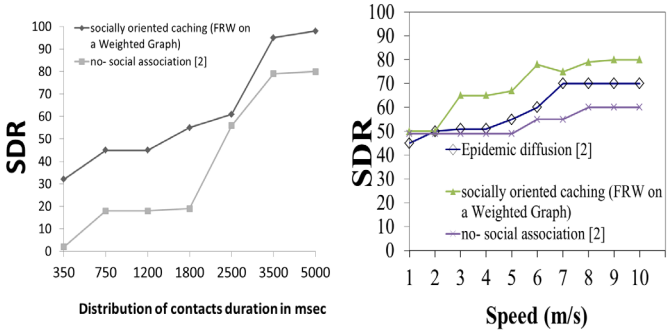


Figure 2(a). The Successful packet Delivery Ratio (SDR) with the distribution of contacts' duration in msec. Figure 2(b). The SDR with the speed of each device comparing three different schemes.

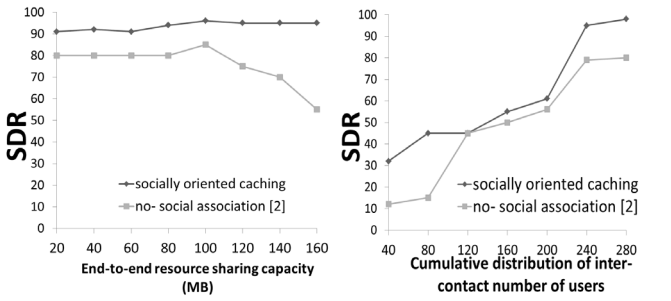


Figure 3(a)-(b). The Successful packet Delivery Ratio (SDR) with the End-to-end resource sharing capacity (MB) and Cumulative distribution of the number of users that are experiencing inter-contact.

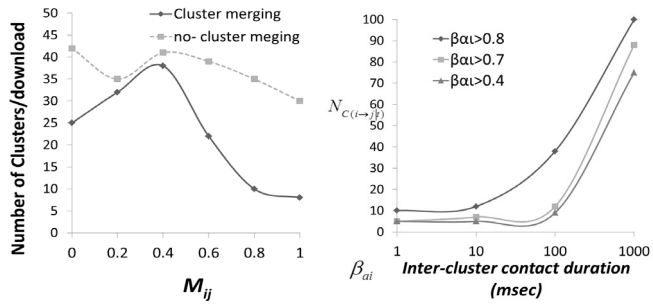


Figure 4(a)-(b). Number of Clusters/download with the Interaction-Mobility estimation of the M_{ij} between peers, and the social centrality values with the successfully downloaded chunk capacity files over the total file capacity with respect to the inter-cluster duration.

Figures 3(a) and (b) show the Successful packet Delivery Ratio (SDR) with the End-to-end resource sharing capacity (MB) and Cumulative distribution of the number of users that are experiencing inter-contact. Figure 3(b) shows that if users will outsource the requested resources using the interaction model, then the associated SDR increases significantly and the number of completed files are also increasing. Figure 4 shows the number of Clusters/download with the Interaction-Mobility estimation of the M_{ij} between peers. If the social centrality of the node increases, which means that the node has greater likelihood of its spatial and temporal location then the SDR increases and the inter-cluster contact time is reduced. Figures 5(a) and (b) show the total download time with the contact duration for complete downloads and the volume of the outsourced capacity with the contact duration for complete downloads. In Figure 6, the SDR with the number of delay-deadlined transmissions when a node requests resources is shown. By using socially-oriented caching the

 ISBN: 978-1-61208-224-0

SDR is kept high when the associated interaction parameters are greater than >0.4 and the social centrality parameter is >0.6.

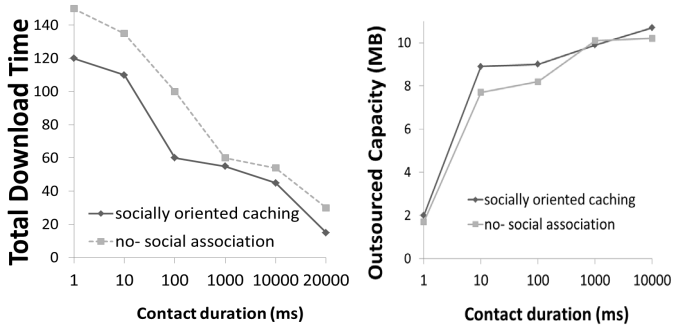


Figure 5(a)-(b). Total download time with the contact duration for complete downloads and the volume of the outsourced capacity with the contact duration for complete downloads.

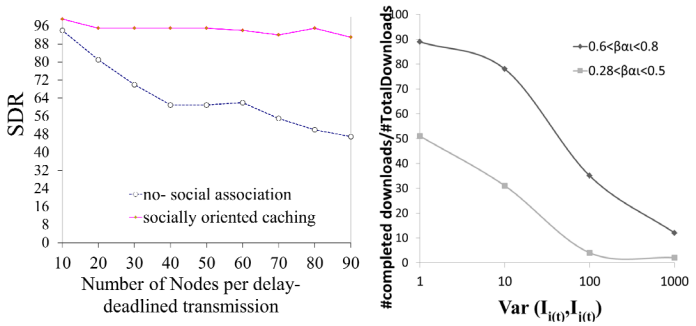


Figure 6(a). SDR with the number of nodes per transmissions with delay-deadlines. Figure 6(b). Successfully completed downloads with the interaction parameters and the social centrality measures.

V. CONCLUSIONS AND FURTHER RESEARCH

This work considers the probabilistic social interactions in order to assign available resources to communicating nodes according to a combined mobility model and the users' social relations. The collaborative resource sharing is achieved through the opportunistic socially-oriented caching model. Based on users' mobility and the associated probabilistic variation based on time and location, the social-based selective replication enables the cache-and-forward outsourcing model to fill the trade-offs between user's mobility, and reliable file sharing. The scheme outperforms from other existing schemes due to the social model which enables on-demand requested file availability. Examination for the effectiveness of the proposed scheme is performed through simulation taking into consideration the offered reliability by the collaborative-social caching replication scheme within the mobility context. Experimental results show that by introducing interaction parameters to mobile users while sharing resources on-the-move, the reliability increases significantly. Next steps and on-going work within the current research context will be the expansion of this model into a file sharing platform where on-the-move the users can share resources in real-time.

REFERENCES

[1] J. Tang, M. Musolesi, C. Mascolo, V. Latora, and V. Nicosia, Analysing Information Flows and Key Mediators through Temporal Centrality Metrics. In Proceedings of 3rd Workshop on Social Network Systems (SNS 2010). Paris, France. April 2010, pp. 256-262.

[2] C. X. Mavromoustakis, "Collaborative optimistic replication for efficient delay-sensitive MP2P streaming using community oriented neighboring feedback". The Eighth Annual IEEE International Conference on Pervasive Computing and Communications (PerCom 2010), Mannheim, Germany March 29-April 2, 2010.

[3] http://www.foursquare.com, last accessed Oct. 2011, pp. 105-110.

[4] C. X. Mavromoustakis and H. D. Karatza, "A Gossip-based optimistic replication for efficient delay-sensitive streaming using an interactive middleware support system", IEEE Systems Journal, IEEE USA, Vol. 4, no. 2, June 2010, pp. 253-264.

[5] S. Toumpis and A. Goldsmith, "Capacity regions for wireless ad hoc networks", IEEE Transactions on Wireless Communications, Vol. 2, No. 4, July 2003, pp. 736-748.

[6] D. Grobe Sachs, C. J. Hughes, S. V. Adve, D. L. Jones, R. H. Kravets, and K. Nahrstedt, "GRACE: A Hierarchical Adaptation Framework for Saving Energy", Computer Science, University of Illinois Technical Report UIUCDCS-R-2004-2409, Februrary 2004.

[7] D. Kliazovich and F. Granelli, "A Cross-layer Scheme for TCP Performance Improvement in Wireless LANs", Globecom 2004, IEEE Communications Society, pp. 841-844.

[8] M. Conti, G. Maselli, G. Turi, and S. Giordano "Cross layering in mobile Ad Hoc Network Design", IEEE Computer Society, February 2004, pp. 48-51.

[9] C. X. Mavromoustakis, and H. D. Karatza, "On the Performance Analysis of Recursive Data Replication Scheme for File Sharing in Mobile Peer-to-Peer Devices Using the HyMIS Scheme". IEEE International Parallel & Distributed Processing Symposium (IPDPS), Rhodes Island, Greece, April 25-29, 2006, pp. 46-53.

[10] Y. B. Ko and N. H. Vaidya. Flooding-based geocasting protocols for mobile ad hoc networks. Mobile Networks and Applications, 7(6), 2002, pp. 471–480.

[11] T. Hara. Effective replica allocation in ad hoc networks for improving data accessibility. In Proceedings of IEEE INFOCOM, IEEE Computer Society, 2001, pp. 1568–1576.

[12] C. X. Mavromoustakis and H. D. Karatza, "Under storage constraints of epidemic backup node selection using HyMIS architecture for data replication in mobile peer to peer networks" Journal of Systems and Software, Elsevier Volume 81, Issue 1, January 2008, pp. 100-112.

[13] C. Mavromoustakis and H. Karatza, "'Reliable File Sharing Scheme for Mobile Peer-to-Peer Users Using Epidemic Selective Caching'". Proceedings of IEEE International Conference on Pervasive Services (ICPS), Santorini, Greece, July 2005, pp. 169-177.

[14] T Camp, J. Boleng, and V. Davies, "A Survey of Mobility Models for Ad Hoc Network Research". Wireless Communication & Mobile Computing (WCMC): Special Issue on Mobile Ad Hoc Networking: Research Trends and Applications. Vol. 2 (5), pp. 483–502 (2002).

[15] G.F. Lawler, "Introduction to Stochastic Processes". Chapman & Hall. Probability Series, (1995).

[16] J. Scott, "Social Networks Analysis: A Handbook". Sage Publications, London, United Kingdom, second edition, 2000.

[17] Fei Hu, Ali Mostashari, and Jiang Xie, "Socio-Technical Networks: Science and Engineering Design", CRC Press; 1st edition (November 17, 2010), ISBN-10: 1439809801.

[18] C. Ding and H. Xiaofeng, Cluster Merging and Splitting in Hierarchical Clustering Algorithms, Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02), 2002, pp. 132-137.

[19] NS-2 Simulator, at http://www.isi.edu/nsnam/ns/, last accessed on 20/12/2011.

 ISBN: 978-1-61208-224-0

Primary Language for Semantic Computations and Communication without Syntax

Petro Gopych
Universal Power Systems USA-Ukraine LLC
Kharkiv, Ukraine
pmgopych@gmail.com, pmg@kharkov.com

***Abstract*—Recent binary signal detection theory (BSDT), extended by its infinity hypothesis (infinity of common prehistory of universe, life, and mind), is called extended BSDT. Its basic notions underlie BSDT primary language (PL, a hypothetical genuine mathematics used by animals for their internal computations). BSDT PL operates with meaningful words defined as finite binary affixes to infinite binary strings that have common infinite initial parts. In this paper, by an analysis of composite PL words (sentences), it has been demonstrated that meanings of their constituents can only conditionally be related. Meanings of composite words taken as a whole are perceived unambiguously and, under condition that communicated parties have common evolution history (respective infinite strings share their infinite initial part), make possible reliable, in particular non-syntactic, meaningful (semantic) communication. It has also been shown the BSDT neural network learning paradigm, "one-memory-trace-per-one-network", and super-Turing hyper-computations are the mandatory requirements for doing semantic computations and for unambiguous understanding of meanings of finite symbolic messages. Numerical and empirical evidences of some PL predictions and potential PL applications are briefly discussed.**

Keywords-context; infinity; meaning; subjectivity; categorization; complexity; super-Turing computations.

I. INTRODUCTION

Many researchers believe the problem of linguistic meaning cannot adequately be solved without solving the problem of consciousness. Taken together, recent binary signal detection theory, BSDT [1], and its atom of consciousness model, AOCM [2], give a chance of finding a new solution to the problem of meaning. The result is a primary language, PL [3], implementing John von Neumann's idea of a low-level "primary language *truly* used by the central nervous system," and structurally "essentially different of those languages to which our common experience refer" [4, p. 92]. BSDT PL [3] is thus a low-level language of symbols (spike patterns) probably used by the nervous system for its internal computations. It may also be used as a precursor to higher-level (including natural) languages that may be built with its help.

In this paper, basic BSDT PL's assumptions (new infinity hypothesis and meaning and subjectivity defined with its help) and some formalism details (computations with infinite binary strings that share infinite initial part) are briefly summarized. Within the PL framework for describing the meanings of components of composite words, the notion of conditional meaning for PL words of different meaning complexity [3] is introduced and on its ground their meaning ambiguity is discussed. For the first time, it will be explained in which way the rigid certainty of meanings of given-level (given meaning complexity) PL words and inevitable ambiguity of relations between meanings of different-level (of different meaning complexity) PL words ensure the richness of PL semantics, and how the PL provides a possibility of reliable *communication without syntax* (or even without any language at all) between animals of the same (or relative) species. For the success, it is needed to fulfill major requirement of BSDT PL infinity hypothesis – for different communicators infinite strings describing the meaning of a finite symbolic message must share their infinite initial part or, in other words, communicators must share significant part of their evo-devo history. Non-syntactic communication and communication without any language at all are the problems of great importance for linguistics, cognitive sciences, and artificial intelligence because their study informs us about the dynamics of language as a population phenomenon, bodily forms of signaling, and about a cognitive and bodily infrastructure for social interaction [5].

Computational, neuroscience, and psychological evidences of some PL predictions are discussed (the latter becomes possible because there are elements of psychology, i.e., meanings, in the background of PL mathematics). We conclude the network learning paradigm "one-memory-trace-per-one-network" and super-Turing hyper-computations are the mandatory requirements for successful semantic computations and for unambiguous understanding of the meanings of finite symbolic messages. It is claimed that, in living organisms (where meanings of communicated messages are crucially important), Turing and super-Turing computations are the everyday, routine, ubiquitous practice.

The paper consists of Sections I to VII and reference list.

II. BSDT PL INFINITY HYPOTHESIS

Extended BSDT, eBSDT, is the BSDT extended by new *infinity hypothesis* [2, 3] implying the infinity of common prehistory of the universe, life, the mind, language, and society and, according to which, main the eBSDT quantities – meanings of finite-in-length symbolic messages – are defined as infinite symbolic strings that have common "in the past" infinite initial parts. BSDT PL [3] and BSDT AOCM [2] are grounded on the eBSDT and closely related because *meanings* are interpreted as *subjective* experiences of respective feelings (qualia) and vice versa [2].

The leading idea is to equate a *real-world physical device* devoted to the recognition of particular meaningful

 ISBN: 978-1-61208-224-0

symbolic (binary for certainty) message originated from a thing of the world, complete binary *infinite on a semi-axis description* of the story of creation of this device in the course of evolution from the beginning of the world until now, and *the meaning* of the message under consideration [3]. Under the things of the world, we understand any inanimate objects, animate beings, and any relations between/within them. If in an arbitrary chosen place to split infinite on a semi-axis binary string just introduced into two parts then its finite and infinite fractions could be thought of respectively as *the name* of a thing and *the context* in which this name appears. Such a generalization allows the defining of an infinite but countable number of meaningful finite binary messages (PL words) that would name all the known (and unknown but conceivable) things of the world.

Because of the common co-evolution of all the things of the world meanings of their names are to be described at a given moment by different infinite strings with *common infinite initial parts*. The main feature of these strings is that, having at their ends different finite-in-length fractions, they share their infinite initial part (the text written in it does not matter here). Thanks to this property, the lengths of infinite PL meaning descriptions which differ, if they are indeed different, may explicitly be compared (Section III B). Since the meaning of a name is simultaneously the physical device designed to recognize exactly this name, name's meaning is the property of perceiving organism (sensory agent) and, through it only, of the thing whose name is currently under consideration. The same is also the reason why a name's meaning is simultaneously the animal's respective internal (psychological) state or its current *subjective* "first-person" experience or quale [2, 3].

In order for the AOCM/PL to be able to do *semantic computations* (i.e., to operate explicitly with meaningful strings of infinite length), specific *super-Turing* techniques and the implementation of real-world super-Turing physical devices are required. These are BSDT ASMs (abstract selectional machines [6]), AOCM/PL's building blocks devoted to processing separate meaningful messages or PL words/names. Because of our infinity hypothesis, the ASMs, AOCM, and PL should be based on notions that appeal to an extent to psychology (to meanings of names) and, for this reason, are *beyond* the scope of traditional mathematics.

III. ELEMENTS OF BSDT PL FORMALISM

A. Meaningless and Meaningful Words

All the conceivable *meaningful* PL expressions are defined as infinite spinlike (with components ±1) binary strings $c_{xi}x_j^i$ of the same infinite length, $l(c_{xi}x_j^i) = l(c_{x0}) = \aleph_0$ bits ($\aleph_0$, Georg Cantor's aleph) or of the same *meaning complexity* [3]. They constitute (are the members of) an ultimate or proper class S_{cx0} (the set of strings of the length $l(c_{x0})$ that is not a member of any other set [7], $c_{xi}x_j^i \in S_{cx0}$; the term "proper class" may intuitively be interpreted "as an accumulation of objects which must always remain in a state of development" [8, p. 325]). Given-level (Section III B) $c_{xi}x_j^i$ are uniquely specified (marked/labeled) by their right-most fractions, i-bit strings x_j^i (x_j^i is an affix added to the c_{xi}, c_{xi} is common infinite context for all the x_j^i of the length i with different arrangements of their ±1 components; $i = 0, 1, 2, \ldots$ and $j = 1, 2, \ldots, 2^i$). The number of elements (the cardinality) of the fraction of the S_{cx0} that comprises all the $c_{xi}x_j^i$ with x_j^i not longer than i bits is the sum $\sum 2^k = 2^{i+1} - 1$ ($k = 0, 1, \ldots, i$). Consequently, between naturals in their usual order and all the elements of the whole S_{cx0}, $c_{xi}x_j^i$, a one-to-one correspondence can be established. That means the S_{cx0} is *countable* and its cardinality, $|S_{cx0}|$, equals $\aleph_0$. On the other hand, $|S_{cx0}| = 2^{i+1} - 1$ with $i = \aleph_0$; that is, if the 1s that are inessential in this expression are omitted, it will be the famous formula for the size of the Cantor's continuum.

An affix x_j^i may simultaneously be treated either as the ijth i-length binary string, message, computer code/ algorithm, vector in i-dimensional binary space (i-BS), point in the i-BS, element of the set of 2^i points of the i-BS, PL word or PL name. Depending on the current context, these terms will further be used interchangeably.

A word/name x_j^i is the *meaningless* fraction of a meaningful string $c_{xi}x_j^i$; i.e. such a name gets its meaning from its context and from itself, $M(x_j^i) = c_{xi}x_j^i$. Different x_j^i specify all the conceivable strings $c_{xi}x_j^i$ and at the same time represent all the conceivable mathematical expressions as i-length binary strings – that is, they provide complete (non-Gödelian) arithmetization of these expressions by ordinals/naturals (x_j^i may be treated as ordinals/naturals written down in binary notations). For this reason, x_j^i are also the ijth eBSDT Gödel's numbers, $G_{ij}^x = x_j^i$, enumerating themselves and meaningful strings $c_{xi}x_j^i$. The same x_j^i are also the ijth partial Gregory Chaitin's Ω, $\Omega_{ij}^x = x_j^i$ (the ijth halting probabilities [9] for arbitrary binary computer codes not longer than i bits running on the ijth Chaitin's self-delimiting computers or, we hypothesize, on respective BSDT ASMs). x_j^i as well as G_{ij}^x and Ω_{ij}^x are *random* and *incomputable* because they are randomly selected from 2^i different binary i-lengh strings and, then, assigned to things to be named [3]. In other words, x_j^i, G_{ij}^x, and Ω_{ij}^x provide irreducible descriptions (specifications) of these things. The totality of given-level (Section III B) values of x_j^i, Ω_{ij}^x, or G_{ij}^x is the totality of given point-of-view *irreducible* descriptions of all the things of the known world [3].

If string variable x^i consists of variables u^p and v^q then $x^i = u^p v^q$, $i = p + q$; x^i is a string template of i empty cells needed to produce the strings x_j^i by filling these cells in +1s and −1s; $u^p v^q$ is a concatenation of u^p and v^q. The values of variables x^i, u^p, and v^q are respectively the strings x_j^i, u_r^p, and v_s^q that are the members of sets S_{xi}, S_{up}, and S_{vq} whose cardinalities are respectively $|S_{xi}| = 2^i$, $|S_{up}| = 2^p$, and $|S_{vq}| = 2^q$; if $p \le q \le i$, $S_{up} \subseteq S_{vq} \subseteq S_{xi}$. Composite set/space S_{xi} may also be interpreted as either the S_{up} whose vectors are colored in 2^q colors or the S_{vq} whose vectors are colored in 2^p colors. If so, p and q are the measures of discrete "colored" non-localities of vectors in spaces S_{vq} and S_{up}, respectively [3]. Three-dimensional blue-and-red binary space ("colored Boolean cube") has earlier independently been used for representing the Boolean functions of one-dimensional cell automata, e.g., [10, ch. 6]. The rainbow of colors in finite-dimensional binary spaces here discussed is a direct generalization [3] of the two-color case [11].

 ISBN: 978-1-61208-224-0

B. Categories, Meaningful Words of Different Levels

The form $C(x^i) = c_{xi}x^i$ (it is a concatenation of string c_{xi} and string template x^i) defines a *category* (notion or concept) of meaningful names $c_{xi}x_j^i$. If to *dynamically* fix p left-most components of an x_j^i as a particular u_r^p, then $c_{xi}x_j^i = c_{xi}(u_r^p v_s^q) = (c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ where $c_{xi} = c_{up}$, $x_j^i = u_r^p v_s^q$ (i.e., x_j^i is a composite string), $c_{vq} = c_{up}u_r^p$, $x_j^i \in S_{xi}$, $u_r^p \in S_{up}$, and $v_s^q \in S_{vq}$. Infinite strings $c_{up}u_r^p \in S_{cu0}$ and $c_{vq}v_s^q \in S_{cv0}$ are the members of different ultimate classes, S_{cu0} and S_{cv0}, but, because of our infinity hypothesis implying that $c_{xi} = c_{up}$, the lengths of $c_{up}u_r^p$ and $c_{vq}v_s^q$ are comparable and the former is $l(c_{xi}x_j^i) - l(c_{up}u_r^p) = i - p = q > 0$ bits shorter (has smaller meaning complexity) than the latter (note, $S_{cv0} = S_{cx0}$ and $l(c_{xi}x_j^i) = l(c_{vq}v_s^q)$; infinite words [12] of automata theory have no such properties and remain within the framework of traditional mathematics). The form $C(v^q) = (c_{up}u_r^p)v^q = c_{vq}v^q$ with values $(c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ provides a temporal sub-categorization of members of the category $C(x^i) = c_{xi}x^i$. Since, under condition $c_{xi} = c_{up}$, $c_{xi}x_j^i$ and $c_{up}u_r^p$ are of different lengths (have different meaning complexities), we refer to their affixes as names of different *levels*: the level of x_j^i is zero, the level of u_r^p (if it is a fraction of x_j^i) is $q = i - p$, i.e. the number of "ignored" bits that differentiate the length of $c_{up}u_r^p$ from the length of $c_{xi}x_j^i$ (q also defines discrete colored 2^q-state non-locality of u_r^p; as $l(c_{xi}x_j^i) = l(c_{vq}v_s^q)$, v_s^q is also a zero-level name). *Only zero-level names get definite meanings* (see Fig. 1), namely x_j^i in $c_{xi}x_j^i$ or v_s^q in $(c_{up}u_r^p)v_s^q$; the string u_r^p has no definite meaning in $(c_{up}u_r^p)v_s^q$ but it gets a strictly defined meaning as a right-most (zero-level) fraction of $c_{ur}u_r^p$ (if u_r^p is not a fraction of any composite string). A right-most (zero-level) item of a composite meaningful name is called the "*focal*" item (it occupies dynamically created "focus of attention"), a composite name's non-focal item produces a focal name's "*fringe*" (by analogy with fringes of memory and consciousness [2, 3]) or its short-range immediate context.

All the PL's meaningful strings are defined on a semi-axis (i.e. they are "one-side infinite"), have the lengths $\aleph_0$ bits (i.e. they are countable), and must *always* be arranged in a way when they *share their infinite initial part,* the length of which is again $\aleph_0$. Since infinite meaningful strings are arranged in such a way, their beginnings (bit-by-bit common infinite initial part) are always the same but their end-points may not coincide and one of these strings may in general be a number of bits longer or shorter than the other (in other words, they may have larger or smaller meaning complexity [3]). Therefore the strings that are of the same infinite length in the sense of Cantor (that are countable) may be of different infinite length (meaning complexity) in the sense of the BSDT PL. The level of a PL name is the measure of such a difference or the *relative* measure of complexity of meanings; absolute measure (the length of a meaningful string taken separately) is useless for comparing meaning complexities because all meaningful strings taken separately have the same length, $\aleph_0$. Consequently, the notions of a name's meaning complexity and a name's level (relative measure of its meaning complexity) exist in the framework of the BSDT PL only and have their roots in its infinity hypothesis. Meaning complexity embraces given the context Shannon-type ensemble complexity (the length of x_j^i in bits) specifying a name's statistical properties and Kolmogorov-type algorithmic complexity (the length in bits, $\aleph_0$, of computer program, $c_{xi}x_j^i$, that gives complete irreducible infinite description of the ASM that selects the x_j^i) specifying the complexity of devices selecting the names of given ensemble complexity [3].

