



STAKE 2010

SEMANTIC TECHNOLOGY AND KNOWLEDGE
ENGINEERING CONFERENCE 2010



Damal Beach Resort, Kuching, Sarawak, Malaysia **26 - 30 July 2010**

**Proceedings of the Second Semantic Technology and
Knowledge Engineering Conference (STAKE 2010)**



KNOWLEDGE
TECHNOLOGY

Organized by
Centre of Excellence in
Semantic Technology
and Augmented Reality



STAKE 2010
SEMANTIC TECHNOLOGY AND
KNOWLEDGE ENGINEERING 2010

www.stake2010.com

Editors: Muhammad Ahtisham Aslam & Dickson Lukose ISBN: 978-983-41371-7-5

© MIMOS BERHAD

Proceedings of the 2nd Semantic Technology
and Knowledge Engineering Conference
(STAKE 2010)

Knowledge Technology Week 2010 Sponsors



Kerajaan Negeri Sarawak



TABLE OF CONTENTS

STAKE 2010

PREFACE	i
ACKNOWLEDGEMENTS	iii
CONFERENCE ORGANIZERS	v
PROGRAM COMMITTEE	vi

KEYNOTE

SPEAKER

Leveraging Artificial Intelligence and Semantic Technologies to Enhance Knowledge Processes in Knowledge Services Applications	Prof. Dr. Eric Tsui	Vii
If a text would know that it is read!?	Prof. Dr. Andreas Dengel	viii

Session 1 – Information Extraction

Sentiment Analysis Using Maximum Entropy and Support Vector Machine	1
Formulating Strategic Directions for Indigenous Knowledge Management Systems	11
Cultural Heritage Knowledge Discovery: An Exploratory Study of the Sarawak Gazette	20
Large-Scale Semantic Text Understanding	28

Session 2 – Semantic Web

Semantic Arabic Search Tool	40
A Semantic Tool to Enhance Digital Libraries	51
Logical Quiz Solving Based on Ontologies	64
Semantic Image Annotation and Retrieval for Sport Images	74

Session 3 – Bioinformatics

INFORMant™: Semantics Knowledge base for 3D Visualized Digital Human	86
An Integrated Health Ontology System	93
Using the Spiral Process Model to Develop a Medical Knowledge Base	101

PREFACE

Semantic technology is emerging and getting popularity very rapidly due to its capability of making the data machine understandable and process able. Large industry players are adopting semantic technologies to capture and engineer the domain knowledge. It is because that organizations are aware that their competitiveness and future success depend on their awareness to the latest developments in the field and on solutions to support management of knowledge.

Knowledge technologies are developed from Artificial Intelligence (AI) and I co-located the STAKE 2010 with Malaysian Joint Conference on Artificial Intelligence (MJCAI 2010) to bring together all stakeholders, including experts, novices, knowledge Engineers, AI experts and other interested parties from around the world.

The Semantic Technology and Knowledge Engineering (STAKE 2010) Conference is dedicated to latest scientific trends in semantic technology and knowledge engineering. It is co-organized by the Center of Excellence in Semantic Technology at MIMOS BHD and the Center of Excellence in Augmented Reality, University Malaysia Sarawak (UNIMAS). STAKE 2010 is second conference of its series. Initially STAKE 2009 emphasized more on applications in the domain of semantic technologies and knowledge engineering but this time is organized as a full conference where professionals can share their ideas and technology trends. One good attraction in STAKE 2010 for participants is that it provides a joint platform both for industry and academia. Researchers from academia will share their latest research and industry people will contribute their research experience from industry and also demo their semantic based applications. Community building is fostered in special interactive events designed around particular theme. Deliberately frequent breaks throughout the conference and social events in the evenings provide excellent opportunities for meeting and networking with researchers and practitioners from all over the world.

We encourage PhD students to participate in the PhD Students Symposium, specifically designed to facilitate current PhD candidates to present to the panel of experts, their research, and challenges they are facing. This will be an excellent opportunity to get feedback on their research from an international panel of experts.

STAKE 2010 offers its attendees selected contributions reviewed by an international Program Committee. The presentations and publications cover amongst others the following areas:

- Information Extraction
- Semantic Web
- Bioinformatics

We have also invited two prominent researchers in Semantic Technology and Knowledge Engineering to delivery their keynote address during the conference. They are:

- Professor Eric Tsui (Hong Kong)
- Professor Andreas Dengel (Germany)

At the end I would like to wish all the success to this conference, and all the other activities organized throughout this Knowledge Technology Week (KTW 2010).

Best Wishes and Thank You.

Dickson Lukose (General Chair)
Muhammad Ahtisham Aslam (Program Chair)
Edmond Ng (Local Organization Chair)

MIMOS BHD
Kuala Lumpur, MALAYSIA
28th – 30th July 2010

ACKNOWLEDGEMENTS

Initially, the Conference on the Semantic Technology and Knowledge Engineering (STAKE) was inclined towards industry applications and demos. This year, due to the importance of the area, it was decided to organize this event as a full conference aiming at publishing quality research papers in the area of semantic technology and knowledge engineering.

Arranging a successful conference is always a team effort. Many people have contributed in the successful execution of this conference. First of all I would like to thank our beloved President and CEO of MIMOS BHD, Dato' Abdul Wahab Abdullah for his continuous support and attention to make this event successful and productive. We could not have done this without his support and motivation. I would also like to thank to Dr. Muhammad Ahtisham Aslam for his dedicated work in successful execution of this conference. Dr. Ahtisham worked as a Program Committee Chair for the STAKE 2010. He made a Program Committee of more than twenty five international experts and got all submitted papers reviewed from at least four reviewers and conducted the final selection and rejection of submitted papers.

Organizing and advertising the STAKE conference specifically and Knowledge Technology Week (KTW 2010) generally, was another important job to be done to get right people at right time. I would like to thank Dr. Ahtisham, Dr. Ahsan Abdullah, Dr Mohammad Reza Beik Zadeh, Dr Karthigayan Muthukaruppan, Mr. Benjamin Min Xian Chu, Mr. Daniel Bahls, Prof Patricia Anthony, Dr. Edmond Ng, Mr. Tan Yew Seng, Mr. Kow Weng Onn, Mr. Arun Anand and Rokiah Bidin for informing and updating local and International universities, research institutions and industry people about this event. I would like to thank all Knowledge Technology/AIC people at MIMOS BHD for their help in organizing this conference. I would also like to thank Mr. Mohsen for continuously updating and managing the STAKE Web site to provide the visitors with right information.

Invited talks are one of the important and attractive events of every conference. I shall take this opportunity to thank Prof. Dr. Eric Tsui and Prof. Dr. Andreas Dengel for delivering interesting invited talks during the STAKE 2010 conference.

Many thanks to Dr. Mohammad Reza Beik Zadeh for arranging an interesting and informative PhD Symposium, Mr. Daniel Bahls for arranging interesting Tutorials and Mr. Tan Yew Seng for arranging workshops. I would also thank to the Local Organizing Committee: Dr. Edmond Ng, Mr. Alvin Yoe Wee and Mr. Lai Weng Kin for smooth organization of this event.

I would also like to extend my appreciation to all Program Committee Members for evaluating submitted papers based on their technical strength, resulting in publishing quality research papers in conference proceedings. I would like to say special thanks to those Program Committee Members who bared the overburden of reviewing papers during the review process. At the same time I would like to thank all authors for their value able contributions to the STAKE 2010 and hope that their research outcomes will open doors for further research and development for new researchers.

My sincere thanks to Ms Rokiah Bidin for all her work in dealing with administrative matters, travels, hotel bookings, and all other logistics related to managing the conference dinners, transportations for the delegates and many other items.

Finally, I wish all the participants of this conference and other co-located events to enjoy the Malaysian hospitality.

Best Wishes

Dickson Lukose (General Chair), MIMOS BHD.

CONFERENCE ORGANIZERS

Conference General Chair:

- Dickson Lukose (MIMOS BHD)

Program Committee Chair:

- Muhammad Ahtisham Aslam (MIMOS BHD)

Local Organizing Chairs:

- Edmond Ng (UNIMAS)
- Alvin Yoe Wee (UNIMAS)
- Lai Weng Kin (MIMOS BHD)

PhD Symposium Chair:

- Mohammad Reza Beik Zadeh (MIMOS BHD)

Tutorial Chair:

- Daniel Bahls (MIMOS BHD)

Committee Secretaries:

- Rokiah Bidin (MIMOS BHD)
- Azean Ahmad (UNIMAS)
- Doris Francis Harris (UNIMAS)

Local Organizing Committee:

- Edmond Ng (UNIMAS)
- Alvin Yoe Wee (UNIMAS)
- Lai Weng Kin (MIMOS BHD)

PROGRAM COMMITTEE STAKE 2010

- Soren Auer (University of Leipzig, Germany)
- Jun Shen (University of Wollongong, Australia)
- Sung-Kook Han (University of Wonkwang, Korea)
- Ahsan Abdullah (MIMOS, Malaysia)
- Yudith Cardinale (Simón Bolívar University, Venezuela)
- Sim kim Lau (University of Wollongong, Australia)
- Mian Muhammad Awais (Lahore University of Management Sciences, Pakistan)
- Bernhard Schandl (Universität Wien, Austria)
- Mohammad Reza Beik Zadeh (MIMOS, Malaysia)
- Sebastian Tramp (University of Leipzig, Germany)
- Olga Streibel (Freie Universität Berlin, Germany)
- Markus Luczak- Rösch (Freie Universität Berlin, Germany)
- Olaf Hartig (Universität Berlin, Germany)
- Ioan Toma (University of Innsbruck, Austria)
- Armando Stellato (Rom University, Italy)
- Johannes Keizer (United Nations Food and Agriculture Organization, Italy)
- Ahsan Morshed (United Nations Food and Agriculture Organization, Italy)
- Thomas Baker (Stanford University, USA)
- Kow Weng Onn (MIMOS, Malaysia)
- Arun Anand Sadanandan (MIMOS, Malaysia)
- Soon Lay Ki (Multimedia University, Malaysia)
- Sebti Foufou (Qatar University, Qatar)
- Syed Malek Fakar Duani Bin Syed Mustapha (Universiti Tun Abdul Raza, Malaysia)

Leveraging Artificial Intelligence and Semantic Technologies to Enhance Knowledge Processes in Knowledge Services Applications

Keynote Speaker

Prof. Dr. Eric Tsui

(Hong Kong Polytechnic University, Hong Kong
Eric.Tsui@polyu.edu.hk)

A knowledge worker in the new economy often has to carry out non-linear work which typically manifests in the form of a random series of knowledge processes. Such processes include, but not limited to, search, classify, discover, share and learn. Despite major advancements in various IT tools to support these processes, integrated systems that are underpinned by AI and ST to perform dedicated knowledge services still do not exist. This talk will comprehensively report on efforts by the Knowledge Management Research Centre (KMRC) at HKPolyU to leverage on the above-mentioned technologies to perform search, classification, knowledge discovery, navigation, and learning, as well as applications for delivering knowledge services for individual knowledge workers and organizations. Examples of such services include automated knowledge audit, computational narrative simulation of scenarios, and an Intellectual Property Management System (IPMS).

About Prof. Dr. Eric Tsui

Eric Tsui joined Computer Sciences Corporation (CSC) in 1989 after years of academic research in automated knowledge acquisition, natural language processing, case-based reasoning and knowledge engineering tools. His research was supported by grants and scholarships from Arthur Young, Rank Xerox, CSC, Graphic Directions, and the Australian Research Council. He was also a gratis visitor to Microsoft Research in February, 2000. Appointed as the Professor of Knowledge Management at the Hong Kong Polytechnic University under the President's Distinguished Professionals Scheme in September 2002, he joined the university as a full time staff in March 2005. In the past 5 years, he has delivered numerous public and custom-designed knowledge management and technologies workshops. He has also consulted for many government departments and private organizations in Australia, Hong Kong, Singapore, Malaysia, and Brunei.

If a text would know that it is read!?

Keynote Speaker

Prof. Dr. Andreas Dengel

(German Research Center for Artificial Intelligence, DFKI GmbH
Andreas.Dengel@dfki.de)

Intelligent systems paying attention to their users and observing what she or he is currently working for are becoming more and more important. Correlations between human actions and document contents are essential for building "electronic information butlers" that are able to anticipate what may be wanted in a given context. Due to the limitations of today's computer interfaces in user observation, we have to consider new approaches for connecting human thoughts and relevance of information.

In this talk, I like to propose a new but highly promising research field where gaze data collected during reading is used to measure attention while reading. Fixation points and saccades are taken to determine reading behavior and for calculating contextual relevance of content while working with documents. Based on these findings, a text may be enriched by invisible mark-ups augmenting a simple alphanumeric document with a virtually multimedia world behind the text. I first give an introduction into the new options for combining the brain and computer-resident information via the eyes and then show the new options and potentials of this emerging research field for application such as learning, classification or search.

About Prof. Dr. Andreas Dengel

Professor Andreas Dengel is Managing Scientific Director at the German Research Center for Artificial Intelligence (DFKI) in Kaiserslautern. In 1993 he was appointed Professor at the Computer Science Department of the University of Kaiserslautern where he holds the chair Knowledge-Based Systems. Since 2009 he also holds a Honorary Professorship (Kyakuin) at the Dept. of Computer Science and Intelligent Systems, Graduate School of Engineering of the Osaka Prefecture University. From 1980 to 1986 Andreas studied Computer Science and Economics at the University of Kaiserslautern. He subsequently worked at the Siemens research lab in Munich and at the University of Stuttgart where he completed his doctoral thesis in 1989. In 1991 he worked as a guest researcher at Xerox Parc in Palo Alto.

Andreas is a member of the IT-Summit Working Group on service and consumer-oriented information technology consulting the German government on questions of future IT strategies. 2008 he co-founded the Institute on Document Analysis and

Knowledge Science (IDAKS) at the Osaka Prefecture University in Japan. Andreas is an advisory board member of the Center of Excellence on Semantic Technologies at MIMOS in Kuala Lumpur, Malaysia, the NEC Computers and Communication Innovation Research Labs (CCIL), and the Int. Conference on Document Analysis and Recognition (ICDAR).

Andreas is and was program/technical chair of international conferences, such as ICPR, ICDAR, DAS, ICFHR, KES, KI, and KM. He is an editorial board member of international journals, like IJDAR and Future Internet. Moreover, he is founder or initiator of several successful start-up companies, In 2005 he received a Pioneer Spirit Award for one of his start-up concepts. He is co-editor of various international computer science journals and has written or edited 8 books and is author of more than 160 peer-reviewed scientific publications. He supervised more than 120 PhD and master theses.

In 2004, Andreas Dengel has been elected a Fellow of the International Association for Pattern Recognition (IAPR) and has been honored for his work several times. Most prominent prizes are the ICDAR Young Investigator Award, the Nakano Award, the Technical Communication Award of the Alcatel SEL Foundation, the Multi-Media Award as well as a Document Analysis Systems Achievement Award he received at Princeton University. His main scientific emphasis is in the areas of Knowledge Management, Semantic Technologies, Information Retrieval, and Document Understanding.

Sentiment Analysis Using Maximum Entropy and Support Vector Machine

Hemnaath Renganathan¹, Boon Wee Low¹

¹ePulze Sdn Bhd, C-41-2, Block C, Jaya One, No 72a, Jalan Universiti, 46200, Petaling Jaya, Selangor Darul Ehsan, Malaysia.
{hemnaath, boonwee}@epulze.net

Abstract: This paper presents a study on a hybrid machine learning technique that classifies sentiment of sentences by recognizing patterns and usage of words. The system uses Maximum Entropy (Maxent) and Support Vector Machine (SVM) with a Bag-of-Words model. Movie review comment sentences taken from Cornell University (Pang/Lee ACL 2005) were used as training dataset. Sentences will be classified according to a negative or positive perspective. We propose an approach of running linguistic rules before extracting text features as a Bag-Of-Words model. Maxent probability values will then be gathered from the Bag-Of-Words model to be passed into SVM for overall classification.

Keywords: Sentiment Analysis, Maximum Entropy, Support Vector Machine, Bag-Of-Words, Lemma, Unigram

1 Introduction

Vast amount of information available in the online environment: blogs, social websites, review sites, forums, has encouraged much analysis that makes use of these information to draw conclusions on public opinions. In this paper, we attempt an analysis of our own: which is to build a sentiment classification engine that classifies sentences within a text as positive or negative. Our classification will be focus on movie reviews and comments.

Sentiment, involves the emotion and state of mind of an author when he/she writes a text. It is difficult to extract sentiment because there is a lot of subjectivity, perceptions, culture and environmental factors to be taken into account. Due to this, achieving good accuracy for a sentiment classifier still proves to be a great challenge.

2 Related Work

There has been a lot of work done in sentiment analysis using different techniques and methodologies. However, most are often performed at a higher document level rather than sentence level. We believe this is because document level text captures

more sense and pieces of information hence leading to better performance of polarity detection. Furthermore, with sentence level analysis, much pre-processing, such as anaphora resolution and semantic analysis has to be performed before the actual sentiment detection begins.

Among those who performed document level sentiment analysis include [1], [3], [5], [12], and [14]. Pang et al. [14] experimented on 3 machine learning classifiers (Support Vector Machine, Maximum Entropy, and Naïve Bayes) at document level. They also tested on different Bag-Of-Words feature combinations such as unigram, unigram + bigram, and unigram + Part-of-Speech (POS) tag. Their results claim that Support Vector Machine (SVM) using unigram as features has the best performance in sentiment detection.

Boiy et al. [3] also performed document level classification with unigram and their results show SVM with highest performance. However, when subjectivity classification is done pre-sentiment analysis, their results show Maximum Entropy (Maxent) with better performance. Pang et al. [14] future work added subjectivity detection as well, to avoid processing of objective sentences and results show an increase in accuracy. A subjectivity classifier has been built but not included within the experiment because our training and testing datasets consist of subjective sentences only.

Extended work by Mullen and Collier [12] proves that within the 1-gram scale, unigram words are outperformed by lemmas, which are the root words. This may be because lemma reduces the number of Bag-Of-Words features without losing much sense or information.

In terms of sentence level classification, Sasha et al. [16] has built a hybrid sentiment classification engine composed of a polarity lexicon and Maxent. They implemented Maxent over lexicon to exploit local and global context of a sentence. They score and rank their sentences using their sentiment lexicon and use these scores as features for their entropy model. Other interesting features [16] planted in the model include scores of neighbouring sentences, user provided ratings, and also the pre-labelling of sentences as positive or negative based on the user rating. Addition of the user rating and pre-labelling increased their overall sentiment over 3 – 5%.

Other sentence level sentiment techniques include the usage of compositional semantics with sentiment lexicon as shown by [4] and [11]. Choi and Cardie [4] particularly show that semantics with sentiment lexicon perform better than any other standard learning methods with Bag-Of-Words. The optimum solution however said by [4] is to integrate semantics into learning so that the algorithm can identify the full grammar formalities of the language. What we propose in our system is to use compositional semantics before the standard learning takes place. It is different in a way that we run linguistic rules over a sentence first before the learning or prediction takes place.

3 Methodology

As mentioned earlier, our method of sentiment analysis is based upon linguistic rules and machine learning. In this section, we explain our approach using Bag-Of-Words

feature, together with the crafting of linguistic rules such as negation and conjunctions. The manner of which we constructed a hybrid machine learning model consisting of Maxent and SVM will also be explained in this section.

3.1 Feature Selection

Feature Selection here basically refers to the attributes required for the learning engine to classify an input correctly; in our case, positive or negative sentiment. Features can be said to be clues or markers that the learning algorithm can utilize to predict or calculate the score for a particular class. The performance in learning is directly influenced by the capabilities of the features to recognize a class. For example, in the case of sentiment analysis, the better the ability of features to detect a sentiment class (positive / negative), the better the performance of the overall sentiment classification engine. Therefore, as mentioned by Kuat and Sasa [9], the idea behind selecting good features is to capture the desired properties of a class.

Each sentence in our case is recognized as an event that is in the form of a feature vector (Event = $f_1, f_2, f_3 \dots F_N$). The number of features per sentence is not fixed since the entropy engine allows the flexibility of variable features. In this section, we will explain the feature candidates that are able to detect a sentiment class (positive / negative) effectively.

Bag-Of-Words Model. What we implement in our system uses a Bag-Of-Words model. This means each word within a sentence is represented as a feature. This leads to a unigram B-O-W model where each word stands alone as an independent feature and a sentence therefore comprises of these independent features. A sentence with 10 words equals to a sentence with 10 features. For example, “we had a good time.” = Event = {“we”, “had”, “a”, “good”, “time” “.”}. Each sentence then has n number of features of which Maxent supports. This method of feature selection is similar to most other sentiment detection experiments conducted, [14], [16],[9], [12] to name a few.

Using B-O-W, the total number of features should cover the entire dictionary of words in a language. However, doing so will cause performance to degrade and some domain specific and new words may not be recognized. So, what we did is to only cover words that are within the training dataset, hence achieving good system performance as well as covering domain specific words.

We modified our B-O-W model to only cater for words tagged as Nouns “NN”, Adjectives “JJ”, and Adverbs “RB” as we believe those words are the most important in extracting sentiment clues. Furthermore, we extended our model to only incorporate hapax legomenas, which are words that occur only once within a body of text. For example, a phrase “as good as it gets” will be represented by {“good”, “it”, “gets”}.

Lemma vs. Unigram. Unigrams are represented by any individual words in a sentence. Lemma on the other hand represents root words. For instance, a unigram “entertaining” has a lemma of “entertain”. Lemma is still a unigram in a sense that it is still represented as a singular word but it is in root form. Unigram is a specialized manner of a lemma because it includes words that are in past tense, present tense and future tense. Therefore, the number of features for unigram is a lot more compared to lemma. Lemma can be extracted from a unigram using Word Sense Disambiguation (WSD) and WordNet lexicon; however we have not a WSD module, so we assume that among lemma candidates for a given unigram, the shortest length lemma is the ideal choice. We conduct experiments with results in the following chapter to find out whether unigrams or lemmas perform better in sentiment detection.

3.2 Compositional Semantics

The main logic here is to extract the correct meaning or phrase from a given sentence using grammar rules, and analyzing the manner of which parts of a sentence is combined. Our semantic module is still in early stage of implementation and can be considered naïve because it does not fully cover the entire linguistic formalities. However, the addition of these rules has helped increase the sentiment detection accuracy. Based on experiments conducted with and without semantic rules, we found that there is an increment in accuracy when rules were applied. On a test of 550 sentences using hybrid machine learning and lemma, without rules, an accuracy of 79.1% was achieved. With rules however, the accuracy increased up to 84.3%. Similarly, on a test with 1000 sentences, without rules we achieved 76.5% while with rules an accuracy of 81.4% was as result. Among certain rules or formalities covered by the module are Negation, and Conjunctions.