Composite names $x_j^i = u_r^p v_s^q$, the values of $x^i = u^p v^q$, are thought of as PL *sentences*. If so, the value of a focal string variable, e.g. v_s^q, corresponds to a sentence's feature/ attribute that is currently in the focus of attention; its fringe, e.g. u_r^p, is the fringe of an animal's memory or consciousness. A composite name's "holophrasical" presentation, e.g. x_j^i, corresponds to the perception/ understanding of a sentence as a whole whereas its serial presentation (e.g., a sequence of v_s^q with $1 \le q \le i$) represents the sentence's serial perception/understanding as a sequence of its meaningful fractions or "words". Given the context, different zero-level names are *synonyms* naming the same thing in different ways (e.g., x_j^i is one of 2^i synonyms defined given the context c_{xi} and v_s^q is one of 2^q synonyms defined given the context c_{vq}); if $c_{vq} = c_{xi}u_r^p$ (i.e., if v_s^q is a focal fraction of compound name $x_j^i = u_r^p v_s^q$), both types of synonyms describe the same thing but of different points of view (grounds for the understanding). Any paraphrase of PL sentences (other choice of their "focal" fractions) cannot change their whole meanings and in that sense BSDT PL lacks "compositional semantics" [3].

As composite words are treated as PL sentences, the set of rules defining the relations of meaning of a composite word to meanings of its constituents represent the PL syntax. If internal structure of PL words/sentences is ignored and their meanings are only perceived as a whole then communication with their help do not appeal to PL syntax and, consequently, is carried out *without syntax*.

C. Meanings, Subjective Experiencies (Qualia) and Truths

Given the c_{xi}, each string x_j^i is selected by its BSDT ASM(x_j^i) intentionally designed in the course of evolution and tuned in the course of its individual development exactly for this purpose [6]. An infinite, symbolically-written, complete description of evo-devo prehistory of designing this real-world physical ASM(x_j^i) is the *explicit meaning* of x_j^i, $M_{\text{expl}}(x_j^i) = c_{xi}x_j^i$. The running of ASM($x_j^i$) itself in its real-world physical form is the *implicit meaning* of x_j^i, $M_{\text{impl}}(x_j^i)$, or internal, "mental" or psychological representation of the thing named by the x_j^i [2, 3]. As $M(x_j^i) = M_{\text{expl}}(x_j^i) = M_{\text{impl}}(x_j^i)$, the meaning of x_j^i given c_{xi} is animal's being in a specific psychological state which is a "quale" (subjective "first-person"/private experience or feeling) of this meaning [2, 3]. In particular, the meaning of a category of names is a set of respective qualia.

The name x_j^i is true if its meaning, $M(x_j^i) = c_{xi}x_j^i$, is true or, in other words, if strings c_{xi} and x_j^i are correctly adjoined to each other. If there is no such correct correspondence, meaningful name is false. Since the cardinality of S_{cx0}, $|S_{cx0}| = \aleph_0$, is infinite, the number of PL truths is also potentially infinite and, for any meaningful string, its truth value

 ISBN: 978-1-61208-224-0

$T(c_{xi}x_j^i)$ certainly exists (it is either "true" or "false"). Each true meaningful name, e.g. $c_{xi}x_j^i$, names by definition the *i*th *real-world* thing given to an animal through its *ij*th psychological state or, in other words, through the activity of physically implemented real-world ASM(x_j^i) [2, 3]. Thus, for meaningful names, the truth is the norm and the falsity is an anomaly caused, e.g., by an animal's dysfunction or disease. In any case, there is *no* lie and *no* liar paradox – a source of Kurt Gödel's incompleteness which does not hold for PL *meaningful zero-level* names, $c_{xi}x_j^i$. This inference is caused by the fact that PL name meanings are always the ones that animals/humans *actually* keep in mind. It is also the reason why BSDT PL works so well as a primary language (to survive, an animal does not lie to itself) [3].

As truth values $T(c_{xi}x_j^i)$ are never communicated together with x_j^i, they should always be discovered in the process of decoding (*understanding*) the received names x_j^i and confirmed by checking their correspondence to the reality or, more directly, to an animal's respective psychological state. At the same time, a zero-level name's fringe items (names), due to their non-locality, have no meanings but only *conditional meanings* (Section IV and Fig. 1). Hence, for PL names of different levels (meaning complexities) relations between their meanings remain fundamentally ambiguous. This vagueness is a BSDT PL counterpart to Gödel's incompleteness (axioms, theorems, and meta-mathematical expressions for which Gödel's results hold are, in our terms, an infinite fraction of infinite in number *meaningless* strings x_j^i) [2,3].

IV. BSDT PL MEANING AMBIGUITY

It is assumed that, in a meaningful string $c_{xi}x_j^i$, its context c_{xi} and its name x_j^i describe respectively the *static* part of the ASM(x_j^i) selecting the x_j^i (its "hardware" already fixed in the course of evolution) and the *dynamic* part of the ASM(x_j^i) (its "software" designed in the course of the hardware's adaptive learning and development). The length of x_j^i in bits, *i*, defines the number of now essential (explicitly considered) features of the *i*th thing named by the x_j^i; the *j*th arrangement of ±1 components of x_j^i is the *j*th PL description of this *i*th thing (e.g., the value +1 or –1 of a component of the x_j^i may mean that the respective feature is included to, +1, or excluded from, –1, the consideration). The complexity of meaning of the name x_j^i reflects the meaning complexity of the physically implemented real-world ASM(x_j^i), not the complexity of the thing named by x_j^i.

If $x_j^i = u_r^p v_s^q$, strings $c_{up}u_r^p$ and $(c_{up}u_r^p)v_s^q = c_{vq}v_s^q$ describe given the context, $c_{xi} = c_{up}$, an ASM(u_r^p) and ASM(v_s^q) that may for a time period dynamically be created from the ASM(x_j^i) that in turn is the product of a similar process described by the string $c_{xi}x_j^i$. ASM(u_r^p) and ASM(v_s^q) are "virtual" ASMs (i.e. temporally designed for) selecting the names u_r^p and v_s^q of the *p*th and the *q*th "virtual" things (i.e. of temporally highlighted/allocated fractions of the *i*th composite thing named by its *ij*th composite name x_j^i); in other words, virtual ASMs highlight the *pr*th and *qs*th "partial" meaningful fractions of the *ij*th description of the *i*th thing. Composite names essentially enrich the PL semantics but raise the problem of comparing the meanings of names selected by ASM(u_r^p), ASM(v_s^q), and ASM(x_j^i).

Zero-level names x_j^i and v_s^q (v_s^q is a part of $x_j^i = u_r^p v_s^q$) name given the context *the same* thing in the same way but from different points of view defined by their contexts (static for x_j^i, c_{xi}, and in part dynamically created for v_s^q, $c_{vq} = c_{up}u_r^p$; Fig. 1(a)). v_s^q is selected by the ASM(v_s^q) that is "virtual" with respect to the ASM(x_j^i), here $c_{xi} = c_{up}$ and for x_j^i and v_s^q their common infinite context is c_{xi}. Thus, ASM(x_j^i) can temporally serve as ASM(v_s^q) but in any case the same thing is under the consideration and the meaning of x_j^i, $M(x_j^i) = c_{xi}x_j^i$, and the meaning of v_s^q, $M(v_s^q) = (c_{up}u_r^p)v_s^q = c_{vq}v_s^q$, may unambiguously be related ($c_{vq}v_s^q$ is simply a variant of $c_{xi}x_j^i$).

If, given the context, $c_{xi} = c_{up}$, names x_j^i and u_r^p are both at the level of zero, then their meanings are to be of *different* proper classes and should have *different* meaning complexities (meaning complexity of x_j^i is $l(c_{xi}x_j^i) - l(c_{up}u_r^p) = i - p = q$ bits larger than that of u_r^p; see Fig. 1(a) and (c)). This means they describe *different* things from the same point of view or the same thing at *different* stages of its evolution. The names x_j^i (Fig. 1(a)) and u_r^p (Fig. 1(c)) are respectively selected by present-stage-of-evolution ASM(x_j^i) and *q*-stages-back-in-evolution ASM(u_r^p) and refer to animals of evolutionary different species. Meaningful string $c_{up}u_r^p$ and respective part of $c_{xi}x_j^i = (c_{up}u_r^p)v_s^q$ may coincide bit by bit but even in this case meanings of x_j^i and u_r^p may only *conditionally* be related to each other and 2^q additional conditions (strings v_s^q in Fig. 1(a)) are required to uniquely establish their correspondence.

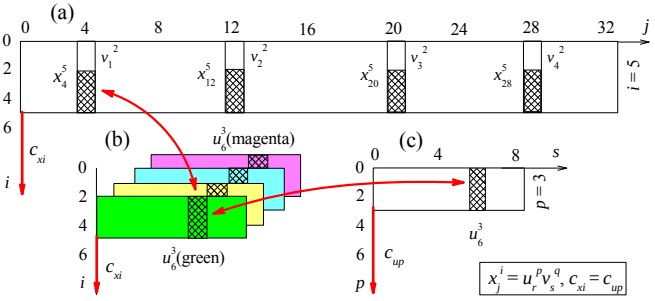


Figure 1. Comparing given the context different-level BSDT PL meaningful names of different proper classes: (a) zero-level names, (b) colored zero-level names corresponding to names in (a), (c) zero-level names that are predecessors to names in (a) and counterparts to names in (b).

If u_r^p is a *q*-level fringe of zero-level focal string v_s^q and they are both fractions of $x_j^i = u_r^p v_s^q$ (Fig. 1(a)) then u_r^p has *no* meaning (Section III B). But it could get a definite *conditional meaning* if one supposes that u_r^p is conditioned by the color of a zero-level name $u_r^p(color)$ selected by a respective *q*-stages-back-in-evolution ASM (zero-level x_j^i in Fig. 1(a) are unambiguously related to zero-level names $u_r^p(color)$ in Fig. 1(b)). $u_r^p(color)$ and u_r^p conditioned by one of *q* colors, though, have conditional but certain meanings. But once colors are deleted (only uncolored strings are used in computations) one-to-one correspondence between x_j^i and $u_r^p(color)$ disappears and, instead of it, we obtain 2^q-state uncertainty between the x_j^i and u_r^p and between the meaning

 ISBN: 978-1-61208-224-0

of x_j^i and conditional meaning of u_r^p (Fig. 1). The origin of the conditioned relationships just explained between meanings of names of different meaning complexities is the properties of ultimate/proper classes caused in turn by BSDT infinity hypothesis. Of this follows that famous Burali-Forti paradox "there can be two transfinite (ordinal) numbers, *a* and *b*, such that *a* neither equal to, greater than, nor smaller than *b*" [13, p. 157] means in our terms that meanings of PL names, whose meaning complexities differ in *q* bits, can only be compared with 2^q-state uncertainty (Fig. 1).

In Fig. 1, panel (a) demonstrates zero-level names x_j^i and v_s^q of meaningful strings $c_{xi}x_j^i$ and $(c_{xi}u_r^p)v_s^q$ ($x_j^i = u_r^p v_s^q$; $i = 5$, $p = 3$, and $q = i - p = 2$); the rectangle has the height i and the width $|S_{xi}| = 2^i = 32$, the ijth bar of the height i in the jth horizontal position designates the name $x_j^i = G_{ij}^x = \Omega_{ij}^x$; bars $x_4^5 = u_6^3 v_1^2$, $x_{12}^5 = u_6^3 v_2^2$, $x_{20}^5 = u_6^3 v_3^2$ and $x_{28}^5 = u_6^3 v_4^2$ that correspond to four colored highlighted bars in (b) are also highlighted (u_6^3 is q-level fringe of zero-level v_s^q that is a focal fraction of x_j^i); substrings v_1^2, v_2^2, v_3^2, and v_4^2 may encode the colors of colored strings u_r^p(*color*) in (b). Panel (b) shows conditioned zero-level names u_r^p(*color*) = G_r^p(*color*) = Ω_r^p(*color*) corresponding to names x_j^i in (a) (the word *color* is a parameter, names u_r^p(*color*) are selected by conditioned q-stages-back-in-evolution ASM and conditionally name the things unconditionally named by the x_j^i); equal-in-size rectangles colored in $|S_{vq}| = 2^q = 4$ colors consist of $|S_{up}| = 2^p = 8$ bars of the height p; uncolored bars in (a) and respective colored bars in (b) (e.g. u_6^3(*green*) and x_4^5) denote different descriptions of the same thing. Panel (c) displays uncolored zero-level (focal) names u_r^p of meaningful strings $c_{up}u_r^p$ (they name evolutionary predecessors of the thing named by the x_j^i); the prth bar of the height p in the rth horizontal position (it is shaded) designates $u_r^p = G_{pr}^u = \Omega_{pr}^u$ for the case $u_6^3 = G_{3,6}^u = \Omega_{3,6}^u$. In (a), (b), and (c), strings that are numerically equivalent to the u_6^3 are shaded in the same way; contexts (they are shown as bold arrows) are equal to each other bit by bit, $c_{xi} = c_{up}$. Uncolored and colored names name real-world and conditioned ("virtual") things, respectively. A bijection, $x_j^i \leftrightarrow u_r^p$(*color*), exists between names in (a) and names in (b); it may be e.g. $x_{28}^5 \leftrightarrow u_6^3$(*magenta*) or $x_4^5 \leftrightarrow u_6^3$(*green*). A bijection also exists from names u_r^p in (c) to given-color names u_r^p(*color*) in (b), e.g. $u_r^p \leftrightarrow u_r^p$(*green*) (once it is established, other conceivable bijections, e.g. $u_r^p \leftrightarrow u_r^p$(*magenta*), become impossible). If colors are deleted, these bijections (they are indicated as curved bidirectional arrows) disappear producing, instead of 2^q-state (4-state in (b)) discrete colored non-locality of vectors u_r^p, 2^q-state (4-state in (b)) uncertainty (degeneracy) of meaning relations between names in (a) and (b), in (b) and (c), and in (a) and (c). Infinite strings $c_{xi}x_j^i$ and $c_{up}u_r^p$ are like Burali-Forti's "transfinite ordinals" *a* and *b* mentioned above.

V. NUMERICAL AND EMPIRICAL BSDT PL VALIDATIONS

Semantic computations produce meaningful results of meaningful data. BSDT PL computations are exactly of this type because we imply that the meaning (infinite context) of any finite mathematical expression is always taken into account when any formal operations (computations) are being done on it. For this reason, elements of psychology (meanings) are always involved in semantic computations and their completely formal (i.e. independent on meanings) descriptions become strictly speaking impossible. Thanks to this fact it does become possible to verify the methods of proposed PL mathematics by methods of psychology and neuroscience or, in other words, by comparing PL computations with internal computational mechanisms that are actually in live animals/humans.

A. Solving Communication Paradox

In Section III B and Section IV we saw that only zero-level names whose internal structure (the manifold of their possible focal and fringe constituents) is ignored or, in other words, only those PL sentences that are presented *without syntax* and perceived "holophrasically" have given context unambiguous meanings. This fact and the fact that meanings of PL meaningful names are the ones that animals/humans *actually* keep in mind [3] make the BSDT PL an appropriate tool for the description of communication without syntax (or without any language at all) that is typical for animals and human infants, e.g. [5] and references therein. We hypothesize: communication without syntax (it is exhibited as an animal's basic/inherent behaviors that truly reflect its respective inner states or behavioristic part of animal's cognition) suffices to support *the simplest* animal sociality.

Since complete meaningful descriptions of PL names, $c_{xi}x_j^i$, are fundamentally *infinite*, during any finite time period they can never be communicated in full even in principle while in fact many times a day everybody observes in others and experiences him/herself successful meaningful information exchanges. This *communication paradox* [2, 3] speaks of everyday, routine, ubiquitous use of super-Turing computations in human meaningful and socially-important communication. The communication paradox can be solved by appealing to BSDT infinity hypothesis [2, 3] and the technique of BSDT ASMs that are super-Turing devices with programmatic and computational processes that are completely separate in time (ASMs do not waste their computational resources on serving themselves and, for this reason, are faster than *universal* Turing machines [6]).

But, dividing the programming and program running is insufficient to overcome the communication paradox. To cope with it, let us additionally assume that the ASM-transmitter and the ASM-receiver share in full their prehistory, i.e., let they were designed, implemented in a physical form, and learned beforehand to perform the same meaningful function – selecting the same finite binary message x_j^i given the same infinite context c_{xi}. If it is, and not in any other case, the meaning of x_j^i, $c_{xi}x_j^i$, is equally encoded, decoded, interpreted and *understood* by both parties and for both parties, the value of its truth, $T(c_{xi}x_j^i)$, is the same. For this reason, and because the name's meaning is simultaneously a psychological state an animal experiences producing as well as perceiving this name in meaningful information exchange, the transmitter and the receiver are to be exactly physically, structurally, and functionally equivalent (are to be "mirror" replicas or "clones" of each other).

 ISBN: 978-1-61208-224-0

Several important PL predictions come out.

B. Coding by Synaptic Assemblies

Where meanings are essential (e.g., in living organisms) BSDT network learning paradigm "one-memory-trace-per-one-network" [14] must be widespread in practice and, in particular, any memory for meaningful records must be built of the number of networks that coincides with the number of records to be stored in memory. This paradigm is not consistent with the usual desire of designers and engineers to store in a network as many memory traces as possible but it is well supported by recent empirical neuroscience finding of coding by synaptic assemblies [15, 16]. In these experiments, mice were trained to perform new motor tasks and in living animals changes in the number of synaptic contacts associated with learning new skills were measured. In complete accordance with the BSDT assumption [14] that each new memory trace should be written down in an always new separate network (synaptic assembly), it turned out "that leaning new motor tasks (and acquiring new sensory experiences) is associated with the formation of new sets of persistent synaptic connections in motor (and sensory)" brain areas [17, p. 859].

C. Super-Turing Computations by Mirror ASMs

To ensure correct understanding of meanings of finite symbolic communicated messages, the ASM-transmitter and ASM-receiver that are the mirror replicas of each other need to be used. Mirror ASMs implement *meaningful super-Turing computations*: for the transmitter and the receiver, they ensure the use of the same infinitely long "boundary conditions" c_{xi} needed to perform Turing-type computations, which have been programmed beforehand, with finite-length strings x_j^i, e.g., as in [14]. Mirror ASMs *physically* divide infinite meaningful message to be processed into infinite, c_{xi}, and finite, x_j^i, parts and take the former into account as their exactly identical "hardware" and "software", designed and *physically* implemented beforehand in the course of animal evolution and development. Thanks to this trick to correctly understand the meaning of the $c_{xi}x_j^i$ it is enough to correctly transmit, receive, and decode the x_j^i only. Mirror ASMs also explain why meaningful communication without syntax is successful only between animals of the same (or relative) species: such animals are *a priori* equipped with the same "hardware" and "software" that fix the common infinitely long context needed to finish meaningful super-Turing computations of current interest over a finite time period. The picture described is well supported by the empirical finding and studying of mirror neurons – the ones that are active when an animal behaves or only observes respective behaviors of others; see e.g. [18, 19] and numerous references therein. The ASM/mirror-ASM computational system just described and the neuron/mirror-neuron circuitries already observed [18, 19] may respectively be treated as theoretical and real-world implementations of super-Turing machines with infinite inputs, which until now have been hypothetical, e.g. [20], that are to be capable of computing with infinite strings or, what is the same, with real-valued/continuous quantities.

D. Knowing Memory Performance without Knowing Memory Record

Since BSDT PL employs a non-Gödelian (but envisaged by Gödel [21]) arithmetization by ordinals/naturals x_j^i and since these ordinals/naturals are given context *randomly* chosen to name the things to be named, the following effect has been predicted [3]. By examining in an experiment an ASM hierarchy (neural subspace [14]) that generates the meaning of a trace x_j^i, all the parameters describing the ASM selecting the x_j^i may successfully be found but the content of x_j^i – specific randomly-established arrangement of its ±1 components – will always remain unknown. If it is, then, for example, the content of a particular given-length memory record does not affect memory performance and cannot empirically be found. This rather surprising prediction has been corroborated well by numerical BSDT analysis [22] of receiver operating characteristics (ROCs, functions providing memory performance) measured in groups of brain patients and control healthy subjects. In such a way the idea of non-Gödelian BSDT arithmetization of meaningful mathematical expressions by ordinals/naturals *randomly* chosen given the context of a set of ordinals/naturals with their given upper limit has numerically and empirically been substantiated.

VI. Examples and Potential PL Applications

Any formal axiomatic system (FAS, it comprises all its axioms and theorems) is supposed to represent an *infinite fraction* of meaningless finite binary strings x_j^i related to particular proper class, S_{cx0} (Section III A). For this reason, FAS computations are also PL computations and numerous available computational results e.g. in physics or biology may be treated as their examples performed given a context defined formally and informally. A separate infinite PL string that gives a meaning to a finite symbolic message x_j^i (e.g., a formula written in binary notations) includes descriptions of the FAS formalism needed to derive it and of the problem that gives it physical sense. This picture is another representation of formal and informal knowledge from e.g. a book and gives nothing new, except of drawing attention to the fact that manipulations with numbers are meaningless until a giving-the-meaning context is added.

The situation changes dramatically once one wants to communicate this formula's *meaning* to someone else. Let us consider a lecturer in a lecture room. In the beginning, he and his students have different knowledge on the formula of interest and students cannot correctly understand its meaning. The lecturer's aim is to give them a piece of additional knowledge and, in this way, to equalize, for all of them, the context of understanding this formula/message. At the end of the lecture, for the lecturer and for his students, infinite PL strings describing specific knowledge should become bit-by-bit equivalent not only "in the past" but also "in the present", and the formula's meaning should be understood by all the parties in the same way [2]. If it is not, a misunderstanding arises. How, for members of a social (semantic) network, the difference in their previous knowledge influences on understanding meaningful messages may empirically be estimated as described in [23].

 ISBN: 978-1-61208-224-0

In this example, non-syntactic messages represent a very small fraction of general flow of information. Among them it may e.g. be the fact that the lecturer is walking when he gives his talk. This non-syntactic and even non-language message (bodily signal) is effortlessly understood by everyone who is in the room because all people are members of the same species and have the same *innate* bodily infrastructure (in particular mirror neuron system) to produce and perceive/understand walking. Humans/animals produce and perceive such message automatically with practically no chance of misunderstanding because its meaning is provided, for all of them, by *innate* infinite PL strings of the same length that are equivalent "in the past".

The range of potential PL applications covers everything where meanings are important. If they become inessential, the PL can be reduced to traditional mathematics (a FAS).

VII. CONCLUSIONS

BSDT PL provides a framework that is sufficient to perform principal semantic computations and based on them communication without syntax. BSDT PL seems also to be sufficient to explain the discrete computational part of intelligence of animals of poor sociality and, consequently, to design the discrete computational part of intelligence of artificial devices (e.g., robots) or computer codes mimicking the behavior of such animals. At the same time, the PL is unable to explain the mechanism of dividing its names (sentences) into focal and fringe components and, consequently, of directing an animal's attention to particular thing – we hope it may be done by methods beyond the BSDT. To explain/reproduce the "attentive" part of animal intelligence in a biologically-plausible way and to design the "attentive" part of the intelligence of intelligent robots, analog (e.g., wave-like) computational methods similar to those that are used in real brains are most probably required.

In contrast to formal languages that are in end the products of a *finitely* defined calculus, BSDT PL is a calculus of finite binary strings (spike patterns or "symbols") with *infinitely* defined contexts. It is grounded on 1) the BSDT [1] providing the technique of encoding/decoding in binary finite-dimensional spaces (BSDT ASMs [6] implement PL's inference rules) and 2) the new infinity hypothesis [2, 3] providing the technique of super-Turing (semantic) computations with infinite binary strings that share their infinite initial part. BSDT PL is the simplest language of its kind and has great potential for designing the adequate models of higher-level languages, including in perspective the natural languages of humans. At the same time, meaning ambiguity of different-level BSDT PL names that have been established as their fundamental property raises many intriguing problems to be solved in the future.

REFERENCES

[1] P.M. Gopych, "Elements of the binary signal detection theory, BSDT," in New research in neural networks, M. Yoshida, H. Sato, Eds. New York: Nova Science, 2008, pp. 55-63.

[2] P. Gopych, "BSDT atom of consciousness model: The unity and modularity of consciousness," in ICANN-09, LNCS, vol. 5769, C. Alippi, M.M. Polycarpou, C. Panayiotou, G. Ellinas, Eds. Berlin-Heidelberg: Springer, 2009, pp. 54-64, doi:10.1007/978-3-642-04277-5_6.

[3] P. Gopych, "On semantics and syntax of the BSDT primary language," in Information models of knowledge, K. Markov, V. Velychko, O. Voloshin, Eds. Kiev-Sofia: ITHEA, 2010, pp. 135-145, http://foibg.com/ibs_isc/ibs-19/ibs-19-p15.pdf <retrieved: February, 2012>.

[4] J. von Neumann. The computer and the brain. New Haven: Yale University Press, 1956.

[5] N.J. Enfield, "Without social context?" Science, vol. 329, Sep. 2010, pp. 1600-1601, doi:10.1126/science.1194229.

[6] P. Gopych, "Minimal BSDT abstract selectional machines and their selectional and computational performance," in IDEAL-07, LNCS, vol. 4881, H. Yin, P. Tino, E. Corchado, W. Byrne, X. Yao, Eds. Berlin-Heidelberg: Springer, 2007, pp. 198-208, doi:10.1007/978-3-540-77226-2_21.

[7] W.V. Quine. Set theory and its logic. Cambridge, MA: Harvard University Press, 1969.

[8] L. Pozsgay, "Liberal intuitivism as a basis of set theory," Proc. Sym. Pure Math., vol. 13, part I. Providence, Rhode Island: AMS, 1971, pp. 321-330.

[9] G. Chaitin. The limits of mathematics. Singapore: Springer, 1998.

[10] K. Mainzer. Thinking in complexity, 5th ed. Berlin: Springer, 2007.

[11] P. Gopych, "BSDT multi-valued coding in discrete spaces," in CISIS-08, ASC, vol. 53, E. Corchado, R, Zunino, P. Gastaldo, Á. Herrero, Eds. Berlin-Heidelberg: Springer, 2009, pp. 258-265, doi:10.1007/978-3-540-88181-0_33.

[12] D. Perrin and J.-É. Pin. Infinite words. Amsterdam: Academic Press, 2004.

[13] H. Poincaré. Science and method. London: Thomas Nelson and Sons, 1914.

[14] P. Gopych, "Biologically plausible BSDT recognition of complex images: The case of human faces," Int. J. Neural Systems, vol. 18, Dec. 2008, pp. 527-545, doi:10.1142/S0129065708001762.

[15] T. Xu, X. Yu, A.J. Perlik, W.F. Tobin, J.A. Zweig, K. Tennant et al., "Rapid formation and selective stabilization of synapses for enduring motor memories," Nature, vol. 462, Dec. 2009, pp. 915-919, doi:10.1038/nature08389.

[16] G. Yang, F. Pan, W.-B. Gan, "Stably maintained dendritic spines are associated with lifelong memories," Nature, vol. 462, Dec. 2009, pp. 920-924, doi:10.1038/nature08557.

[17] N.E. Ziv and E. Ahissar, "New tricks and old spines," Nature, vol. 462, Dec. 2009, pp. 859-861, doi:10.1038/462859a.

[18] G. Rizzolatti and L. Craighero, "The mirror-neuron system," Ann. Rev. Neurosci., vol. 27, 2004, pp. 169-192, doi:10.1146/annurev.neuro.27.070203.144230.

[19] C. Keysers, J.H. Kaas, V. Gazzole, "Somatosensation in social perception," Nat. Rev. Neurosci., vol. 11, June 2010, pp. 417-428, doi:10.1038/nrn2833.

[20] T. Ord, "The many forms of hypercomputations," App. Math. Comp., vol. 178, July 2006, pp. 143-153, doi:10.1016/j.amc.2005.09.076.

[21] K. Gödel, "Remarks before Princeton bicentennial conference on problems of mathematics," in The Undecidable, M. Davis, Ed. New York: Raven Press, 1965, pp. 84-88.

[22] P. Gopych and I. Gopych, "BSDT ROC and Cognitive Learning Hypothesis," in CISIS-10, AISC, vol. 85, Á. Herrero, E. Corchado, C. Redondo, Á. Alonso, Eds. Berlin-Heidelberg: Springer, 2010, pp. 13-23, doi:10.1007/978-3-642-16626-6_2.

[23] J. Krüger, C. Krüger. On communication. An interdisciplinary and mathematical approach. Dortrecht, The Netherlands: Springer, 2007.

 ISBN: 978-1-61208-224-0

Intelligent LED Lighting System with Route Prediction Algorithm for Parking Garage

Insung Hong, Jisung Byun, and Sehyun Park
School of Electrical and Electronics Engineering, Chung-Ang University
Seoul, South Korea
axlrose11421@wm.cau.ac.kr, jinsung@wm.cau.ac.kr, shpark@cau.ac.kr

***Abstract*—Various LED applications have been developed and implemented in diverse spots because of LED's characteristic. However, some specific places such as a parking lot are required to be studied to increase energy efficiency. In this paper, the proposed LED lighting systems for a parking lot provides energy efficient management by turning on and off LED lights according to vehicle's movement through a route prediction algorithm.**

Keywords-LED, ZigBee, Route prediction, Energy efficiency, Parking lot

I. INTRODUCTION

In recent years, usage of LED lights has been increasing. As interest of energy saving broadens, light power consumption, which are one of the highest power consumption, becomes a big issue. At this point, LED lights will be the most practical answer to decrease power consumption in a home or building. Moreover, LED technology has been researched by many companies and research centers with advanced countries as the center, and now it can be substituted for the existing lights. There are various types from low power to power LED, and it is expected that it will influence lighting markets seriously.

LED has a variety of advantages compared with the existing lights. First of all, it is easy to interwork with other electronic modules such as sensors and communication module to provide new services, and can be controlled more elaborately because LED is a kind of electronic components. For example, fluorescent lights are difficult to control their brightness. Even though it is available to control, an additional control module is needed. However, LED can be handled by PWM (Pulse Width Modulation) or current control to change its brightness easily. Furthermore, it has a low power characteristic, whose power consumption is much lower than the existing lights, and a heating problem has been improved compared to the past.

By using the characteristics of LED, various LED applications are now developed. LED lighting systems now provide a lighting function and are united with ICT (information and communications technology). The intelligent lighting control system [1] provides improved user-oriented services based on wireless sensor networks, and pattern recognition about user activities and profiles. In reference [2] and [3], it shows similar services by using location-based living patterns. These researches decrease power consumption in a lighting system and service response time.

Usage of LED lighting systems has been spread gradually from normal lights in a building to parking lot. According to the characteristics of a parking lot, although movement of people is less than other places, lighting systems are always needed in this place. A parking garage requires more lights than other places, and a certain level of brightness should be maintained. In even an outdoor parking lot, lights are needed at night. Although lights are required for most of the day, frequency of the light usage is very low so this causes inefficient power consumption. By replacing the existing light with LED light, it is assumed to save power consumption as described before. Therefore, we use the route prediction algorithm to save power consumption. There are various route prediction algorithms [4] – [6] but we design the route prediction algorithm by using motion detection sensors which are included in the light controller.

In this paper, we design an intelligent LED lighting system which is suitable for a parking lot and implement it to verify its energy efficiency compared with the existing lights. Moreover, this system includes two types of sensors, illumination and motion detection, and ZigBee communication, and analyzes vehicle movement to turn on and off the minimum number of LED lights.

In Section II, we describe the hardware structure and middleware of this system are described. Section III presents the energy efficiency of this system, and it is compared with other light systems. Section IV describes the analysis of the experimental result and concludes this paper.

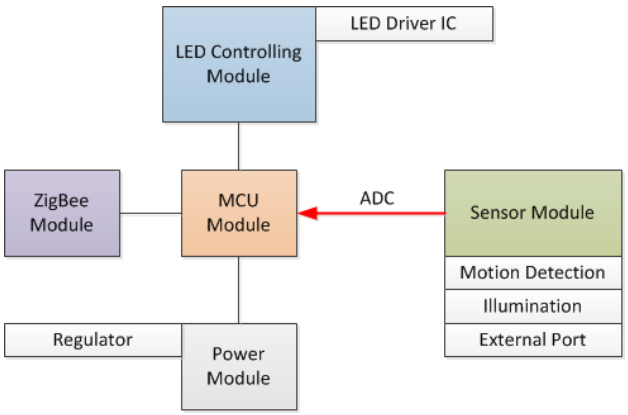
II. DESIGN OF INTELLIGENT LED LIGHTING SYSTEM

For the intelligent LED lighting system, the proposed system in this paper is composed as follows. Each light in a parking lot has a lighting controller in ZigBee-based sensor network. Each lighting node includes an illumination sensor and motion detection sensors. Based on the two sensor information, the lighting system decides to turn on or off the lights in the expected route where a car will enter to save needless power consumption.

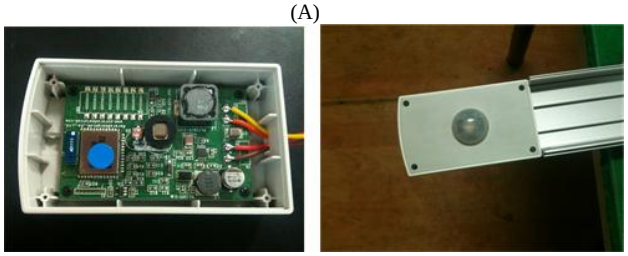
 ISBN: 978-1-61208-224-0

A. *Overrall of the LED Lighting Controller*

The overall structure of this controller is similar to the Light Enabler [2]. Basically, the controller consists of five parts, MCU module, LED control module, power module, sensor module, and ZigBee module.



(A)



(B)

Figure 1. (A) Hardware Composition of Lighting Controller (B) LED Lighting Controller and Motion Detection Sensor

The basic hardware structure of this controller consists of the five parts, and each part is controlled by the MCU module. In particular, the sensor module is designed to change the types of the sensors in this system, and the system performs its function according to the values of the two sensors, the illumination and motion detection sensor.

B. *ZigBee-based Sensor Network Structure*

Based on the ZigBee network, each node sends and receives data. The basic idea is that sensed data in the sensor module, illumination and motion detection, are transmitted to the neighbor or entire nodes. Each node decides to process the sensed data after analyzing the packets based on an allocated address. Each node is included in the assigned group, and this group can be changed according to the structure of a parking lot like figure 2. There is the Group Router node in the group including the certain number of nodes like figure 2, and the only two types of nodes, the Group Router and Coordinator node, are able to communicate with each other to minimize collision in the ZigBee network. The normal nodes can communicate with the Group Router node in the same group, not in different groups or the coordinator node.

In a network initialization time, the Coordinator node checks status of every node in each group by receiving initialization packets from each Group Router node. The most important role of the Coordinator node is to collect sensed data and the status of nodes and control each group based on the data.

The Group Router node performs similar functions like the Coordinator node but the functions are limited in the single group. For example, if a normal node in the same group detects a motion, the Router Node receives this data from that node, and transmits control packets to other nodes in the same group. That is, the Router Node collects the sensed data in the same group, and decides to send the data to the Coordinator Node based on the internal policy.

Types of controls by the Coordinator and Router Node are as in the following. There are the two controls, the in-group-control and group-unit-control, and the Coordinator decides one between the controls according to kinds of sensed information and internal policy in the middleware.

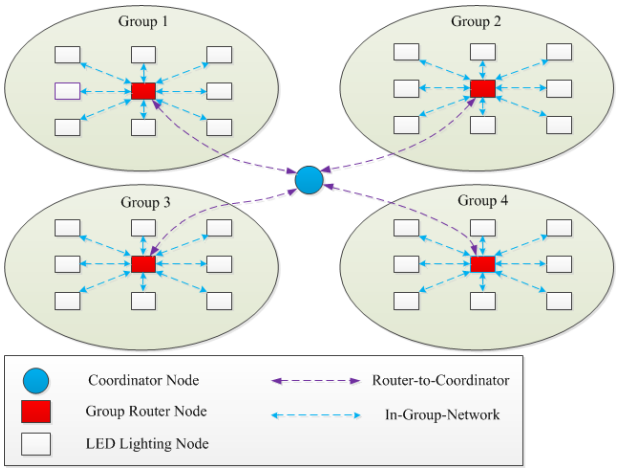


Figure 2. ZigBee Network of the proposed system

C. *Middleware Structure*

As described above, there are the two control types of the LED lighting node by the Coordinator and Router nodes.

- In-group-control: the control which is performed in a group based on the collected sensed data
- Group-unit-control: the control which can affect multiple groups by the Coordinator node

The most difference between the two types of the controls is the scope to affect group units. For example, if there are nine LED lighting nodes including one Group Router node, the eight nodes can communicate with only the Group Router Node. Therefore, if one node detects a motion, it sends motion detection data to the Router node, and the Router node determines to process the data and send this event to the Coordinator node. The algorithm of the in-group-control is relatively simple. Based on the critical value which can be set initially or changed by the Coordinator node, if the sensor value of motion detection

 ISBN: 978-1-61208-224-0

exceeds the critical value, the Router node performs the in-group-control to turn on or change brightness of lights in the same group. The designed sensor modules are an illumination sensor and motion detection sensor, but the critical value affects only the motion detection sensor. A PIR (Passive Infrared) sensor is used to detect movement, and the MCU module reads an analog value changed by the sensor module. The illumination sensor influences a brightness of the LED light. This sensor cannot turn on or off the light but the maximum or minimum brightness of the light is changed according to the illumination sensor.

Figure 3 (A) shows the essential structure of the middleware. The illumination manager and motion detection manager collect sensed data from the sensor module, and the two types of data are sent to the sensor value manager and the LED brightness control manager respectively. The LED brightness control manager changes the minimum and maximum brightness of the light based on the illumination sensor, and the sensor value manager determines to control the light by comparing the critical value with the sensed data. Through this process, the Router node decides to perform the in-group-control.

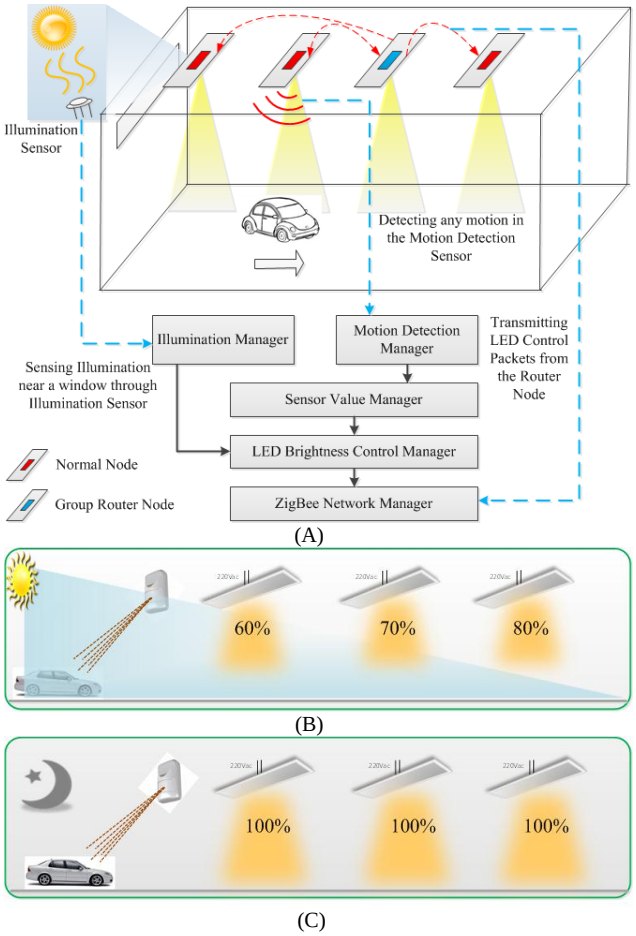


Figure 3. (A) Middleware of the LED Lighting System (B) Brightness Control in day time (C) Brightness Control in night time

D. Route Prediction

We described the middleware structure about the simple operation based on collected data. However, in an actual environment such as a parking lot, this simple operation could decrease energy efficiency and make functional problems. Therefore, we design the LED lighting system with the LED controller in consideration of the floor plan of a test bed, interworking with each LED light group, and a pattern recognition by a motion detection to maximize system efficiency.

Figure 4 (A) shows the test bed floor plan of this system. We implement the total 48 LED light nodes in the test bed, and a red line box in figure 4 (A) means one group. Moreover, each group is assigned according to drive routes. If the LED lighting node cannot detect any motion, it controls its brightness at the minimum level considering the illumination sensor, and changes to the maximum level if detecting a motion. The red circle points are motion sensing modules including the sensor module and ZigBee module so that the system can notice going in and out of vehicles by using the sensing modules. The green circle points are the Coordinator nodes which manage 24 nodes respectively.

As described above, the in-group-control is used when the light node detects unexpected motion mostly but the group-unit-control is used to control the brightness of a drive route according to the vehicle's movement when a vehicle enters a parking lot. For example, if a vehicle enters a parking lot, the Coordinator node notices the movement of the vehicle and turns on lights near the vehicle. This passive LED light operation is widely used in commercialized systems. In this paper, we propose the more advanced system with better energy efficiency based on pattern recognition.

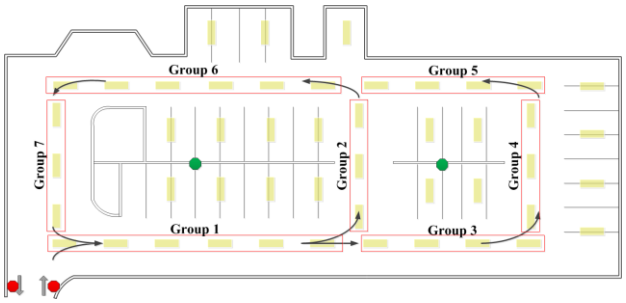
In figure 4 (B), a vehicle is in the route A, and it could choose the router B or C after passing the route A. The normal lighting operation generally turns on both route B and C while the vehicle is in the route A, or turns on one of them after it enters the route. The former causes inefficient energy consumption because one path is not used, and in the latter a driver is difficult to secure a clear view. In this paper, we develop the algorithm with the LED lighting node to improve energy efficiency. Comparing figure 4 (A) with (B), the route A has the total six LED lighting nodes with a motion detection sensor. Each node sends a motion detection data to the Router node when it senses movement. By collecting this data, the Router node can infer a velocity of a vehicle.

$$Inst\_Velocity = \frac{Distance}{(Time\_Node_n - Time\_Node_{n-1})} \quad (1)$$

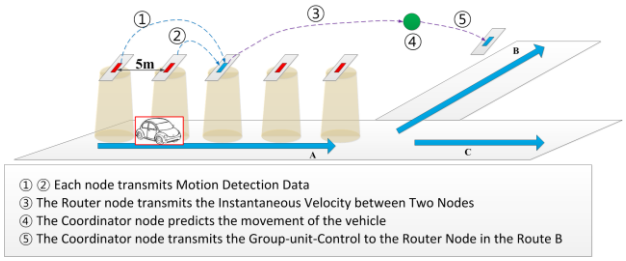
$$Avg\_Velocity = \sum_n Inst\_Velocity_n / n \quad (2)$$

Equation (1) and (2) describe the instantaneous and average velocity, and the first one is the instantaneous velocity between two nodes. The Router node checks the time when receiving a motion detection data and figures out the

velocity if having two successive motion detection data like equation (1). The distance factor in equation (1) is set 5 meters in this test bed but it can be changed according to different environment. The instantaneous velocity is sent to the Coordinator node, and equation (2) is generated in the Coordinator node by using the collected data.



(A)



(B)

Figure 4. (A) The floor plan of the tes bed (B) Description of the route prediction

Through the two velocity information, the Coordinator node predicts whether the vehicle enters route B or C. If the velocity of the vehicle is constant and over the certain value which is set in the route A, the Coordinator node turns on the LED lights in the route C. However, if the velocity is constant under the value, it turns on the lights in the route B. Because it is difficult to drive at a high speed in a parking lot, this system can save inefficient energy consumption.

To improve this algorithm to predict drive path, the system uses the two velocity information as follows. For example, the Coordinator node uses the three velocity information, the initial, final instantaneous and average velocity. Therefore, if the final instantaneous value is larger than the initial one, the Coordinator node regards this movement as a direct drive. As described above, in case of maintaining constant velocity, it is considered as a direct one. On the other hands, if the initial velocity is larger than the final one, it is regarded as a turning drive. However, according to driving habits of drivers, some vehicles would maintain constant and slow velocity under the certain value, and it is hard to predict which route a driver chooses. In the case, the Coordinator node turns on both routes' lights which the driver can choose to handle the exception case.

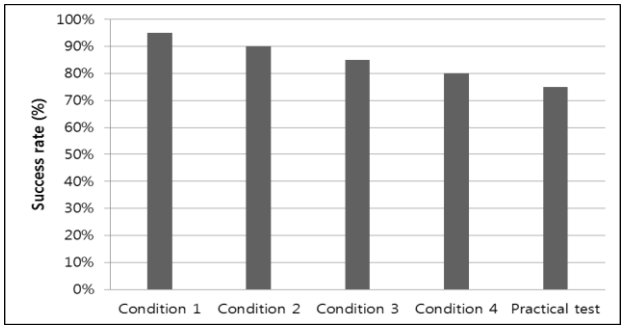
Furthermore, there are the two motion sensing modules in the entrance so the system can recognize the situation a vehicle enters or gets out a parking lot. If the vehicle enters and no movement is detected during 30 seconds, the system recognizes it as a completion of parking and turns on all the lights at the half brightness to help a driver to find and move to an entrance. However, in case of getting out, the system does not turn on the lights.

III. IMPLEMENTATION

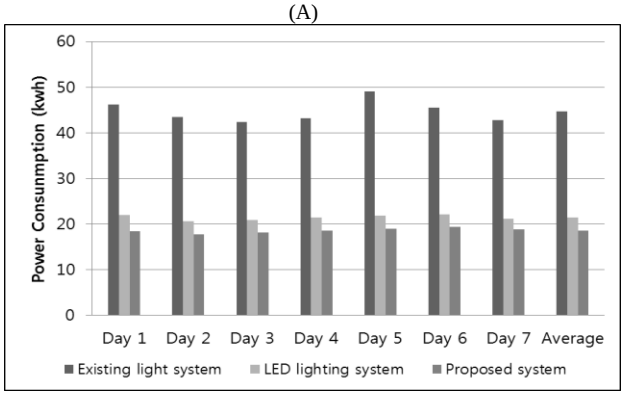
We implement the proposed system in the test bed like figure 4 (A) and figure 6, and verify the energy efficiency and accuracy of the drive route prediction. The following table shows the conditions of the test bed.
In this test bed, the average number of vehicles using the parking lot is approximately thirty, and we limit the number of driving vehicles to only one at the same time.

TABLE I. EXPERIMENTAL CONDITION

Unit	*Specification*
Number of Used LED lights	48
Area of the test bed	1800 m²
Number of available parking rooms	40
Cars per day	25~30



(A)



(B)

Figure 5. (A) Accuracy of route prediction (B) Comparison of energy efficiency in the three types of the lighting system

The first experiment is about the accuracy of the prediction. We choose four types of driving patterns, and each driving way is repeated 20 times.

1) Drive to stay constant 20km/h

 ISBN: 978-1-61208-224-0

2) Drive to accelerate velocity from 20 to 25km/h

3) Drive to stay constant 20km/h and slow down speed near a corner

4) Drive to stay constant 15km/h

5) Practical test for a day

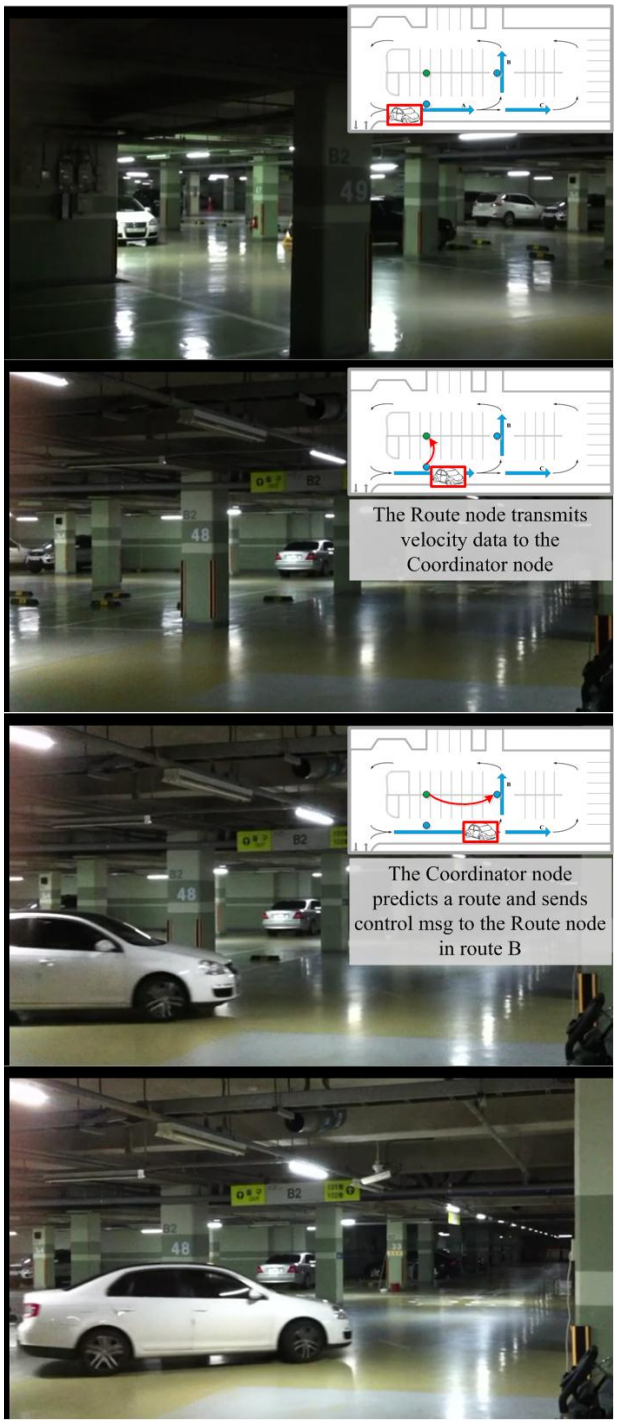


Figure 6. Implementation in the test bed

Each driving way is repeated 20 times, and, in addition, we test the proposed system for practical users. As a result of this, the practical test shows higher rates of exceptional events than the other tests.