Handling Negation. Negation can overturn sentiment polarity of feature words especially when using unigrams. This can indirectly affect the polarity of an entire sentence. At feature (word) level, if a negation is detected, “NOT_” will be appended as prefix to the feature. To determine whether a feature word is negated, an automated method is performed that would move in reverse from the feature position. The process will move backwards in search of negation triggers. The process is similar to that performed by [9] and [5]. The list of supported negation includes adverbs features such as “not”, “barely”, “hardly”, “rarely”, etc. Other negations from different Part-Of-Speech tags and phrases such as “lacks in” or “lacks of” will be replaced as “not” in pre-processing.

The manner, of which the backward search is performed, is by searching for negation triggers until it comes across a stop point. A stop point comes from the presence of a backward Noun, Verb, or Adjective. The only exception here is for Adjectives, where a stop point of a backward Verb is ignored. Other backward tagged features such as Determiners and Pronouns are not considered stopping points so the algorithm will continue searching for triggers. A negation will negate itself if there is any overlap. Sample of the negation algorithms works are shown in Table 1.

Table 1. Negation Case Samples

Sentence Input	Negation Output
That/NN is/VB/ not/RB a/DT bad/JJ thing/NN	That/NN is/VB NOT_bad/JJ thing/NN
The/DT movie/NN does/VBZ not/RB seem/VB entertaining/JJ	The/DT movie/NN does/VBZ NOT_seem/VB NOT_entertaining/JJ

Conjunction Rules. Conjunction, similar to negation plays an important role in determining a sentence sentiment. Where negation is more involved in the polarity of individual words, conjunction directly relates to the polarity of the entire sentence. Therefore, it is important to implement conjunction rules to determine the exact meaning of statements expressed by the writer. Conjunctions that we detect include “but”, “although”, “however”, “while”, etc. Some of the applied conjunction rules are shown in Table 2.

Table 2. Conjunction Rule Case

Rule Explanation
<p>(1) “but” Conjunction</p> <p style="text-align: center;">(phrase A . . . but . . . phrase B)</p> <p>Example: (The story was alright) but (the acting was terrible.)</p> <p style="text-align: center;"><i>If (phrase B < length threshold) :</i></p> <p>Analyze and reverse polarity of phrase A for sentence sentiment</p> <p style="text-align: center;"><i>Else :</i></p> <p>Cut Off phrase A and analyze phrase B for sentence sentiment</p> <p style="text-align: center;">Similar algorithm applies to “however” conjunction.</p>
<p>(2) “although Conjunction</p> <p style="text-align: center;">(although . . . phrase A , phrase B)</p> <p>Example: although (the acting was good), (the story was terrible.)</p> <p>Cut off phrase A and analyze phrase B for sentence sentiment</p> <p style="text-align: center;">(phrase A . . . , although . . . phrase B , phrase C)</p> <p>Example: (the movie was great), although (it could be better), (I really enjoyed it.)</p> <p>Cut off phrase B and analyze phrase A + C for sentence sentiment</p> <p style="text-align: center;">Similar algorithm applies to “despite”, “while”, “though” conjunctions.</p>

It is important to note that some conjunction tagged words when paired with other words such as “anything but” does not trigger an overturn of polarity within a sentence. For instance, “the movie is anything but good” means “the movie is not good”. There is no switch in polarity within the sentence. Such situations are treated before the conjunction rules are applied. Phrases like “anything but” is replaced by “not”, “not only . . . but also” is replaced to “not_only . . . “but_also” with underscore, so the detection of the negation “not” and conjunction “but” is nullified.

3.3 Machine Learning Classification

This section carries discussion on the methods of supervised learning that will be part of the sentiment classification. First is Maxent, a classification technique that takes into account as much feature uncertainty as possible. The other supervised learning algorithm used for experimenting and testing purposes is SVM that draws conclusion from training inputs through the usage of n-dimension hyper plane modelling.

Maximum Entropy Classification. Maximum Entropy, Maxent, is a probability distribution machine learning algorithm. The probability as always returns a distributional range of 0 to 1. Biasness is reduced when using maximum entropy because the algorithm takes into account all manner of uncertainty. The algorithm is also flexible to adapt to various constraints. We use entropy to estimate the probability of a sentence as positive or negative sentiment. We focus on using maximum entropy to learn the conditional distribution pattern from labelled training data sentences.

Using only maximum entropy, we have to manually specify a probability threshold limit that will distinguish two sentiment polarity classes: positive / negative. We assume that if the probability output < 0.5 the input sentence is negative in sentiment, whereas if the probability output ≥ 0.5 , the input sentence portrays a positive sentiment.

The features for the entropy model, as mentioned, are individual words. An entropy event consists of a sentence that has n-features depending on n-number of words. Maximum Entropy works on an independent assumption on variables so words can be easily added as features without worry of feature overlapping.

The sentence events within the training data are tagged as positive or negative respectively so that the algorithm can learn from the features and draw its own conclusions. It is important to note that we do not use recall on our testing data. The version of maximum entropy that we used for our experiment comes from openNLP SharpEntropy.

Maximum Entropy Classification + Support Vector Machine (SVM). In this section, we explain the use of two classification engines together as a hybrid. The works of the maximum entropy are as explained above. We will experiment the presence of Support Vector Machine (SVM) combined with the entropy probability to see if the sentiment detection performance is improved.

SVM is a classification technique using kernel principal and Gaussian analysis. It is supervised learning method that will be applied to classify sentences. The SVM algorithm is based on the statistical learning theory while entropy is based on probability learning theory. Many experiment works such as from [12] and [14] show that SVM combined with unigram perform the best.

We extend the usage of SVM together with the entropy probability to classify the sentences as positive or negative. The probability output from the entropy model would be fed into SVM as a feature to predict the final sentiment of the sentence. We realise that many input sentences that consist of both positive and negative words tend to sit in the middle probability range of 0.45 – 0.55. Sitting in this range, it is difficult

to completely determine what polarity the sentence portrays. Therefore we pass on the entropy training probability values into SVM, so that it will draw a better segregation on sentences that lie within the 0.45 – 0.55 range. We specify proper training parameters for SVM for modelling that will be explained in section 4.

4 Experiment

We evaluated the accuracy for movie reviews using 2 comparative analyses: unigram vs. lemma, entropy vs. entropy SVM. Entropy SVM basically means a hybrid classification using both Maxent and SVM. We would like to see whether the combination of SVM together with Maxent raises the performance of sentiment detection. The entropy model is used in all experiment cases.

Lemma as mentioned is the root word for a specific unigram. In one test case, we replace the unigram features with lemma. Lemma is retrieved using P-O-S tag and WordNet lexicon. For syntax tagging and tokenization we used the open source module openNLP. For entropy, Sharp Entropy third party modelling was used. The version of SVM comes from libSVM.

For training, the movie review sentence dataset from Cornell University NLP section was utilized, while, for testing, we manually gathered sentences from IMDB and Rotten Tomatoes. There are 2 test sets, a total of 1100 sentences as set A and 2000 sentences as set B. Each test set consist of equal amounts of positive and negative sentiment sentences.

Without SVM, we manually set the entropy threshold to 0.5 for discriminating positive and negative sentences. With SVM however, we simply pass in the bulk of training probabilities into the engine so that it can find its own discriminate. For SVM training, libSVM cost is set as 128 and gamma as 2.

4.1 Experiment: Unigram vs. Lemma

In this experiment, we compare the usage of unigram features and lemma features to see which returns better performance. We will test these features with and without the presence of SVM to get a better understanding of the results.

4.2 Experiment: Entropy vs. Entropy SVM

In this section, we will experiment whether the addition of SVM to create a hybrid model increases the performance of sentiment detection. Without SVM, an input sentence will be formatted and passed into the entropy model. The entropy model will be manually set to discriminate ≥ 0.5 probability as positive and < 0.5 as negative. With SVM however, the discriminate settings need not be given as the SVM engine will deduce it's on discriminative value.

5 Result

Table 3 and 4 below shows the sentiment precision results acquired from the two test sets. Results are expressed as percentage of sentences classified correctly. Accuracy of lemma vs. unigram features and also on the entropy and hybrid classification engine (entropy SVM) is shown.

Table 3. (Set A) 1100 Test Sentences for Lemma and Unigram

<i>Classifier \ Feature</i>	<i>Lemma</i>	<i>Unigram</i>
<i>Entropy</i>	83.6%	81.5%
<i>Entropy SVM</i>	84.3%	82.0%

Table 4. (Set B) 2000 Test Sentences for Lemma and Unigram

<i>Classifier \ Feature</i>	<i>Lemma</i>	<i>Unigram</i>
<i>Entropy</i>	81.1%	80.1%
<i>Entropy SVM</i>	81.4%	80.3%

From the results seen above, we can conclude that lemma performs better than unigram in sentiment performance as also claimed by [12]. There is also an increase in sentiment accuracy if not minor, with the implementation of a hybrid model of entropy and SVM. Using SVM to deduce a threshold value for entropy probabilities works better than setting a manual hand coded value.

6 Discussion

From the experiment results show in section 5, it is proven that lemma along with entropy SVM provide the best performance for sentiment detection. Most current research such as from [5], [9], [13], [16] have used unigram as their main feature for supervised learning but our experiments have shown that lemma provides a credible if not better solution as a feature. Lemma, as mentioned, transforms words from their past, present, future tenses back into their general root form. We believe that lemma performs better than unigram because it increases credibility while significantly reducing the number of Bag-Of-Words features. Increasing credibility of features allows the supervised learning algorithm to draw better conclusions.

The presence of SVM to support Maxent has resulted in better accuracy in sentiment detection. This proves that SVM does help in solving the entropy probability issue of sentences that lie in the range of 0.45 – 0.55. Setting a manual threshold without SVM is not as good compared to using a hybrid of both, although it can be said that the difference is not very significant (less than 1%). However, it is an improvement nevertheless.

The addition of linguistic rules is vital in getting good performances on sentiment detection. The rules that we added as pre-processing as increased the overall detection of up to 5%. We can assume that if more linguistic rules are added, the better the performance will be. However, as mentioned by Choi & Cardie [4], the idea should be allowing the system to learn the rules by itself using learning techniques. When a system has learnt most, if not all linguistic rules, detecting sentiment should not be very difficult thereafter. This is the approach we would like to take on as future enhancement along with the addition of WSD.

Word Sense Disambiguation (WSD) is important especially when using words as features. It is important to know exactly what a word means. Especially if a word is subjective, we hypothesize that removing it before pre-processing would result in better performance for sentiment detection. According to Rentoumi [17], a word with its literal meaning is more subjective compared when used metaphorically. For instance, “he is such an animal”, the word “animal” here does not refer to the literal meaning of creatures, but more metaphorically as a brutal person. So providing a bias consideration on words that are used as metaphors might increase sentiment detection, especially in opinion reviews. These are all work in progress for enhancements.

We would also be branching into much semantic work in order to retrieve the subject of the sentiment. Determining sentiment is not enough unless there is a subject of which it belongs to. For example, “John did badly in his exams”, can be easily seen as a negative sentiment; however the sentiment should be directed towards the subject “John”. This requires further involvement in other NLP areas such as Anaphora Resolution, Entity Name Recognition, etc.

7 Conclusion

In this paper, we present a method of classifying sentiment using a hybrid supervised learning model; Maxent with SVM. We also worked on using lemma instead of unigrams for features and that showed encouraging results. Addition of linguistic rules before pre-processing also demonstrated well in sentiment performance. Our future work as mentioned in section 6, include word sense disambiguation, anaphora resolution, subject / object detection, and entity name recognition. These will allow us to not only detect sentiment, but to provide an interface for portraying sentiments to their subjects.

Our approach will be on developing domain specific sentiment engines. Examples of domains include movies, politics, finance, etc. Moving from general sentiment analysis into domain based analysis will remove a lot of misunderstanding of meaning. For example, “sinful” chocolate cake (food) is actually a good thing but “sinful” in the domain of politics indicates negative. Many adjective and verbs that mean positive sentiment for movies and restaurants may prove to be negative when used upon say, finance or political sentences. We believe this direction of research will lead us to promising results in the area of sentiment analysis.

References

1. Alekh, A.; Pushpak, B.: Sentiment Analysis: A New Approach for Effective Use of Linguistic Knowledge and Exploiting Similarities in a Set of Documents to be Classified. I.I.T Mumbai, India (n.d.)
2. Berger, A.; Stephen, A.D.P.; Vincent, A.D.P.; A Maximum Entropy Approach to Natural Language Processing. In: Computational Linguistics, Vol. 22(1), pp. 39-71. Massachusetts, USA (1996)
3. Boiy, E.; Hens, P.; Deschacht, K.; Moens, M.: Automatic Sentiment Analysis in On-line Text. In: Proceedings ELPUB 2007 Conference on Electronic Publishing. Vienna, Austria (2007)
4. Choi, Y.; Cardie, C.; Learning with Compositional Semantics as Structural Inference for Sub sentential Sentiment Analysis. In: Proceedings of EMNLP 2008, the Conference on Empirical Methods in Natural Language Processing. Honolulu, Hawaii (2008)
5. Gindl, S.; Liegl, J.: Evaluation of Different Sentiment Detection Methods for Polarity Classification on Web-based Reviews. In: Computational Aspects of Affectual and Emotional Interaction (CAFEEi). Patras, Greece (2008)
6. Hang, C.; Vibhu, M.; Mayur, D.: Comparative Experiments on Sentiment Classification for Online Product Reviews. In: Proceedings of the 21st International Conference on Artificial Intelligence. Boston, Massachusetts (2006)
7. Jaochims, T.; Granka, L.; Pan, B.; Hembrooke, H.; Gay, G.: Accurately Interpreting Clickthrough Data as Implicit Feedback. In: SIGIR'05. Salvador, Brazil (2005)
8. Kim, S.M.; Hovy, E.: Determining Sentiment of Opinions. In: Proceedings of the COLING Conference. Geneva (2004)
9. Kuat, Y.; Sasa, M.: Sentiment Analysis of Movie Review Comments. Massachusetts Institute of Technology (2009)
10. Liu, B.: Sentiment Analysis and Subjectivity. To Appear: Handbook of Natural Language Processing, Second Edition. Chicago, US (2010)
11. Moilanen, K.; Pulman, S.: Sentiment Composition. In: Proceedings of Recent Advances in Natural Language Processing. Borovets, Bulgaria (2007)
12. Mullen, T.; Collier, N.: Sentiment Analysis using Support Vector Machines with Diverse Information Sources. In: Proceedings of EMNLP 2004, pp. 412-418. Barcelona, Spain (2004)
13. Nipun, M.; Shashikant, K.; Priyank, P.: Sentiment Identification Using Maximum Entropy of Movie Reviews. Department of Computer Science, Stanford (n.d.)
14. Pang, B.; Lee, L.; and Vaithyanathan, S: Thumbs Up? Sentiment Classification Using Machine Learning Techniques. In: Proceedings of EMNLP 2002, the Conference on Empirical Methods in Natural Language Processing, pp. 79-86. Philadelphia, US (2002)
15. Popescu, A.M.; Etzioni, O.: Extracting Product Features and Opinions from Reviews. In: Proceedings of Human Language Technology and Conference on Empirical Methods in Natural Language Processing. Vancouver, Canada (2005)
16. Sasha, B.G.; Hannan, K.; McDonald, R.; Neylon, T.; Reis, A.G.; Reynar, J.: Building a Sentiment Summarizer for Local Service Reviews. In: WWW Workshop on NLP in the Information Explosion Era. New York, USA (2008)
17. Rentoumi, V.: Sentiment Analysis Using Word Sense Disambiguation Techniques. In: Software and Knowledge Engineering Laboratory. Athens, Greece (n.d.)

Formulating Strategic Directions for Indigenous Knowledge Management Systems

Tariq Zaman, Narayanan Kulathuramaiyer, Alvin Yeo

Universiti Malaysia Sarawak (UNIMAS), Sarawak, Malaysia
tariqzaman@lawyer.com, {nara, alvin}@fit.unimas.my

Abstract: In modern organisational structures knowledge management practices consist of knowledge generation, capture, sharing and application. The organisations emphasize on codification and documentation of implicit knowledge and transform it to explicit form. Indigenous communities however have much less codified knowledge relying mainly on oral and tacit form. The communities have their own processes of storage, leveraging, sharing and applying knowledge which is different from knowledge management processes of corporations and research organizations due to the oral and tacit structures of these processes. In this paper we present a model for formulating strategic directions for an indigenous knowledge management system. We have designed a knowledge management assessment tool for Indigenous Knowledge Management Systems (IKMS) which has been tested in remote community in Bario, Sarawak. On the bases of our assessment of IKMS, community capacity and resources, we have developed a strategic map for IKMS in Bario. This work serves as an extension to the previous literature on designing the Balanced Scorecard for IKMS.

Keywords: Indigenous Knowledge, Balanced Scorecard, Indigenous Knowledge Management System, Traditional Knowledge.

1 Introduction

In existing literature, the term indigenous knowledge, traditional knowledge, traditional ecological knowledge, local knowledge and indigenous technical knowledge are used interchangeably. In addition, some of the commonly asserted characteristics of indigenous knowledge include the following: it is generated within communities; it is location and cultural specific; it is a basis for decision making and survival strategies; [generally] it is not systematically documented, it covers critical [issues: such as] primary production, human and animal life, natural resources management[;] it is dynamic and based on innovation, adaptation and experimentation, and it is oral and rural in nature [1]. Indigenous knowledge, which has generally been passed from generation to generation by word of mouth, is in danger of being lost unless it is formally documented and preserved [2]. The rapid change in the way of life of indigenous people has largely accounted for the loss of Indigenous Knowledge (IK). Younger generations underestimate the utility of

indigenous knowledge systems (IKS) because of the influence of modern technology and education [3]. Over the last two decades there has been a great increase in interest in Indigenous Knowledge (IK) from a variety of groups including development agencies, researchers, governments and corporate world. An increasing number of cultural heritage institutions in the western world are exploring digitisation as a means of preservation and/or improving access and knowledge of their collections. The World Bank's 'Indigenous Knowledge for Development Program' [4] and UNESCO's 'Best practices on Indigenous Knowledge' [5] are the examples. These initiatives are focusing on creation of databases of indigenous knowledge in the same systematic way as western knowledge. In any case, the objective of databases is typically twofold. They are intended to protect indigenous knowledge in the face of myriad pressures that are undermining the conditions under which indigenous people and knowledge thrive. Second, they aim to collect and analyse the available information, and identify specific features that can be generalised and applied more widely in the service of more effective development and environmental conservation [6]. So these organisations focused on IK as a corpus of facts rather than IK as a system. IK as a system has a much broader understanding of Indigenous people as they place themselves in relation to the environment in which they live. Dr. Gada Kadoda while addressing the Unisa community during the 2010 CSET African Scholar Programme highlighted the issue of the lack of indigenous knowledge systems theories written for research purposes. She added that, "In creating a shift from the reliance on the Western knowledge systems to the indigenous knowledge systems, we have to start from what we do not have" [7].

On the basis of the current debate between IK as corpus of fact and IK as a system our main research questions are, Is there any existing IKMS in indigenous communities? And if 'yes'; How the IKMS deal with the community knowledge assets? This research is limited to the first question. We developed the assessment tool for IKMS and proposed methods for assessment of community capacities, resources and skill. The strategic direction and strategic map is based on the results of the assessments using these tools.

2 From Assessment of IKMS towards Strategic Direction: The proposed model

2.1 Assessment of indigenous knowledge management system

Bukowitz and Williams suggested a knowledge management diagnostic (KMD) tool to gauge the KM efforts of an ordinary business and research organisation according to the knowledge management process oriented model [8]. It is based on the "KM Process Framework", which consists of seven KM activities get, use, learn, contribute, assess, build/sustain, and divest (see Fig.1). The four activities "get, use, learn and contribute" designate the daily routine in dealing with knowledge. The other three activities "assess, build/sustain and divest" are attributed to the strategic planning of the organisations' knowledge management. KMD tool is used in many

studies to learn about the KM efforts of an organisation, also when these efforts were not called “KM” [9]. This is one of the attractions for selecting this tool in our research. The indigenous communities don't refer their activities as knowledge management practices but they have a very strong system of transferring knowledge from one generation to another. Ruddle (1993) examined the traditional ecological knowledge for sites in Venezuela and Polynesia. He examined that by the age of 14, children were competent in household tasks, cultivation (plant identification, harvesting), seed selection, weeding, animal husbandry, fishing and hunting [10]. The original KMD contains 140 questions, 20 questions for each of the seven knowledge management processes. The respondents are expected to choose from three options of whether the statement is strongly, moderately or weakly descriptive of the organisation. The more strongly the statements in the section are descriptive of the organisation, the higher is the score. For calculating the score, the following formula is used as described by Bukowitz and Williams for knowledge management diagnostic (KMD);

Number of S responses which stands for strong: $S \times 3 = A$ (A represent the result after multiplication)

Number of M responses which stands for Medium: $M \times 2 = B$ (B represent the result after multiplication)

Number of W responses which stands for weak: $W \times 1 = C$ (C represent the result after multiplication)

Number of Ms: $M \times 2 = B$ (B represent the result after multiplication, M for Moderate)

Number of Ws: $W \times 1 = C$ (C represent the result after multiplication, W for Weak)

Accumulated Point Score = Z (Z represents the result of $A+B+C$)

Maximum total point score = 12

Percentage score = () % of each section

When this tool is applied in researches conducted in developing countries, the researchers found that the KMD was based on several assumptions that might not necessarily be relevant due to the nature of their organisations and structures. Many questions were left unanswered, especially in the strategic processes of assess, build and sustain, and divest. As a result of this finding, the researchers decided to modify the original KMD using the response rates to each of the questions and whether the question could be considered relevant to research organisations [9]. The indigenous communities also faced the problem of lack of proper structure in terms of knowledge management. No single person or group was explicitly assigned to be responsible for enhancing and supporting knowledge management activities within the community.

On top of that large numbers of knowledge assets are in tacit and implicit form. So we also modified the standard KMD and combined the seven KM processes in three categories knowledge utilization (Use, get and contribute), knowledge accumulation (learn, assess and update), knowledge construction (build, divest and innovation) (Fig.2).



Fig. 1. Bukowitz & Williams KM process model. **Fig. 2.** Proposed IKMS process model.

2.1.1 Research Method

We carried out our study in Bario in a remote rural community, located on the island of Borneo, close to the border between Kalimantan and Sarawak, Malaysia. Flying to Bario, is the only practical way to get there. The road to Bario has been recently completed and it is 14 hour bone shaking ride by all accounts to the nearest town Miri. Bario comprises of 12 longhouses with a population of around 1,000 people. The majority of people are Kelabits, one of the smallest ethnic groups in Sarawak, and are mainly farmers. Bario was selected because of its geographical isolation and the progressive nature of the community. The Kelabits of Bario generate income from fragrant Bario rice, tourism and Homestay programs.