In the second experiment, we compare the energy efficiency of this system with others. To do this, we measure three types of power consumption by using a wattmeter in the test bed for a week, the fluorescent lights, LED lights with simple operation, and the proposed system. The figure 5 (B) shows that the proposed system saves 10 percent power consumption beside the simple operation.

IV. CONCLUSION

In this paper, we design and implement the intelligent LED lighting system for a basement garage having specific characteristics different to other places. By implementing the system, the LED lighting system with the route prediction saves approximately 60 percent power consumption compared to the existing fluorescent lights and 13 percent consumption beside the LED lighting system with the simple operation.

However, in the first experiment for the practical users, the system could not predict a driver's route and considers it as an exceptional event at higher rates than the others. It is assumed that this is caused by various drivers' patterns and size of vehicles. To complement this error, the algorithm should be improved to include driving patterns and complement response and processing time of the ZigBee network.

We plan to expand this system in a larger parking lot. This new test bed has more intersections and vehicles so the improved sensor management to handle more sensors is required. We also plan to design a management application for administrator to manage the whole lighting system.

ACKNOWLEDGMENT

This research was supported by the MKE(The Ministry of Knowledge Economy), Korea, under the HNRC(Home Network Research Center) –ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency (NIPA-2010-C1090-1011–0010) and by the Human Resources Development of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government Ministry of Knowledge Economy (20104010100570).

REFERENCES

[1] M. Pan, L. Yeh, Y. Chen, Y. Lin, Y. Tseng, "A WSN-Based Intelligent Light Control System Considering User Activities and Profiles," *IEEE Sensors Journal*, vol. 8, no. 10, pp. 1710-1721, Oct. 2008

[2] Y. Uhm, I. Hong, G. Kim, B. Lee, S. Park, "Design and implementation of power-aware LED light enabler with location-aware adaptive middleware and context-aware user pattern," *IEEE Trans. Consumer Electronics*, vol. 56, no. 1, pp. 231-239, Feb. 2010

 ISBN: 978-1-61208-224-0

[3] Z. Hwang, Y. Uhm, Y. Kim, G. Kim, S. Park, "Development of LED smart switch with light-weight middleware for location-aware services in smart home," *IEEE Trans. Consumer Electronics*, vol. 56, no. 3, pp. 1395-1402, Aug. 2010

[4] A. H. Networks, V. Namboodiri, S. Member, and L. Gao, "Prediction-Based Routing for Vehicular," *Networks*, vol. 56, no. 4, pp. 2332-2345, 2007.

[5] R. Simmons, B. Browning, Y. Zhang, and V. Sadekar, "Learning to Predict Driver Route and Destination Intent," *Proceedings of the IEEE*, pp. 127-132, 2006.

[6] K. Park, M. Bell, I. Kaparias, and K. Bogenberger, "Learning user preferences of route choice behaviour for adaptive route guidance," *Engineering and Technology*, pp. 159-166.

 ISBN: 978-1-61208-224-0

Proactive Assistance Within Ambient Environment

Towards intelligent agent server that anticipate and provide users' needs

Hajer Sassi
LIFL Laboratory – University of Lille 1
xBrainLab-USILINK
59655 Villeneuve d'Ascq Cedex – France
hajer.sassi@ed.univ-lille1.fr

José Rouillard
LIFL Laboratory – University of Lille 1
59655 Villeneuve d'Ascq Cedex –France
jose.rouillard@univ-lille1.fr

***Abstract*—User needs are expanding and becoming more and more complex with the emergence of newly adopted technologies. As a result, the convergence of smart devices, having the capability to communicate as well as sharing information and ensuring user need satisfaction, leads to profoundly change the way we interact with our environment. They should provide an adaptive assistance in both reactive and proactive mode and new communication methods focused on multimodal and multichannel interfaces. However, most of existing context-aware systems have extremely tight coupling between applications' semantic and sensor's details. So, the objective of our research is to implement an approach which can support the ability to reuse sensors and to evolve existing applications to use new context types. In this paper, we illustrate our approach for proactive intelligent assistance and we describe our architecture based on three principal layers. These layers are designed in order to build applications which can increase the welfare of the user situated in intelligent environment.**

Keywords-Intelligent Interfaces; ubiquitous computing; human-computer interaction; proactive assistance; multimodal interfaces; multi-channel interfaces.

I. INTRODUCTION

Ambient Intelligence (AmI) aims at insuring the comfort of users in theirs daily tasks based on context information. In our life, we often repeat usually the same tasks. For example, seeing the weather forecast before going outside, consulting agenda to verify appointments, control children tasks, etc.

User searches to delegate a majority of these daily tasks to her intelligent environment in order to decrease her responsibilities. As a consequence, she wants to satisfy her needs without any explicit intervention through the capability of the intelligent environment to perceive user's personal environment in order to resolve her daily tasks. Therefore, AmI follows the goals of Ubiquitous Computing, a paradigm that was first suggested by Weiser in the early 1990s. His vision was to increase the welfare of a user situated in a computer everywhere environment by supporting human assistance in an intimate way [1].

One research domain that requires the computer- everywhere model of ubiquitous computing is that of the "intelligent environment" [2]. In this domain, a wide range of simple information (e.g., light sensor, audio/video sensor, temperature sensor, google calendar, information from the web, etc.) and composite information (e.g., presence sensor and preferences of users) can be collected from heterogeneous sensors in order to determine automatically users' needs based on their context's information.

In this context-aware domain, many ad hoc systems exist in order to be able to perform an adaptive assistance. However, these systems present two main limits: the difficulty to develop due to the requirements of dealing directly with sensors and the difficulty to evolve because the application semantics are not separated from the sensor details (also rules).

So, building applications, depending on context-aware which can support reuse sensors and new context types stays hard tasks, which covered many context-aware features.

As said by Dey in his thesis "context has the following properties that lead to the difficulty in use "[3]:

- Context is acquired from non-traditional devices (i.e., not mice and keyboards), with which we have limited experience. For example, tracking the location of people or detecting their presence may require Active Badge devices [4], floor-embedded presence sensors [5] and video image processing…
- Context must be abstracted to make sense to the application; Active Badges provide IDs, which must be abstracted into user names and locations.
- Context may be acquired from multiple distributed and heterogeneous sources. Detecting the presence of user in a room reliably may require combining the results of several techniques such as image processing, audio processing, floor-embedded pressure, etc.
- Context is dynamic; changes in the environment must be detected in real time and applications must change behavior to constant changes.
- Context information history, as shown by context-based retrieval applications [6, 7]; context history can be used to recognize user's activities and to fully exploit the richness of context information.

 ISBN: 978-1-61208-224-0

These difficulties prevent to build context-aware applications the ability to support reuse of sensing technologies in new applications and evolution to use new context in new ways. In this paper, we present a system which can support new context types and evolve dynamically according to user's preferences.

This document is organized as the following: First, we describe some previous context-aware applications. Second, we present our research problematic and how we proceeded to resolve it. Third, we describe our proposed architecture and an illustrative example. Lastly, we state our future works and conclusions.

II. RELATED WORK

Weiser's vision in his article "The Computer for the 21st Century" [8] is to serve people's daily tasks through an intelligent environment which should acts invisibly and unobtrusively in the background and freeing users from tedious routine tasks in order to reduce users' responsibilities.

Ubiquitous computing aims to integrate each intelligent entity that can be identified and provide information about user's context such as sensors which can provide immediate information according to user's situation. Thus, user's goals and desires can be anticipated from the interaction context which is defined by Dey [9] as "any information that can be used to characterize the situation of an entity. An entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including the user and application themselves".

Many projects are developed around context aware. In 1998, Coen created the Intelligent Room MIT [2]. This is a conference room equipped with 12 cameras, 2 video projectors, display devices, microphones and loudspeakers. The goal of this room is to interact with different form of modality. In the field of home automation, Mozer created the Adaptive House [10] which is an intelligent home equipped with 75 sensors in order to provide information such as temperature, ambient light, door's and window's situations. Adaptive house has also the capability to manage energy. Microsoft has also created the project named EasyLiving [11], which calculates user's position and propose service depending on his position. Each of these projects illustrates convincing results from different uses cases which proposed. On the other hand ubiquitous computing aims to change ordinary interfaces by intelligent interfaces in order to let user feeling natural communication on many levels (complexity, size, and portability). In the 70's, the technology-driven focus on interfaces was slowly changed and in the 80's the new field of Human-Machine Interaction (HCI) appeared. With the appearance of new technologies such as data mining, machine learning, speech/voice recognition, facial recognition and omnipresent computing, the basic technology based on ordinary interfaces can difficultly use. Consequently, the interaction human-machine should change the way we interact with the ambient environment by providing new intelligent interfaces able to adapt its behavior according to user's situation. Around 1994 until 1996, intelligent agents, practical speech recognition and natural language applications appeared. However, since then intelligent user interfaces evolve slowly. On the other hand, implementing and maintaining interfaces, which should be at the same time proactive and intelligent, is still far from easy.

III. RESEARCH QUESTIONS

The inference of user's requirements or proactive assistance is a very delicate problem, which we have chosen to explore through the following question, "proactive assistance: why, when and how to use it?"

The first question "Why" has for objective to search how can proactive assistance reduces user's responsibilities. As we know, we have many boring routine tasks and we search to delegate more of them to our intelligent environment in order to have more time for other more complex tasks. Thus, by the capability of the intelligent environment to perceive environment and user's habits, system based on proactive assistance could anticipate users' needs without any explicit request. The second question "When" is devoted to determine the adequate time; when intelligent environment decide to communicate user's need. Once intelligent environment determines user's needs, it should interpret user's real situation in order to decide if service can be communicated. However, the last question "How" is interested to adapt the way we interact with our environment. Depending on context, our system should find the adequate modality (text, speech and gesture) and channel (Internet and phone channel) according to user's situation.

IV. PROPOSED ARCHITECTURE

To respond to our research questions, we have chosen to implement an architecture based on three principals layers (see Figure 1), which can communicate between them throw two different modes: the push and the pull modes, which are used, in our system, to provide reactive and proactive interactions. Each layer has for role to provide a service to the layer above in order to resolve user's needs. However, the mechanism of adaptation is shared between the second and third layer.

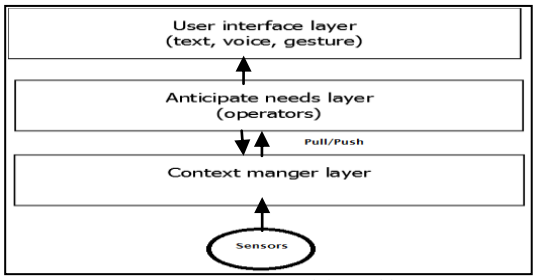


Figure 1. Context model's architecture

 ISBN: 978-1-61208-224-0

A. *Context Manager Layer*

To build systems able to act differently according to context awareness, intelligent environment should perceive and control sensors networks regularly through the "context manager layer". This layer should communicate with heterogeneous sources in order to collect information and register them in the database [12, 13, 14]. This layer is based on context provider and context repository. It controls the behavior of sensors and saves new issues values (static, temporary and dynamic information) in context repository. It should also communicate directly with the second layer in order to publish information even before context repository registers information in database for later use. An example of sensors that we used to collect information is a Radio Frequency Identification (RFID) reader accompanied with RFID tags. When RFID reader detects an RFID tag (see Figure 2 and Figure 4), it firstly determines the user's name in order to salute him/her (see Figure 3 and Figure 5) and secondly calculates the number of persons at home.

Figure 2. Bob's RFID tag.

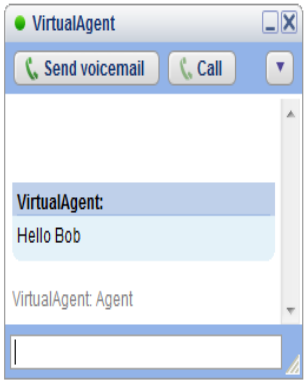


Figure 3. Agent detects Bob's RFID tag and welcomes him on Gtalk.

Figure 4. Marc's RFID tag.

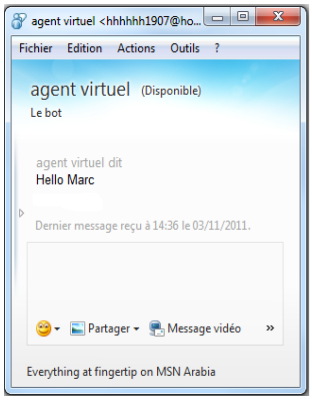


Figure 5. Agent detects Marc'sRFID tag and welcomes him on MSN.

To gather user's information (current activity and preferences), we have chosen to ask some questions according to the user's context as follows:

Firstly, our system have not any information about user, it learns user's information by asking a set of questions which are triggered depending on the context.

1) Case one: we create an xml file which contains some questions grouped by theme.

```
<?xml version="1.0" encoding="utf-8"?>
<questionnaire id="1" theme="Tv">
  <question format="input" nom="FrequenceTv" dataType="xsd:string">
    <ennonce>Do you often watch TV? (yes, no, sometimes)</ennonce>
  </question>
  <question format="input"  nom="SeriePreferee" dataType="xsd:string">
    <ennonce>What is your favorite television series? (put "0" if you do not
have)?</ennonce>
  </question>
  <question format="input"  nom="EmissionRegardee" dataType="xsd:string">
    <ennonce>Which type of emission you like watching? (put "0" if you do not
have)?</ennonce>
  </question>
  <question format="input"  nom="EmissionPreferee" dataType="xsd:string">
    <ennonce>What is your favorite broadcast? (put "0" if you do not
have)?</ennonce>
  </question>
</questionnaire>
```

Figure 6. TV questionnaire

According to context, system tries to collect user's knowledge. It triggers a questionnaire (see Figure 6) depending on user's situation (e.g., user is watching TV), and it stores responses in the database, thanks to a natural language multimodal dialog. As we can see in Figure 6, we have chosen four questions about user's frequency of watching TV, her favorite series, her favorite category of emission and its title. Based on answers given by user, system will infer new decision related on her preferences such as send notification when program TV contains user's favorite category of emission.

2) Case two: User can also enter data through a software entity (e.g., Website, Google calendar, Face- book, etc.) and provide access to system which can use this software in order to more help user. This layer distinguishes three types of information: the static information, the temporary information and the dynamic information. Static information remains unchanged during the process of learning (e.g., name, age, etc.). Temporary information can be sometimes changed (e.g., preferences, taste, etc.). However dynamic information changes frequently (e.g., location, mood). All these types of information are stored in a database in order to be used later.

B. *Anticipate Needs Layer*

In our research, we are based on "context manager layer" in order to anticipate user's services. In this layer, we try to exploit stored data context manager by associating a set of adaptive operators. Actually, we distinguish three types of operators:

1) Conversion operator: the context manager stores a data in initial format, after that "anticipate needs layer" tries to adapt this format in order to associate a meaning manageable by the system. For example: when temperature sensor sends the raw data "2", the conversion operator interprets this value as "it's cold" or "it's hot", according to the real situation of the user.

 ISBN: 978-1-61208-224-0

2) Extract operator: in many cases our system integrates logical sensors such as google calendar, RSS stream, etc. However these sources provide imprecise information. Therefore, the mission of this operator should extract only relevant information. Example: extract just the minute from the current time.

3) Coupling operator: in other cases, system should aggregate various and heterogeneous (logical and/or physical) data. Thus we propose a coupling operator which tries to collect many data in order to "understand" non-trivial situations. For example detecting the location of users in a living room requires gathering information from multiple sensors throughout the intelligent home. It should also, in many cases, combine the results of several techniques such as image processing, audio processing, floor-embedded pressure sensors, etc., in order to provide valid information.

C. User Interface Layer

In a ubiquitous environment, the behavior of services does not depend just on explicit user interaction but also on environment's perceptions. Combing these two sources of information, system can better respond to user's expectations. Our system has to provide an adaptive way of interaction according to the user's situations. The "user interface layer" should be able to define the context and choose the best way to interact by selecting the appropriate modalities and channels.

Our work tackles the ability of ambient computing to permit context-aware interactions between humans and machines. To do so, we rely on the use of multimodal and multi-channel interfaces in various fields of application such as coaching, learning, health care diagnosis, or home automation.

Using a multimodal approach allows users to employ different kinds of modalities (keyboard/mouse, voice, gesture, etc.) in order to interact with a system. The synergistic multimodality is quite natural for humans, but very difficult to implement, mainly because it requires some sharp synchronizations. Fusion mechanisms are used to interprets inputs (from user to machine) while fission mechanisms are used to generate outputs (from machine to user).

Using a multi-channel approach allows users to interact with several channels choosing the most appropriate one in order to exchange with an entity. Such channels could be, for instance, plain paper, e-mail, phone, web site. For the moment, our prototype supports text, speech and gesture as inputs and text and speech as outputs. Once system anticipates user's need through the second layer, "user interface layer" communicates with "context manager layer" in order to check information related to user's situation (e.g., user location, user status, etc.)

In our approach, the influence of the context appears in both second and third layers. The context is used, firstly, to anticipate user's needs and secondly to find the appropriate way of interaction depending on user's situation.

V. SCENARIO

In order to ensure the communication with user anyway she is, we decided to work on multi-channel interfaces and we have chosen to use two types of channels which are: internet channel and phone channel [15].

A. Internet Channel

To demonstrate the identified requirements, a scenario is given in the following. The scenario is about Mr. Marc's favorite TV show. The smart home of Mr. Marc is initially equipped with a standard set of context sensors: in-house location, time, number of persons, favorite show and favorite channel. When our system detects that Marc is connected, it salutes him ("Hello, Marc") and starts to dialog and interact with him (see Figure 7).

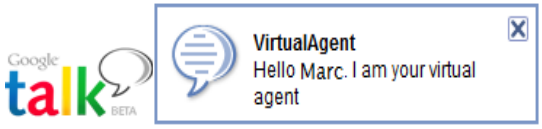


Figure 7. User is logged on

Then, the system checks time, our TV service and user's preferences concerning TV shows. If program TV contains user's preferences show, agent calculates the remaining time from the start date of the show and decides to send this information to the "User Interface Layer". Afterward, this last layer sends a request to the "Context layer manager" in order to determine user's situation. For example, at the office, the system will provide this service using a classical text modality by sending information which contains the title of the show, the time of diffusion, the remaining time and the following question: "Thank you for answering by "YES" or "NO" (see Figure 8).

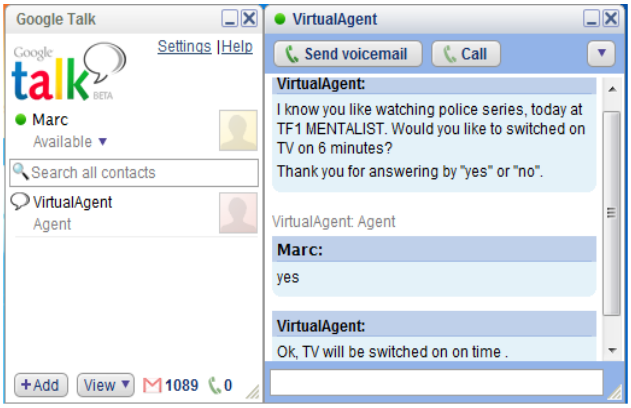


Figure 8. Agent notify user about her best show

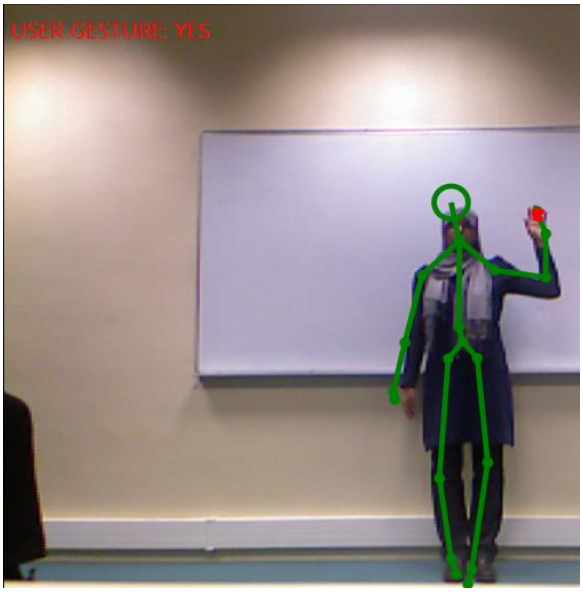


Figure 9. User's gesture response "yes"

If the user responds "YES" using either a keyboard (see Figure 8), a voice recognition or a gesture through a Kinect sensor (see Figure 9), the agent turns on TV in the appropriate time. In that scenario, by executing this action, the system sends, after six minutes, a new text message to the user, telling that the TV is switched on TF1 channel (see Figure 10), but it can also in other situations (e.g., user at home) communicate the same service using a more natural modality such as the vocal one (speech synthesis).

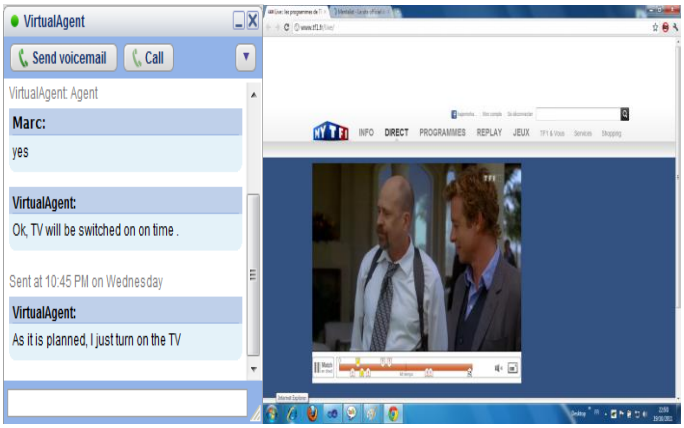


Figure 10. Agent notifies user that the TV is switched on

If the user responds "NO" (to a question such as "Would you like to watch that show now?", see Figure 8), our agent tries to understand the reason, and asks the following question "Are you still interested by this category of show" in order to understand her motivations. If the user responds also "NO", the agent updates this information in the database (the user is no more interested by this TV show).

As a motion sensing, we have chosen to use the Kinect sensor which can used to interpret specific gestures by using an infrared projector, camera and a special microchip to track the movement of individuals in three dimensions. To implement gesture recognition, we firstly define a set of constraints to describe gesture (the joint, the distance, etc.), and secondly, we associate to this gesture a specific event.

In our scenario we have chosen as joint the head, the left and the right hand. If the user raises her left hand, the system interprets this gesture as "NO" and if she raises her right hand, the system interprets it as "YES". Afterward, our system behaves as for the text modality. For the voice recognition, we also used the Kinect sensor's capabilities to recognize human voices. So, user can response by saying "YES" or "NO" vocally and system analyzes this response according to the grammar defined previously.

The goal of using many modalities such as text, voice and gesture is to let the user choose, according to her situation, the most adequately modalities.

B. Phone Channel

As we said in previous sections, we tried to provide proactive intelligent interfaces which can associate different types of modalities with different channels. However, when the user is disconnected from the internet network channel, and if the agent has important information to communicate to her, it should find a new way of communication to reach her wherever she is (home, office, outside, etc.). So, as second channel of communication that can be interesting in our work, we have chosen the phone channel, which allows our system to communicate with people when they are disconnected from the internet. This step is very important in our research; it ensures the continuity with the user by sending for example a Short Message Service message (SMS) as illustrated with Figure 11.

Figure 11. Sending SMS through phone channel to reach disconnected user

 ISBN: 978-1-61208-224-0

VI. CONCLUSION AND OUTLOOK

In this paper, we have proposed the notion of proactive assistance as a solution to increase the productivity and the welfare of the user situated in intelligent environments. So, we have presented an approach model based on three principal layers: "context manager layer", "anticipate needs layer" and "user interface layer". Each one has a specific functionality: the first one communicates with heterogeneous sensors in order to collect context's information, in real time. The second layer tries to adapt collected information to anticipate user's needs. Afterward and depending on person's situations, "user interface layer" chooses the appropriate way of interaction throw the capabilities of the system to support multimodal and multi-channel interfaces; it can manage text, voice and gesture modalities as inputs, and text and/or speech as outputs. We have realized a prototype based on the architecture layers described below. This prototype, about TV show preferences, illustrates our approach and implements proactive services which can adapt themselves depending on each user's situation. We have also implemented other services (using Google Agenda, weather forecast, Phydgets sensors, etc.) which are not described in this paper.

In the very close future, we envisage an evaluation with users by proposing a set of proactive services in order to study users' behavior and our approach capabilities to manage many users simultaneously. In the medium-term, we want to focus on how system can react when context anticipates more than one need in the same time or when several triggers are at the origin of the same need. We have already a theoretical solution for the first problem; we will add a priority ponderation to user's desires. Moreover, the second problem is being studied and we should obtain quickly solutions in order to respond to users' expectations.

REFERENCES

[1] M. Weiser, Some computer science issues in ubiquitous computing, Communications of the ACM (1993), Vol 36(7).

[2] M. H. Coen, Design principles for intelligent environments, Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence (1998), American Association for Artificial Intelligence, Madison, Wisconsin, United States, pp. 547-554.

[3] A. K. Dey, Providing architectural support for building context-aware applications. Thesis, 2000.

[4] R. Want, A. Hopper, V. Falcao, and J. Gibbons, The active badge location system. ACM Transactions on Information Systems 10(1): pp. 91-102. January 1992.

[5] R. J. Orr and G. D Abowd, The Smart Floor: a mechanism for natural user identification and tracking. In the Proceedings of the Conference on Human Factors in Computing Systems (CHI 2000), pp. 275-276, The Hague, Netherlands, ACM. April 1-6, 2000.