For assessment of each of the KM processes, we selected a set of variables. Standard forms of variables do not always accurately reflect the situation of indigenous communities, particularly as resources and intellectual property are shared commodity. So the variables [need to] be modified on the base of indigenous peoples' inherent values, traditions, languages, and traditional orders/systems, including laws, governance, lands, economies etc [11].

From KMD tool we selected the questions relating to our variables and where necessary, modifying the tool accordingly. In response of each question the community shared their experience of managing their collective knowledge. Snowball sampling was used to recruit subjects for this study.

Fifteen respondents from Bario were selected from different indigenous communities of Bario. The respondents include farmers, religious leaders, school teachers, tourist guides, members of community council (JKK), and women entrepreneurs our results (Table 1.) are based on the responses of our subjects/respondents.

Table 3. The results of IKMS assesment from Bario

No.	Variable.	Strong/Moderate/Weak
Section 1- knowledge utilization (Use, get and contribute)		
1	Community recognition of required knowledge.	Weak
2	Have recognition to individual and collective knowledge.	Moderate
3	Have well established practices of stakeholders' involvement in decision making.	Weak
4	Collaborate with other communities and government for development.	Weak
5	Participate in strategic networks and partnerships.	Strong
Section 2- knowledge accumulation (learn, assess and update)		
6	Have mechanism for sharing knowledge.	Weak
7	Use external knowledge.	Strong
8	Protection of knowledge assets.	Weak
9	Acceptance to new technologies	Strong
10	Have recognition of knowledgebase as asset	Strong
Section 3- knowledge construction (build, divest and innovation)		
11	Community supports new technologies.	Strong
12	Community Promote s team building and group activities for mutual learning.	Strong
13	Acknowledgment to individual contributions.	Weak
14	Have ability to outsource skills and expertise.	Weak
15	Participation in research groups for acquiring new knowledge.	Weak

From the survey results (table 1), the gaps were identified in sub domains of indigenous knowledge management processes for the Bario community. The results show that the Bario community has systems for the knowledge management although some of the features were found to be weak and needed improvements. So instead of looking only at indigenous knowledge as corpus of fact we evaluated the existing systems of knowledge management in these communities and subsequently explored interventions and strategic directions for strengthening the weak components.

2.2 Exploring community capacity and resources

While formulating the strategic direction for a community, the focus should not be limited to the assessment of IKMS. The other factors that need to be taken into consideration include the capacity of the community and available resources. The community capacity [represents] the combined influence of a community's commitment, resources and skills that can be deployed to build on community strengths and address community problems and opportunities [12]. Capacity building in this respect is not limited to economic development but also offers a foundation for making good decisions about the stewardship of a region's natural, human and cultural resources, indicating the way of life can be maintained and improved over time. In addition, the indigenous communities have a close relationship with the environment, where they live in harmony with the natural resources. So in case of formulating the strategic directions for indigenous knowledge management it is very much important to explore the currently available resources and consequently measure capacity of the community. We adopted the assessment method developed by International Institute of Rural Reconstruction for exploring the resources and measuring the capacity of the community in relation to IK. Their manual outlines more than 30 different recording and assessment methods drawn from participatory appraisal, anthropological, sociological and community organizing approaches [13].

2.3 Formulating the strategic directions for IKM in Bario

While formulating the strategic direction (Table 2) for IKMS in indigenous community, we considered the KM processes identified to be weak together with the capacity of the community and resources. In the case of Bario, we have already assessed the IKMS situation with the help of KMD tool where the results showed that the knowledge utilization process needed more focus then followed by a focus on the knowledge construction and knowledge accumulation processes. Fig. 3 contains the strategy map for IKMS in Bario. When we formulated the strategic direction we noted that the processes tend to be supporting each other to achieve the overall goal "maximizing the benefits from indigenous knowledge assets".

Table 4. Strategic Direction for IKMS in Bario

IKMS Processes	Strategic Directions.
Knowledge utilization (Use, get and contribute)	<ul style="list-style-type: none">• Identify knowledge gaps and address.• Develop collaborative decision making process.• Setting of common goals and objectives.
Knowledge accumulation (learn, assess and update)	<ul style="list-style-type: none">• Focus on sustainable transfer of knowledge; strengthen CoPs etc.• Improve situational understanding.

Knowledge construction (build, divest and innovation)	<ul style="list-style-type: none"> • Recognition of individual role in IKMS. • Develop partnerships. • Leadership development.
---	---

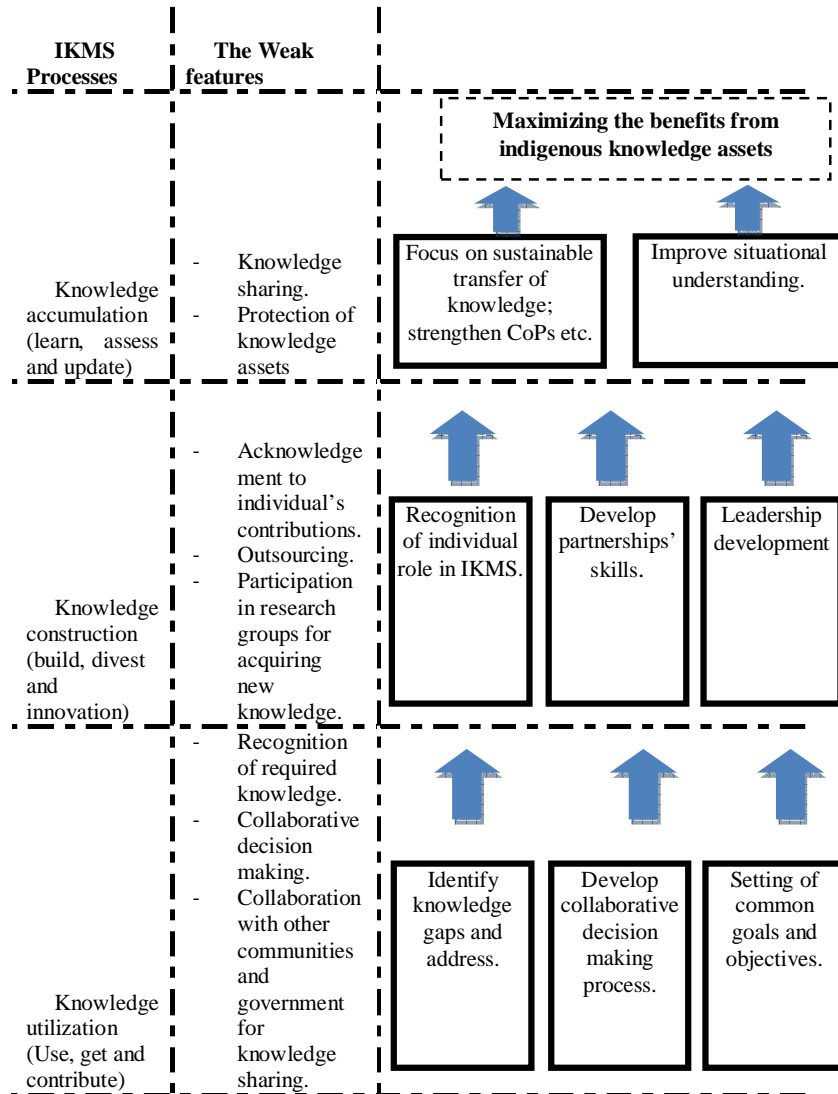


Fig 3. Strategy map for IKMS in Bario

3 Conclusion

It is an irrefutable fact that with the passage of time we are rapidly losing indigenous knowledge, so while designing the development intervention we also need to focus on indigenous system of managing the community's knowledge. Thus far, as we are successful in analysing the situation of IKMS in one indigenous community. Our future research includes the comparative study of the proposed tool and in carrying out the interventions.

References

1. Hagar, C. (2003). Sharing Indigenous Knowledge: To Share or Not to Share? That Is the Question. *Bridging the Digital Divide: Equalizing Access to Information and Communication Technologies*. Nova Scotia.
2. Ngulube, P. (2002). Managing and Preserving Indigenous Knowledge in the Knowledge Management Era: challenges and opportunities for information professionals. *Information Development* , 8.
3. Ulluwishewa, R. (1993). Indigenous knowledge, national resource centres and sustainable development. *Indigenous Knowledge and Development Monitor* , 1 (3), 11-13.
4. The World Bank Group. (2010). *Indigenous Knowledge for Development*. Retrieved 06 17, 2010, from Indigenous Knowledge: <http://web.worldbank.org/WBSITE/EXTERNAL/COUNTRIES/AFRICAEXT/EXTINDKNOWLEDGE/>
5. UNESCO. (2010). *Best practices on indigenous knowledge*. Retrieved 06 2010, 17, from Database of best practices on indigenous knowledge: <http://www.unesco.org/most/bpindi.htm>
6. Agrawal, A. (2002). Indigenous knowledge and the politics of classification. *International Social Science Journal* , 187-297.
7. UNISA. (2010). *Reverting to indigenous knowledge systems*. Retrieved 06 18, 2010, from Unisa Online: <http://www.unisa.ac.za/default.asp?Cmd=ViewContent&ContentID=23551>.
8. Bukowitz, W., & Williams, R. (1999.). *The Knowledge Management Fieldbook*. London: Pearson.
9. Okunoye, A., Innola, E., & Karsten, H. (2002). Benchmarking Knowledge Management in Developing Countries: Case of Research Organizations in Nigeria, The Gambia, and India. *Third European Conference on Knowledge Management*, (pp. 625-637). UK.
10. Grenier, L. (1998). *Working with Indigenous Knowledge: A Guide for Reserachers*. Ottawa: Internation Development Research Centre (IDRC).

11. UNPFII. (2008). *Resource Kit Indigenous People Issues*. Secretariat of the United Nations Permanent Forum on Indigenous Issues/DSPD/DESA United Nations publication.
12. The Aspen Institute. (1996). *Measuring Community Capacity Building*. Queenstown, MD.
13. IIRR. (1996). *Recording and Using Indigenous Knowledge*. Silang, Cavite 4118, Philippines.

Cultural Heritage Knowledge Discovery: An Exploratory Study of the Sarawak Gazette

M.O.Rosita, R.Fatihah, K.M.Nazri, Alvin W.Yeo and Daniel Y.W.Tan

Department of Information Systems,
Faculty of Computer Science & Information Technology,
Universiti Malaysia Sarawak (UNIMAS),
94300 Kota Samarahan, Sarawak

{morosita, rfatihah, kmnazri, alvin}@fit.unimas.my
{leonard_tan84}@yahoo.com

Abstract. Sarawak Gazette has been used as a source of reference in social science research and is considered as one of the important repository of Sarawak's history, government and politics, people and their way of life, landscape, flora and fauna. In order to preserve this valuable source of information, the documents have been converted into digitized format. With the digitized format, further work can be done in order to allow not only text searches but also semantic searches. The objectives of this study are to identify structure features of the Sarawak Gazette and to find out whether the data can be represented in the form of an ontology. This is done by applying a text mining approach that is used to extract knowledge from the digital archives and to come out with a knowledge representation in the form of an ontology. This paper reports the preliminary findings of the project as well as the goals and challenges.

Keywords: ontology, semantic, historical archives, digital archives, cultural heritage, and knowledge discovery.

1 Introduction

In this paper, we take the definition of *cultural heritage* as national heritage or heritage is the legacy of physical artifacts and intangible attributes of a group or society that are inherited from past generations, maintained in the present and bestowed for the benefit of future generations [5]. According to UNESCO [6], the term "cultural heritage" covers several main categories of heritage which is tangible cultural heritage which means movable cultural heritage (paintings, sculptures, coins, manuscripts, etc.), immovable cultural heritage (monuments, archaeological sites, and so on), underwater cultural heritage (shipwrecks, underwater ruins and cities and so on) and intangible cultural heritage (oral traditions, performing arts, rituals, and so on). Cultural heritage also include natural heritage (natural sites with cultural aspects such as cultural landscapes, physical, biological or geological

formations, and so on). The need for digital cultural heritage resources is increasing with the aim to manage, preserve and include meaning within the information.

As a result, digital libraries and web-based systems using semantic web technology have been developed to meet this need. There has been research which has resulted in systems of cultural heritage contents based on the technologies of Semantic Web [2] and [4].

Although the research in this area has been done, different forms of ontological systems have been used in different digital libraries. Ontology is defined as an explicit specification of conceptualization [11]. Basically, this definition is accepted as a definition of what ontology is for the Artificial Intelligence research community. Our intention is to preserve the documents, specifically the Sarawak Gazette, provide public access to the documents and also to allow for quick searching and information retrieval.

Sarawak Gazette has been used as a source of reference in social science research and is considered as one of the important repository of Sarawak's history, government and politics, people and their way of life, landscape, flora and fauna. It was first published in August 1870. Its objectives of this monthly articles were "to provide those Europeans who reside at Outstations with concise statements of official business and other matters of public interest..."; and "to serve as a recognized report of the condition of the various residencies under the Sarawak Government, in their relations to the natives and to the trading interests which most of them possess, for circulation in other countries and settlements." [1].

Sarawak Gazette was crucial during the colonial times as it actually study the surrounding environment; social state, economy and politics of the communities. Besides that, the articles also consist of law and order, conditions of life in the various residencies or districts, ethnic relations, general comments on the country, landscape, and social life [1]. Later, it became the annual report and reference for the colonial administration. The gazette reflects different patterns of administration from various districts and residencies.

At present, the Sarawak Gazette is only kept at the museum in its original form, and very few Sarawakians or Malaysians know that Sarawak Gazette exists. Hence, this paper describes the main goals and challenges of the exploratory study of Sarawak Gazette, existing technologies, methodology and results of the initial preliminary work.

2 Related Work

There has been concern about digital preservation in the library community for many years [8], but a serious and active interest is a relatively recent phenomenon [9]. In 1996, the Commission on Preservation and Access (CPA) and the Research Libraries Group (RLG) in the USA published a joint report on *Preserving digital information* which identified problems, made recommendations and suggested areas for further research [9].

Recently, libraries and museums are applying digital technologies to make their materials available and accessible via the Internet. The aim is to facilitate automated

information search and retrieval of these materials. Historical archives begin doing similar digitization activities as well to provide flexible options compared with the paper-based form. The collections are even scattered everywhere, for example there are some older issues are kept by the Sarawak Museum, Sarawak State Library (Pustaka Sarawak) and even in the National Library of Australia. Thus, it is difficult to search for the required information. For example, for the social scientists, searching for information is tedious and time consuming as this requires them to manually read and search through the documents.

Some of these older issues are rather delicate, deteriorated and the public cannot access them. Furthermore, these historical archives contain wealth of information which is useful to social scientists (anthropologists, political scientists, historians) but the public should also need to be educated and make known of their cultural heritage. Hence, there is an urgent need to preserve and provide access not only to interested researchers but as well to the public. Apart from that, the experts from the social scientists as well as from the languages could play a vital role on providing detail information from the gazette. For example, some local words in those days might no longer been used today or the word have been replaced with new spelling or word.

Therefore, this paper will look on the possible solution of preserving documents, which involves using Sarawak Gazette and providing access to the specific users and the public and also quick searching and information retrieval method. With the digitised format, further work can be done in order to include meaning into these archives. By providing meaning, the usability of the digital libraries can be greatly improved through the use of semantic techniques [7].

There are many digital libraries exists. For example the Greenstone Digital Library software is to preserve the documents and also provide access to the public. This has been implemented by the New Zealand digital library (a digital library which digitized some of the copies in which that document will be published online and can be downloaded by the public).

3 Explanatory Approach

There are 2 methods for this process, manual extraction that involves human and another involves semi-automated method using KAON. Several documents were used (Sarawak Gazette articles) and extraction was done with the aid of KAON. The output of this stage is a recognized ontology and ontology relationships in the format of XML.

From the original articles, the document needs to be refined. This structure corresponds closely to RDFS; the one exception is the explicit consideration of lexical entries. The separation of concept reference and concept denotation, which may be easily expressed in RDF, allows providing very domain-specific ontology without incurring an instantaneous conflict when merging ontology—a standard request in the Semantic Web. In this section, we provide details of the digitization and pre-processing process of the Sarawak Gazette.

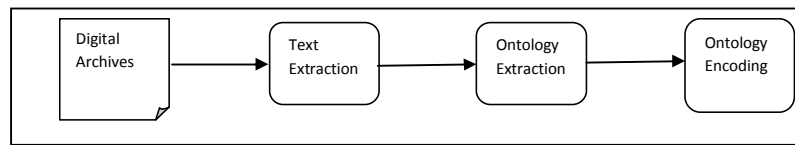


Fig 1. Text Processing Approach

i) Text extraction

The digital archives were available in the format of PDF files where each file contains several pages (images not text). The files later were run through an Optical Character Recognition (OCR) machine to extract and stored in TXT files. Once the document has been converted into the desirable format (XML), the next process is conducted. Generally, ontology are metadata add to the value of XML data by making it easier to interpret the relationships within the data. There is a way organizing XML documents according to domain-specific ontologies with better-chosen element names. Nonetheless, it cannot be estimated that all documents for a given domain will actually or directly respond to a standardized XML schema for three reasons: 1) author standardized terminology, 2) evolving ontologies, and 3) overlapping domains [13].

ii) Ontology extraction

In this stage, the content checking of TXT files and necessary modifications are done manually if there are any OCR errors. The entire ontology which would create a new sub-domain of the ontology is modeled. Thereby, ontology learning techniques partially rely on given ontology parts [13].

iii) Ontology encoding

There are differences between the semantic structures and metadata. In the structured metadata, the metadata encoding will reflect the definition structure, which means it also indirectly reflects the semantic structure that defines the fields. For the interoperability, the systems of structured metadata that apply the well defined encoding and directly reflect the semantic structure.

Moreover, Karlsruhe Ontology and Semantic Web framework (KAON) is an open source software and an ontology editor that supports entity extraction process. KAON was used to create ontology to store the events to enhance future search options. The ontology is encoded using KAON, an open source tool that provides services for ontology and metadata management, as well as interfaces needed to create and access Web-based semantic applications. KAON includes a comprehensive tool suite allowing easy ontology creation and management and also provides a framework for building ontology-based applications. It builds on available resources and provides tools for the engineering, discovery, management and presentation of ontology and metadata [3].

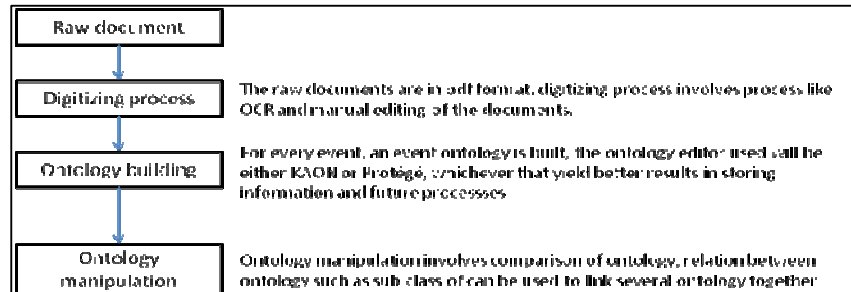


Fig 2. Process flow of ontology construction and information manipulation

4 Challenges and Limitations

The approach was applied to the collections of the Sarawak Gazette. The structures features of the Sarawak Gazette were identified; almost all articles have these four parts such as “Header” title of the article, date and age number, “Body” that contains the details of an issue and ordinance, “Appendix” and “Footer”.

Several challenges were faced when implementing this study and will be discussed in the section below.

High OCR errors

Due to the high rate of OCR errors that is only 60% recognizable characters, 50 articles have been successfully ‘cleaned’ so far. Only 80% of these articles were cleaned and 40% is only readable by human eye and is badly damaged. If the actual article is not badly damaged, the process can take up to 1 day per article (shown in Fig.3).

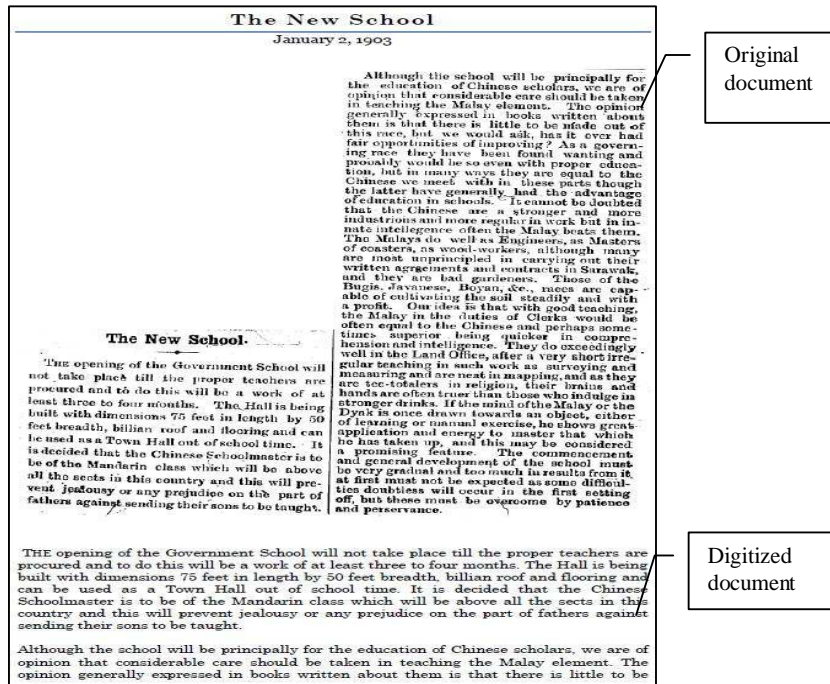


Fig. 3. Sample of original scanned image and digitized document.

Spelling changes, obsolete names, grammatical errors

Problems in spellings especially the words used in the 1800s and early 1900s where there are variations of spellings were faced by researcher and this again has delayed the time of preparing the desirable format data. Example, such as the word “dyak” is similar with the new word “dayak”.

Another challenge is obsolete names, for example a place named “Simanggang” is no longer exists but now is known as “Sri Aman”.

In order to accommodate for this phenomenon, processes need to develop a lexicon based on the old terms used from existing resources books or to extract automatically using named-entity recognition approaches.

5 On-going Work

Another way to organize the collections is based on these three concepts which are “Issues” that covers public announcements and Orders, “Location” and also “Time Line” or Year. There were also other problems encountered such as creating relevant scenarios or in other words what user wants to know or analyze from a social science

perspective or other perspectives as well. For an example, whether a murder occurred in a place named Sarikei is related with another murder in other location in that period.

Also, the first occurrence of the word say Dyak in the Sarawak Gazette would indicate the existence of such natives. The location of this instance also indicates the first time the government recognizes the ethnic group, and future occurrences and locations can be tagged to identify the migration patterns /distribution of the ethnic groups.

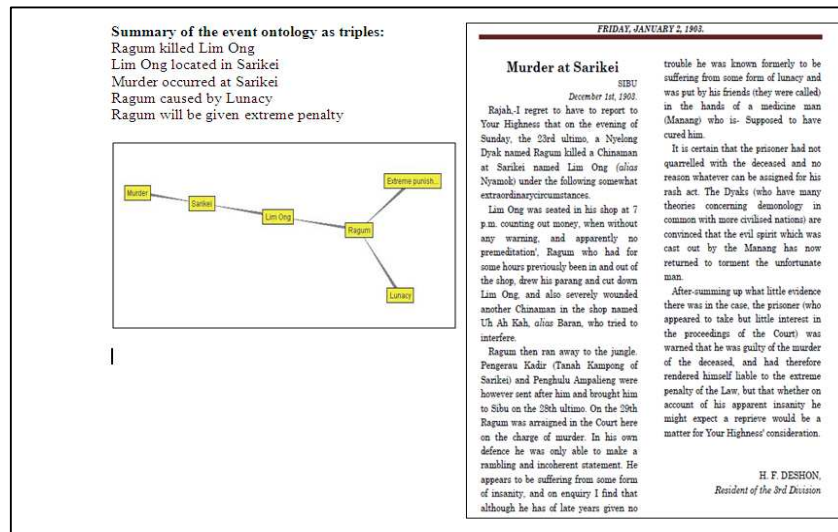


Fig 4. Sample event's result using KAON

6 Conclusion & Future Work

This exploratory study has shown that there are possibilities to have semantically related data within the historical archives. User close interventions in this case, social scientists are needed to identify relevant scenarios and interactions.

There are a lot of technologies for ontology but most of it had to be bought to use it. Nonetheless, the limitation about this is that not much ontology editor support semi-auto ontology building, as 3rd party information is needed as well.

The objective of this paper was to do exploratory study of the Sarawak Gazette which involves preserving the documents and provide access to the public. This can be done through the explicit specification of conceptualization – ontology- to quick searching and information retrieval method about Sarawak Gazette. The first result of the exploratory study has shown that there are possibilities to have semantically related data within the historical archives. It also can be use to build the ontology and extract relevant relationships. User close interventions in this case, social scientists

are needed to identify relevant scenarios and interactions. More tools are required to process text—named entity relationship. Building the ontology requires effort and validation. Expert refinement from the social scientists will be beneficial to identify possible linkages and to solve the problem in context.

Acknowledgements. The authors would like to thank Associate Professor Noburo Ishikawa, Kyoto University, and Professor Rashid Director of East Asian Studies, Universiti Malaysia Sarawak in supporting this project.

References

1. The Sarawak Gazette, <http://www.faradalemedia.com/sg/home.html>.
2. Hyvönen E., Mäkelä, E., T. Kauppinen, Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkari, P., Laitio, J., Nyberg, K.: CultureSampo—A National Publication System of Cultural Heritage on the Semantic Web 2.0, Lecture Notes in Computer Science, The Semantic Web: Research and Applications.
3. Bozak E., Ehrig, M., Handschuh, S., Hotho, A., Maedche, A., Motik, B., Oberle, D., Schmitz, C., Studer, R., Stumme, G., Sure, Y., Staab, S., Stojanovic, L., Stojanovic, N., Tane, J., Volz, R., Zacharias, V.: KAON - Towards a large scale Semantic Web.
4. Ahonen E., Hyvönen E., Publishing Historical Texts on the Semantic Web: A Case Study, IEEE International Conference on Semantic Computing, (2009).
5. Cultural Heritage Definition: Wikipedia, http://en.wikipedia.org/wiki/Cultural_heritage
6. Cultural Heritage Definition: United Nations Educational Scientific and Cultural Organization (UNESCO), <http://portal.unesco.org/culture>
7. Hyvönen E., Mäkelä, E., T. Kauppinen, Alm, O., Kurki, J., Ruotsalo, T., Seppälä, K., Takala, J., Puputti, K., Kuittinen, H., Viljanen, K., Tuominen, J., Palonen, T., Frosterus, M., Sinkkilä, R., Paakkari, P., Laitio, J., Nyberg, K.: CultureSampo—Finnish culture on the semantic web 2.0. thematic perspectives for the end user, Proceedings Museums and the Web 2009, Indianapolis, USA, April 15-18 (2009).
8. Future goals in digital archiving, <http://www.ukoln.ac.uk/services/elib/papers/other/jisc-npo-dig/foreword.pdf>.
9. Waters, D and Garrett, J.: Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by the Commission on Preservation and Access and the Research Libraries Group. Washington, DC, <http://www.ukoln.ac.uk/services/elib/papers/other/jisc-npo-dig/foreword.pdf>
10. Preserving Digital Information, Report of the Task Force on Archiving of Digital Information, commissioned by The Commission on Preservation and Access And The Research Libraries Group, (1996).
11. Ontologies and Semantic Web, <http://www.obitko.com/tutorials/ontologies-semantic-web/what-is-ontology.html>.
12. Ian H. Witten, Building Digital Library Collections with Greenstone, New Zealand Digital Library Project, (2007).
13. Chaudhri, A., R. Zicari, and A. Rashid, *XML Data Management: Native XML and XML Enabled DataBase Systems*, Addison-Wesley, USA, (2003).
14. Alexander Maedche and Steffen Staab, Ontology Learning for the Semantic Web, http://www.aifb.uni-karlsruhe.de/~sst/Research/Publications/ieee_semweb.pdf.

Large-Scale Semantic Text Understanding

Benjamin Chu, Fadzly Zahari and Dickson Lukose

MIMOS Berhad, Technology Park Malaysia, 57000 Kuala Lumpur, Malaysia
{mx.chu, fadzly.zahari, dickson.lukose}@mimos.my

Abstract. Semantic Text Understanding involves linguistic based processing of text and transforming it into a conceptual representation of its meaning. In this paper, we introduce a potential system for Text Understanding that is highly scalable that takes as input unstructured text and produces conceptual graph representation of its meaning. Several experiments were conducted to gain explanatory insights into its scalability and performance. Varied test scenarios and configuration setups were experimented. Through these experiments we will demonstrate how the configuration of the system deployed affect the scalability and performance.

Keywords: Semantic Text Understanding system, Natural Language Processor, Semantic Text Interpreter.

1 Introduction

Statically-orientated linguistic text processing has been studied in the Information Retrieval (IR) field for some time. The development of fast algorithms for text processing is critical in helping a user locate relevant materials among the retrieved documents as quickly as possible. Extracting actionable insight from large highly dimensional data sets, and its use for more effective decision-making, has become a pervasive problem across many fields in research and industry. When the amount of data increases, both in terms of size and dimensions, it is becoming harder to make accurate interpretations while retaining the main features of the data. In the real world, many information retrieval tasks are difficult because of high data dimensionality and the lack of annotated examples to train the retrieval algorithm, thus resulting in the performance of IR algorithms often unsatisfactory.

In recent years, much effort is focused in web mining and text mining, especially for Information Gathering, Intelligence Management and Information Consolidation initiatives. Much of these work is focused on mining unstructured text, be it on the web or otherwise. Producing the semantic representation of the unstructured text is the goal of these initiatives. To this end, explicit knowledge representation formalism is essential. One such representation formalism is Conceptual Graphs [2]. It is a knowledge representation formalism based on Semantic Networks and the Existential Graphs of C. S. Peirce. There are many recent developments in semantic representation formalisms. Among them include RDF [29], RDF-S [30], and OWL [31].

The ability to represent the semantic of the text is only one of the aspects of Semantic Text Understanding System (STUS). The most crucial aspect of such a system is the algorithm for processing these sentences and producing corresponding conceptual structures. Further to that is the aspect of scalability and performance. In this paper, we focus our attention to the scalability and performance of such a system. To this end, this paper is outlined as follows. Section 2 describes related work on Natural Language Processing; Section 3 introduces the architecture of the STUS, while Section 4 describes the Graphical User Interface (GUI). In Section 5, we will outline the systems configurations for the experiments to be conducted, while in Section 6 we discuss the results from these experiments. Finally, Section 7 concludes this paper by overall outcome, and discusses future research directions.

2 Related Work

Many studies have been carried out in the field of Natural Language Processing (NLP), especially in the aspects of text understanding. We focus our review on systems that utilizes conceptual graphs as its knowledge representation scheme [1, 9, 10, 12]. The initiative by Hensman [10] utilizes a combination of syntactic and semantic information from WordNet [22] and VerbNet [23]. The system utilizes a labeling algorithm to annotate semantic roles in a sentence before conceptual graph representation are generate. On the other hand, the initiative by Boytcheva et al. [11] resulted in a system called CGExtract that utilizes a component names Parasite [25] which performs the morphological, syntactic and semantic analysis of the input sentence. The effort by Petermann [12] implemented an NLP system called CoKEMan to analyze and process a collection of reference manuals into conceptual graph representation by using both syntactic and semantic processing.

Extracting meaningful information from unstructured text is the essential challenge addressed by most of these initiatives. In developing these systems, the researchers have addressed several computational linguistic challenges including lexical, morphological, syntax and semantic processing. When dealing with semantic processing, there are several fundamental challenges. Among them include the coverage of the knowledge based used by the system for processing the “text”, limitations on the algorithms and rules for conducting the semantic processing, and the knowledge representation scheme used. In our review, we focused on systems that utilized conceptual graphs as their knowledge representation scheme. To build a Text Processing system based on conceptual graph, we will require a system (can also be a library of a Software Development Kit) that is will enable us to represent and manipulate knowledge in the form of conceptual graphs. There already exist several such tools within the research community. Among then include Extendible Graph Processor [18], PROLOG-CG [13], Notio [14], Amine [15], CGKEE [16], CharGer [17, 21], Cogitant [19] and CGPro [20]. All these systems are focused on implementation of various algorithms for graph manipulations (e.g., join, maximal join, projection, etc.). Some efforts were directed to visual editors, and liner representation of conceptual graphs. To-date, there is no commercially available

Conceptual Graph Processor or Editor¹. In addition, none of the above listed system has an efficient graphs storage and retrieval mechanism. To our knowledge, none of them are of commercial strength. Also, none thus far have been demonstrated the ability to efficiently handle billions of graphs (where real world applications, particularly in the Semantic Web require systems with these capabilities). These are challenges to overcome to secure large scale commercial adoption of conceptual graphs.

All the text processing systems we reviewed above are focused on algorithms and rules associated to generating semantic representation of the meaning of the “text” in conceptual graphs. None are focused on scalable architecture to produce ultra high performance. Scalability still remains a challenge that needs to be resolved. In the following section, we will describe the architecture of the Semantic Text Understanding System (STUS), which is designed for extremely high degree of scalability and performance.

3 Architecture Overview

The STUS is implemented using the Semantic Technology Platform (STP) that was developed. The STP contains a large number of web-service enabled components, plug-ins and tools that can be utilized to deploy sophisticated applications in a Service Oriented Architecture (SOA) environment [27]. Web-Service is enabled via SOAP [28]. Fig. 1 depicts the architecture of the STUS. This system is designed to be used by many people on the web. The Graphical User Interface (GUI) developed for this system is written in Java Swing thus it is executable from the web browsers.

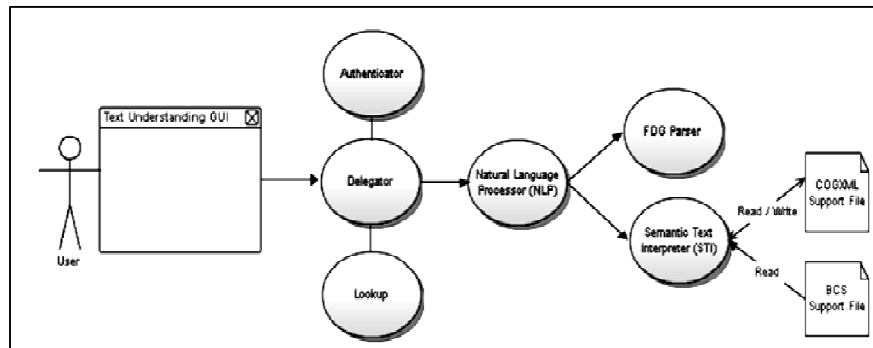


Fig. 1. Architecture of the Semantic Text Understanding System
© 2009-2010 MIMOS Berhad. All Rights Reserved.

¹ FDGP provides a full analysis of texts by showing how words and concepts relate to each other in sentences for analytic applications to understand text beyond the level of words, phrases and entities: also their interrelations (such as events, actions, states and circumstances) that constitute the "story" of the text [8].

As shown in Fig. 1, there are 6 different components used to configure the STUS. They are: Delegator, Authenticator, Look-Up Server, Natural Language Processor (NLP), FDG Parser (FDGP), and Semantic Text Interpreter (STI). Delegator is responsible for handling incoming request from authenticated clients, identifying corresponding free components that can respond to the request, and sending the request to the designated component. Authenticator is responsible for performing the authentication of a client. The Look-Up Server is where all instances of each type of components are registered. It maintains the status of each of the components that make up the application. The NLP component is responsible for accepting an incoming “text”, finding a freely available FDGP, sending the “text” to the FDGP, collecting the output of the FDGP, then finding a freely available STI and passing output of the FDGP to the STI, collecting the output of the STI (a set of conceptual graphs), and finally sending these graphs to the Delegator, who will then pass it to the client that initially sent the “text”. The Client Application will then display the received conceptual graph in linear form. The FDGP² is responsible to perform syntactic analysis on incoming “text” and produce “syntax structures”. The STI is responsible to performing semantic analysis on the incoming “syntax structure” and produce one or more “semantic representation”. In this case the semantic representation is in the form of Conceptual Graphs. Fig. 2 depicts the interactions between these components as described above.

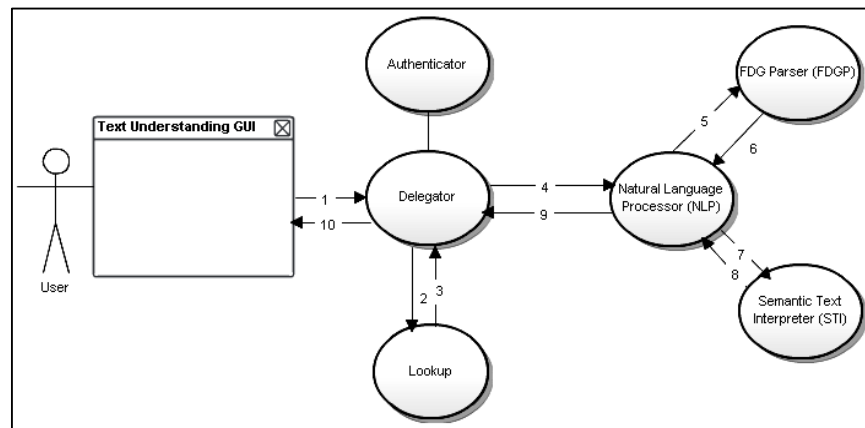


Fig 2. Interactions between components in the STUS
© 2009-2010 MIMOS Berhad. All Rights Reserved.

² FDGP provides a full analysis of texts by showing how words and concepts relate to each other in sentences for analytic applications to understand text beyond the level of words, phrases and entities: also their interrelations (such as events, actions, states and circumstances) that constitute the "story" of the text [8].

4 Graphical User Interface

The GUI for the STUS, as depicted in Fig. 3, is implemented in SWING, and is designed to run as a standalone application from any machines (running a JVM), or via Web Browsers. It is designed for the user to send single sentence at a time, multiple sentences at a time, or enable the user to select a text file containing very large number of sentences (for example, a document, or even a text corpus).

There are three main panels on the GUI. First panel is the “Sentence Output Panel” where the sentence that is being process will be displayed. The second panel is the “FGDP Output Panel” where the output of the FDGP is displayed (i.e., output from the syntax analysis). Finally, the third panel is the “CG Output Panel” where the full CG or partial CG that was generated by the STUS will be displayed. The content of these 3 panels in Fig. 3 is for processing the following sentence:

“George-Lucas drives to London with Francois-Truffaut because he has to meet Quentin-Jerome-Tarantino for an important lunch on Friday.”

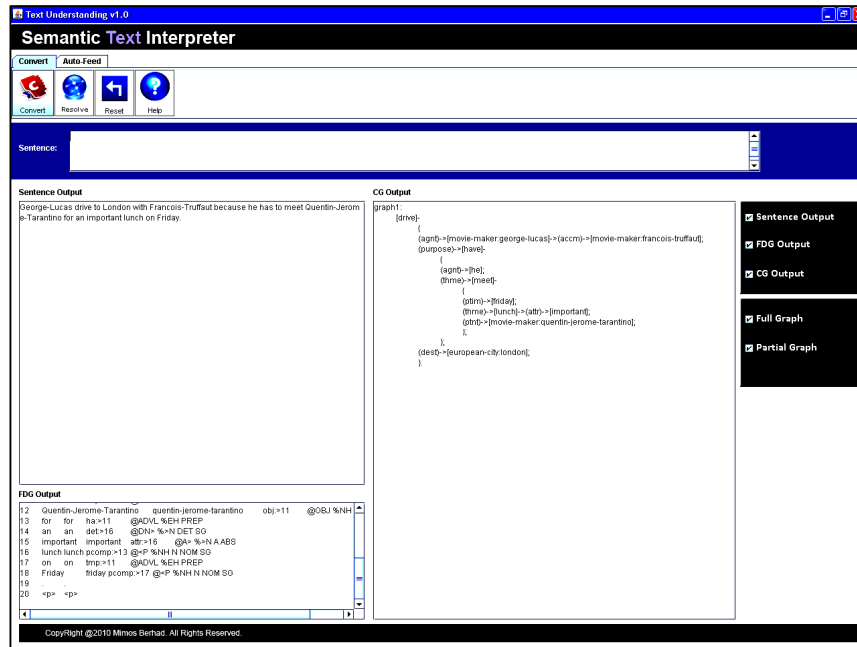


Fig. 3 Graphical User Interface for STUS
© 2009-2010 MIMOS Berhad. All Rights Reserved.

The user can also dictate to see only “Full Graphs” or “Partial Graphs”. That is, not all sentences will be processed by the STUS to produce a fully connected graph. Due

to numerous reasons³, STUS will produce one or more partial graphs to represent of partial meaning of the sentence. The GUI provides the option for the user to view only the full graphs, or to also view the partial graphs.

5 Experiment Objective

5.1 Scalable Parameters

Scalability is to handle growing amounts of work in a graceful manner [3]. In this system the size of the given document (number of sentences in a text) identifies the scale of the input and the correctness of the output. Meanwhile, the response time of the system specify the performance of the system. In this section the focus is on the scalability of the system with respect to the number of sentences.

The Fig. 1 shows that the STI and the FDGP are involved in the text understanding process. The performance of the STI component is measured based on the time it needs to process a single sentence and return the corresponding Conceptual Graphs (CGs). The processing time is largely depends on the sentence complexity.

We intend to evaluate the scalability and performance of the STI and FDGP by increasing the number of sentences as well as increasing the number of users. To support large-scale request architecture was designed using multiple server environment and Service Oriented Architecture. This architecture can provide service to large number of users' inputs. The system architecture can handle large number of sentences (in a bucket) given by each user in a shorter time. In order to illustrate the scalability of this architecture, a configuration for system performance measurement test bed is explained in the following section.

5.2 Measurement Configuration Set up

The measurement starts from the time when user send the input sentence(s) via GUI until the STI output in Conceptual Graph (CG) representation is displayed. As previously mentioned the NLP component controls and handles activities between FDGP and STI. NLP and FDGP are multi-threaded, while STI is single-threaded. The rest of this section describes the measurement set ups in details.

5.2.1 Load Measurement Set ups

³ The FDGP would not produce the complete "syntax structure" for a sentence if it is not grammatically correct and/or if some words are not recognized by the parser. On the other hand, it is also possible that the STI was not able to produce the conceptual graph as the algorithms and rules in STI is not sufficient to handle the syntax structure of the sentence and/or certain individuals (proper nouns) are not found in the knowledge base used by STI. A detailed discussion of this matter is beyond the scope of this paper.

SoapUI [24] which is an Open Source Web Service Testing Tool for Service Oriented Architecture has been used to evaluate the load testing for 300 concurrent users against the architecture described in Section 3.

A load test in the SoapUI can be executed using a number of strategies depending for how each of the test cases is executed. As shown in Table 1, a simple strategy is selected for a load test with configurable delay. We have selected 300 threads to simulate the concurrent users and set for each of the test delay to be 1000 milliseconds. Since the random parameter we have set its value to zero, the random setting will be ignored and run without any delays.

Table 5. SoapUI Pro 3.5 Configuration Settings

Load Test Setup	
Threads	300
Strategy	Simple
Test Delay	1000
Random	0.0

5.2.2 Environment Configuration Scenarios

The measurement is tested on several configurations where Linux CentOS 64-bit and Window XP 32-bit are the environment setup for our testbed. Table 2 shows the setup required for the load test being carried out where the results are shown in Table 5. As shown in Table 3 and Table 4, there are two scenarios being configured for the experiment tests.

In Table 3, there are total of five client desktops used in this testbed configuration. One of the client machines is used as the location for the FDGP server and 1 Delegator and 1 Lookup for the SOA framework. The other four client machines are configured for each of the components required: 2 NLP, 2 FDGP, and 2 STI. Total instances for this setup are 8 NLP, 8 FDGP, and 8 STI. The end performance results are shown in Table 6.

Table 4 describes the setup for two machines in the Grid network and a local machine for FDGP server. Grid Server 1 is configured with 1 Delegator and 1 Lookup for the SOA framework whereas 10 web instances for each for NLP and STI are setup on Grid Server 2. The end performance results are shown in Table 7.

Table 2. Load Test Configuration

Machine ID	Quantity	Components
Grid Server 1	1	20 NLP, 20 FDGP, 20 STI
Grid Server 2	1	4 Delegator, 4 Lookup
Grid Server 3	1	20 NLP, 20 FDGP, 20 STI
Local Server 1	1	40 FDGP

Table 3. Experiment Configuration Test Setup A

Machine ID	Quantity	Components
Client Desktop	4	2 NLP, 2 FDGP, 2 STI
Local Server 1	1	1 FDGP, 1 Delegator, 1 Lookup

Table 4. Experiment Configuration Test Setup B

Machine ID	Quantity	Components
Grid Server 1	1	1 Delegator, 1 Lookup
Grid Server 2	1	10 NLP, 10 STI
Local Server 1	1	10 FDGP

The testing experiment for scalability measurement was carried out base upon the above setting and the results are further discussed in the following section.

6 Experiment Results

We have experimented the system from three perspectives: 1) Performance of FDGP to return syntactic information for input text, 2) Load test for concurrent user performance, 3) Conversion of text to CG. Following sections describe each test in detail.