[6] M. Lamming, and M. Flynn, Forget-me-not: intimate computing in support of human memory. In the Proceedings of the FRIEND 21: International Symposium on Next Generation Human Interfaces, pp. 125-128, Meguro Gajoen, Japan. 1994.

[7] J. Pascoe, Adding generic contextual capabilities to wearable computers. In the Proceedings of the 2nd IEEE International Symposium on Wearable Computers (ISWC'98), pp. 92-99, Pittsburgh, PA, IEEE. October 19-20, 1998.

[8] M. Weiser, "The computer for the 21st century", Scientific American, 265(3):94–104, September 1991.

[9] A. K. Dey, and G. D. Abowd, Towards a better understanding of context and context awareness. In Proceedings of the Workshop on the What, Who, Where, When and How of Context-Awareness, affiliated with the CHI 2000 Conference on Human Factors in Computer Systems, The Hague, Netherlands. New York, NY: ACM Press, 2000.

[10] M. C. Mozer, The neural network house: an environment that adapts to its inhabitants. Proceedings of the American Association for Artificial Intelligence Spring Symposium on Intelligent Environments, AAAI Press (1998), Menlo Park , CA, pp. 110-114.

[11] B. Brumitt, B. Meyers, J. Krumm, A. Kern, and S. Shafer, EasyLiving: technologies for intelligent environments. Handheld and Ubiquitous Computing (2000), pp. 97-119.

[12] A. K. Dey, D. Salber, M. Futakaw and G. D. Abowd, An architecture to support context-aware applications. Submitted to UIST, 99.

[13] G. Chen, M. Li, and D. Kotz, Data-centric middleware for context-aware pervasive computing. Journal of Pervasive and Mobile Computing, Volume 4, Issue 2, pp. 216-253, 2008.

[14] J. Hong, The context fabric: an infrastructure for context-aware computing. CHI 2002, April 20-25, 2002, Minneapolis, Minnesota, USA. ACM 1-58113-454-1/02/0004.

[15] J. Rouillard, J-C Tarby., X. Le Pallec R. Marvie, Facilitating the design of multi-channel interfaces for ambient computing, The Third International Conferences on Advances in Computer-Human Interactions, ACHI 2010, St. Maarten, Netherlands Antilles, pp. 95-100.

 ISBN: 978-1-61208-224-0

Semantic Analysis of Medical Images Using Fuzzy Inference Systems

Norbert Gal[1], Vasile Stoicu-Tivadar[2]
Department of Automation and Applied Informatics,
"Politehnica" University of Timisoara,
Timisoara, Romania
E-mail[1]: norbert.gal@aut.upt.ro
E-mail[2]: vasile.stoicu-tivadar@aut.upt.ro

***Abstract*— Medical images carry information in a special data format about an organ and the related pathologies. Physicians must decipher this information from the image. This paper suggests a framework for analyzing the image on the semantic level using linguistic data. The framework implements several numerical algorithms to extract the physical features of the existing objects. The semantic interpretation for identifying the organs and possible pathologies from the found physical features is done using fuzzy inference systems. The clinical testing of the framework is a work in progress but the laboratory results are promising.**

Keywords-medical image processing; fuzzy systems; linguistic interpretation

I. INTRODUCTION

Medical images are one of the most basic and common medical diagnosis tools. The most common medical image formats are the X-ray images, the computer tomography images (CT), magnetic resonance images (MRI), the ultrasound images, angio CT, PET (Positron emission tomography) scan, and so on. For a trained eye they can describe accurately the internal organs of a human being and indicate the presence of pathologies. The first step in any medical image interpretation is the segmentation of the image. The trained eye can segment and analyze the image on the cognitive level. In contrast computers need specific algorithms for this task from the first step down to the last. The first step generally in image analysis is the segmentation process. By segmenting the image, the interested areas are separated from the background of the image. The next step is the feature extraction from the found objects. The interested physical features are the shape descriptors, the size of the object, its histogram and the location of the object. These concepts are used in medical image analysis where the objects in the medical image represent different organs and associated pathologies.

In the past years, several medical image analysis frameworks were developed [1]-[3]. One of the shortcomings of the frameworks is the lack of possibility to interpret the segmented object on the cognitive level. This paper proposes a framework that can analyze the object from the image on the semantic level. The physical features of the objects which are recorded using numerical values are transformed into "spoken language". For this process, fuzzy algorithms are implemented [4]. By using fuzzy inference systems coded in XML files [5], the framework can be adapted for different situations. Section two describes the methods which were used to extract numerical and semantic data from the medical image as well as several of the inference rules. Section three presents the case study of the pilot experiment and the first results.

II. USED METHODS

The coding of the information in medical images varies from one image type to another. The majority of the images use shades of gray to represent the reflectivity of the scanned object. For a computer to decode the information, specific methods must be used. The first step is to separate the objects found in the image from the background. The second step is to extract the numerical data from the image segments. These numerical data represent the physical features of the object. For analyzing the image on the semantic, level the numerical data must be converted in to linguistic data. This is done using a fuzzy inference system. This part will present the methods used by the framework to create linguistic results from regarding the data found in the image.

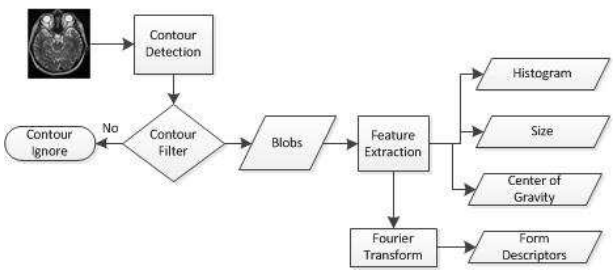


Figure 1. Segmentation workflow

 ISBN: 978-1-61208-224-0

A. Segmentation Process

The segmentation process is a mixed segmentation process. Figure 1 shows the diagram of the segmentation

Figure 2. The segmented blob, the contour and the masked image

process. First, the image is segmented using a manual segmentation which selects the area of interest. To the selected zone of interest, first the Otsu thresholding algorithm is applied [6]. Statistics are calculated for the two classes of intensity values (foreground and background) that are separated by an intensity threshold.

The criterion function is $\sigma_{Bi}^2 / \sigma_T^2$ for every intensity, i = 0,..,I-1, where σ_{Bi}^2 is the between-class variance, σ_T^2 is the total variance and I = 256, the maximum of the intensity gray level. The intensity that maximizes this function is the optimal threshold. By using this technique a mask image is created and is applied to the selected zone of interest.

The masked image is subjected to the second part of the segmentation process.

The first step in the interpretation process was to identify the threshold value for the segmentation. The formula for computing the threshold value is:

$$T_{\text{sh}} = \text{H}_{sg} + k \tag{1}$$

T_{sh} is the threshold value for which the segmentation process takes place. H_{sg} is the global histogram of the medical image, and k is a factor to control the threshold value. k is defined empirically. The best results are obtained for k = 8-10. These values were determined empirically using several test images.

The second part of the segmentation process is an automatic segmentation which records the existing objects from the masked image. The found objects are called “blobs” shown on Figure 2. These blobs have several physical features. These features are the area recorded in pixels, the location recorded by the center of gravity, the histogram on 3 channels and the bounding rectangle.

The shape of the blob is recorded using the Fourier transformation which provides the frequency components of the contour. These frequency components can be used to identify the shape of the organ.

B. Feature Extraction

The characteristics that are useful for the interpretation of the object are the shape of the object, the histogram value, the size of the object and the location of the object in the image.

For classifying the shape of the object, the Fourier transform can be used [7]. Each pixel from the contour of the object can be located by its Cartesian coordinates. These two coordinates are used to write the complex function of the object:

$$f(r) = x_r + j\, y_r \tag{2}$$

where r = 1… N-1 pixels. Fourier transform gives us the frequency components that make up the outline. This representation reduces the problem of analyzing the shape outline from 2D to 1D. The one-dimensional discrete Fourier transform of the function $f(r)$ is:

$$F(n) = 1/N \sum_{r=0}^{N-1} f(r) * e^{-j2\pi 2\pi r} \tag{3}$$

Excluding F (0), Fourier components do not depend on location of the analyzed shape, thus providing an efficient way to classify contours. To produce a more reliable shape, you can use a 'low pass' filter on the Fourier components, to remove the fine special structures. By computing the Fourier components of a closed contour and by ignoring the first component, the remaining components can be used for contour identification.

The size of the object is represented by the pixels found inside the blob. At this step, the physical size of the blob is measured in units of pixels. For a correct size determination the pixel size must be known.

In the case of DICOM (Digital Imaging and Communications in Medicine standard), images the pixel size can be computed from:

$$PixelSize = \frac{DOFV}{512} \tag{4}$$

The DFOV (display field of view) settings are 16, 20 and 50 for pediatric, head and whole body acquisition, respectively. The typical size of a CT image is 512 x 512 [8]. In case of non-DICOM medical images a segmented fuzzy inference system is proposed which is presented in section 3.

The color detection algorithm is based on determining the mean and variance of the pixel [9]. These methods were modified to take in account only one color plane as the majority of the medical images are in gray scale. The interesting features are the mean or average level of the gray level in the image and the variance or the contrast of the colors.

First, the image is converted into a gray scale image. For each pixel in the image the following algorithm is used:

$$G(x,y) = 0.2989R(x,y) + 0.587G(x,y) + 0.114B(x,y) \tag{5}$$

The coordinates of the pixels are noted using (x, y) is a sub-image of a specific size centered in (x, y). The mean value of the sub-image can be computed using:

 ISBN: 978-1-61208-224-0

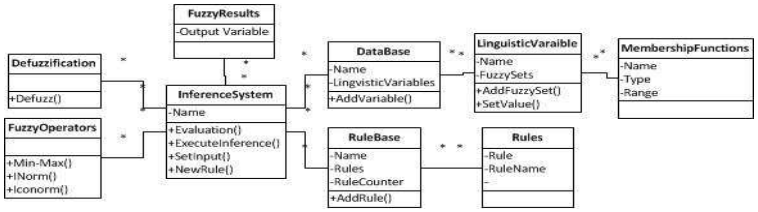


Figure 3. UML Class diagram of the fuzzy inference system

$$G_{ms} = \frac{1}{N} \sum_{(s,t)\varepsilon S_{xy}} G(s,t) \quad (6)$$

The variance is:

$$G_{vs} = \frac{1}{N} \sum_{(s,t)\varepsilon S_{xy}} \left| G_{ms} - G(s,t) \right| \quad (7)$$

The N is the total number of pixels in the sub-image. When all the values were calculated a color descriptor vector can be computed:

$$V = [L, W, G_{ms}, G_{vs}] \quad (8)$$

L and W represent the length and width of the sub-image. This color descriptor contains enough information for suitable color recognition.

The shades of gray represent solid and fluid objects. Than the black objects have a higher absorbance ratio then the white objects. This means that black objects are soft objects and the white spots represent solid objects like bones or calcifications.

The location of the object is recorded using the Cartesian coordinate system in pixels of the center. If the position of the object related to other objects (if it is vital e.g., in case of a tumor) it must be verified if the center of the second object is inside or not of the boundary of first objects and if the boundaries overlap each other or one boundary is inside or not of the other.

C. The Interpretation Process

The found physical features are converted to linguistic features using the fuzzification process. The linguistic features or fuzzy variables are used to identify the selected object. The selected object can be an organ or a malformation of the organ. The object can represent other malformations caused by other malformations or they can indicate other pathologies. The fuzzy variable is a collection of several linguistic features of the same type.

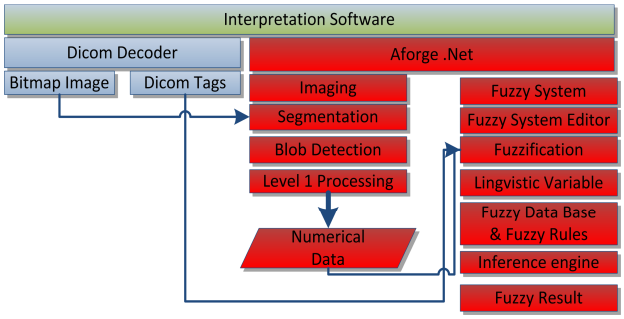


Figure 4. Framework Arhitecture

By using several fuzzy variables, which cover all the physical feature types, an inference engine is created. The suggested framework implements a rule based expert system. The rules are created in such a manner that they will cover as many possible options as they can.

These rules are very similar to the natural language communication. They can be compared with some clear instructions coming from one person to another. In their general form they have an antecedent and a consequence separated by the "THEN" statement. The antecedent is a conjunction of several fuzzy terms (using the statement IS) and several logical operators (AND, OR, NOT) between them.

For example: "*IF Size IS Long AND Histogram IS Black AND Size IS Small AND Location IS Head THEN Diagnosis IS Normal-Ventricle*"

The software implementation of the framework is done using visual C# programming language and AForge open source programming framework [10]. The framework was modified to load the fuzzy inference system from an XML (eXtensible Markup Language) file. This permits an easy and fast reconfiguring of the inference system.

III. TESTS AND RESULTS

The proposed architecture of the framework, shown on figure 4, is constructed to allow the interpretation of any type of medical image which is in the DICOM format [11], by applying minimal changes to the major parts of the interpretation system.

The framework is constructed from the combination of two open source software. The first open source software decodes the DICOM files and separates the DICOM TAGs (metadata referring to the patient and to the imaging device) from the image itself. The second open source software implements the segmentation and feature extraction process, as well as the interpretation software. The UML Class diagram of the interpretation process is presented in the figure 3. The interpretation software was modified to permit the loading of the fuzzy inference system from an XML file and the editing of the inference system as well. New rules and linguistic variables can be created. This permits an easy change and adaptation of the inference system, without stopping or decompiling the framework.

 ISBN: 978-1-61208-224-0

For the initial testing of the framework and the fuzzy inference system, we have considered the case of a 74 year old male with a confirmed malignant brain tumor.

For the initial numerical data acquisition, magnetic resonance image sets were chosen in the axial plane from the head section. Each image slice has a thickness of 5 and the space between the slices is 6.5. For testing the framework and the inference engine, the middle slices of the acquisition set were selected, slice 6 to slice 9. These slices contain the most useful data for building an inference system and a rule base.

The physical features of the found objects are presented in the table below:

TABLE I. OBJECT FEATURES

Object Name	Slice nr	Histogram	Size mm^2	Location (x , y)
Damaged ventricle	6,7,8,9,	132,150, 144, 112	6.14, 6.8, 6.8, 5,	250-118,
Normal Ventricle	6,7,8,9,	27, 29,-	2.5, 2.9,-	162-228 and 262-228
Brainmater	6,7,8,9,	76, 85, 82, 75	5.4, 4.8, 7, 3,15	Inside the cranium
Fragments	6,7,8,9,	54, 53, 49,43	0.5, 1, 0.9	Inside the cranium
Bones	6,7,8,9,	150, 155, 1.92	4.5, 3.6, 1,25	Inside the cranium

Using these numerical values the following linguistic variables and fuzzy rules were created:

a) Histogram: has a range from 0 to 255, where 0 from 50 is for the Black label, 40-70 stands for Dark, 70-90 DarkGray, 90-150 is Gray, 150-200 LightGray and 180 to 200 stands for White.

b) Size: has a range from 0 to 10, where 0 from to 1 is for the VerySmall label, 1-5 stands for Small, 4-7 Normal and 6 to 10 stands for Big.

c) Location: has a range from 0 to 512 on two axes (horizontal and vertical) to position the investigated objects center of gravity.

d) Shape: describes the shape of the invetigated object making use of the Fourier descriptors.

e) The Fuzzy Rules: in total 13 fuzzy rules were created to correctly identify the malformations and the soraunding tissues. Several rules that recognize the ventricular malformation are presented below:

<Rule12>IF Size IS Long AND Histogram IS Black AND Size IS Small AND Location IS Head THEN Diagnosis IS Normal-Ventricle</Rule12>

<Rule13>IF Size IS Round Histogram IS Gray AND Size IS Big AND Location IS Head THEN Diagnosis IS Abnormal-Ventricle</Rule13>

The first results are promising. Because the testing of the framework is a work in progress, the case of only one patient was tested. For the image set from which the numerical data was collected to build the inference system and the fuzzification process had a success rate of 100%. For the other image sets from the same patient but from the other image acquisition planes, coronal and sagittal the success rate dropped by 10%. New unforeseen situations had appeared for which new rules hade to be made.

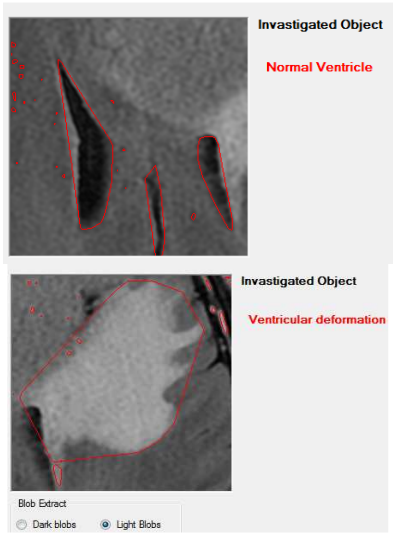


Figure5. A normal ventricle and a ventricular deformation

IV. CONCLUSIONS

The clinical testing of the framework is a work in progress. The numerical and the semantic level of the framework are completed. The used software architecture offers a high level modularity and adaptability to the framework. The framework can be adapted to new and different situations, new ideas and new conditions associated to a diagnosis. Unlike neural networks, where the learning mechanism is based on training data sets, this system permits the medical staff to create the inference rules and to debug the system in case of an error.

The first tests have promising results. For the image sets from the axial data acquisition plane the inference system had a 100% rate of success for identifying the tissue types, the organs and the malformation produced by the brain tumor. When testing the inference system for the other data acquisition planes the accuracy has dropped due to minor differences between the numerical data obtained from the axial plane and the other planes. However these tests were done in laboratory conditions to test the primary functions of the framework. The next step is to test the framework in real life clinical conditions using heterogeneous image sets from different clinical domains. A more complex testing of the framework to build the rules for the other body sections is scheduled for April 2012.

The described method has the advantage to decrease the number of clinical errors in imagistic interpretation. The medical staff will have a suggestion about the diagnostic, in this way reducing the stress level for patients and for the medical staff too.

 ISBN: 978-1-61208-224-0

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/88/1.5/S/50783 (2009) of the Ministry of Labor, Family and Social Protection, Romania, co-financed by the European Social Fund – Investing in People.

REFERENCES

[1] S. Li, T. Fevens, and A. Krzyzak, "A SVM-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets," *International Congress Series 1268, doi:10.1016/j.ics.2004.03.349*, pp. 207– 212, 2004

[2] J. Rivera-Rovelo and E. Bayro-Corrochano, "Medical image segmentation, volume representation and registration using spheres in the geometric algebra framework," *Pattern Recognition 40*, pp. 171 – 188, 2006.

[3] Y.M. Zhu and S.M. Cochoff, "An object-oriented framework for medical image registration, fusion, and visualization," Computer Methods and Programs in Biomedicine 82, pp. 258–267, April 2006.

[4] L.A. Zadeh, "Fuzzy sets", Information and Control, 8 pp. .338-353,1965.

[5] N. Gal and V. Stoicu-Tivadar, "XML as a Cross-Platform Representation for Medical Imaging with Fuzzy Algorithms," EFMI Special Topic Conference, EFMI STC, Lasko, Slovenia, April 2011, pp. 83-86.

[6] X. Xu, S. Xu, Lianghai Jin, and E. Song, "Characteristic analysis of Otsu threshold and its applications", *Pattern Recognition Letters*, Volume 32, 7, 1 May 2011, pp. 956-961.

[7] W. Duan, F. Kuester, Jean-L. Gaudiot, and O. Hammami, "Automatic object and image alignment using Fourier Descriptors", *Image and Vision Computing*, Volume 26, Issue 9, 1 September 2008, pp. 1196-1206.

[8] V.S. Smith, L.G. Shapiro, D. Hanlon, R.F. Martin, J.F. Brinkley, A.V. Poliakov, G.A. Ojemann, and D.P. Corina, "Evaluating Spatial Normalization Methods for the Human Brain," *27th Annual International Conference of the Engineering in Medicine and Biology Society*, 2005. IEEE-EMBS 2005, April 2006, pp. 5331-5334.

[9] N. Wang, and G. Wang, "Shape Descriptor with Morphology Method for Color-based Tracking" *International Journal of Automation and Computing* 04(1), pp. 101-108, 2007.

[10] http://www.aforgenet.com/ <retrieved: 01, 2012>

[11] O. S. Pianykh, "Digital Imaging and Communications in Medicine (DICOM)", 2008 Springer-Verlag Berlin Heidelberg, ISBN 978-3-540-74570-9.

 ISBN: 978-1-61208-224-0

Clinical Decision Support Based on Topic Maps and Virtual Medical Record

Valentin-Sergiu Gomoi, Daniel Dragu, Vasile Stoicu-Tivadar
Department of Automation and Applied Informatics,
"Politehnica" University of Timisoara,
Timisoara, Romania
e-mail: valentin.gomoi@aut.upt.ro, daniel.dragu@aut.upt.ro, vasile.stoicu-tivadar@aut.upt.ro

***Abstract*— Clinical decision support (*CDS*) systems are limited by the lack of access to medical data. This paper presents a system that takes advantage of different standards (*HL7 CDA, Arden Syntax*) in order to have access to various data sources. The system consists of units used for retrieving, in a standardized format, medical data, an inference module and a data manager part, that connects all the systems components.**

Keywords – HL7 standards; clinical decision support; Topic Maps.

I. INTRODUCTION

Using the medical decision support systems (*CDSS*) brought an improvement of healthcare act [1]. The outputs of these systems usually are free text recommendations, alarms and reminders for medical purposes. The first step in obtaining these is the interaction of two major types of actors: medical staff and IT specialists. In order to obtain computer interpretable "medical knowledge", the results of the discussions between the two types of actors are then put in a computer interpretable format, followed by discussions, verifications, implementation and updates (as result of incompleteness, new research, and so on) [1] [2]. Based on patient medical data (e.g., patient data: blood pressure, heart rate, etc.) and taking the advantage of having the medical knowledge in a computer interpretable format, the use of an inference engine lead to new medical recommendations. The development and implementation of *CDS* "include members with different expertise, including medical informatics" [2].

Advantages of using *CDS* are: faster the implementation of new medical knowledge, integration of local databases, decrease of costs, view patient data in a graphic manner for each step which should be followed, to avoid reading a vast amount of data regarding each step of a narrative medical guideline [1]. Multiple *CDS* (clinical decision support) solutions have been developed: *Asbru* [3], *GLARE* [4], *GASTON* [5], *Egadss* [6], *Gello* [7], etc.

Studies about statistical data in the establishment of these support systems in healthcare are presented in [2]. One of the statistics shows that the use of *CDS* increased physicians' performances in 64% of the studies and regarding the patient outcomes 13% of studies established a benefit. In [8], it is stated that in over 90% randomized controlled, clinical practice improved, based on the use of *CDS*.

There is a major challenge of these systems, the interoperability between them and other medical information systems (*EHR*, *EMR* or different medical local databases) [2]. New steps in solving this problem are made by *HL7* group by developing a data model that wants to become a standard for the representation of medical knowledge for *CDS* (clinical decision support), the name of the data model is *virtual Medial Record* (*vMR*).

During the time, those they work with data, information and knowledge were confronted with the need of a data model able to represent their real words into machine-readable formats. There were developed a lot of data models claiming that they are the most eligible for some specific tasks, and they were, but they also have limits, and sometimes, the models were designed with a high machine-readability, but they lack in human-readability.

In this paper, we present a solution that implements the proposed *vMR* and other standards as: *Topic Maps* (used for the representation of the *vMR*), *HL CDA* (*Health Level 7 Clinical Document Architecture*), *Arden Syntax* (formalism for the representation of medical rules). This project is focused on the development of an application that has to store data from different medical documents and to develop a connection with *CDS* based on documents in a *vMR* format/vocabulary to improve the healthcare act by allowing the access to more vast and relevant clinical information in order to generate more accurate clinical recommendations.

This paper is structured as follows. System architecture and used technologies are reveled in section two. Different standards and the benefits they bring are presented in section three. Section four illustrates aspects of implementing the *vMR* with the help of *Topic Maps* and the connection of the *vMR* with an existing solution. Conclusion and future research directions are presented in section five.

II. SYSTEM ARCHITECTURE

The project has as main components: *Data manager*, *Interface*, *Inference engine*, *HL7 CDA Component* and *TM-vMR* (for the representation of medical information with the

 ISBN: 978-1-61208-224-0

help of *Topic Maps* technologies and implementing *vMR* data model). All these components are further presented in the next sections. A first model was presented in [9]; in this paper a more complex and detailed architecture is revealed (software technologies, implementation of various modules or the interaction between modules). In Figure 1, an overview of the system architecture is presented. In order to implement this architecture and the different standards, several tools are used:

- *.Net* with *C#* for the development of the *Interface*, *Database* services and *Data Manager*,
- *Java* for the inference part, using *Arden Syntax* formalism for the representation of medical rules
- *Topincs* (*PHP* based) for the implementation of the *vMR* data model with the help of *Topic Maps (TM)* resulting the *TM-vMR* module

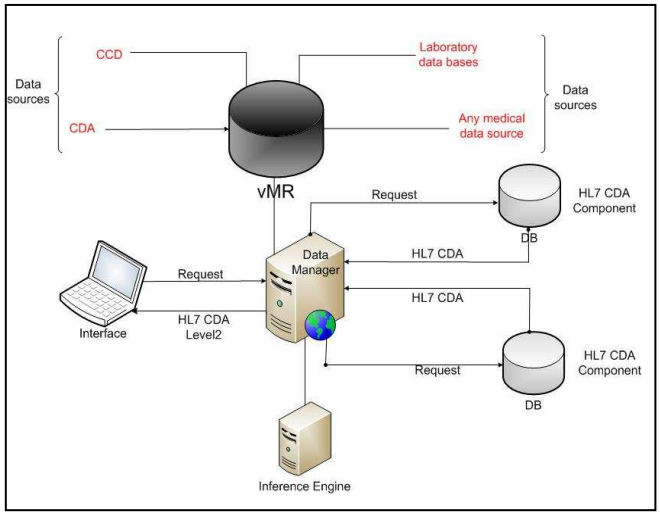


Figure 1. System Arhitecture

Concerning *TM-vMR* the domain which has to be represented was already sketched. The concepts and the relationships between them were defined in *vMR DAM (Domain Analysis Model)* [10]. They were further used to create a *TM* schema that allows populating the topic map in a schema-driven way. The software used to represent that domain is *Topincs*, "a software for rapid development of web databases" [11].