6.1 FDG Parser (FDGP)

This section describes experimental results (as shown in the Table 5) of time taken by FDGP component to return syntactic information of the input text back to NLP component. The first column (*Test Sentences*) of the table shows the number of input sentences, the second column (*Individual*) shows time taken by the FDGP to return syntactic information to NLP component when each sentence is passed to FDGP individually. The third column (*Bucket Sentence*) shows the time taken by the FDGP component when more than one sentences are sent together as a bucket of sentences. Table 5 clearly shows that the time taken to return the syntactic information when sentences are passed in a bucket is much lesser than the time taken when sentences are sent to FDGP individually.

Table 5. FDGP Results

Test Sentences	Individual Sentences (s)	Bucket Sentence (s)
10	4	1
100	21	1
1000	240	8

6.2 Load Test

Performance (i.e. supporting multiple users at a time) is one of the important issues that we have addressed in the work done. We have tested the system for performance and it can support more than 300 concurrent users without effecting its time of response. We have used the SoapUI (i.e. a tool for performance testing) to test the system. Results obtained from SoapUI are shown in the Table 6.

Test Step (i.e. *invoke* as shown in the Table 6) sets the startup delay (in milliseconds) for each thread, setting to the delay value to 0 will start all threads simultaneously), *min*, *max* and *avg* are the shortest and longest and average time for the test step, *last* means last time for the test step, *cnt* shows the number of times the test step has been executed, *tps* is the number of transactions per second for the test step, *bytes* shows the number of bytes processed, *bps* is the bytes per processed, *err* shows the number of assertion errors for the test step) and finally *rat* is the error ratio. From the first iteration in Table 6 shows that the average time taken by the system to reply 300 concurrent users is 1356.5 milliseconds.

The first experimental result (as shown in the first row of the Table 6) shows that the invoke test for 300 concurrent users, total number of request sent altogether is (*cnt*) 38386. The minimum (*min*), maximum (*max*) and average (*avg*) time per request taken by the system to reply these requests is 75, 24886 and 1356.5 milliseconds respectively. Time taken to reply for the last request is 8086.

Table 6. Load Test Statistical Result

Test Step	min	max	avg	last	cnt	tps	bytes	bps	err	rat
Invoke	75	24886	1356.5	8086	38386	126.38	422100	138973	0	0
Invoke	9	31388	1364.8	8233	38290	126.23	421072	138821	0	0
Invoke	166	37159	3304.4	10657	21078	68.85	231253	75547	0	0
Invoke	110	38827	3332.8	15694	20988	67.78	230261	74366	0	0

6.3 Text to CG

The third set of experiments is to test the system efficiency (i.e. time taken by system) to generate the CG from input text. We have tested the system in two different configuration setups (i.e. setup A & B as discussed in the Section 5.2.2) to verify the scalability and performance of the system. Experimental results of the system in the setup A are shown in the Table 7. The first column of the Table 7 (i.e. *Data File Name*) is the name of the source file passed to the system. *No. Sentence* is the number of sentences contained by the source file. *Full Graph* is a complete CG with fully connected nodes. *Partial Graph* is also a CG but with incomplete nodes. *FDGP Error* shows the number of sentences that could not be converted to CG. *Total Graph* is basically sum of *Full* and *Partial Graphs* generated by the system. First experimental result (1st row of the Table 7) shows that the time taken by the system to process the input text file consisting of 100 sentences is 80 seconds and the total number of graphs generated is 94.

Table 7. Results of text to CG transformation with configuration setup A

Data File Name	No. Sentence	Full Graph	Partial Graph	FDGP Error	Total Graph	Time(s)
<i>set-100.txt</i>	100	31	63	6	94	80
<i>sample-597.txt</i>	597	132	228	237	360	471
<i>reut2-007-v7.txt</i>	3914	563	587	2764	1150	2583
<i>reut2-003-v0.txt</i>	4205	605	674	2926	1279	2609
<i>reut2-004-v0.txt</i>	4262	582	660	3020	1242	2891

Experimental results of the system test in the setup B are shown in the Table 8. Experimental results clearly show that the time taken by the system to convert a text file of 100,597 and 3914 sentences to CG reduced from 80,471, 2583 (Table 7) to 4, 22 and 140 (Table 8) seconds respectively. The approach of sending sentences to FDGP in a bucket resulted in great reduction of time to extract syntactic information from input text but even then single threaded interaction between NLP and STI was big difficulty need to be resolved to acquire better performance and scalability. By enabling the NLP component to be multi-threaded for the interaction with multiple instances of STI will greatly reduce the time as depicted in Table 8.

Table 8. Results of text to CG transformation with configuration setup B

Data File Name	No. Sentence	Full Graph	Partial Graph	FDGP Error	Total Graph	Time(s)
<i>set-100.txt</i>	100	31	63	6	94	4
<i>sample-597.txt</i>	597	132	228	237	360	22
<i>reut2-007-v7.txt</i>	3914	563	587	2764	1150	140
<i>reut2-003-v0.txt</i>	4205	605	674	2926	1279	165
<i>reut2-004-v0.txt</i>	4262	582	660	3020	1242	172

7 Conclusion

The STUS comprises of several web-service enabled components: NLP, STI, and FDGP. Following this architecture, the separation of responsibilities leads to a solution that is (1) high performing, and (2) scalable. The high performance and scalability characteristics are possible due to the multi-threading of the NLP component, which enables it to take advantage of utilizing the entire available STI components concurrently. This enables the NLP component to freely exploit the hardware and run at optimal speed.

We profiled the performance of the system based on the proposed architecture on several test-bed configurations and presented the results. In order to conduct a comprehensive and equitable evaluation of performance, different test scenarios are used and different hardware configurations are employed.

The results of the experiments demonstrate that the performance of the STUS show dramatic improvements when we are able to take advantage of the SOA framework

(meaning that when we were able to use all the available STI components in the system, we see increase in performance).

Even though the performance of the STUS is good, and the SOA architecture allows us to further increase its performance, there are still many areas that need improvement. Particularly in the semantic processing of the unstructured data, where we are looking into anaphora resolution and word sense disambiguation to improve the quality of the generated conceptual graphs. The other area of improvement that we are considering is to remove the FDGP from this system, and to enhance the STI to perform semantic analysis directly on the unstructured text, rather than on the “syntax structure” produced by the FDGP⁴.

Acknowledgements. Authors of this paper would like to extend their heartfelt thanks to Dr Ahtisham, Dr Karthigayan and Dr Beik from MIMOS Bhd., for their invaluable contribution.

References

1. Sowa, J.F., Way, Eileen C.: Implementing a semantic text interpreter using conceptual graphs. In: IBM Journal of Research and Development, Vol. 30 (1), pp. 57–69, 1986.
2. Sowa, J.F: Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley (1984).
3. Luke, E.: Defining and measuring scalability. In: Scalable Parallel Libraries Conference, pp.83–86, IEEE Press (1994).
4. Xiao, L., Hu, B., Croitoru, M., Lewis, P. and Dasmahapatra, S.: A knowledgeable security model for distributed health information systems. In: Computers and Security, Vol. 29 (3). pp. 331–349, 2010.
5. Delugach, S.: Towards building active knowledge systems with conceptual graphs. In: 11th International Conference on Conceptual Structures, ICCS’03, pp.296–308, Springer (2003).
6. Baget, J., Carloni, O., Chein, M., Genest, D., Gutierrez, A., Leclère, M., Mugnier, M., Salvat, E.: Towards Benchmarks for Conceptual Graphs Tools. In: Conceptual Structures Tool Interoperability Workshop, CS-TIW’06, pp. 72–86, Aalborg University Press (2006).
7. Boytcheva, S., Angelova, G.: Towards Extraction of Conceptual Structures from Electronic Health Records. In: 17th International Conference on Conceptual Structures, ICCS’09, pp.100–113, Springer (2009).
8. FDG Parser: <http://www.connexor.eu>
9. Spyns, P., De Moor, G: A Dutch medical language processor. In: International Journal of Bio-Medical Computing, pp. 181–205, 1996.
10. Hensman, S.: Construction of Conceptual Graph representation of texts. In: Proceedings of Student Research Workshop at HLT-NAACL, Boston, pp. 49–54, 2004.
11. Boytcheva, S., Dobrev, P., Angelova, G.: CGExtract: Towards Extraction of Conceptual Graphs from Controlled English. In: Conceptual Structures: Extracting and Representing Semantics, Contr. ICCS-2001, pp. 89–102.
12. Petermann, H.: Natural Language Processing and Maximal Join Operator. In: 4th International Conference on Conceptual Structures, ICCS’96, pp.100–114, Springer (1996).

⁴ As seen in the results of the experiments, almost 2/3 of the sentences are rejected by the FDGP, mainly because these sentences violate the grammar used by the parser to syntactic analysis.

13. PROLOG-CG, <http://prologpluscg.sourceforge.net>
14. Southey, F., Linders, J.G.: Notio – A Java API for Developing CG Tools. In: Conceptual Structures: Standards and Practices, Vol 1640, Lecture Notes in Artificial Intelligence, pp. 262–271, Springer-Verlag (1999).
15. Kabbaj, A., Janta-Polczynski, M.: From Prolog++ to Prolog+CG: A CG Object Oriented Logic Programming Language. In: Conceptual Structures: Logical, Linguistic, and Computational Issues, Vol. 1867, Lecture Notes in Artificial Intelligence, pp. 540–554, Springer-Verlag (2000).
16. Lukose, D.: CGKEE: Conceptual graph knowledge engineering environment. In: 5th International Conference on Conceptual Structures, ICCS'97, pp.598–602, Springer (1997).
17. CharGer, <http://charger.sourceforge.net>
18. Tsui, E., Garner, B., Lukose, D.: Extendible Graph Processor. In: 5th International Conference on Conceptual Structures, ICCS'97, pp.594–602, Springer (1997).
19. Cogitant, <http://cogitant.sourceforge.net>.
20. H. Petermann, L. Euler, and K. Bontcheva.: CGPro – a Prolog Implementation of Conceptual Graphs. In: Technical Report, University of Hamburg, FBI-HH-M-251/95, 1995.
21. Delugach, H.: Charger: Some Lessons Learned and New Directions. In: Working with Conceptual Structures: Contributions to ICCS 2000, G. Stumme, Ed. Aachen, Germany, pp. 306–309, Shaker Verlag (2000).
22. WordNet, <http://wordnet.princeton.edu>
23. VerbNet, <http://verbs.colorado.edu/~mpalmer/projects/verbnnet.html>
24. SoapUI, www.eviware.com
25. Parasite, http://www.co.umist.ac.uk/sta_/ramsay.htm
26. MIMOS Berhad: www.mimos.my
27. Service Oriented Architecture (SOA): <http://java.sun.com/developer/technicalArticles/Webservices/soa/>
28. Simple Object Access Protocol (SOAP): <http://www.w3.org/TR/soap/>
29. Resource Description Framework (RDF): <http://www.w3.org/TR/2004/REC-rdf-primer-20040210/>
30. Resource Description Framework Schema (RDF-S): <http://www.w3.org/TR/rdf-schema/>
31. Web Ontology Language (OWL): <http://www.w3.org/TR/owl-ref/>

Semantic Arabic Search Tool

Majdi Beseiso¹, Abdul Rahim Ahmad¹, Jamilin Jais²

¹ Universiti Tenaga Nasional (UNITEN),
Km 7, Jalan Kajang-Puchong, 43009 Kajang, Selangor, Malaysia.
{abdrahim, majdi}@uniten.edu.my, jamilin@gmail.com

² Imam Muhammad Ibn Saud Islamic University,
Riyadh, Saudi Arabia.
jamilin@gmail.com

Abstract. Since the advent of Semantic Web in late 90's, numerous websites took advantage from the capabilities provided by the semantic web technologies, such as the language support, techniques used, and semantic search capabilities. These capabilities encompass smart reasoning over facts and numbers, semantic search and facts and numbers interoperability. However, most semantic web technologies are dedicated to processing Latin scripts, thus the niche for providing Arabic script support in these technologies. This paper describes the design and implementation of an Arabic semantic web retrieval engine named SemARAB that employs semantic ontology. SemARAB enable users to search based on keyword semantic through an easy to use visual search interface. To provide an effective retrieval and to tackle problems in Arabic language processing, the tool was built based on semantic similarity between concepts of specific ontology and content-based similarity for different resources. The model is implemented for searching on the "electronic commerce" domain, and evaluation is done based on this.

Keywords: Arabic Semantic Technology, semantic web, semantic, ontology.

1 Introduction

Semantic refers to the study of meaning. In the perspective of software technology, semantic means the utilization of domain or field knowledge to further improve the software to make it adaptive, user friendly, efficient and intelligent. Through the use of expertise or knowledge, software and other applications can be enhanced or developed to their maximum functionalities and performance (Lee, 2004). Thus, semantic web helps the development of software's that are more convenient, efficient and adaptive to use. Various technologies have been developed to support web-based communication, such as using complex "language" codes and information storage systems.

One result of this is the keyword-based search engines as an Internet search tool. However, there is still this problem of how to make information search easier and more precise to achieve. Although the role of keyword-based search engines in facilitating information storage, dissemination and search had been recognized, there

have been some concerns regarding their use (Antoniou & Van Harmelen, 2004). Search engines are hampered with the problems of “high-recall, low precision”, “low recall, no precision,” sensitivity of results to vocabulary, and that result are produced in single pages. In short, search engines are limited to producing search results with general meanings. In addition, the results are not ‘machine accessible,’ meaning that they cannot be passed on for further processing by subsequent software tools.

One solution to these problems is the Semantic Web Technology. It is a systematic initiative implemented by members of the World Wide Web Consortium to improve the quality of services in the Internet, the process of which is conducted in layers and focuses in improving the use of explicit metadata, inferential agents, ontologies and logic. Semantic Web technology is vital since it particularly helps reduce “information overload,” link “stovepipe systems and enrich poor content aggregation (Daconta et.al, 2003). The technology also reduces blockages in information networking by linking different systems and improving the way program content is being populated. The technology was conceptualized by Tim Berners Lee in 1990 to create a “world wide web” of documents linkable throughout the world. This facilitates “meaning transmission” rather than just document transfer (Orr, 2005).

In applying Semantic web technology, developers have to ensure that the original meanings of the texts are integrated into the language system being used in the Internet by using ontologies. These ontologies correlate concepts and ideas, and define rules in logically correlating these concepts and ideas (Orr, 2005). There are currently development standards and specifications in which tools are developed for creating and developing Semantic Web applications. Most tools are used to express knowledge using language specifications such as Extensible Markup Language (XML), Resource Description Framework (RDF), RDF Schema, and Web Ontology Language (OWL). The list of tools just keeps growing every day.

This paper propose an Arabic semantic search tool named SemARAB that automates the process of information retrieval from Arabic web resources that is relevant to a user’s query. SemARAB is a search paradigm which aims to increase effectiveness of a retrieval system by identifying an additional semantic layer for the results returned by the search engine. This tool model is based on semantic similarity between concepts from a specific ontology and content-based similarity for different resources. The proposed tool addressed the problems of semantic search for Arabic data, Arabic language processing and the absence of resources in Arabic language annotated with semantic metadata. This paper is organized as follows; section 2 describes the challenges in the development of semantic tools in Arabic language. Section 3 describes the proposed tools architecture, followed by the tools’ GUI description in section 4. The sub modules in the tool are described in section 5 followed with the testing domain and experiments in section 6 and 7 respectively. Conclusion is derived in section 8.

2 Semantic Web Challenges, Future and Requirements

Semantic web is the future of the World Wide Web. The web today is designed for human utilization only, not for computers and machines. Semantic Web will add more

functionality to web pages and make contents, information, data, and other documents available for automatic processing by the computer itself. Semantic Web has many benefits such as consistent interconnection and minimal error messages, and will be distributed to everyone just like the ordinary Web. It can be accessed by corporations and individuals (Berners-Lee, 2001).

A study on the trend of Semantic Web by Cardoso (2007); indicated that majority of the use of ontologies for building knowledge is in education, followed by computer software, government services, business and life sciences. There are also some use of ontologies to represent knowledge in communication, media, and healthcare. Ontologies are also used for sharing of information among other users and software agents. This means that the data must be understood by humans and the computers. Other purposes of ontologies can be for code generation, data integration, data exchange, document annotation, information retrieval and many more.

The primary goal of Semantic Web is to describe information that is understandable by machines. With opportunities for the Semantic Web to be fully used worldwide, there are challenges that wait such as the development of ontologies, formal semantics of Web language, and the trust and proof models (Lu, 2002). Ontologies are the most essential part of developing Semantic Web. There are different aspects of this part which include ontology representation languages, ontology development, ontology learning approaches, and ontology library systems. These aspects manage, adapt, and standardize the ontologies (Lu, 2002).

In a simple definition, a Semantic Web is a Web that is enhanced to be able to clearly understand the contents and structure of a web page. It makes web search engines and web browsers capable of reading and processing web contents to provide better services to users. It will enable computers and other machines to process ideas and solve difficult optimization problems.

2.1 Arabic Language Challenge

Arabic is the language of millions of people in the countries of Middle East and Northern African countries. This has become their official language. In fact, it has become the religious language among Muslims throughout the world. Arabic language is very important in the lives of the Muslims as it is the language of the Qur'an. Unfortunately, there have been little researches or studies of incorporating the language in connection with computers (Tjoa, A., et al, 2006). Thus, the proposed study of using the Arabic language as framework for semantic web is a new and motivating topic that will benefit millions of people who are using the language in their everyday lives. Writing in Arabic is done from right to left, in contrast to the English language. Arabic morphological analysis is a very complex task because the language is highly inflectional and derivational (Hammo et al, 2002). There are 28 letters in the Arabic alphabet while the English language has only 26 letters (Adetunji & Ahmed, 2008).

The Semantic Web will soon conquer the world of technology. But the tools and languages are not ready for Arabic language users. The rise and development of the Semantic Web must address the challenges of the complicated Arabic language. Al-Khalifa and Al-Wabil (2007), raised concern over the existing Semantic Web tools

and applications in supporting the Arabic language. Since the technology would become available worldwide, it must prepare itself for the hundreds of different languages to support. There are issues concerning the Semantic Web that pose challenges to the Arabic language and its users. There is lack of Arabic language support in the available and current Semantic Web tools, lack of existing Arabic Semantic Web applications, and limited support for Arabic research on Semantic Web technologies (Al-Khalifa, 2007).

A common problem with Semantic Web tools is the processing and encoding of Arabic texts. Currently, there are a lot of applications that can encode Arabic scripts. To solve the problem, Semantic Web tools should focus on one encoding schema for Arabic such as Unicode (Al-Khalifa, 2007). Another problem with the Semantic web is that 49 percent of the ontologies in the ontology libraries are written in English. This is in connection with the lack of Web tools and software development applications that cater to the Arabic language and its users. If there are tools and applications that support Arabic, it is only very limited and still lacking the full functionality and processing (Al-Khalifa, 2007).

2.2 Semantic Web Future and Requirements

A great and clear example of what the Semantic Web can offer is the “plug-and-play” technology. In the past years, installation and configuration of software or other applications are manually done in the computer. The “plug-and-play” technology lets the user to just plug any device in the computer and it will automatically be read. The same thing goes with the Semantic Web. Computer by itself will be the one to process the things needed by the user. The Semantic Web is an example of how fast technology can change. It is not just a tool for people to use but it can assist in evolution of knowledge as well.

The Semantic Web, when open for public use world wide, would be exposing new concepts and would let everyone exchange expressions. Its use of a unified logical language will enable the computer to connect the world to a universal Web. By this connections and links, humans can have access to a wide array of knowledge and ideas. In this way, all people can live together, work together, and learn together. It opens a lot of possibilities for the next generation of technology users. That is why it is very important for the Semantic Web tools and applications to prepare themselves to support all languages such as Arabic in order to fulfill the Semantic Web goal of connecting the world into one network.

The structure of the Semantic Web has three layers namely the metadata, schema, and logical layers. The tool considered for the metadata layer is the RDF and RDFS for the schema layer. But the RDFS lacks formality and one of the challenges of the Semantic Web is to reconstruct it to a formal semantics. Formal semantics can lessen the confusion and other problems attributed to the development of Semantic Web using languages and tools (Lu, 2002).

3 Proposed Tools Architecture

The proposed tool, *SemARAB* employs the Semantic Web to improve results of the search engine by understanding more about what the user is trying to find information about. *SemARAB* provides the user with easy to use interface just to determine the keywords and the type of object he is looking for, the *SemARAB* use the search engine to get the results, filter and ranks them based on the ontology to show documents referring to the chosen denotation. Figure 1 shows the different sub-modules of *SemARAB*.

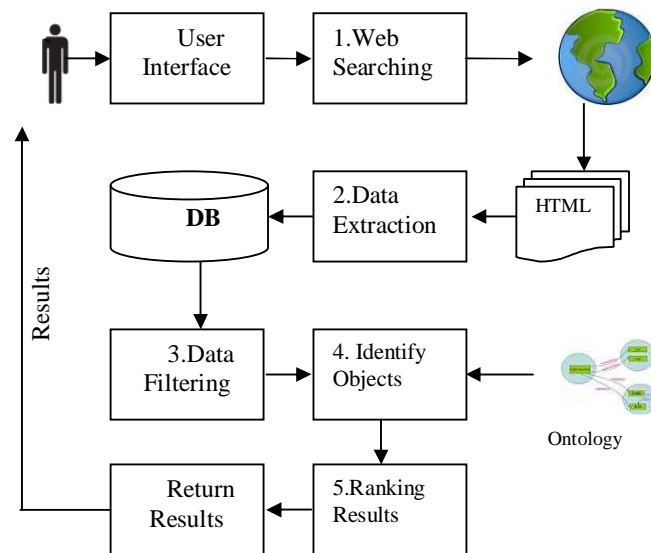


Figure 1: Modules of SemARAB

In this section we will go over these modules in some details:

1. Web Search use general search engine to find all documents URLs related to user query, based on user keyword's and the concepts ontology. For example if we need to search for "John" as organization, the query will run the search engine with keywords of John plus all related concepts to organization from the ontology.
2. Data Extraction: Extracting data from HTML documents is not a trivial way. Several research groups have focused on the problem of extracting data from HTML documents. For our approach we are extracting the content of the web data for the first 100 search results provided by the search engine.
3. Data Filtering look for keywords provided by user plus all concepts from the related ontology, if there are at least one, the document will be kept for

document ranking. Data filtering support Arabic language processing for the concepts search similarity.

4. Identify Objects: based on similarity measurement between the concepts ontology and the extracted documents, the module start to create different instances.
5. Ranking Results determine the frequency of ontology concepts which exists in the different extracted documents.

4 Graphical User Interface

The *SemARAB* GUI (see Figure 2) allows the user to search for an object in five areas (Persons, Organization, Sales, Operations, and various) based on our Arabic ontology for e-commerce.



The screenshot shows the SemARAB GUI. At the top, it says "SemARAB" in blue. Below that, "Semantic Arabic Search for E-Commerce" is written. There is a "Search for" label followed by a white text input field. To the right of the input field is the Arabic text "ابحث عن". Below the input field, there are five radio button options: "Persons (الأشخاص)", "Sales (المبيعات)", "Operations (العمليات)", "Organization (المنظمات)", and "Various (متنوعة)". The "Persons" option is selected. At the bottom of the form are two buttons: "Search" and "Reset".

Figure 2 : The *SemARAB* GUI

Form-based queries are popular methods used for websites. They allow the user to fill out a form and specify all kinds of search criteria. The form will be processed to generate the query, which will be executed to retrieve the data specified by the user. The advantage of the option boxes is that the user has to choose the type of objects he wants to retrieve. Usually these types come from the concepts ontology.

5 Expressiveness of SemARAB Tools Submodules

This section discusses the expressiveness of two important modules; the data filtering module and identifying objects module.

Data filtering module is responsible to measure Arabic concepts similarity between results returned by the search engine and the concepts from our built ontology. The proposed module is divided into the following four stages.

1. Tokenization: each result from the search engine must be tokenized and calculate the frequency words from the searched keywords or ontology so every sequence of character has a space before and after it can be used as token in Arabic language. Word is defined in Arabic language as pronoun (harf), verb (fe'l) or noun (esem). The most problem in Arabic language is the pronoun conjunct with nouns and verbs at the beginning or end of them with some changes is the structure of the word.
2. Remove stop words: stop word is a type of word which is repeated frequently in Arabic language but without any importance meaning for the similarity measurement. Figure 3 below describe the process involve in this stage:

- Convert the encoding to Unicode
- Remove punctuation (:, , /, \, ?, !, “, *), Remove diacritics (ˆ ˘ ˙ ˚ ˇ ˛ ˜ ˝ ّ َ ِ ُ ُ ُ َ ِ ُ), none letters, ال, ل, فـال, لـل
- Replace اُ, اِ, اِي with اُ, ع with ي, ة with ة

Figure 3: Process of word normalization

3. Stemming: Arabic is a Semitic language and its basic feature is that most of the words are built up from and can be analyzed down to its roots. The exceptions to this rule are common noun and particles. Morphological analysis was developed by Khoja and Garside (1999), which first peels away layer of prefix and suffixes, then checks a list of patterns and roots to determine whether the reminder could be a known root with a known pattern applied. If so, it returns the root. Otherwise, it returns the original word, unmodified. This system also removes terms that are found on a list of 168 Arabic stop words.

Figure 4 shows the Components of Data Filtering & object Modules.

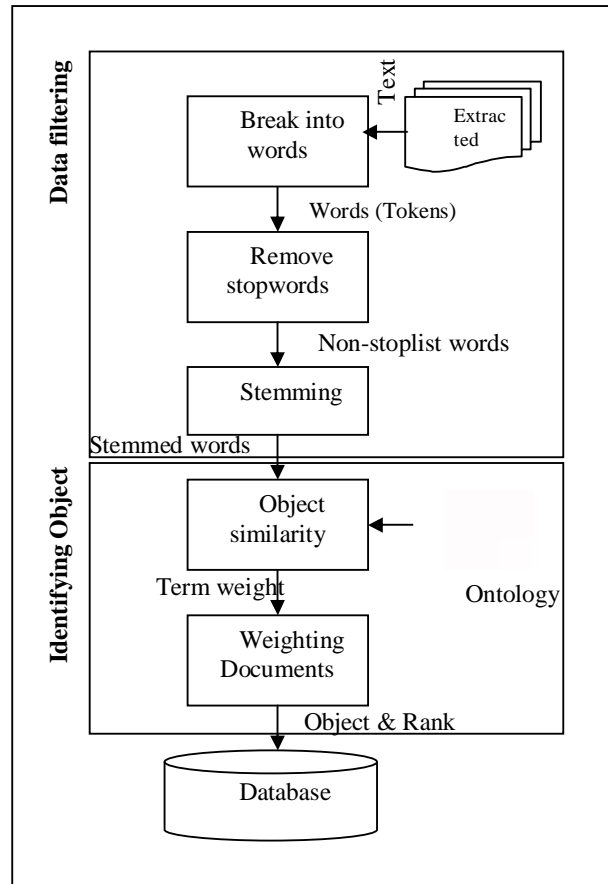


Figure 4: Components of Data Filtering & object Module

Identifying Object Module: This module will determine the related objects of our concepts ontology in the document that we had extracted before. Therefore, it will compare between the documents words and the ontology concepts for e-commerce. Then, it will store the ontology concepts which are found on the document. The

The **Identifying Object** module is divided into the following two stages:

1. **Object Similarity:** cosine similarity is used to measure similarity between the extracted document and the ontology to identify the objects.
2. **Weighting:** The weighting process is the most important process because it gives a rank reflecting the importance of the words.

6 SemARAB Testing Tool Domain & Ontology

SemARAB uses the Google search engine to provide the traditional text search results. And the following ontology for e-commerce:

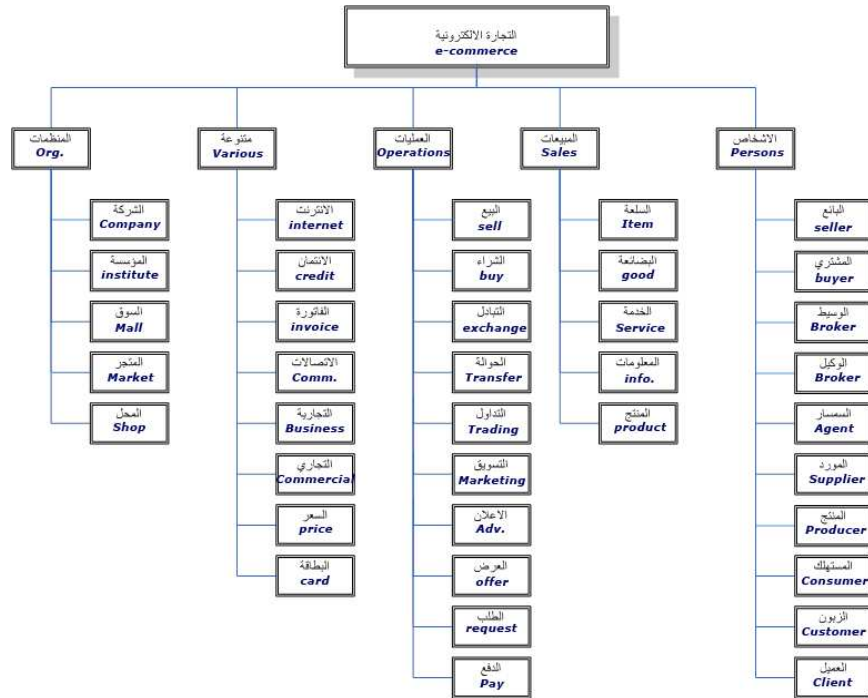


Figure 5: E-commerce Arabic ontology

Ontology is used in semantic web to decide and determine the relationships between terms, and the meaning of these terms (Maedche and Staab, 2001). Ontology can offer a good approach to represent concepts properties, and relationships between these concepts for a specific domain in a semantic way. Currently there are no available ontologies for any domain in Arabic language. We build our ontology for “e-commerce” domain in Arabic language by analyzing more than 100 e-commerce web sites such as souq.com and adabwafan.com in order to collect and determine the structure for the ontology.

7 Experimenting SemARAB Tool

In web search engines such as Google and Bing, users query the web data by entering few keywords (bag of words) in the engine's search bar and then the system computes response by matching the keywords against the index. The use of the "bag of words" in search engines cannot fully represent the users' intent and is insufficient to pose semantic queries (Diao *et al.*, 2000).

SemARAB differs from normal search engines in the sense that it uses concepts ontology as an additional layer for semantic web to enhance search relevancy and return less number of unrelated results. It provides user with an easy use of GUI to construct queries and explores the results. The GUI gives the user the ability to determine in which branch of ontology to search. Table-1 below shows the capability of *SemARAB* and some search engines to answer user queries and return the more relevant results.

Table 1: *SemARAB* vs. Web Search Engines

Parameter name	SemARAB		Google		Bing	
	Relevant	Irrelevant	Relevant	Irrelevant	Relevant	Irrelevant
Ahmad as a person	68 %	32 %	35 %	65 %	32 %	68 %
HP laptop as a sales	90 %	10 %	81 %	19 %	80 %	20 %
AIOthaim as an organization	74 %	26%	33 %	67 %	33 %	67 %
1999 SR as a various	65%	35%	37 %	63 %	29 %	71 %

Notice that the search engines like Google and Bing return too many irrelevant results and do searching for all attributes in the web documents, while *SemARAB* tool returns only the requested data that matches the user's query.

8 Conclusion

In this research we investigated the role of conceptual ontology in web search and information retrieval. We implement a semantic Arabic search tool called *SemARAB*. In its current form, *SemARAB* is well suited for a number of semantic web applications. Future improvements to the propose system came from the facts that *SemARAB* system is domain-dependent and the system in its current form is restricted to answering queries pertaining to a particular ontology. In order to operate on a different domain, user should create new ontology.

References

1. Adetunji, B., & Ahmed, A., (2008). Basic Arabic National Open University of Nigeria. Nigeria: National Open University of Nigeria. Retrieved July 16, 2009 from <http://www.scribd.com/doc/15032666/Basic-Arabic-httpal3arabiyablogspotcom>
2. Al – Khalifa, H., & Al – Wabil, A., (2007). The Arabic language and the semantic web: Challenges and Opportunities. The 1st int. symposium on computer and Arabic language.
3. Antoniou, G. & Van Harmelen, F. (2004). A Semantic Web Primer. Cambridge, United States: MIT Press
4. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The Semantic Web." Scientific American May 2001. Print.
5. Cardoso, J. (2007). The Semantic Web Vision: Where are We? Intelligent Systems In Intelligent Systems, Vol. 22, No. 5, pp. 84-88.
6. Daconta, M., L. Obrst, K. Smith. 2003. The Semantic Web: The Future of XML, Web Services, and Knowledge Management. John Wiley, Inc., June, 2003.
7. Diao Y., Lu H., Chen S., and Tian Z. (2000), toward learning based web query interface, Proceeding of the 26th international conference on very large databases, Cairo, Egypt, 317-328.
8. Hammo B., Abu-Salem H., and Lytinen S.(2002), QARAB: A Question Answering System to Support the Arabic Language, Workshop on Computational Approaches to Semitic Languages.
9. Herman, Ivan. "W3C Semantic Web FAQ." W3C Semantic Web. 27 June 2008. Web. <http://www.w3.org/RDF/FAQ>.
10. Lee, J., (2004). Introduction To Semantics Technology. IBM T.J. Watson Research Center. Retrieved July 13, 2009 from <http://www.alphaworks.ibm.com/contentnr/introsemanitics>
11. Lu, Shiyong, Ming Dong, and Farshad Fotouhi. "The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications." Information Research 7.4 (2002). Print.
12. Orr, B., (2005). The Semantic Web: It's the Latest Version of the World Wide Web... ABA Banking Journal, 97
13. Tjoa, A., Andjomshoaa, A., Shayeganfar, F., & Wagner, R., (2006). Semantic Web Challenges and New Requirements. IFS, Vienna University of Technology, Vienna, Austria
14. S.Khoja, R. Garside, (1999) Stemming Arabic Text, <http://www.comp.lancs.ac.uk/computing/users/khoja/stemmer.ps>, 1999
15. Maedche A., Staab S., "Ontology Learning for the Semantic Web," IEEE Intelligent Systems, vol. 16, no. 2, pp. 72-79

A Semantic Tool to Enhance Digital Libraries

Herbert Lee¹, Winnie Lam², Keith C.C. Chan³ and Eric Tsui⁴

¹ Quantum Cybertech Ltd

² Automated Systems Ltd

³ Department of Computing

⁴ Department of Industrial & System Engineering
The Hong Kong Polytechnic University

{herbertl}@gmail.com, {Eric.Tsui}@polyu.edu.hk

Abstract. A digital library is a place that store huge amount of information that encompasses almost every domain of knowledge. Researchers often find it takes too much of their time to locate the information that they need. Most libraries today don't provide enough semantic tools to help users for their information research. Majority of digital library search tools are built on index search which is not based on semantics. This kind of index search doesn't design for exploratory purpose and that knowledge seeking activities of a digital library are mostly exploratory in nature. In recent years, a new generation of semantic digital library (SDL) has emerged. A study of today's libraries with such semantic capabilities has been thoroughly discussed. The paper presents different views of a SDL and points out some important features that a SDL should process. Finally, the paper has introduced ANATOM, a semantic search and annotation tool that can provide ordinary digital libraries with semantic capability. ANATOM is powered by universal domain ontology that represents almost every domains of knowledge. Millions of digital artifacts in different domains can be automatically associated with concepts that are most relevant. Related information can be graphically displayed, retrieved and semantically ranked. Complex and natural language query can be supported. All these features of a digital library are the wish list of most library users. The paper has made an evaluation on this semantic tool and has made recommendation of the future development which can enhance current digital library to a new level of usage.

Keywords: Semantic digital library, ANATOM.

1 Introduction

A digital library is a portal for information, an access point to knowledge. The term "Digital Library" refers to a very broad meaning which spans from digital artifacts and metadata repositories, archives, and content management systems to very complex systems that provide digital library services to research and practice communities (Dagobert Spergel 2009). The Digital Library is not just a digitized

collection with information management tools. It is a series of activities that brings together collections, services and people in support of the life cycle of knowledge management on creation, dissemination, use and preservation of data and information.

Another purpose of the digital library is to improve teaching and assignments through the incorporation of library material. Information that is stored in these library systems increases exponentially as our knowledge has built up within time. Searching for the right information within a library system becomes a complex task. The search mechanism of a digital library can be accessed through their directory structure or via to a full-text index. These search facilities are syntactical in nature. It can't support a query such as "list all documents that are associated with the benefit of semantic digital library", especially when the term "semantic digital library" can't even be found in the documents. On the other hand, a majority of knowledge seeking tasks, especially for researchers, are exploratory in nature. In here, the searchers do not aim at object selection in the search process. Instead, the goal is to increase their knowledge in the journey rather than have a particular object to select at the end of the process. However, present search mechanism in digital library doesn't design to ease knowledge navigation. This is where the introduction of semantics to digital library can be useful. Semantic information is represented by metadata attached to each object and by one of more ontologies to provide semantic context for searches. The addition of such flexibility in search and knowledge navigation capabilities to a library system can be referred to a semantic digital library (SDL); despite there are other views of what a SDL should be.

2 Examples of a new generation of digital library

In the past few years, a few research projects have been proposed on a new generation of digital library (Bekaert et al. 2005, Lutzenkirchen 2002, Candela L. and Pagano 2007). Greenstone is a popular digital library designed to provide librarians with the ability to create and publish heterogeneous collections of digital contents on the Web like text, images, videos and e-books. Storage of XML-based documents has been proposed in Greenstone and each content item can be described using metadata compliant with the Dublin Core standard (Bainbridge et al. 2001, Witten et al. 2000).

D-Space (Tansley et al. 2003) is a digital library aimed at providing long-term preservation of heterogeneous contents with the aim to improve some of the limitations in Greenstone. Authors usually submit their documents to the system and define metadata for them and, for such reason, D-Space is referred to as an author oriented digital library. D-Space introduces a multi-roles approach to content publishing by identifying the authors and organizations that provide the contents, librarians that perform content validation and users that are interested in content retrieval. Content-based workflows can be customized in order to cope with the needs of specific organizations and to delegate different tasks to different stakeholders.

In order to provide a flexible and reusable solution to data preservation and organization, the Fedora project (Lagoze et al. 2005) explored a service-oriented approach to data interoperability in digital libraries by designing and developing a distributed architecture for contents publishing, aggregation, and retrieval. Composite

information is obtained by aggregating physical contents, viewed as bit-streams, located worldwide into the Fedora repositories. Fedora allows content editors and archivists to define semantic connections between archived contents which are treated as a set of physical contents. Other works related to content preservation in digital libraries are described in Bekaert et al. (2005) and Lutzenkirchen (2002). In particular, the aDORe project adopts the MPEG-21 DID content representation model in order to provide preservation and retrieval of heterogeneous multimedia contents.

The above mentioned systems are centered on contents, defined as binary resources enriched by metadata devoted to preservation, storage and retrieval purposes but not intended for data structuring. Preservation and evolution of a data model in those approaches are implemented as a low-level mechanism, where data are processed as bit-streams instead of instances of well-defined structures (i.e., XML Schema). On the other hand, a few more sophisticated research projects have been focused on improving the effectiveness of digital libraries in cultural heritage by moving towards a deeper semantic representation of the stored data, integrating ontologies and tools devoted to content annotation (Woroniecki et al. 2007).

CultureSampo is a platform aimed at combining and accessing heterogeneous archives of cultural heritage related contents. Each metadata schema used to represent data are mapped onto a shared ontology, the ONKI ontology, in order to provide semantic interoperability between contents. This semantic enrichment leads to new approaches to information access. CultureSampo introduces a perspective-based access to contents, where each perspective is represented by a subset of semantic features of the stored contents, such as temporal or geographical information. CultureSampo provides a set of functionalities required for content publication, annotation and retrieval. Content retrieval is exploited by means of both relation-based and semantic features-based approaches. Collaborative content generation in accordance to the Web 2.0 philosophy has been introduced into the CultureSampo infrastructure. In order to improve the amount of semantic information added to the contents, such tasks have also been partly automated by introducing domain independent annotation agents based on common thesauri and ontologies.

Interoperability of cultural heritage datasets and schemas between different platforms available on the Web has also been exploited by the AMA, Archive Mapper for Archeology project, (Eide et al. 2008) as a part of EPOCH. The tools developed during the AMA project are aimed at providing semi-automated mapping of cultural heritage custom data to the CIDOC-CRM, a formal ontology devoted to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information.

CCHO: Cantabria's Cultural Heritage Ontology (Hernandez et al. 2008) is aimed at effectively integrating cultural heritage data in the region of Cantabria. Contents have been properly annotated by using the CCHO and, as in CultureSampo, can be browsed by a semantic-based search engine according to several perspectives like geographical maps, historic event timelines and semantic relations between items.

In contrast to the previously described projects, which are based on a wide and formally defined ontology as in CIDOC-CRM, the OCHRE - Online Cultural Heritage Research Environment project, adopts an approach based on a lightweight, extendable and general ontology called the Core Ontology. This ontology covers the domain of cultural heritage by means of a small set of highly general concepts and relationships in order to grant a higher level of abstraction. The OCHRE's ontology

can be extended and refined for each different project according to the amount of specialized semantic information required to characterize a given collection.

Digital libraries specialized in cultural heritage management have been further improved by integrating social practices, like social and collaborative tagging, arising from the Web 2.0 experience. Tags and annotations can be provided either manually by the users or in a semi-automatic way. Contents can be semantically enriched in order to improve the effectiveness of both navigation and retrieval tasks.

The CHI system, designed and developed by the RHCe (Regional Historic Centre Eindhoven) (van der Sluijs and Houben 2008) is another example of the effectiveness of integration between the Web 2.0 approach and a cultural heritage devoted digital library. CHI is devoted to the storage and access of photo and video archives concerning cultural heritage. A specific set of metadata has been assigned to each of these archives. Users can search, browse and visualize the collections hosted by the RHCe by accessing them according to different dimensions, each one identified by a specific set of metadata as in CultureSampo. However metadata could refer either to a specific domain ontology (e.g., OWL time ontology used to represent the temporal dimension) or to a user defined set of keywords (tags) assigned to a specific resource by the users in a collaborative way.

Finally, annotations regarding to a specific content item could be harvested and collected from the network by looking at metadata used by different platforms and users that describe the same contents (e.g., the metadata assigned to the same painting into two different collections, hosted by different digital libraries devoted to cultural heritage). The HarvANA, Harvesting and Aggregating Networked Annotations system, (Hunter et al. 2008) uses a RDF model to represent tags/annotations and OAI-PMH to harvest the tags/annotations of a specific content item from a network of heterogeneous digital libraries (e.g., a book characterized by a specific ISBN code).

3 Views of a semantic digital library

After a thorough discussion of some examples of a new type of digital libraries, it is not hard to visualize that the trend of introducing Semantic Web technology into digital libraries is getting popular. The following are different perspectives from some scholars on semantic digital library (SDL).

Dagobert Soergel (2009) gives a very thorough view of the necessary functions and characteristics of a semantic digital library. His view of a SDL is to be supported by a Knowledge Organization System (KOS). KOS is a system for organizing knowledge into a structural form. He projects a vision of digital library that integrates access to materials with access to tools for processing materials and supports users who are individuals and communities through functions for selection, annotation, authoring and collaboration. With semantic support, digital libraries can perform complex searches for documents and automate functions for groups and individuals.

In the view of Sadeh & Walker (2003), library portals take the approach of the Semantic Web technology in their proposed model – MetaLib. MetaLib consists of a unified interface for users to interact with the system, a Universal Gateway for the system to interface with heterogeneous of library resources, and a Knowledgebase

that provides support for the two mentioned interface modules. The heart of MetaLib is the Knowledgebase which consists of a collection of metadata about different resources and the rules of accessing them.

Others view SDL should emphasize social functions in order to reflect the human side of semantics. O'Reilly promotes Web 2.0 as a platform that provides power to the users. It induces social and collaborative actions by inviting individual and community users to socialize and communicate in a free and opened environment. Social networking services (SNS) allow a user to create and maintain an online network of social friends and such services are integrated into digital library in a unified social semantic information space as proposed by John G. Breslin (2009).

Kurk et al. (2009) proposes JeromeDL as a social semantic digital library where users can bookmark interesting digital artifacts in semantically annotated directories. Users can create new knowledge by commenting the content of the blog posts, wikis and forums as well. These bookmarks, annotations and knowledge can be shared within a social network. In addition, JeromeDL provides innovative browsing, filtering and navigation solutions.

This marks a new era in digital libraries which integrate semantics and social functions into the library services. The challenges are integrating information from different metadata sources, providing interoperability with other systems and delivering more robust semantic search and social functions. The Semantic Web, as introduced by Tim Berners Lee, is an extension of the current web in which information is given well-defined meaning, enabling computers and people to work in cooperation. In other words, Web 3.0 introduces semantics to blogs, wikis, search, forums, communities and social networks, and in this case semantic digital library.

4 A tool to enhance digital libraries with semantic functions

The technology that underpins the Semantic Web is ontology. The Semantic Web technology (SWT) has been introduced by Tim Berners-Lee in Scientific American Magazine in 2001 (Bernard-Lee 2001). Ten years have been elapsed and yet SWT hasn't massively been adopted. Even Berners-Lee himself has admitted in the International Semantic Web Conference in 2009 that SWT hasn't been massively explored. Part of the issue is the inherent complexity of the concept of the Semantic Web. Even simple sets of data linked by RDF, which was one simple component of his grand vision, "is still remarkably difficult as a paradigm shift," he said. One of the main reasons why SWT hasn't prevailed is the difficulty in building ontology that powers semantic search.