The implementation of this architecture will allow:

- to create the basis for a further integration of medical knowledge sources that are encoded in various technologies,
- the connection of any *vMR* compatible *CDS* to a medical web database,
- extensibility for *vMR DAM*, given by the possibility to connect it to virtual any other electronic knowledge source, through the *TMDM (Topic Maps Data Model)*
- capacity of representing "anything whatsoever". TM can be viewed as an envelope for any other knowledge representation, and there are studies about how to represent the most important of these *KM* (Knowledge Representation) technologies with the help of *TM* technology [12],
- to offer a package of services especially designed for *CDS* developers,
- to create a tool that permits users to use a database without the need of knowledge about the technology that database stands on,
- to bring a semantic technology, with unlimited applicability within the domain of information representation,
- development of a collaborative solution for the capturing of medical information.

III. USE OF DIFFERENT STANDARDS

The reasons for developing a *CDS* that is implementing the different standards for the communication and knowledge representation are presented in the next paragraphs.

A. *vMR*

As stated in [2] [13] [14] [15], important reasons for limitation of implementing *CDS* in medical units are: the use of different models, the lack of a standard representation of clinical information and terminologies associated that are used in medical institutions. To meet these needs, the *Working Group HL7 CDS vMR* initiated the *vMR* project, which had as objective to support the development of scalable and interoperable *CDS*, by establishing a "standard model to represent clinical information inputs and outputs that can be transmitted between systems *CDS* and other medical systems." [15]. This model contains 131 medical data elements. The *vMR* data model appeared as a necessity for the interoperability between different *CDS* and data sources. This data model allows the representation of a large range of information concerning the patient. In the development of the *vMR* model (the patient data and the requirements to be integrated in) 22 institutions from 4 different countries, have been involved (representing 20 *CDS* systems). *vMR* data model is in process of becoming a standard for the representation of medical knowledge used in different *CDS* systems in order to solve the problem of interoperability [16]. This model is not mandatory to be used as the unique data source for *CDS* but it can be used for the interoperability of the existing system. Other representations (e.g., *HL CDA*) of medical data can be used in order to make the *CDS* more adaptable to a local context. The *vMR* will be represented with the help of *Topincs* (a *Topic Maps* open source software) [11]. The access to medical data will be made (through the *Data Manager*) with the help of different service already existing in this software and new ones which are to be developed.

 ISBN: 978-1-61208-224-0

B. Topic Maps

TMs deliver the right information in the right context to the right person at the right time [17]. A technology that can represent "anything whatsoever" [18]; such a technology is very useful as a response to one of the main requests of this project, to integrate many medical knowledge sources at the input of *CDS* systems. *Topic Maps* can "be applied to any application domain you can think of" [17].

One of the most important challenges in topic map authoring is to keep the level of "subjectivism" of the topic map as low as possible. That means that any topic map author leaves a personal "fingerprint" on the final representation, firstly depending on his/ her individual knowledge about the represented domain and his/her ability to view, conceptualize and represent subjects within the application domain.

"Topic Maps is a technology for encoding knowledge and connecting this encoded knowledge to relevant information resources. Topic maps are organized around topics, which represent subjects of discourse; associations, representing relationships between the subjects; and occurrences, which connect the subjects to pertinent information resources." [18]. *Topic Maps* has as attributes: semantic, semantic web, extensibility, flexibility, high representative power, envelope for any other *KR* (knowledge representation) technologies, human readability and computer readability, standard, XML-based syntax as interchange format, smart navigation, subject-centric approach for *KR*, rapid information retrieval, source integration, open vocabulary, possibility to get different views of the same assertion. A presentation of how the *TM* technology can interact with *CDS* can be found in [19], where is explained the way in which the use of such a technology can improve *CDS*.

C. Other standards

Semantic Web technologies are used to create data stores on the Web, build vocabularies, and write rules for data handling. XML (Extensible Markup Language) and ontology (e.g., Web Ontology Language) are two components of the Semantic Web. In our case the XML is a *HL7 CDA* message in XML format. The *HL7 Clinical Document Architecture (CDA)* is a document markup standard that specifies the structure and semantics of clinical documents for the purpose of exchange. A *CDA* document is a defined and complete information object that can include text, images, sounds, and other multimedia content. *CDA* documents are encoded in Extensible Markup Language (XML). The clinical content of *CDA* documents is defined in the *RIM (Reference Information Model)* [20] [21].

The *Arden Syntax* is a clinical guideline formalism accepted as an official standard by *HL7* (textual language and intuitive). It is freely available, a mature and actively maintained open standard. This is the reason why *Arden Syntax* is used instead of other guideline formalisms as *Proforma, GLIF (GELLO), Asbru*, etc. [6] [22].

IV. System Components and Implementation of Different Standards

An overview of the CDS systems shows a large number of approaches (*Asbru, Proforma, GLARE*). Almost all CDS allow medical guidelines and protocols to be generated, edited, verified, executed (reasoning based on medical rules and medical databases) and visualized. In order to represent the medical knowledge there are used different technologies: rule based technology (*Arden Syntax*) or task network (*Asbru, Proforma, EON*), unique for all of them [3]. The various CDS depend very much on the medical local databases sometimes they being developed around them (databases). All of these approaches are usually hard to be deployed in different medical units, as stated in section two. In order to overcome these gaps we propose a system which brings the advantage of using well known standards (as main inputs) and also implements a very powerful knowledge representation technology (*Topic Maps*).

In this section, the main components of the system are presented.

A. Existing CDS

To obtain information from different database that can be a laboratory or radiology database and then represent the information in *HL7 CDA (HL7 Clinical Document Architecture)* format and send it to the decision system, the *HL7 CDA Component* has been developed. The *HL7 CDA Component* implementation is made in *Visual Studio .Net 2008*, in *C#* language (as a web service). The databases for the current activity are on *SQL Server 2008*, but the solution is similar for *Oracle* or *MySQL*

The inferring engine is based on Egadss open source solution [6] [22]. In order to have a standardized communication interface between databases and "recommendation generator" - *Egadss* has as inputs *HL7 CDA* (*Level 3*) standard messages as XML files, where the patient data is represented (XML retrieved from the *HL7 CDA Component*). Another standard used by *Egadss* is *Arden Syntax*, which is a clinical guideline formalism accepted as an official standard by *HL7* group, being used for the representation of medical rules. The result of the inference is *CDA Level 2* document, containing the medical recommendations [6] [22].

The communication between all the components of the *CDS* is based on web services, representing de *HL7 CDA Components*, decision making system or the interface. To manage the connection and the order in which the different web services are called, a *Data Manager* was developed. *Data Manage* has the roles to respond at different requests that come from the main components of the system (interface, medical data source, inference engine). In order to realize this, three communication channels are opened (see Figure 1), with: *Interface*, *HL7 CDA Component* and

 ISBN: 978-1-61208-224-0

the *inference engine (Egadss).*The interface allows the medical staff and the patients to visualize the steps of the protocols, medical information regarding a patient; different alarms and they can insert feedback concerning the recommendations. The interface is implemented using *ASP.Net* platform with C#. A more detailed description can be found in [23]. Beside the use of *HL7 CDA documents* other sources can be added to the system through the *Data Manager*. One of these sources can be a *virtual Medical Record (vMR)* that implements the specifications from [10]. The *vMR* is implemented with the help of *Topic Maps*.

B. TM-vMR

The implementation of this module is based on the *vMR* data model and *Topic Maps* technology. The representation of the data model is realized with the help of the *Topincs* open source software.

Implementation steps:

- the strategy used to convert *vMR* terms into the *Topic Maps* constructs is to create a topic type for every *vMR* class within the *vMR DAM* atomic terms; the attributes of a class become occurrence types for the topic type corresponding to that class (Figure 2).
- to model all relationships within the *vMR DAM*
- populating the topic map can be done manually, but a tool for automatic data integration into the topic map is being to be developed.
- evaluate the results, by connecting the database with a *CDS*

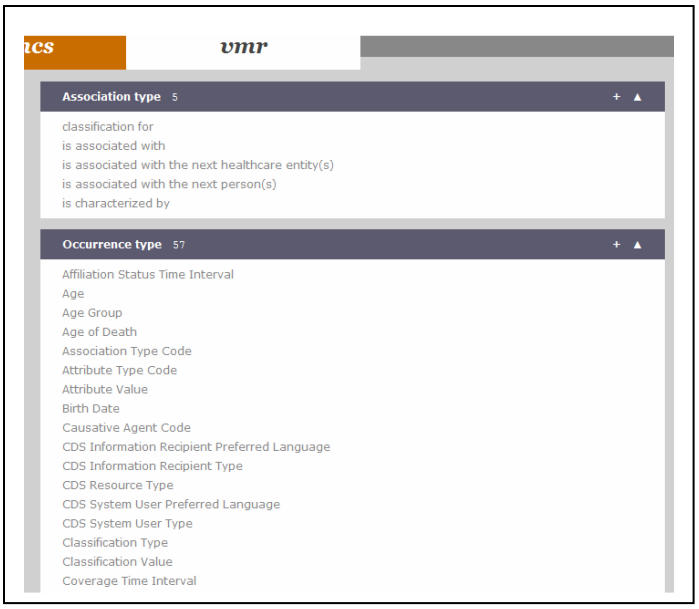


Figure 2. TM-vMR representation

To achieve the possibility of custom development of the *Topincs-based* application, the author of the topic map schema has to define the serialization names for all terms within the schema. In this way, a programming interface is automatically provided by the software, allowing the topic map author to use the exposed methods, in order to program the behavior of the application in such a way to get a better response to the requests of the project. To ease the rapid discovery of a topic map construct that a serialization name represent, a notation convention for serialization names was used. The exact form of the name of any *vMR* data model term was used for naming topic map types within the schema. Serialization name keeps the name but normalization was required:

- for topic types: all capital letters will be converted in lower-case, and the spaces between words will be replaced with minus sign "-";

- for any other topic map construct within the schema, the correspondent serialization name is created by adding a prefix which has the role to show what type it represents. (This convention was made for the further development of the project)

C. CDS connection to TM-vMR module

The connection between the existing system and the *TM-vMR* represented with the help of *Topincs* open source is realized with the help of the web services. These services are consumed from the *Data Manager*. This is a client server architecture where the server is the *TM-vMR* and the client is the *Data Manager*.

NuSOAP is the technology used for the development of web services in *PHP* for the access to the *TM-vMR* ontology. These services allow the work with the patient data that is represented in the *TM-vMR* ontology trough the help of a topic map objects which in "*Topincs*" is called "*tobject*". The *tobjects* allow the insertion, deletion, modification and many other types of special functions to work with patient data from the *TM-vMR*. The web services allow the connection of the model with the existing CDS trough the *Data Manager*. The *Data Manager* calls the web services from the *vMR*, based on the data needed for a certain patient (based on patient ID) for a set of medical rules in order to generate new medical recommendations.

Regarding privacy/integrity issues our system should achieve the *HIPAA* (Health Insurance Portability and Account-ability Act) requirements; the first step was implementing the communication over *HTTPS*.

V. CONCLUSIONS AND FUTURE WORK

Regarding the contribution the implementation of the different standards and the use of *Topic Maps* in the presented system lead to: integration of medical knowledge sources that are encoded in various technologies, extensibility for *vMR* data model and connection of any *vMR* compatible *CDS* to a medical web database.

This implementation allows an easier integration of the system (compared with systems that do not implement *HL7* standards) in different medical units allowing the connection with various types of data sources. With the help of the

 ISBN: 978-1-61208-224-0

mentioned technologies the *vMR* was represented (Figure 2). Different web services are developed in order to have access to the different elements of the *TM-vMR*.

In this early stage of the development, the topic map can be used only for browsing through the elements of the schema (topic types, association types, constraint types).

Further developments of the system consist in: the implementation of other web services for a better interrogation and manipulation of the medical knowledge from *TM-vMR*; the development of an automatic way to integrate the medical sources in the *TM-vMR;* use of smart cards for the authentication of different actors (for data privacy).

The implementation of presented system will help the medical staff to increase the quality of medical care by: reducing the variation in medical practice, giving more efficient treatments and using new medical knowledge in current clinical practice.

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU 107/1.5/S/77265, inside POSDRU Romania 2007-2013 co-financed by the European Social Fund – Investing in People and the strategic grant POSDRU/88/1.5/S/50783, Project ID50783 (2009), co-financed by the European Social Fund – Investing in People, within the Sectorial Operational Programme Human Resources Development 2007-2013.

REFERENCES

[1] K. Rosenbrand, J. van Croonenborg, and J. Wittenberg, ”Guideline Development”, in Computer-based Medical Guidelines and Protocols: A Primer and Current Trend, A. ten Teije, S. Miksch, and P.J. Lucas, vol. 139, pp. 3 -21, 2008.

[2] A. Latoszek-Berendsen, H. Tange, H.J. van den Herik, and A. Hasman, “From clinical practice guidelines to computer-interpretable guidelines. A literature overview.”, in Methods of Information in Medicine (2010), vol. 49, Issue 6, pp. 550-570, 2010.

[3] P. Clercq, K. Kaiser, and A. Hasman, ”Computer-Interpretable Guideline Formalisms”, in Computer-based Medical Guidelines and Protocols: A Primer and Current Trend, A. ten Teije, S. Miksch, and P.J. Lucas, vol. 139, pp. 22-43, 2008.

[4] P. Terenziani, S. Montani, A. Bottrighi, G. Molino, and M. Torchio, “Applying Artificial Intelligence to Clinical Guidelines: The GLARE Approach”, in Computer-based Medical Guidelines and Protocols: A Primer and Current Trends, A. ten Teije, S. Miksch, and P.J. Lucas, vol. 139, pp. 121-139, 2008.

[5] P.A Clercq, J. A. Blomb, A. Hasman, and H.M. Korstenb, “GASTON: an architecture for the acquisition and execution of clinical guideline-application tasks”, in Med Inform Internet Med, pp. 247-63, 2000.

[6] J. H. Weber-Jahnke and G. McCallum, “A light-weight component for adding decision support to electronic medical records”, in Hawaii International Conference on System Sciences, Proceedings of the 41st Annual, ISBN: 978-0-7695-3075-8, pp. 251 - 251, 2008.

[7] M. Sordo, A.A. Boxwala, O. Ogunyemi, and R.A. Greenes, “Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support.”, in Studies In Health Technology And Informatics, Volume: 107, Pages: 164-168, 2004.

[8] K . Kawamoto, C.A. Houlihan, E.A. Balas, and D.F. Lobach, “Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success” in *BMJ*, pp. 765-768, 2005.

[9] D. Dragu, V. Gomoi, and V. Stoicu-Tivadar, “Automatic generation of medical recommendations using topic maps as knowledge source”, in 6th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), SACI2011, ISBN: 978-1-4244-9108-7, pp. 593 – 597, 2011.

[10] Virtual Medical Record (vMR) for Clinical Decision Support – Domain Analysis Model, http://wiki.hl7.org/images/archive/6/6b/20110729073300!HL7vMR_vMR_Domain_Analysis_Model_2011_Sept_Ballot.pdf, accessed in 05.12.2011.

[11] R. Cerny, “Topincs: A software for rapid development of web databases”, http://www.cerny-online.com/documents/Topincs%20-%20A%20software%20for%20rapid%20development%20of%20web%20databases.pdf, accessed in 21.12.2011

[12] L. M.. Garshol, “Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all”, Journal of Information Science, Vol. 30, pp. 378-391, 2004.

[13] J.A. Osheroff, J.M. Teich, and B. Middleton, E.B. Steen, A. Wright, D.E. Detmer, “A roadmap for national action on clinical decision support”, in J Am Med Inform Assoc., pp. 141-145, 2007.

[14] K. Kawamoto, “Integration of knowledge resources into applications to enable clinical decision support: architectural considerations.” in Greenes RA, Clinical Decision Support: the Road Ahead. Boston: Elsevier, pp. 503-538, 2007.

[15] K. Kawamoto et al., “Multi-National, Multi-Institutional Analysis of Clinical Decision Support Data Needs to Inform Development of the HL7 Virtual Medical Record Standard”; in AMIA Annu Symp Proc. pp. 377–381, 2010.

[16] Virtual medical record, http://wiki.hl7.org/index.php?title=Virtual_Medical_Record_%28vMR%29, accessed in 23.11.2011.

[17] H.H. Rath, White Paper –The TM Handbook, http://www.sts.tu-harburg.de/~r.f.moeller/lectures/anatomie-i-und-k-system/empolistopicmapswhitepaper_eng.pdf, accessed 21.12.2011

[18] ISO/IEC 13250-2:2006 (the Topic Maps Data Model) – accessed in 23.11.2011.

[19] D. Dragu, V. Gomoi, and V. Stoicu-Tivadar, “Topic Maps as knowledge base toautomatically generate medicalrecommendations”, in 2011 IEEE 9th International Symposium on Intelligent Systems and Informatics (SISY), ISBN: 978-1-4577-1975-2, pp. 459 - 464, 2011.

[20] B. Blobel , K. Engel, and P. Pharow , “HL7 Version 3 Compared to Advanced Architecture Standards”, *Methods of Information in Medicine* , pp. 343–353, 2006

[21] HL7 Clinical Document Architecture, Release 2.0, HL7 version 3 Interoperability Standards, Normative Edition 2009, Disk 1 – Standards Publication, 2009.

[22] I. Bilykh, J. H. Jahnke, G. McCallum, and M. Price, “Using the clinical document architecture as open data exchange format for interfacing EMRs with clinical decision support systems”, in Proceedings of the 19th Symposium on Computer-Based Medical Systems (CBMS’06), pp 855-560, 2006.

[23] V. Gomoi and V. Stoicu-Tivadar, “A new visualization solution for medical computer based protocols”, in Proc. of 9th International Conference on Information Comunication Technologies in Health 2011 (ICICTH-2011), pp. 82-89, 2011.

 ISBN: 978-1-61208-224-0

Measuring the Interoperability Degree of Interconnected Healthcare Information Systems Using the LISI Model

Mihaela Vida*, Lăcrămioara Stoicu-Tivadar*, Elena Bernad**,
*Faculty of Automatics and Computers, University "Politehnica" of Timişoara,
Timişoara, Romania
** Department of Obstetrics and Gynecology, University of Medicine and Pharmacy "Victor Babes"
Timisoara, Romania

Email: mihaela.vida@aut.upt.ro , lacramioara.stoicu-tivadar@aut.upt.ro, ebernad@yahoo.com

***Abstract*—Due to the diversity of information systems in healthcare and the need of accessing data in a ubiquitous and pervasive manner, the interoperability issue has grown in importance. This work presents how the Levels of Information System Interoperability model can be applied to study the interoperability degree in order to interconnect healthcare information systems. This work presents an algorithm adapted for healthcare information system, which can determine the message exchange rate between healthcare information systems. The analysis is done looking at a hospital department (obstetrics-gynecology), general practitioner offices, radiology departments and laboratories that work together and have different information systems. This algorithm computes the interoperability degree from the technical interoperability point of view. A tool which calculates automatically the technical interoperability of a healthcare information system, based on the proposed algorithm, is under development. The benefits resulting from the calculus of the interoperability degree are reflected in the assessment of the status of informatization and degree of intercommunication in a certain healthcare environment. Also, it is helpful for software developers to know what is expected from a good application for the domain.**

Keyword-LISI; HL7 CDA; CCD; interoperability; healthcare information system.

I. INTRODUCTION

Increased life expectancy and the consequent increase in the prevalence of chronic illnesses pose serious challenges to the sustainability of the national health systems in Europe.

Seamless care is the desirable continuity of care delivered to a patient in the healthcare system across the spectrum of caregivers and their environments. Healthcare services have to be continuous and carried out without interruption such that when one caregiver ceases to be responsible for the patient's care, another one takes on the responsibility for the patient's care. Such a paradigm poses serious problems regarding the interoperability between healthcare information systems.

Interoperability is the ability of two or more systems or components to exchange information and use the information that has been exchanged [1].

Interoperability might be provided at different levels. These interoperability levels can start from simple data exchange and meaningful data exchange with agreed vocabulary to functional interoperability with agreed communication application behavior, or finally, a service-oriented interoperability [2].

Communication between different systems and their components in a complex and highly dynamic environment must fulfill some requirements: openness, scalability, flexibility, portability, distribution, standard conformance, service-oriented semantic interoperability and appropriate security and privacy services. This communication is based on a standard (e.g., HL7 version 3, HL7 Clinical Document Architecture). [3]

The Electronic Healthcare Record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting, including information about: patient demographics, progress notes, problems, medications, vital signs, medical history, immunizations, laboratory data and radiology reports [4].

The paper presents a particular environment studying the communication level between healthcare information systems for hospitals, laboratory, radiology and general practitioner offices. The major problem is that the healthcare information systems do not communicate directly one with the other making it impossible to create an electronic medical record seamless and following a timeline. The work gives an image of the current situation for which the analysis is made.

If the degree of interoperability between healthcare information systems can be evaluated, it will have benefits for assessing the status of informatization and degree of intercommunication in a certain healthcare environment and also for software developers to know what is expected from a good application for the domain.

 ISBN: 978-1-61208-224-0

Also, it is important to improve the interoperability of healthcare information systems and add more information to Electronic Health Record (EHR).

In section two, is presented the standards used in healthcare information systems communication. Section three presents the interoperability study where is described the LISI model and it is measured the degree of technical interoperability and, Section four concludes the paper solutions.

II. STANDARDS USED IN HEALTHCARE INFORMATION SYSTEMS COMMUNICATION

One of the mandatory criteria to ensure the interoperability between the healthcare information systems is to use a standardized communication. In the next paragraphs, a system architecture and the standards used for communication between components is presented.

A. System architecture using standards

Figure 1 presents the system architecture using standardized communication. The system consists of three healthcare information systems for the obstetrics-gynecology, for radiology and for analysis laboratory communicating using the HL7 CDA and a healthcare information system for the general practitioner office which communicates with the hospital departments using the CCD (Continuity of Care Document) standard [5], [6].

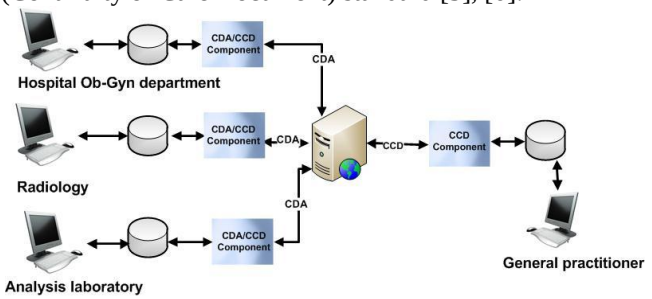


Figure 1. System architecture

The developed two Components, one for the CDA/CCD standard and the second for the CCD standard, give the possibility to extract data needed from the medical unit database (from obstetrics-gynecology, radiology, laboratory or general practitioner database). The two Components are developed in Visual Studio .NET 2008, using C# language. For the moment, the Components can extract data from a SQL Server database, but in the future will be generalized in order to extract data from different databases types. A connector was developed that extracts from XML in CDA/CCD format the data and inserts it into the proper fields and tables in database.

B. Using HL7 CDA (Clinical Document Architecture)

The HL7 CDA is a document markup standard that specifies the structure and semantics of clinical documents. The developed HL 7 CDA Component extracts the data from a local database and presents it as an HL7 CDA standard message. The CDA derives its content directly from the HL7 Reference Information Model (RIM) and therefore is specially design to integrate current HL7 technologies. The common architecture can be adapted for progress notes, radiology reports, discharge summaries, transfer notes, medications, laboratory results and patient summaries. The CDA is an XML document that consists of a header and body [7].

The HL7 CDA/CCD standard uses Logical Observation Identifiers Names and Codes (LOINC). This is a universal code system for identifying laboratory and clinical observations [8], adapted in this case for the Romanian healthcare system.

An XML in CDA format, as an example of a message from a lab, is presented in Figure 2.

```
<!--******************Labs Section******************-->
<component>
  <section>
    <code code="11502-2" codeSystem="2.16.840.1.113883.6.1"
     codeSystemName="LOINC" displayName="Labs" />
    <title>Blood test</title>
    <entry>
      <observation classCode="OBS" moodCode="EVN">
        <code code="19180-9" codeSystem="2.16.840.1.113883.6.1"
         codeSystemName="LOINC" displayName="beta-HCG" />
        <effectiveTime value="20110402" />
        <value xsi:type="PQ" value="15000" unit="mUI/mL" />
      </observation>
    </entry>
  </section>
</component>
```

Figure 2. CDA laboratory result

The CDA contains LOINC codes, which are used for representation of the laboratory results (e.g., LOINC code 19180-9 is used for beta-HCG analysis) and also the analysis value (in Figure 2 the beta-HCG value is presented - 15000 mUI/mL). All the LOINC codes used in this CDA message are adapted for Romanian healthcare systems.