Numerous articles have been published explaining the ontological engineering process is complex, difficult, time consuming and resources demanding. In order to enable SWT to prevail, the process of building and maintaining ontology has to be simple and cheap (Maedche & Staab 2001). To make the issue more complicate, the ontology that is needed to support SDL has to be very broad in scope, unless the library concerns only a very specific domain. In this case, the ontology needs to represent almost every domains of knowledge in order to power, for example, a

university digital library. Building & maintaining ontology of such size is beyond the effort and expertise of a library development project team.

An alternative approach is to choose a suitable ontology from a third party. In this paper, an ontology, ANTOM, is chosen for the task. ANTOM (Automated Ontology Manager) is a lightweight domain ontology, consisting of 7 million concepts which represents most of the general domains of knowledge. Ontology of such size is not readily available worldwide. It is an interesting project for evaluating the performance of such SDL prototype, since ANTOM is readily integrated with an advance semantic search and annotation engine. ANTOM doesn't need to replace any of the system modules in the digital library, not even the original search engine. It works in parallel with the digital library through an alternate search interface that can be invoked by the user when he chooses to perform a semantic search. After the system is properly configured, ANTOM will perform a semantic indexing of all the documents in the digital library repository. Documents which are associated with the concepts represented by ANTOM's ontology are automatically annotated with these concepts. The indexing process will take awhile depending on the number of documents to be parsed by the system. When this function has been completed, the user can invoke the semantic search function through the digital library interface. ANTOM home page will appear and the user can perform all the additional functions provided by this alternate search engine. Fig. 1. illustrates the simple integration of a general purpose digital library with ANTOM.

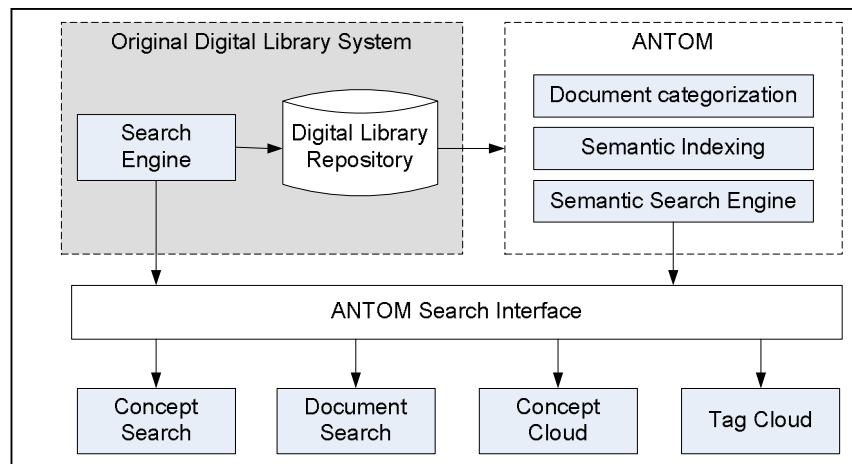


Fig. 2. Integration of ANTOM with a digital library system.

The home page for ANTOM is shown as in Fig. 2. At time of evaluating the added semantic feature of this digital library prototype, only about 3,000 documents which are related to knowledge management, semantics and e-learning are used.

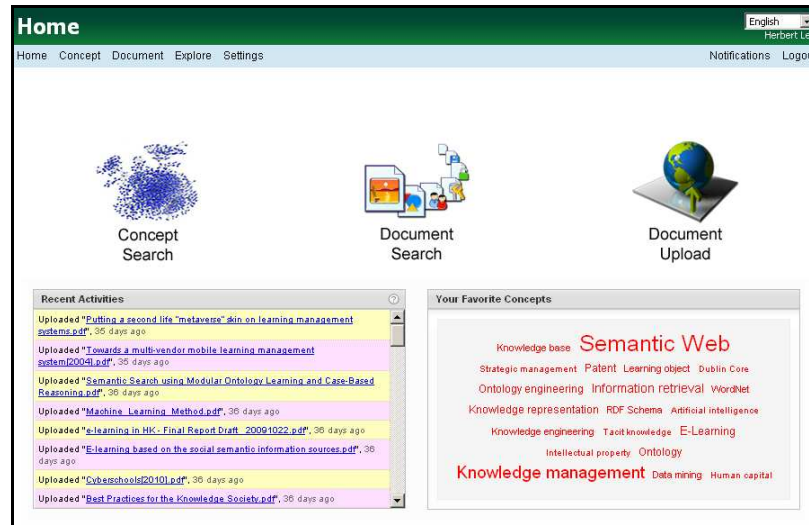


Fig. 2. ANATOM home page.

5 ANATOM: Application Features

1. Concept Search

One of the valuable feature in ANATOM is it can support either search by documents or by concepts. In research work, most library search activities are exploratory in nature, i.e. the user performs a knowledge seeking task instead of looking for a specific object. The best way for this kind of exploratory activity is to provide a map to ease knowledge navigation. ANATOM provides the user a concept map in which the user can navigate indefinitely through the nets of related concepts (Fig. 3). This feature is very versatile because ANATOM practically “knows” every domains of knowledge. The user can input almost any concept name, and the system will be able to display the related concept map.

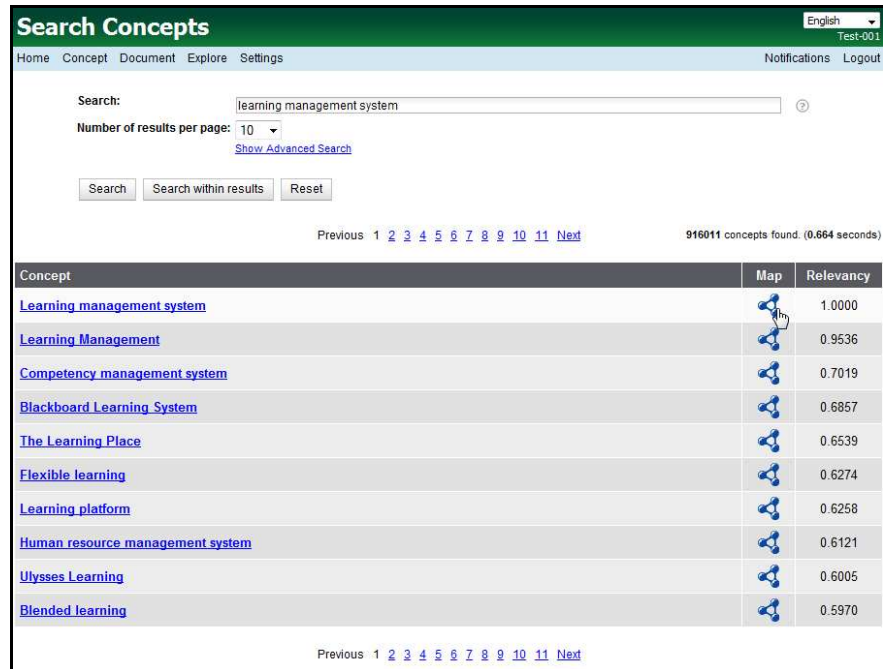



Fig. 3(a). Search for concept “Learning management system” and choose the map icon  of a concept in the list will display the concept map as shown in Fig. 3(b).

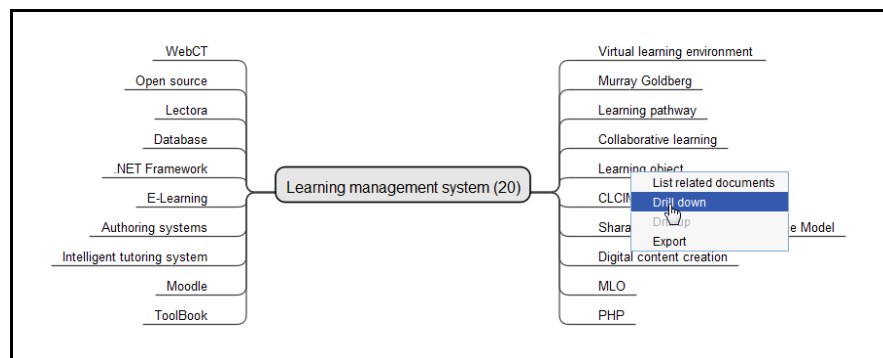


Fig. 3(b). To navigate through the concept map, click on the siblings (“Learning object”) and choose “Drill down”.

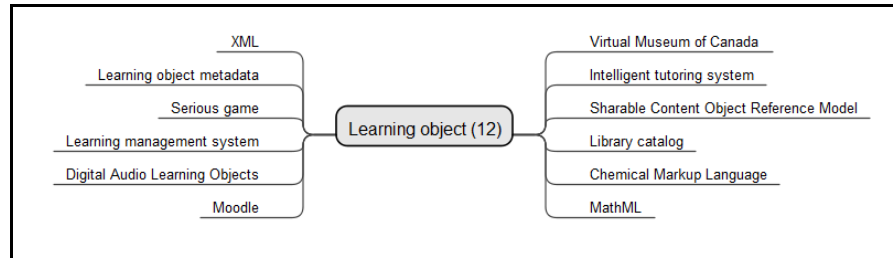


Fig. 3(c). The chosen sibling concept becomes the centre of focus. The navigation process can be indefinite.

2. Automated document categorization

When the documents in the digital library repository are under the initial process of semantic indexing, each document is automatically populated to the associated concepts. The user can retrieve documents associated with a particular concept during the navigation process (Fig. 4). Alternatively, the user can search document through the document search manual (keyword search). All the related documents listed in order of relevancy ranking are displayed.

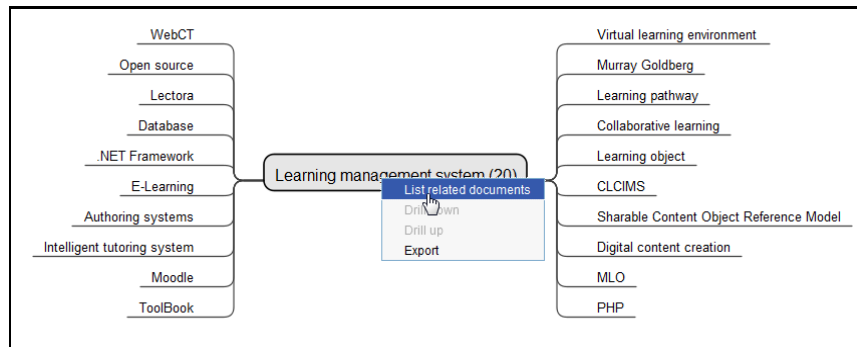


Fig. 4(a). Select a concept and choose to retrieve documents related to that concept.

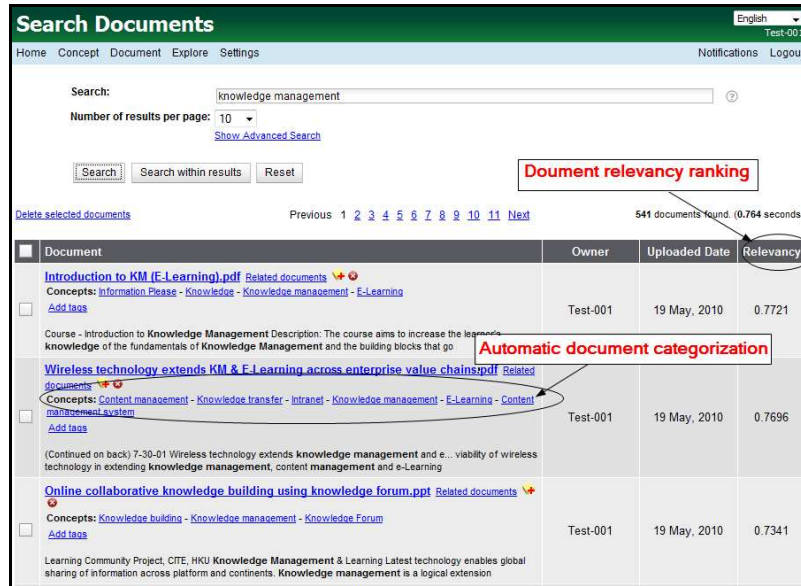


Fig. 4(b). Documents retrieved are listed in order of relevance.

Another special semantic feature is the display of concept cloud in addition to tag cloud. Concept cloud may not be familiar to most people. In here, concept cloud refers to the concepts that are relevant to a set of documents. They are obtained at time of parsing all documents in the digital library repository. The concept cloud in here represents all the relevant concepts that are related to the corpus. The displayed concepts in the home page window are those that have passed a certain relevancy threshold. The concept cloud constitutes the domains of the corpus, i.e. the 3,000 documents that are used (Fig. 5).

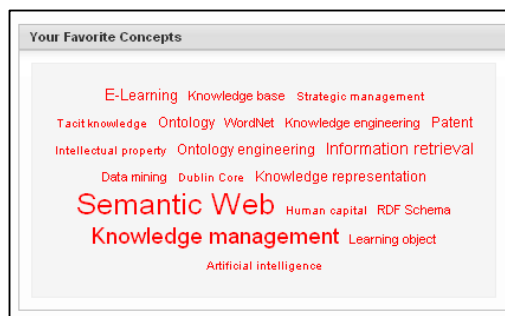


Fig. 5. Concept cloud that represents the entire domain of the documents in the repository.

3. Search of related documents for a document in context

In addition to the support of various complex queries, ANTOM can search for articles that are most related to a particular document of the user's choice. This is a very versatile feature which can save the library user considerable time from reading hundreds of related articles to find out which ones are most related to the one in context. Fig. 6(a) and Fig. 6(b) illustrate this feature.

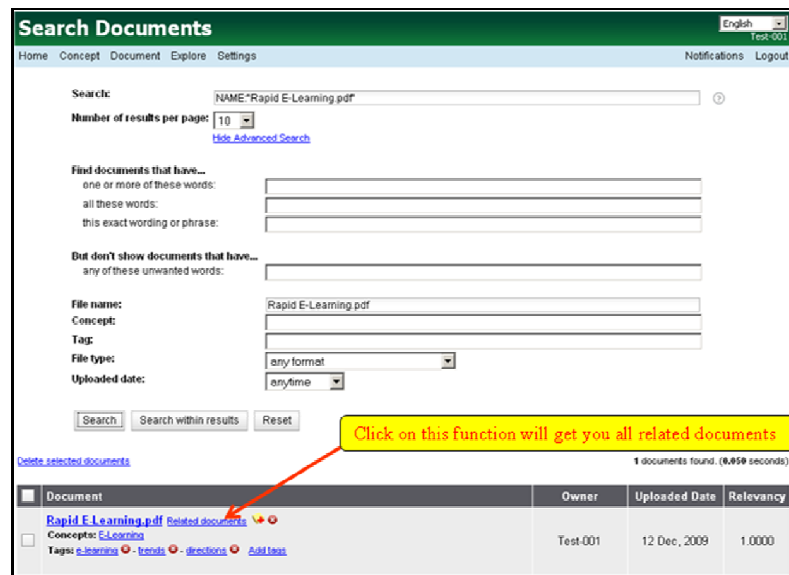


Fig. 6(a). Retrieve a document by its document name and try to find associated documents

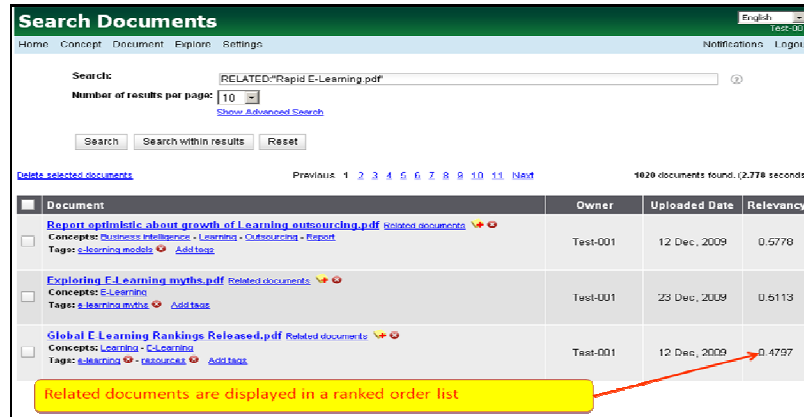


Fig. 6(b). Click on "Related documents" will display a ranked list of relevant documents

6 Conclusion

A new generation of digital library is emerging which incorporate Semantic Web technology and social functions into to the library portal. Semantic digital library is especially useful in research in which exploratory search is the primary activities in the search process. Ontology is the technology that underpins semantic search and there is a need to provide ontology with a wide range of scope in order to power SDL that represents a diversity of knowledge domains. The paper has demonstrated such ontology through a preliminary evaluation of ANTOM⁵ that integrates to a general digital library to provide semantic features. The added semantic functions are indeed helpful to digital library users. In addition to these features, there are other suggested developments that can promote a digital library to even higher level of usage. Many digital library development projects seem to neglect the personalization feature. To extend the idea even further, a new generation of digital library may include a personal learning environment for every user. In this respect, a new meaning of digital library will include semantic & social functions, personalization and e-learning. This grand vision of digital libraries would probably draw the attention of academia and developers to further explore their potential.

⁵ For those who are interested to try out ANTOM, information for getting a trail account can be located in these links: <http://www.box.net/shared/lxslui34qp> and <http://www.box.net/shared/mva4cvty8e>

References

1. Bainbridge, D., Buchanan, G., Mcpherson, J., Jones, S., Mahoui, A., and Witten, I. (2001).: Greenstone: A platform for distributed digital library applications. In *ECDL '01: European Digital Library Conference*, pp. 137–148, Berlin. Springer-Verlag.
2. Bekaert, J., Liu, X., and Van de Sompel, H. (2005).: aDORe: A modular and standards-based digital object repository at the Los Alamos National Laboratory. In *JCDL '05: Joint Conference on Digital Library*, pp. 367–367. ACM.
3. Berners-Lee, T., Hendler, J. & Lassila, O. (2001).: The Semantic Web, *Scientific American* 284(5):34.
4. Breslin, J.G.: Social Semantic Information Spaces. In *Semantic Digital Libraries*, pages 55-68, Springer-Verlag (2009).
5. Candela L., Castelli, D. and Pagano, P.: A reference architecture for digital library systems: Principles and applications. In *Digital Libraries: Research and Development, 1st International DELOS Conference*, pp. 22–35 (2007).
6. Soergel, D.: Digital Libraries and Knowledge Organization. In *Semantic Digital Libraries*, pp. 9-39, Springer-Verlag (2009).
7. Eide, O., Felicetti, A., Ore, C., D'Andrea, A., and Holmen, J.: Encoding cultural heritage information for the semantic web. In *Procedures for Data Integration through CIDOC-CRM Mapping, EPOCH Conference on Open Digital Cultural Heritage Systems*, pp. 1–7 (2008).
8. Hernandez, F., Rodrigo, L., Contreras, J., and Carbone, F.: Building a cultural heritage ontology for Cantabria. In: Annual Conference of CIDOC (2008).
9. Hunter, J., Khan, I., and Gerber, A.: Harvana: harvesting community tags to enrich collection metadata. In *JCDL '08: Proceedings of the 8th ACM/IEEE-CS joint conference on Digital libraries*, pp. 147–156, New York, NY, USA. ACM (2008).
10. Kruk, S.R., Cygan, M., Gzella, A., Woroniecki, T. and Dabrowski, M.: JeromeDL: The Social Semantic Digital Library. In *Semantic Digital Libraries*, pp. 139-150, Springer-Verlag (2009).
11. Lagoze, C., Payette, S., Shin, E., and Wilper, C.: Fedora: An architecture for complex objects and their relationships (2005).
12. Lutzenkirchen, F.: MyCoRe - ein open-source-system zum aufbau digitaler bibliotheken. *Datenbank-Spektrum*, 4:23–27 (2002).
13. Maedche, A. & Staab, S.: Ontology Learning for the Semantic Web, *IEEE Intelligent Systems* 16(2):72-79 (2001).
14. Sadeh, T. & Walker, J.: Library portals: toward the semantic Web. *New Library World* 104(1184/1185):11-19 (2003).
15. Tansley, R., Bass, M., Stuve, D., Branschofsky, M., Chudnov, D., McClellan, G., and Smith, M.: The DSpace institutional digital repository system: Current functionality. In *JCDL '03: Joint Conference on Digital Libraries*, pp. 87–97. IEEE (2003).
16. van der Sluijs, K. and Houben, G.H.: Metadata-based access to cultural heritage collections: the RHCE use case. In *PATCH'2008: Proceedings of the 2nd International Workshop on Personalized Access to Cultural Heritage, workshop at the 5th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2008)*, pp. 15–25 (2008).
17. Witten, I., McNab, R., Boddie, S., and Bainbridge, D.: Greenstone: A comprehensive open-source digital library software system. In *ICDL '00: International Conference on Digital Libraries*. ACM (2000).
18. Woroniecki, T., Gzella, A., Dobrowski, M., and Ryszard Kruk, S.: JeromeDL - A Semantic Digital Library. In *Semantic Web Challenge Co-located with The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference the 2nd Asian Semantic Web Conference*, Busan, Korea (2007).

Logical Quiz Solving Based on Ontologies

Vooi Keong Boo , Patricia Anthony

UMS-MIMOS Center of Excellence in Semantic Agents, School of Engineering and Information Technology University Malaysia Sabah, Locked Bag No. 2073, 88999 Kota Kinabalu, Sabah, Malaysia;

Abstract. One of the advantages of developing semantic technology is to be able to infer a small amount of knowledge to produce a large amount of information. To achieve this, RDF, RDFS and OWL are introduced to represent the relationship between concepts such that inference can occur between the defined knowledge to generate more knowledge. This paper elaborates on the concept of building ontology which could induce some simple logical deductions based on the existing information. The reasoning standard used is the OWL DL which covers most of the vocabulary defined under OWL standard. This paper also demonstrates the use of semantic technology in solving a complicated logical quiz called the Einstein quiz. An ontology based on the quiz will be built by coding the information provided in the quiz into a semantic standard. The completed ontology will be evaluated using Pellet inference engine of OWL DL such that more information can be generated from the existing ontology by the reasoning process. The ability of the inference engine to generate the solution shows the analytical ability of semantic technology.

Keywords: Semantic, Inference, OWL DL, knowledge representation languages, Pellet reasoning

1 Introduction

Semantic technology is a field of study that focuses on knowledge representation that is understood by both human and machine. The representation of knowledge starts with the concept of Resource Description Framework (RDF) which describes knowledge in a data model [1], followed by RDF Schema (RDFS) that connects piece of knowledge defined in RDF data model to a graph of knowledge based on vocabulary such as `rdfs:domain`, `rdfs:range`, etc. The further representation of knowledge can be referred to the standards such as DAML, OIL, DAML+OIL and OWL. [2] shows some example of the knowledge representation based on DAML standard while [3] gives the example on writing knowledge in OIL standard. Both papers also suggest that the use of RDF standard is not enough to represent human knowledge due to the limitation of the vocabulary supported. Besides, [2] also gives an overview on how both DAML and OIL could be merged together to form another standard of DAML-ONT which is further merged into a new ontology language called the DAML+OIL. DAML+OIL eventually evolved to OWL which is the current

standard web ontology language supported under W3C. As the standard evolves, more vocabularies are proposed to represent knowledge. Currently there are existing sets of vocabulary that could represent some simple knowledge and human concepts. A full set of definitions for vocabulary under OWL can be found in [4].