C. Using CCD (Continuity of Care Document)

The Continuity of Care Document (CCD) is an electronic document exchange standard for sharing patient summary information among providers and within personal healthcare records. It summarizes the most commonly needed pertinent information about current and past health status in a form that can be shared by all computer applications, it respects a set of constrains on CDA that define how to use the HL7 CDA to communicate clinical summaries and it is built using HL7 CDA elements [9].

CCD is a combination between ASTM CCR (Continuity of Care Record) and HL7 CDA.

The definition given for CCD by ASTM is: a core data set of the most relevant administrative, demographic and clinical information facts about a patient's health care, covering one or more health care encounters [10].

 ISBN: 978-1-61208-224-0

CCD templates include: header, purpose, problems, procedures, family history, social history, payers, advance directives, alerts, medications, immunizations, medical equipment, vital signs, functional states, results, encounters and plan of care [9].

In the current healthcare information system, the CCD standard for communication is used to support the communication between the hospital departments and the general practitioner's office. The general practitioner sends a request in XML format containing the ID (personal numeric code – CNP, which in Romania is the unique ID for each person) to the hospital department application and the CDA/CCD Component extracts the data from the hospital department database and sends the information in CCD format to the general practitioner office.

In Figure 3, an XML sequence in CCD format is presented containing lab results sent from one of the hospital departments to the general practitioner's office and it is adapted for the Romanian health system.

```
- <component>
  - <observation classCode="OBS" moodCode="EVN">
      <templateID root="11"/>
      <code displayName="Eritrocite" codeSystem="2.16.840.1.113883.6.1" code="11273-0"/>
      <statusCode code="completed"/>
      <effectiveTime>20110515</effectiveTime>
      <value value="5.36" unit="x10^6/uL" xsi:type="PQ"/>
    - <methodeCode codeSystem="2.16.840.1.113883.5.84" code="460179">
      - <referenceRange>
        - <observationRange>
            <text>4.00-5.80 x10^6/uL</text>
          </observationRange>
        </referenceRange>
      </methodeCode>
    </observation>
</component>
```

Figure 3. CCD example

The XML in CCD format contains a laboratory result: erythrocytes, which are codified with LOINC code 11273-0, adapted for Romanian health system and the value of this test result.

III. INTEROPERABILITY STUDY

A. LISI model

LISI (Levels of Information System Interoperability) is a complete, descriptive model of classification with levels of interoperability based on individual, unique project specifications [11].

LISI is a reference model for assessing information systems interoperability. It is used for defining, measuring, assessing, and certifying the degree of interoperability required or achieved between organizations or systems [11].

B. LISI Interoperability Maturity Model

LISI Interoperability Maturity Model has 5 levels [11]. In this paper and previous work [12] these levels are adapted for healthcare informatics systems.

The LISI levers are:

- Level 0 named Isolated (Environment: Manual)
- Level 1 named Connected (Environment: Peer-to-Peer)
- Level 2 named Functional (Environment: Distributed)
- Level 3 named Domain (Environment: Integrated)
- Level 4 named Enterprise (Environment: Universal)

To fit into a LISI level, we studied two types of interoperability: operational and technical. Two scores obtained from analyzing the two interoperability types will result representing the interoperability degree of the studied healthcare information system. A scale corresponding for each LISI level will be considered (e.g., if the scale is 0 the level is Level 0 - Isolated).

C. LISI Scope of Analysis

In Figure 4, the LISI scope of analysis for two HIS systems are presented. The operational interoperability has a semantic understanding. For each XML received in CDA or CCD format a tool will analyze the codes (LOINC or ICD-10-AM) and if all the analyses corresponds to the evaluation criteria then the healthcare information system will receive a score (a scale to 1 – 100). A similar analysis is presented in [13], where SNOMED codes are analyzed. Scoring the technical interoperability it will be possible to appreciate on what LISI level the healthcare information system is situated.

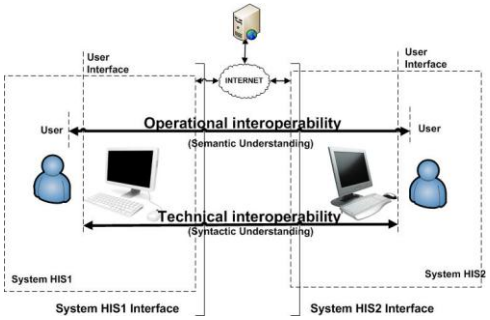


Figure 4. LISI Scope of analysis

The technical interoperability is the condition achieved among electronic system communications when information or services can be exchanged directly and satisfactory between them and their users, includes aspects such as application interfaces, open standards and data integration. If two or more healthcare information systems are capable of communicating and exchanging data, they are technical interoperable. In general, XML or SQL standards provide syntactic interoperability. In this work, an algorithm that determines the technical interoperability is presented.

D. Measuring the degree of technical interoperability

In Figure 5, the studied healthcare system architecture is presented comprising the obstetrics-gynecology department, 2 radiology (1 internal and 1 external) departments, 4 analysis laboratories (1 internal and 3 external), and 1 general practitioner office. The technical interoperability

 ISBN: 978-1-61208-224-0

degree for the obstetrics-gynecology healthcare information system is studied below. This healthcare information system communicates using standards, with the radiology and analysis laboratory using HL7 CDA, and with the general practitioner using CCD.

A scale is proposed to evaluate systems interoperability potential for technical interoperability point of view:

- 0 – 35 points the systems are not interoperable that means that the system is on LISI level 0 or level 1,
- 36 – 65 points the systems are interoperable in some degree that means that the system is on LISI level 2 or 3,
- 66 – 100 points the systems are interoperable that means that the system is on LISI level 4.

To study the interoperability degree an algorithm [11] is applied adapted for healthcare information systems.

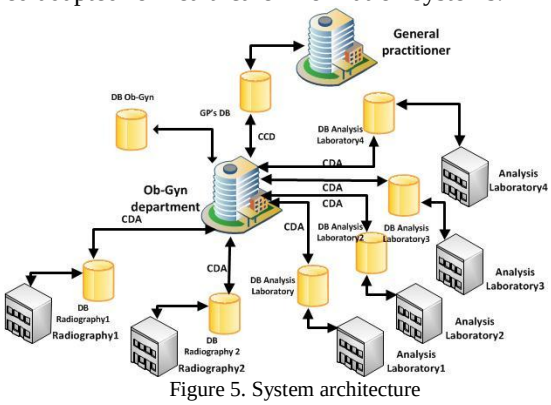


Figure 5. System architecture

For each step of the testing process, points have been associated in order to reflect the interoperability score for the systems. For each steps, a score is allocated; it represents how well the system meets the requirements (e.g., if the system has the possibility to communicate using standards, and for how many system are interconnected).In the next lines, the testing steps are presented:

Step 1 - Analyzing if the functionalities are the same

In order to establish that the system functionalities are the same, the data workflow and management between obstetrics-gynecology, radiology, laboratory and general practitioner was monitored during a week at County Emergency Hospital Timisoara, Romania – Bega Clinic, obstetrics-gynecology department. The referrals were studied and the data sets were identified and based on these, the conclusion was that the system functionalities are the same. The scored obtained at this step is 7/10.

Step 2. – Study the degree in which if the communication is based on the same standards.

We assumed that the messages are transmitted with the help of HL7 CDA standard in laboratory analysis, radiology cases and CCD for the general practitioner. For each case two Components were developed, one which extracts data for creating the CDA and the second to extract data for CCD. The score obtained in this step is 9/10.

Step 3. Analyze if the message data elements are common.

The data elements are common because the ob-gyn department sends referrals to the analysis laboratory, radiology, and the general practitioner office and receives back the same type of documents. All the communication between healthcare information systems presented here is based on CDA and CCD standards. The scores obtained at this step is 7/10.

Step 4. Calculate the connectivity index with the formula:

$$c_i = \frac{k}{n*(n-1)}$$

where: c_i = connectivity index for HIS;
k = number of connection (path between nodes),
n = number of nodes (participating units).

$$k = 7;\ n = 8;\ c_i = 0.16$$

The scored obtained at this step is 2/10.

Step 5. Monitoring the protocols and data flow in departments and analyzing the information exchange.

The ob-gyn department sends an XML file in CDA format to the analysis laboratory, to radiology and in CCD format to the general practitioner office, and so the data flow between the medical units is standardized. The scored obtained at this step is 3/10.

Step 6. Calculate the capacity of the ob-gyn department which is the rate at which data may be passed over time.

$$Q_{eff} = (Q_{\max} - Q_{oh}) * (t_f - t_p)$$

where, Q_{eff} = effective system capacity (data rate); $Q_{\max}$ = maximum data rate; Q_{oh} = system overhead data rate; t_f = time slot duration (unit transmission); t_p = unit propagation time

Another measure is the calculus of the department's overload which occurs when more data must be exchanged than the system is able to transmit. The overload is placed in a queue and it is transmitted when capacity is available.

$$M_{OL} = n_t * \sum_{y=1}^{n_t} (M_q)_y$$

where, M_{OL} = system message overload; n_t = number of transmitting nodes; M_q = messages in queue to be transmitted by node.

The system underuse was calculated, occurring when the system data rate/message load is less than full capacity but messages are waiting in queues to be transmitted.

$$Q_{uu} = M_{OL}, \text{ for } M_{OL} \le (Q_{eff} - Q)$$

$$Q_{uu} = Q_{eff} - Q, \text{ for } M_{OL} > (Q_{eff} - Q)$$

where, Q_{uu} = system underutilization (data rate);
Q = measured/observed data rate

Another parameter calculated was the under capacity of the system, which occurs when messages remain in queues and the system data rate is at the maximum.

$$Q_{uc} = (Q + M_{OL}) - Q_{eff}$$

where, Q_{uc} = system under capacity (data rate)

For the laboratory a maximum number of 300 messages a week were estimated, supported by the system, for the radiology internal department 100 messages a week, for the external department of radiology 80 messages a week, 50 messages for general practitioner.

 ISBN: 978-1-61208-224-0

In order to compute the interoperability score, 2 days were considered for the time of message transmission (T_f) and 4 days for the response time (T_p), because in Romanian health system the patient must wait minim 4 days to receive the laboratory results.

- Ob-gyn->Laboratory = 40 messages / day =>200 messages / week
- Ob-gyn->Radiology intern department = 15 msg / day => 75 msg / week
- Ob-gyn->Radiology extern = 10 msg / day => 50 msg / week
- Ob-gyn-> General practitioner = 10 msg/day => 50 msg/week
- T_f = 2 days
- T_p = 4 days

The results after applying the formulas are:

$$Q_{eff} = 1804;\ M_{OL} = 96;\ Q_{uu} = 96;\ Q_{uc} = -1594$$

The scored obtained at this step is 17/40.

Step 7. Interpreting the result and analyzing the data elements in HIS.

Analyzing all the steps, we concluded that: this type of system architecture benefits of a standardized communication; it is possible to add other healthcare information systems; the systems can be improved a lot; the healthcare information system can support more messages, because after computing the underuse capacity we concluded that more messages can be added without affecting the communication. The scored obtained at this step is 9/10.

Table I represents a summary of the steps analysis.

TABLE I. INTEROPERABILITY SCORE

Steps	1	2	3	4	5	6	7
Ob-gyn points	7/10	9/10	7/10	2/10	3/10	17/40	9/10
Total	54/100 points						

After applying these steps and computing the scores, the result was that the obstetrics-gynecology department has a score of 54 points, which represents a percentage of 54/100, regarding the interoperability potential with the analysis laboratory, radiology and general practitioner from the technical interoperability point of view. This score shows that the healthcare information system for ob-gyn department is ready to communicate to other healthcare information systems, but improvements have to be made.

IV. CONCLUSIONS AND FUTURE WORKS

The paper presents an algorithm adapted for healthcare information systems for assessing the technical interoperability degree of the ob-gyn department healthcare information system. After analyzing these two types interoperability, two scores will result which will show the interoperability degree of a healthcare information system, the degree in which it is ready to easy communicate with other similar ones. If the degree of interoperability between healthcare information systems can be evaluated, it will have benefits for assessing the status of informatization and degree of intercommunication in specific or general environments and the data available for the clinical staff and patients will be more consistent driving to better practice and patient healthcare status, and also will reduce medical errors. This study of interoperability degree will help the physicians to have more information about the patient, for software developer to develop more complex healthcare information systems and the most important is the patient that will benefit of a better treatment.

In the future works, we will analyze the operational interoperability, it will be develop a smart tool using the current study results determining the technical interoperability in an automated way and also a tool for operational interoperability.

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/88/1.5/S/50783, Project ID50783 (2009), cofinanced by the European Social Fund – Investing in People, within the Sectorial Operational Programme Human Resources Development 2007 -2013.

REFERENCES

[1] "IEEE Standard Glossary of Software Engineering Terminology," IEEE Std 610.12-1990.

[2] B. Blobel and P. Pharow, "A Model-Driven Approach for the German Health Telematics Architectural Framework and the Related Security Infrastructure", Studies in Health Technology and Informatics, Vol. 116, IOS Press, 2005.

[3] B. Blobel and P. Pharow, "A Model-Driven Approach for the German Health Telematics Architectural Framework and the Related Security", Vol. 116, IOS Press, 2005.

[4] Electronic Health Record (EHR), http://www.himss.org, Accessedin 10.01.2012.

[5] O. Lupşe, M. Vida, L. Stoicu-Tivadar and V. Stoicu-Tivadar, "Using HL7 CDA and CCD standards to improve communication between healthcare information systems", Proc. 9th IEEE International Symposium on Intelligent Systems and Informatics, SISY 2011, Subotica, Serbia, ISBN: 978-1-4577-1973-8, pp. 453-457, 2011.

[6] M. Vida, O. Lupşe, L. Stoicu-Tivadar and V. Stoicu-Tivadar, "ICT Solution Supporting Continuity of Care in Children Healthcare Services", SACI, pp. 635-639, 2011.

[7] HL7 Clinical Document Architecture, Release 2.0, HL7 version 3 Interoperability Standards, Normative Edition 2009, Disk 1 – Standards Publication.

[8] LOINC (Logical Observation Identifiers Names and Codes), www.loinc.org, Accessed in 10.01.2012.

[9] Healthcare Information and Management Systems Society Electronic Health Record Vendor Association (EHRVA), Quick Start Guide, HL7 Implementation Guide: CDA Release 2 – Continuity of Care Document (CCD), 2007.

[10] J. Ferranti, C. Musser, K. Kawamoto and E. Hammond, "The Clinical Document Architecture and the Continuity of Care Record: A Critical Analysis", Journal of the American Medical Informatics Association, volume 13, no 3, 2006.

[11] M. Kasunic and W. Anderson, "Measuring Systems Interoperability: Challenges and Opportunities", Technical Note CMU/SEI-2004-TN-003, 2004.

[12] M. Vida and L. Stoicu-Tivadar, "Measuring medical informatics systems interoperability using the LISI model", 8th International Symposium on Intelligent Systems and Informatics (SISY), pp. 33 – 36, 2011.

[13] F. Farfan, V. Hristidis, A. Ranganathan, and M. Weiner, "XOntoRank: Ontology – Aware Search of Electronic Medical Records", ICDE '09, pp.820-831, 2009.

 ISBN: 978-1-61208-224-0

Cloud Computing and Interoperability in Healthcare Information Systems

Oana-Sorina Lupşe, Mihaela Marcella Vida, Lăcrămioara Stoicu-Tivadar
Faculty of Automatics and Computers
University "Politehnica" of Timişoara
Timişoara, Romania
Email: (oana.lupse, mihaela.vida, lacramioara.stoicu-tivadar)@aut.upt.ro

***Abstract*—One of the areas with greatest needs having available information at the right moment and with high accuracy is healthcare. Right information at right time saves lives. This work proposes a solution based on cloud computing implemented for hospital systems having as a result a better management, high speed for the medical process, and increased quality of the medical services. Cloud computing technology is still new but promises a revolution in the entire connected areas. At national level, hospital information systems are somewhat rare and not very well managed. Cloud computing allows using the latest technologies at low prices (pay-per-use) and with minimum resources necessary for clients. The paper suggests a model for the architecture of the information systems in two key departments of a hospital: Pediatrics and Obstetrics, and Gynecology using interoperability for better access to information and preparing the system for future connectivity.**

Keywords-cloud computing; HL7 CDA; interoperability; Pediatrics; Obstetrics and Gynecology

I. INTRODUCTION

The most critical area that requires a lot of information, a lot of data and computing power is the healthcare domain. Doctors need, in critical moments, the medical history of patients in real time. Patients are sent to various investigations, supposing a high rate exchange of data between departments of medical units. Doctors need complete medical information of the patients to provide a complete and accurate treatment.

The technology that we chose to solve these problems is cloud computing because the resources are dynamically scaled (doctors can store a lot of medical data when they need) and is used over the Internet as services (doctors can access the medical data when and where they need it). To access this technology one can use a variety of Internet-connected devices which can access programs and development environments offered by cloud computing [1]. The information available at the right moment and location can save lives and significantly decreases the sources of medical errors increasing the quality of life of a patient. Another element used in our proposal to solve the problem of data exchange between medical units is ensuring the interoperability of the developed systems through HL7 CDA Standard [2].

This solution can be improved having a better security system for the medical data and creating a longitudinal data sheet of the patient (medical records for entire life span).

Cloud computing is a technology that could help a vitally important area because it offers a complex infrastructure at low cost and also provides greater computing power to achieve comprehensive health care operations.

In section two, the architecture and characteristics of cloud computing are described. Section three presents cloud computing applied in healthcare. Section four deals with interoperability in Pediatrics, and Obstetrics and Gynecology systems. Section five discusses cloud computing as a solution supporting information systems in a hospital, and Section six concludes the suggested solution.

II. ARCHITECTURE AND CHARACTERISTICS OF CLOUD COMPUTING

Cloud computing, defined by NIST (National Institute of Standards and Technology) [3] is a technology that supports ubiquity, it is convenient, supplies on demand access to the network for sharing computing resources (e.g., networks, servers, storage, applications and services), can be launched and developed quickly with minimal management and without service provider interaction.

The cloud model consists in five essential characteristics, three service models and four models of development (Figure 1).

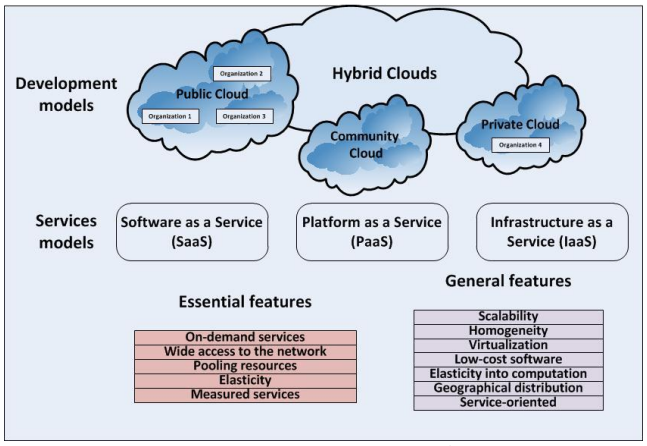


Figure 1. Elements and characteristics of the clouds

 ISBN: 978-1-61208-224-0

A. Essential characteristics

The essential five characteristics of cloud computing are: **On-demand services**: consumers can connect to a website and can use web services to access additional computing resources whenever they need; **Wide access to the network**: web services are based on cloud computing and for this reason can be accessed from any device connected to the Internet; **Pooling resources**: customers can share computing resources with other clients, so these resources can be reallocated dynamically and can be hosted anywhere; **Elasticity**: cloud computing allows users great flexibility that customers can scale systems (and costs) up or down as required; and **Measured services**: cloud computing monitors and records resources usage, which enables customers as payment for use (pay-per-use), a fundamental paradigm for cloud computing [3].

B. Cloud computing architecture

Cloud computing architecture consists of: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS).

Infrastructure as a Service is delivering hardware (servers, network technologies, storage) as a service. It also includes the operating system and virtualization technology for resource management. Currently, the best job profile is Amazon's Elastic Compute Cloud (Amazon EC2) [3]. It provides a web interface that allows customers accessing virtual machines.

Platform as a Service offers an integrated set of software that provides everything that a software developer needs to build an application - an online environment for quick development of web applications using browser-based development tools.

Software as a Service – business applications hosted and delivered as a service via the web. These kinds of applications do not require installation of additional computer programs, the most popular being the e-mail in a web browser [3].

C. Models of development clouds

Cloud computing is offered in four different forms: **Public clouds** – are held by a company selling cloud services to the general public; **Private clouds** – are owned by a single organization and are being used only in that organization; **Community clouds** – belonging to several organizations and allowing access only to those concerned for certain actions; **Hybrid clouds** – a composition of two or more types of clouds (private, public or community) that remain unique entities but are linked by standard technologies that enable portability of applications [3].

For medical applications, the best choice of a model is the private one for reasons of security and data privacy. In a private cloud, medical data can be accessed only by the authorized medical staff.

D. Related elements of cloud computing

The cloud architecture includes as most popular principles virtualization and SOA (Service oriented architecture).

Virtualization is at the core of most cloud architectures. The concept of virtualization allows an abstract representation of logical and physical resources including servers, storage devices, networks and software. The basic idea is to pool all physical resources and their management as a whole meeting the individual demands from these shared resources. In addition to virtualization, service-oriented architectures and web services are considered important in cloud computing.

Service-oriented architectures have components implemented as independent services that can be linked together in a flexible way and can communicate through messages. In cloud computing virtualized IT infrastructures, platforms and developed applications are implemented as services and are made available for use in service-oriented architectures. In public clouds, services are offered over the Internet on standard web protocols and interfaces.

SOA offers positive benefits such as [4]:

- Language-neutral integration: uses XML (eXtensible Markup Language).
- Component reuse: after is creating a web service to achieve an application, it can be reused for other applications which have service like this, and no longer is needed to rewrite code.
- Organizational agility: after building blocks of software which respects the user specification, it is possible to recombine and integrate quickly.
- Using existing systems: enable integration between new and old systems components.

III. CLOUD COMPUTING IN HEALTHCARE

In the medical field, cloud computing offers great potential for quick access to medical information. Health IT infrastructure is very complex and for this reason organization has taken additional measures to protect the patient's private data under HIPAA (Health Insurance Portability and Accountability Act). Maintaining confidentiality and integrity of information stored in all forms, and providing data backup and recovery processes in extreme cases are extremely important in this field. Quick access to medical history of each person at any location can accelerate diagnosis and treatment quality, avoiding complications, increasing quality and saving lives. In addition, cloud computing can help patients to gain access to their medical history from anywhere in the world via the Internet contributing to personalization in healthcare. The healthcare domain needs increased security and privacy levels, meaning that cloud computing technology has to be

 ISBN: 978-1-61208-224-0

more carefully managed in order to achieve this requirements. The matter is less technical and more ethical and legal. Before cloud computing technology can be fully adopted as a structure for health IT, providers must gain the trust of society and to demonstrate that they meet the HIPAA (Health Insurance Portability and Accountability Act) standard [5].

More than ever, healthcare services need cooperation between healthcare units due to high mobility of individuals for work or holidays. It is very important to ensure the availability of medical data to all the locations a patient is present in. Several scenarios and developments are already available in literature and presented in the following.

In [6], a scenario is presented to implement a cloud-based service for ePrescribing: the physician that uses the application is connected to the PHR (Personal Healthcare Record) system and reads a summary of medical history of each patient's records and selects a list of drugs. The application validates the selection of drugs based on their interaction with other drugs, patient allergies and medication history of the patient. If there are not incompatibility alerts, the prescription is stored in data centers of Insurance Organization waiting to be processed in Pharmacies. These systems are stored in a private cloud because in this way the information can be accessed only by authorized persons. Another proposal [6] is implementing a Semantic Wiki for User Training, based on the cloud technology available on demand and implemented on Amazon cloud infrastructure, a flexible, low-cost and scalable platform. Wiki users use the same database to store and read medical information. This solution offers support only for the ePrescribing system and for a cloud-based wiki.

In [7], a model is presented as an integrated EMR (Electronic Medical Record) sharing medical data between medical units. The application is developed on a cloud platform that keeps the EMR system on the form of Software as a Service and can be used by Government, Hospitals, Doctors, Patients, Pharmacies and Health Insurance Organizations, through the Internet. This system allows access to national data sharing; the data center is common to all units. Communication between the data center and the healthcare organizations is done via HL7 messages. All patient data are stored and accessed in the same location over the Internet from any healthcare organizations.

Using cloud computing in medicine results in benefits for the medical units and patients. Several benefits are:

- it is useful in storing medical data (cloud computing is scalable, increasing or decreasing resources, as needed),
- offers remote access (the data can be accessed via the Internet from anywhere),
- allows data sharing between authorized units
- the updates for the medical history of the patient - consultations, prescriptions, hospitalization - are made in real time and are useful for future treatment validation.

IV. INTEROPERABILITY IN PEDIATRICS AND OB-GYN SYSTEMS

4.1 General information about interoperability in cloud

Interoperability is the ability of two or more systems or components (for example two or more medical informatics systems) to exchange information and use the information that has been exchanged [8].

A web service is any service that is available over the Internet or an Intranet, uses standardized XML messaging system and is self-describing, discoverable and not tied to any operating system or programming language [9].

In eHealth is mandatory to use a standardized communication. In presenting the proposed system, one standard is used: HL7 CDA (Clinical Document Architecture).

Cloud computing technology supports interoperability, ensures high availability of resources, systems are "always ON", and available to communicate with other computing systems in the cloud.

4.2 Standard used in healthcare information systems

The HL7 CDA standard is a document markup standard that specifies the structure and semantics of "clinical documents" for the purpose of data exchange [2].

CDA has three levels of document definition: Level 1 (the root hierarchy, and the most unconstrained version of document), Level 2 (additional constrains on the document via templates at the "Section" level), Level 3 (additional constrains on the document at the "Entry" level, and optional additional constrains at the "Section" level) [10].

In Figure 2, a CDA example for the pediatrics healthcare system developed in order to evaluate the proposed architecture is presented. The codes used in Romania are ICD-10-AM and LOINC (translated in Romanian) [11].

```
<section>
  <code code="101155-0" codeSystem="2.16.840.1.113883.6.1"
   codeSystemName="LOINC" />
  <title>Alergii si Reactii Adverse</title>
  <text>
    <list>
      <item>Penicilina - Urticarie</item>
    </list>
  </text>
  <entry>
    <observation classCode="OBS" moodCode="EVN">
      <code code="L50.0" codeSystem="2.16.840.1.113883.6.3"
       displayName="Urticarie" />
      <entryRelationship typecode="MFST">
        <observation classCode="OBS" moodCode="EVN">
          <code code="Z88.0" codeSystem="2.16.840.1.113883.6.3"
           codeSystemName="ICD10" displayName="Alergie la penicilina" />
        </observation>
      </entryRelationship>
    </observation>
  </entry>
</section>
```

Figure 2. CDA example

The CDA example presents that a patient has allergy to penicillin, and it is represented with ICD-10-AM (in allergy case is used L50.0).

CDA documents are encoded in XML. The process is derived from the HL7 RIM (Reference Information Model) and also uses HL7 version 3 data types [10].

 ISBN: 978-1-61208-224-0

The information flow between the components of our model is this: The Pediatrics application sends an XML with the mother ID, baby's ID and the baby's birthdate, the ob-gyn application reads the XML request, identifies the needed data, and converts it to a XML in CDA format and sending it to the Pediatrics department where the data is read and filled in the baby's chart.

V. CLOUD COMPUTING AS A SOLUTION SUPPORTING INFORMATION SYSTEMS IN A HOSPITAL

As cloud computing can support different healthcare information systems by sharing information stored in diverse locations, a solution based on this technology was adopted for our case. A private cloud-based infrastructure was developed for each healthcare unit. To eliminate the drawback of cloud computing represented by weak security we have chosen the private cloud for each unit.

The architecture for the systems in the cloud is presented in Figure 3. All the medical data are stored in a private cloud and all the departments of the hospital can access medical patient data when is needed. In this case, the medical act is performed quickly, and the typing errors reduced, all of this driving to higher quality.

For increased security the suggested solution consists in a private cloud-based architecture where applications and data storage can be found within each private data center of the hospital (one in the Pediatrics hospital and one in the Ob-Gyn hospital). When individual patient data is needed from one department to another – both having different health information systems - it will be transmitted in real time to the proper location using an HL7 CDA message solution.

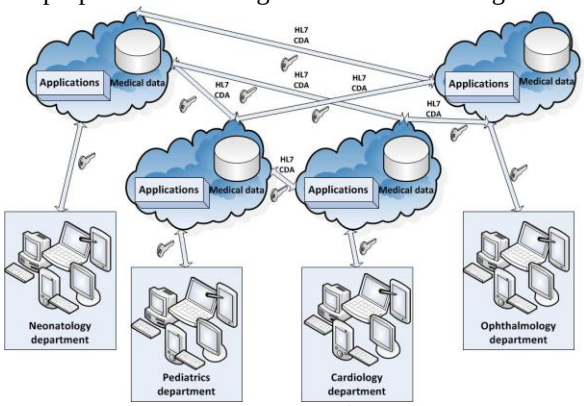


Figure 3. Architecture for hospital system

The solution ensures interoperability of the systems and a clear communication. Also the flexibility of the solution allows the connectivity in the cloud for new systems and devices.

The proposed solution is under development in ASP.NET environment for Windows Azure and the hospital database will be integrated into SQL Azure.

We started with two departments of the hospital: Pediatrics, and Obstetrics and Gynecology, because these are important starting points for the EHR (Electronic Health Record).

First contact with the medical world is starting at birth after which all the information about an individuals' health, immunizations, treatments, problems during pregnancy and all information of the child at birth are stored in the department of obstetrics and gynecology. After birth, the child is taken into care by a pediatrician for monitoring and treatment, if the case.

This is the reason why these two departments are the first departments of a hospital we wanted to give the opportunity to have a better communication and computing power and also more storage space using cloud computing and communications through HL7 CDA. The architecture of the solution is presented in Figure 4.

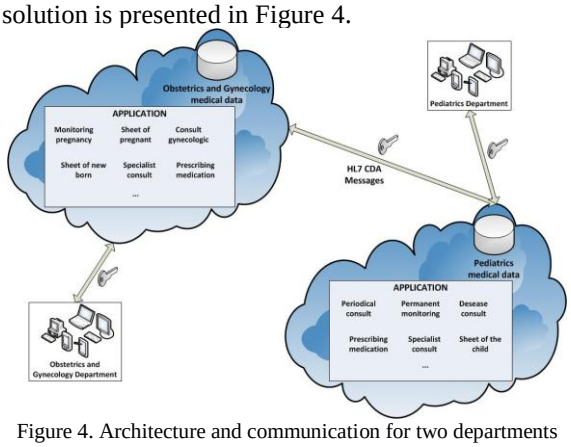


Figure 4. Architecture and communication for two departments

The applications were developed for each department separately (Pediatrics and Ob-Gyn) and also the support for communication in a local network. The next step is to upload the applications on the cloud and interconnect them.

To achieve interoperability we use XML files based on HL7 CDA standard. In Figure 5, the flow of data between the two medical units which exchange information is presented.

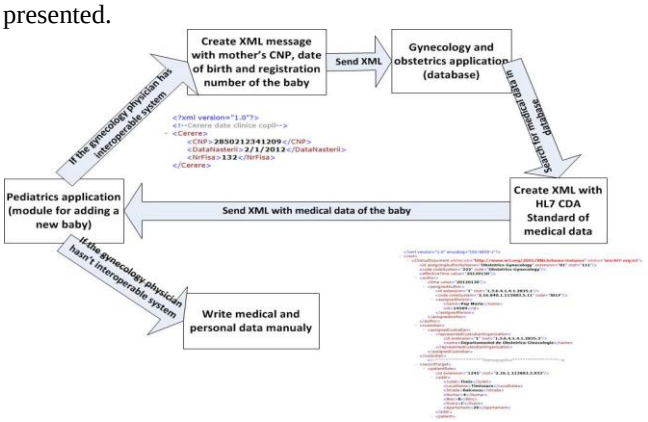


Figure 5. Exchange of medical data between units

The data of a new born child being added in the database of the pediatrician; the physician will be asked if wants to add the data manually or retrieve it from the database of the hospital, which technically is located in the private cloud of

 ISBN: 978-1-61208-224-0

the Obstetrics and Gynecology unit, where the baby was born. When the data acquisition from the Obstetrics and Gynecology unit option is chosen, the Pediatrics application will create an XML file with the PIN (Personal Identification Number) of the mother, date of birth of the child and registration number of the child (every child is registered at birth with a unique identification number in the hospital). The XML file with these data will be sent to the data server from the private cloud of the unit of Obstetrics and Gynecology. When these dates are available in the server, via a specific application it will check the validity of the received message will analyze the request and if the data exists in the server the application will form another XML file which contains the medical data record of the baby from birth until to the day of discharge. These XML file is created in HL7 CDA standard format, and it will be sent to the unit who requested the data.

Once received, the required medical data in XML format, the Pediatrics application will read the XML file and will display the medical records to the location point where the physician adds the patient. The received medical data will be saved in the database server of the private cloud of the Pediatrics unit. The pediatrician will have access to the medical history of the baby from birth and during pregnancy, information important for monitoring and treating the child.

For the applications to communicate better with each other and more effectively, we used the HL7 CDA standard, due to its features structuring the medical data on several levels and with certain codes that can be read by any application that uses these medical standards.

VI. CONCLUSION AND FUTURE WORKS

Using the cloud computing technology a medical act may considerably improve the access to information, which can be done be much easier. The scalability, that is the key of the cloud computing, can offer more resources needed for certain operation at any time.

The collaboration between medical units is an opportunity offered by cloud computing for healthcare staff. With this technology can be checked the availability of a physician, a medical specialist, a product or a service at different times and in different cases. Patients can be guided to appropriate persons or units where they can find what they need. This is a huge benefit for patients and health professionals, increasingly the quality of the medical service. The costs of the IT infrastructure will be cheaper because the medical units will only rent the infrastructure to store medical data as it need and will no longer need the latest equipment for the applications used, managed or maintained. They need only computers or devices with access to Internet.

The private cloud solution ensures the security of data and communication between departments, and messaging is done in a secure way. The application is equipped with a module that verifies the received and sent information.

Future work will improve the security solution (implement HIPAA requirements, using HTTPS) and will evaluate the results through measuring the interoperability degree achieved by the presented solution [12].

ACKNOWLEDGMENT

This work was partially supported by the strategic grant POSDRU/88/1.5/S/50783, Project ID50783 (2009), cofinanced by the European Social Fund – Investing in People, within the Sectorial Operational Programme Human Resources Development 2007 -2013.

REFERENCES

[1] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: state-of-the-art and research challenges," Journal of Internet Services and Applications, vol. 1, pp. 7-18, May 2010.

[2] "HL7 version 3 Interoperability Standards Normative Edition 2009, Based on HL7 v3 Data Types," Release 1, Disk 1 – Standards Publication

[3] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," NIST Special Publication 800-145, September 2011

[4] G. Raines, "Cloud Computing and SOA," Systems Engineering at MITRE, Service-Oriented Architecture (SOA) Series, 2009

[5] P. K. Bollineni and K. Neupane, "Implications for adopting cloud computing in e-Health," Master's Thesis Computer Science, September 2011

[6] D. Papakonstantinou, M. Poulymenopoulou, F. Malamateniou, and G. Vassilacopoulos, "A cloud-based semantic wiki for user training in healthcare process management," XXIII Conference of the European Federation for Medical Informatics (MIE 2011), August 2011, vol. 169, pp. 93-97, doi: 10.3233/978-1-60750-806-9-93

[7] B. Pardamean, and R. R. Rumanda. „Integrated Model of Cloud-Based E-Medical Record for Health Care Organizations," 10th WSEAS International Conference on E-Activities, December 2011, pp. 157-162

[8] "IEEE Standard Glossary of Software Engineering Terminology," IEEE Std 610.12-1990.

[9] E. Cerami, "Web Services Essentials," Third Indian Reprint, O'Reilly Media, Inc., 2007

[10] CDA levels: http://www.corepointhealth.com, Accessed in 20.02.2012

[11] J. M. Ferranti, R. C. Musser, K. Kawamoto, and W. E. Hammond, "The Clinical Document Architecture and the Continuity of Care Record: A Critical Analysis," Journal of the American Medical Informatics Association, vol. 13, pp. 245-252, February 2006

[12] M. Vida, and L. Stoicu-Tivadar, "Measuring medical informatics systems interoperability using the LISI model," 9th IEEE International Symposium on Intelligent Systems and Informatics (SISY 2010), September 2010, pp. 33-36, doi: 10.1109/SISY.2010.5647396

 ISBN: 978-1-61208-224-0

A SNP Prioritization Method Using Linkage Disequilibrium Network for Disease Association Study

Erkhmbayar Jadamba and Miyoung Shin
Bio-Intelligence & Data Mining Lab.,
Graduate School of EECS,
Kyungpook National University,
Daegu, South Korea
e-mail: erkhembayar@knu.ac.kr, shinmy@knu.ac.kr

***Abstract*—The problem of identifying and prioritizing various types of genetic markers including *single nucleotide polymorphisms* (SNPs), which are involved in human diseases such as cancer, is a one of primary challenge in current disease association studies. In this work, we propose a prioritization method, SNPRank that employs linkage disequilibrium (LD) network to improve the prioritization of candidate SNPs in disease association study. For the construction of LD network structure, we defined mutual links between SNPs based on $r^2 > 0.6$, and prioritized such SNPs that are linked to other highly ranked SNPs. For experiments, we applied our method to identify SNP markers associated with prostate cancers. The results showed that the proposed method can improve upon existing approaches by newly finding disease related SNPs which could not be identified by existing approaches.**

Keywords-SNP marker; disease association study; linkage disequilibrium network; SNP ranking.

I. INTRODUCTION

After completion of Human Genome Project in 2003 [1], most of researchers were interested in specific areas which are varied between individuals to individuals. Out of all the genetic variations, a *single nucleotide polymorphism* (SNP, pronounced snip) is known to contribute to 90% of them with being almost uniformly distributed across the genome. The SNP is a DNA sequence variation occurring when a single nucleotide –A, T, G, or C- in genome (or other shared sequence) differs between members of a biological species [2]. In recent disease association study, the presence of certain SNPs is often used as a significant clue to identify gene markers which predispose individuals to specific diseases. That is, some SNPs can be involved in increasing the risk of human disease, although most SNPs are not responsible for causing a particular disease phenotype. Thus, the problem of identifying such SNPs that are associated with disease in humans is a major task of disease association studies.

In this paper, we have overviewed current existing methods such as *single SNP analysis* methods and introduced our new approach in order to solve existing approach problems. In last section, we have showed that by allowing the usage of LD based network construction, SNPRank improves the performance over the state-of-the-art ranking method such as GWAS approach [3].

II. RELATED METHODS

. Most of existing methods use *single SNP analysis*, which include a chi-square test, Fisher`s test and Cochrane-Armitage trend test [3]. In these approaches, candidate SNPs are ranked based on the statistical significance of the test and top few SNPs are chosen to be highly associated with the phenotype.

A. Cochraen- Armitage Trend Test

Cochrane-Artimage test for trend, named for William Cochran and Peter Artimage, is used in categorical data analysis when the aim is to assess for the presence of an association between a variable with two categories and variable with k categories [4].

$$T = \sum_{i=1}^{k} t_i (S_{1i} R_2 - S_{2i} R_1) \quad (1)$$

Trend test statistic can be shown as in (1). In genetic application, the weight t_i can be different according to genetic models described in [3]. In order to test allele is dominant A over allele B, the choice is: t = (1,1,0); if we assume Allele A is recessive to allele B, the choice is: t = (0,1,1).To test whether alleles A and B are codominant, the choice is: t = (0,1,2) [4]. In disease association study, the additive (or codominant) version of the test is mainly used.

However, when number of SNPs are in millions, statistical significance of each SNP would be too small to rely on; this leads to the difficulty in finding significant SNPs in top ranked results. To solve such problems, in this work, we propose a new SNP ranking method, called *SNPRank*.

III. PROPOSED METHOD 'SNPRANK'

The newly proposed method SNPRank is taking some ideas from Google`s popular PageRank [5] algorithm. Adapting this concept in bioinformatics field was firstly attempted on gene expression data analysis with GeneRank [6] algorithm by Morrison et al. in 2005. Here, our method employs *linkage disequilibrium* [7][8] based network structure along with ordinary GWAS test result to produce an efficient prioritization of the SNPs in a disease association study. In particular, SNPRank method attempts to improve ranking results in such a way that relative ranking of a SNP makes it higher if it is linked to other highly connected SNPs.

 ISBN: 978-1-61208-224-0

A. LD Network Construction

Network construction can be summarized into following steps.

- Order candidate SNPs according to chromosome position value
- Calculate LD values (r^2) [7][8] between two SNPs
- Define each SNPs as nodes on network structure
- If r^2 between two SNP is greater than the threshold add the edge between the SNPs to the network
- Build adjacent matrix for SNPRank

Our aim here is to construct a network structure by using correlation between SNPs. The correlation between two SNPs can be estimated by using ***r square* measurements** [7][8], which can be obtained by using (2), between them.

$$r^2 = \frac{D^2}{P_A \times P_a \times P_B \times P_b} \quad (2)$$

where P_A , P_B , P_a , P_b are frequency of each allele and *D is LD measurement* defined by [6]. When the two alleles are not independent, we consider them to be in a state of linkage disequilibrium (LD). When the dependence between SNP is high, the two SNPs are considered to be in a state of high LD. After estimating the LD measurements we constructed network structure and considered each SNPs as nodes in the graph structure. We assumed there that there is an edge between SNPs if r^2 between two SNP is greater than ≥ threshold. We have tried different threshold values in range of (0.2 to 0.9), see Table I. SNPs are presented as a node in network structure. From the network structure, we have built the adjacent matrix(*4*) structure which is used as an input in in our SNPRank.

B. SNPRank

Letting $r_j^{[n]}$ denote the ranking of SNP *j* after the n^{th} iteration, it is defined by

$$r_j^{[n]} = (1-d) \times tr_j + d \times \sum_{i=1}^{N} \frac{w_{ij} \times r_i^{[n-1]}}{\deg_i} \quad (3)$$

Here, tr_j denotes ordinary GWAS test statistic of i^{th} SNP and w_{ij} denotes an element of the adjacent matrix *W* representing LD network on candidate SNPs. In particular, $w_{ij} = w_{ji}=1$ if *i* and *j* are adjacent and $w_{ij} = w_{ji} = 0$ otherwise. Also, $d \in (0,1)$ is a control parameter which is to define the weight of network structure reflected to calculate ranking statistic.
The value d = 0.80 is appears to be used by Google. From previous studies, d = 0.6 gave the best result in GeneRank algorithm in case of gene expression data [5].

$$\deg_i = \sum_{j=1}^{N} w_{ij} \quad (4)$$

Formula (4) indicates the degree of i^{th} SNP. The SNPRank method proceeds iteratively, updating the ranking for *j* th page from $r_j^{[n-1]}$ to $r_j^{[n]}$ according to the formula (3).

IV. Experiments and Results

A. Dataset

For experiments, we have used dataset from GSE [8], which include genotype called data profiles of 20 prostate cancer tumors paired with normal samples for 500568 SNPs. For evaluation, we counted how many *truly disease related SNPs* are in top n-ranked result by using prostate cancer related gene list [9] as gold standard. That is, SNPs are considered biologically meaningful if its associated genes match with any one of *gold standard genes* [10].

B. Results

To obtain better result, we implemented matching process in different ranges of parameter d and r^2. The best improvement of performance was when $r^2 \geq 0.6$, $d = 0.5$ when comparing current approach. We have implemented SNPRank, when $\boldsymbol{r^2} \geq$ in range of [0.4 to 0.9] and $\boldsymbol{d}$ is in range [0 to 1]; if $d = 0$, the ranking returned is based on solely on the absolute value of Cochrane-Armitage test results for that SNP. For $d = 1$, we return the ranking based on Linkage Disequilibrium Network connectivity. By setting d in the range [0 to 1], we interpolate between two extremes.

TABLE I. PERFORMANCE SENSIVITY TO r^2, WHEN D=0.5

d = 0.5	50	100	150	200	250	300	400	500
Cochrane Rank	4	6	10	11	13	16	19	21
SNPRank								
r^2>0.5	3	7	8	11	12	12	18	21
r^2>0.6	4	7	10	13	16	18	20	23
r^2>0.7	4	9	10	11	12	15	17	20
r^2>0.8	4	7	11	11	13	16	19	22
r^2>0.9	4	7	9	11	12	14	18	21

Since the choice of d =0.5 was suggested in original GeneRank algorithm, we have checked performance sensitivity to the choice of $\boldsymbol{r^2}$. In Table I, column heads represent top rank SNPs in range of 100 to 500. We compared how many '*gold standard'* genes are matched in top SNPs in two prioritization method classical Cochrane Rank and new SNPRank. We noted the best performance was when $r^2 \geq 0.6$.To evaluate the performance for novel SNP identification we compared the SNP ids and its associated genes for SNPRank with GWAS Cochrane Test ranking. Comparison was performed for top 50 SNPs to 500 SNPs when $r^2 \geq 0.6$, $d = 0.5$ in Table II.

TABLE II. COMPARISON OF THE EXISTENCE OF PROSTATE CANCER GOLD STANDARD SNPS AND GENES IN SNPRANK AND GWAS RESULTS: O - EXIST , X- NOT EXIST , RED – SNPS NOT IN GWAS RESULT, GREEN – GENES NOT IN GWAS RESULT

		Top 500 SNPs	
SNPs(rs ID)	**Gene Name**	**SNPRank**	**GWAS rank**

rs41488045	NR5A2	O	O
rs41330844	CDH9	O	O
rs17162712	NR5A2	O	O
rs41401450	RNASEL	O	O
rs4261554	CDH8	O	O
rs41498345	HK2	O	O
rs6801782	FHIT	O	O
rs16966932	CDH8	O	O
rs1448988	FGF16	O	O
rs8047093	CDH8	O	O
rs4287583	**CDH8**	O	X
rs231150	**TRPS1**	O	X
rs1019731	IGF1	O	O
rs34011899	**CDKN2A**	O	X
rs41517846	MYC	O	O
rs7194529	CDH1	O	O
rs395920	CDH13	O	O
rs41348046	TRPS1	O	O
rs17098265	PRKCH	O	O
rs10079737	**CDH9**	O	X
rs9936929	**CDH13**	O	X
rs5749939	**MAPK1**	O	X
rs6560010	**DAPK1**	O	X

V. CONCLUSION

In this work, we have addressed the problem of ranking and prioritizing biomarkers called SNPs which are the most common form of genetic variations on the human genome, and they have been widely used as genetic markers for studying common and complex human diseases. The tremendous number of SNPs, which is estimated at more than eleven million, poses new challenges for discovering and ranking procedures associated with such studies. Our purpose is to support effective disease association studies by providing operational prioritization methods for SNP markers based on both their allele frequency information and Linkage disequilibrium measurement. To achieve this purpose , we have proposed a novel integrative approach, SNPRank method, which allows us to combine linkage disequilibrium based SNP connectivities and conventional rank statistics to produce more robust SNP markers in disease association study, compared with traditional methods only based SNP genotype frequency. In particular, with $d = 0.5$ when $r^2 \geq 0.6$ is used, we observed no deterioration and overall improvement over original Cochrane-Armitage test results. Also, our new method SNPRank incorporated with LD network structure was shown to improve GWAS performance by newly identifying some of *truly disease related SNPs*, which include rs4287583, r231150, rs34011899, rs10079737, rs9936929, rs5749939, and rs6560010. In addition, our SNPRank identified new genes (e.g., TRPS1, CDKN2A, CDH9, CDH13, MAPK1, DAPK1) in top ranks, which could not be identified by conventional approach.

VI. FUTURE WORK

The work described in this paper comprises one step toward the goal of identifying disease variants, SNP, which underlying human diseases. For extending the work, we are interested in conducting simulation studies to examine the performance of the proposed method under various genomic experimental conditions, e.g., using the Next Generation Sequencing data. Finally, we mention the main lines of research of prioritizing genetic variation for certain disease will be still remain open for us after finishing this paper. In future, our particular would be using Next Generation Sequencing methods for identifying and prioritizing bio-markers in common and complex human disease.

ACKNOWLEDGMENT

This work was supported by the Korean Research Foundation of Korea (KRF) grant funded by the Korea Government (MEST) (No. 2009-0067724).

REFERENCES

[1] Collins, F. S., M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology." *Science* 300, no. 5617 (Apr 11 2003): 286-90.

[2] Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc.. 22 Jan 2004. Web. 04 Dec 2011. <http://en.wikipedia.org/wiki/Single-nucleotide_polymorphism>

[3] Lewis, C. M, "Genetic Association Studies: Design, Analysis and Interpretation." *Brief Bioinform* 3, no. 2 (Jun 2002): 146-53.

[4] Wikipedia, The Free Encyclopedia. Wikimedia Foundation, Inc.. 22 Jan 2004. Web. 04 Dec 2011. <http://en.wikipedia.org/wiki/Cochran-Armitage_test_for_trend>

[5] Page, Larry, "PageRank: Bringing Order to the Web", Stanford Digital Library Project, talk. August 18, 1997 (archived 2002)

[6] Morrison, Julie, Rainer Breitling, Desmond Higham, and David Gilbert, "Generank: Using Search Engine Technology for the Analysis of Microarray Experiments." *BMC Bioinformatics* 6, no. 1 (2005): 233

[7] Hill, W. G, "Estimation of Linkage Disequilibrium in Randomly Mating Populations." *Heredity (Edinb)* 33, no. 2 (Oct 1974): 229-39.

[8] Barrett, J. C., B. Fry, J. Maller, and M. J. Daly, "Haploview: Analysis and Visualization of Ld and Haplotype Maps." *Bioinformatics* 21, no. 2 (Jan 15 2005): 263-5.

[9] Danford, T., A. Rolfe, and D. Gifford, "GSE: A Comprehensive Database System for the Representation, Retrieval, and Analysis of Microarray Data." *Pac Symp Biocomput* (2008): 539-50.

[10] Castro, P., C. J. Creighton, M. Ozen, D. Berel, M. P. Mims, and M. Ittmann, "Genomic Profiling of Prostate Cancers from African American Men." *Neoplasia* 11, no. 3 (Mar 2009): 305-12.

 ISBN: 978-1-61208-224-0