The advantage of using semantic technology in knowledge representation is in its ability to generate new knowledge based on existing knowledge from a process known as reasoning. Reasoning allows the information in A-box to be expanded and hidden information to be discovered. Currently, the information in the A-box can be generated without much problem since it only refers to the existing knowledge in the database. For example, Wu *et al.* [5] proposed to build ontologies from Wikipedia page with the application of WordNet to create linkage between concepts retrieved from Wikipedia. Völkel *et al.* [6] converts Wikipedia page to a semantic Wikipedia by changing some of the content in Wikipedia page to RDF standard so that it gives meaning to machine. The focus of the research is to convert existing information in WWW to semantic representation. But even with a large amount of data available, the method to process the existing information require a well defined ontology.

For the purpose of building ontology from existing knowledge source, the properties of vocabulary defined under OWL should be studied first. There are some researches that focused on converting an existing knowledge to OWL's form. For example, Doan *et al.* [7] propose a way to describe Pedagogical Resources with OWL standard while Rector *et al.* [8] also proposed the some idea on building ontology with an example of pizza.

By building an efficient ontology, it is possible for an inference engine to answer a user's request even though the information is not recorded in the database. In this paper, a few examples using inference engine to perform simple logical deductions will be presented. This paper also shows the potential for semantic technology, in particular, the knowledge representation standard to solve quiz just like a human being.

2 Simple Logical Deduction from Inference

Semantic technology includes the possibility of performing some simple natural deduction by computer. A simple example is as follow:

Let's say there exist four different characters A, B, C and D. Each character is connected to a different number from a set of 1,2,3,4. As human being, if it is given that:

A to 3, B to 1, C to 2

We could deduce that D will be connected to 4 naturally because the other 3 possibilities of number can be rejected since they have already been assigned. In semantic technology, with the combination of **owl:oneOf**, **owl:cardinality**, **owl:InverseFunctionalProperty**, **owl:differentFrom**, **rdfs:range** and **rdfs:domain**, it is possible to perform this type of deduction by building the ontology about the problem. The ontologies can be defined as follow:

```
:Char
  a owl:Class ;
  rdfs:subClassOf owl:Thing ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :has
    ] ;
  owl:oneOf (:A :B :C :D) .

:has
  a owl:ObjectProperty , owl:InverseFunctionalProperty ;
  rdfs:domain :Char ;
  rdfs:range :Num .

:Num
  a owl:Class ;
  rdfs:subClassOf owl:Thing ;
  owl:oneOf (:num1 :num2 :num3 :num4) .
```

The standard used to define ontologies in this paper is notation3, or n3 standard. For example, the n3 above will reflect the condition of the problem of “Lets say there exist four different characters: A, B, C and D. Each character is connected to a different number from a set of {1,2,3,4}”. The default namespace in this paper is blank for readability purpose. Now, to solve the problem based on the information, the hint should be defined as follow:

```
:A a :Char ;
   :has :num3 ;
   owl:differentFrom :D , :C , :B .
:B a :Char ;
   :has :num1 ;
   owl:differentFrom :D , :C , :A .
:C a :Char ;
   :has :num2 ;
   owl:differentFrom :D , :A , :B .
:D a :Char ;
   owl:differentFrom :C , :A , :B .
```

This problem can be solved with inference engine that supports OWL DL or OWL FULL reasoning. Pellet is an example of inference engine that can solve this problem. The solution is illustrated in Figure 1.

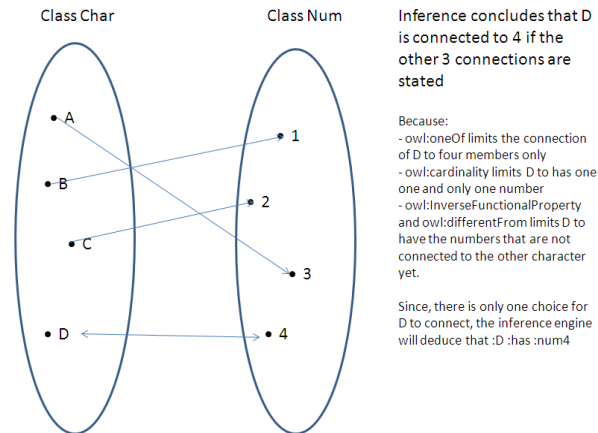


Fig. 3. An example of simple logical deduction based on inference

3 Logical Quiz Solving

Based on this simple deduction process, it can be assumed that the inference engine should be able to solve problem that is based on logical thinking. This section will show an example on how to apply semantic technology to solve a quiz based on logical thinking, called the Einstein quiz.

The Einstein quiz is defined as follow:

- F1. There are five houses in a row and in five different colours.
- F2. In each house lives a person from a different country.
- F3. Each person drinks a certain drink, plays a certain sport, and keeps a certain pet.
- F4. No two people drink the same drink, play the same sport, or keep the same pet.
- H1. The Brit lives in a red house
- H2. The Swede keeps dogs
- H3. The Dane drinks tea
- H4. The green house is on the left of the white house
- H5. The green house owner drinks coffee
- H6. The person who plays polo rears birds
- H7. The owner of the yellow house plays hockey
- H8. The man living in the house right in the centre drinks milk
- H9. The Norwegian lives in the first house
- H10. The man who plays baseball lives next to the man who keeps cats
- H11. The man who keeps horses lives next to the one who plays hockey
- H12. The man who plays billiards drinks beer
- H13. The German plays soccer
- H14. The Norwegian lives next to the blue house

H15. The man who plays baseball has a neighbour who drinks water.

Question: Who has a pet fish?

F1 to F4 are the facts of the quiz while H1 to H14 refer to the hints provided in the quiz. The challenge of the quiz is to figure out who is the person that has a pet fish. Basically to solve this quiz, deduction based on elimination (as described in Section 2) is used. However in this particular quiz, the deduction process is complicated. Thus, this problem is selected to analyze the potential of semantic technology in solving complex problem.

The key point in solving the quiz is to build an ontology that reflects all conditions, and statements of the quiz. The definition of ontologies for the quiz is similar to the problem stated before. For example, the statement **:hasColor rdfs:type owl:InverseFunctionalProperty** means that a different house should not have the same color. While the statement **:Color owl:oneOf (:Yellow :Red :Blue :Green :White :ColorX :ColorY)** means that only seven colours could exist in class Color. The Concept of house in the quiz could be represented by following statements.

```
:House
  a owl:Class ;
  rdfs:subClassOf owl:Thing ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :hasCountry
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :hasDrink
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :right
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :hasPet
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :left
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
```

```
        owl:cardinality "1"^^xsd:int ;
        owl:onProperty :hasColor
    ];
owl:equivalentClass
    [ a    owl:Restriction ;
      owl:cardinality "1"^^xsd:int ;
      owl:onProperty :Play
    ];
owl:oneOf (:House1 :House2 :House3 :House4 :House5 :HouseX :HouseY) .
```

While the concept of game could be stated as follow:

```
:Game
  a    owl:Class ;
  rdfs:subClassOf owl:Thing ;
  owl:oneOf (:Baseball :Billard :GameX :GameY :Hockey :Polo :Soccer) .
```

Finally, the relationship between concepts is connected with the object property as follow:

```
:Play
  a    owl:ObjectProperty , owl:InverseFunctionalProperty ;
  rdfs:domain :House ;
  rdfs:range :Game .
```

Facts F1 - F4 are defined by the combination of all the n3 about concept and relationship above. For all the classes defined, there will be seven instances instead of five as stated in the quiz. This is done to simplify the development of the ontologies without changing the difficulty of the quiz. In the standard quiz, there are five houses however, out of the five houses, only three have a house on both the right and left side while the other two houses do not share this property (The first house on the left hand side does not have a house on the left side and the first house on the right hand side does not have a house on the right side). To solve the problem we add two more houses in which their properties are considered known to ensure that “every house will have a house on the left and right side” and hence all instances could share the same object property.

```
:House1 :right :House2
:House2 :right :House3
:House3 :right :House4
:House4 :right :House5
:House5 :right :HouseY
:HouseY :right :HouseX
:HouseX :right :House1
```

All the instances of house will be connected in a chain and the object property of right and left can be a common property for the class House.

After all the facts are defined, the next step is to write the entire hints provided based on the semantic standard. There are two types of hint that are provided in the

Einstein quiz. The first type is an absolute hint, in which the hint describes the exact information on the houses. For example, Hint 9, "The Norwegian lives in the first house" is an absolute hint since it is a definitive statement to indicate that the person in first house is from Norway. This type of hint can be written in n3 format easily as follows:

```
:House1 :hasCountry :Norw
```

The second type of hint is a non-absolute hint since it only describes some relationship between instances. An example of this hint is Hint 5, "The green house owner drinks coffee". This hint only states that the person who is staying in green house drinks coffee but it doesn't indicate exactly which house is the green house. This type of hint does not allow any conclusion to be made about the five houses. This hint can be stated with vocabulary of owl:equivalentClass by describing how an instance of Class who lives in a green house as follows:

```
:ColorGreen
  a owl:Class ;
  rdfs:subClassOf :House ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:hasValue :Coffee ;
      owl:onProperty :hasDrink
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:hasValue :Green ;
      owl:onProperty :hasColor
    ] .
```

The example coded in n3 means that any instance which has value of green for property of hasColor will be categorized under ColorGreen class. However, at the same time, any instance that belongs to ColorGreen class should have value of coffee for the property of hasDrink. In other word, this means that the instance of house must have the value of green and coffee at the same time. In the Einstein quiz, more than one hint can be grouped into one class with owl:equivalentClass. For example, Hint 4 and Hint 5 also describe the green house. Hence, both hints can be defined under one class as follows:

```
:ColorGreen
  a owl:Class ;
  rdfs:subClassOf :House ;
  owl:equivalentClass
    [ a owl:Restriction ;
      owl:onProperty :right ;
      owl:someValuesFrom :ColorWhite
    ] ;
  owl:equivalentClass
    [ a owl:Restriction ;
```

```

    owl:hasValue :Green ;
    owl:onProperty :hasColor
  ];
  owl:equivalentClass
  [ a owl:Restriction ;
    owl:hasValue :Coffee ;
    owl:onProperty :hasDrink
  ].
    
```

For Einstein quiz, three hints are absolute, which are Hint 8, Hint 9, and Hint 14. These hints could be defined as the following n3:

```

:House1 :hasCountry :Norw    (The Norwegian lives in the first house)
:House2 :hasColor :Blue     (The Norwegian lives next to the blue house)
:House3 :hasDrink :Milk     (The man living in the house right in the centre
drinks milk)
    
```

The other hints are defined based on concept of equivalent class as described before. By providing all information to the ontology of the quiz, the inference engine will be able to solve the quiz just like human being. The solution is based on deductions and eliminations.

By combining Hint 4 and Hint 5, the result is as shown in Figure 2. The inference engine can deduce that House 4's owner lives in green house and also drinks coffee while House 5 is white color because it is the only valid solution.

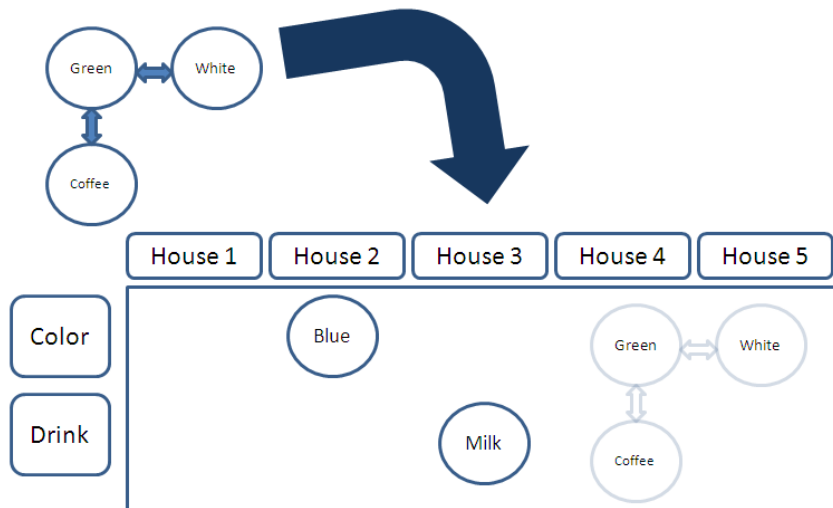


Fig. 2. Reasoning based on Hint 4 and Hint 5, it can be deduced that House 4 is green since it is the only possibility for class ColorGreen to map onto the ontology.

Further reasoning will be generated if more knowledge is provided to the ontology. The Quiz can be solved if all the hints are written semantically and parsed to the ontology. By invoking the inference engine that supports OWL DL to the completed ontology, the result as shown in Figure 3 is obtained. In the final step, the inference engine is able to generate a piece of statement that says "person in house4 has a pet fish" and hence the quiz is solved.

As shown from the result of solving a quiz with inference engine, if the vocabularies are defined properly, semantic technology has the potential to perform complex analysis and deduction similar to human being. In Figure 3, it can be seen that the reasoning process employed by human being and the one described here is exactly the same as the key to solving the quiz manually is to determine the position of the green house first, then find all the colors, the types of drink country and finally the game.

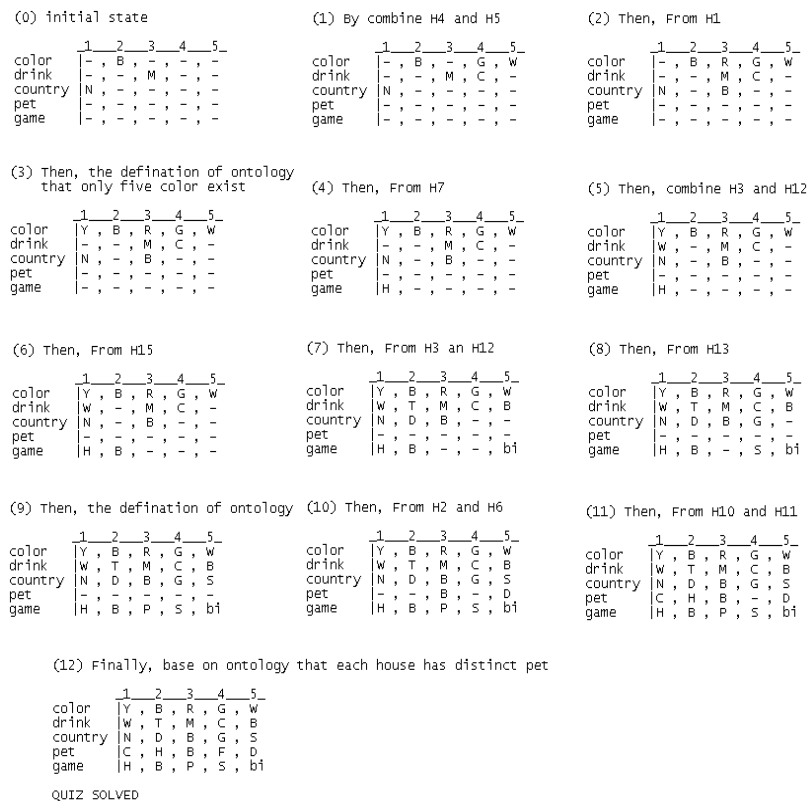


Fig. 3. The reasoning processes that occurred while solving the quiz. Some reasoning orders are not absolute. For example, Process 8 can happen as soon as Hint 12 is evaluated but before Hint 3 is evaluated by the inference engine.

4 Conclusion

The vocabulary provided in the semantic technology represents some general concepts that are understood by human being. By defining the knowledge properly, it is possible for an inference engine to perform some simple logical deductions from the existing knowledge. Without the reasoning capability, semantic technology is just another way to represent data in machine but the reasoning engine allows machine help human being to analyze and solve complex problem just like human being. However, the construction of ontology with semantic vocabulary should be very precise for the analysis to be accurate. Even a single missing statement can cause the deduction process to fail.

Acknowledgements

The first author is supported by Ministry of Higher Education under the Fundamental Research Grant Scheme (FRGS).

References

1. Grigoris Antonious, Frank van Harmelen. *A Semantic Web Primer*: The MIT Press Cambridge, England
2. Deborah L. McGuinness, Richard Fikes, Lynn Andrea Stein, James Hendler. DAML-ONT: An Ontology Language for the Semantic Web. In *SPINNING THE SEMANTIC WEB*, 1st Ed.; Dieter Fensel, James Hendler, Henry Lieberman, Wolfgang Wahlster; Publisher: The MIT Press Cambridge, Massachusetts, London, England, 2005; pp. 65-93.
3. Michel Klein, Jeen Broekstra, Dieter Fensel, Frank van Harmelen, Ian Horrocks. Ontologies and Schema Languages on the Web. In *SPINNING THE SEMANTIC WEB*, 1st Ed.; Dieter Fensel, James Hendler, Henry Lieberman, Wolfgang Wahlster; Publisher: The MIT Press Cambridge, Massachusetts, London, England, 2005; pp. 95-139.
4. Deborah L. McGuinness, Frank van Harmelen, Citing Internet sources URL, <http://www.w3.org/TR/owl-features/>
5. Fei Wu, Daniel S. Weld. Automatically refining the wikipedia infobox ontology. In *Proceeding of 17th international conference on World Wide Web*, Beijing, China, 2008, pp. 635-644
6. Max Völkel, Markus Krötzsch, Denny Vrandečić, Heiko Haller, Rudi Studer. Semantic Wikipedia. In *proceedings of the 15th international conference on World Wide Web*, Edinburgh, Scotland, 2006, pp. 585 – 594
7. Bich-Liên Doan, Yolaine Bourda, Nacera Bennacer. Using OWL to Describe Pedagogical Resources. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Joensuu, Finland, 2004. pp. 916 - 917
8. Alan Rector, Nick Drummond, Matthew Horridge, Jeremy Rogers, Holger Knublauch, Robert Stevens, HaiWang, Chris Wroe. OWL Pizzas: Practical Experience of Teaching OWL-DL: Common Errors and Common Patterns. In E. Motta et al. (Eds.): *EKAW 2004 LNCS*, vol 3257. pp. 63-81. Springer, Heidelberg (2004)

Semantic Image Annotation and Retrieval for Sport Images

Tan Yew Seng, Chan Teck Loong and Dickson Lukose

Artificial Intelligence Center, MIMOS Berhad, Technology Park Malaysia, Kuala Lumpur,
Malaysia
{tan.yseng, tl.chan, dickson.lukose}@mimos.my

Abstract. The availability of image capturing devices and the ease of sharing photos in many Web 2.0 websites have resulted in an exponential growth in the number of images available on the Internet. This calls for the need of an effective image annotation and retrieval system to work with images of broad domain. In this paper, we discuss a method and show how ontologies can be used to support image annotation and similarity-based image retrieval. The paper discusses a supervised learning method for classifying sports images into conceptual classes and subsequently using external resources like Wordnet and ConceptNet to provide controlled vocabulary for image annotation. Our results suggest that conceptually similar images can be retrieved by means of ontological classification and relations with promising results.

1 Introduction

Web 2.0 has seen the emergence of many photograph-sharing websites such as ImageShack, Flickr, Riya, Picasa, etc. It is estimated that over 20 billion images are being shared on ImageShack alone and approximately 2.5 billion photos are being added to Facebook every month [1]. The content of these images is generally broad because they are not confined to a specific domain. This poses a great challenge for online image search and retrieval.

The two popular approaches to image search and retrieval are i) Content-Based Image Retrieval (CBIR) and ii) metadata-based (normally textual descriptions) image retrieval which usually uses tags, keywords, conceptual labels, etc. [2, 3]. CBIR works by retrieving images without considering any external textual metadata. Typically, CBIR considers low-level features, such as color, texture, shape and spatial location to find results from a given set of images with similar features. However, CBIR based approach suffers from the challenges of the semantic gap [4, 5]. On the other hand, in the metadata-based approach, typically a keyword-based query is used similar to the traditional textual information retrieval systems. This approach also suffers from many of the same problems of metadata based textual information retrieval [6]. This problem is further compounded by the free-form nature of social photo tagging in popular online communities, resulting in spelling variants, synonyms and disambiguation problems [7].

In this paper, we investigate one way of using semantic web technologies to address some of the problems associated with image annotation and retrieval. We discuss the use of an ontological annotation approach that uses external ontologies to provide controlled vocabularies for tagging images. The paper is organized as follows, in section 2 we will review the related work. Then in section 3, the system is presented, subsequently in section 4 a brief description of the implementation and a walkthrough of the system is provided. In section 5, we will discuss the results. Finally in the conclusion section, the challenges faced are discussed and summarized.

2 Related Work

Ontology-based techniques [8, 9, 10] and metadata languages [3] contributed to the process of image annotation and retrieval, by providing means for defining class hierarchies. This provided well-defined semantics and a flexible data model for representing metadata descriptions. For example, RDF Schema [11, 12] can be used for defining hierarchical ontology classes and RDF for annotating images according to the ontology. An OWL reasoner can be used to infer new class memberships based on metadata associated with an image. The ontology, together with the image metadata forms an RDF graph and a knowledge base, which can facilitate new semantic image retrieval services.

Hyvonen et al [8] presented a manual ontology-based image annotation system with description logics. The major difficulty pointed out was the extra effort needed in creating the ontology, and the detailed annotations. Popescu et al [9] exploited term hierarchy extracted from Wordnet [13] to form a conceptual structure to organize images, and to control the area in the image database where similar images are searched for. More recently, Chai et al [10] have proposed an ontology-based approach for photo annotation that leverages on text and metadata analysis, with face detection. Kesorn et al [18] proposed a system for sports image annotation based on text analysis and low-level features analysis.

There has been active work in the area of image tagging and the management of tags, in particular for tag recommendations and suggestions to ease the efforts of manual image annotation. For example, Heymann et al [14] and Golub et al [15] explored the idea of tag prediction with and without suggestions from a controlled vocabulary.

In our work, we perform image classification on an image's global features using method proposed by Maree et al [16] in order to map it into an ontological class (conceptual class). In our method, it draws on external resources (Wordnet [13] and Conceptnet [17]) to provide controlled-vocabulary to allow users to tag images semi-automatically, and domain ontology to expand semantic properties of the input image. The annotations are stored in RDFS and OWL notations, to facilitate retrieval by similarity.

3 Overview of the System

The system attempts to address some of the issues of image annotation and retrieval. In this section we present the high-level architecture as shown in Fig. 1 of the system for image analysis, ontology-based annotations and image retrieval.

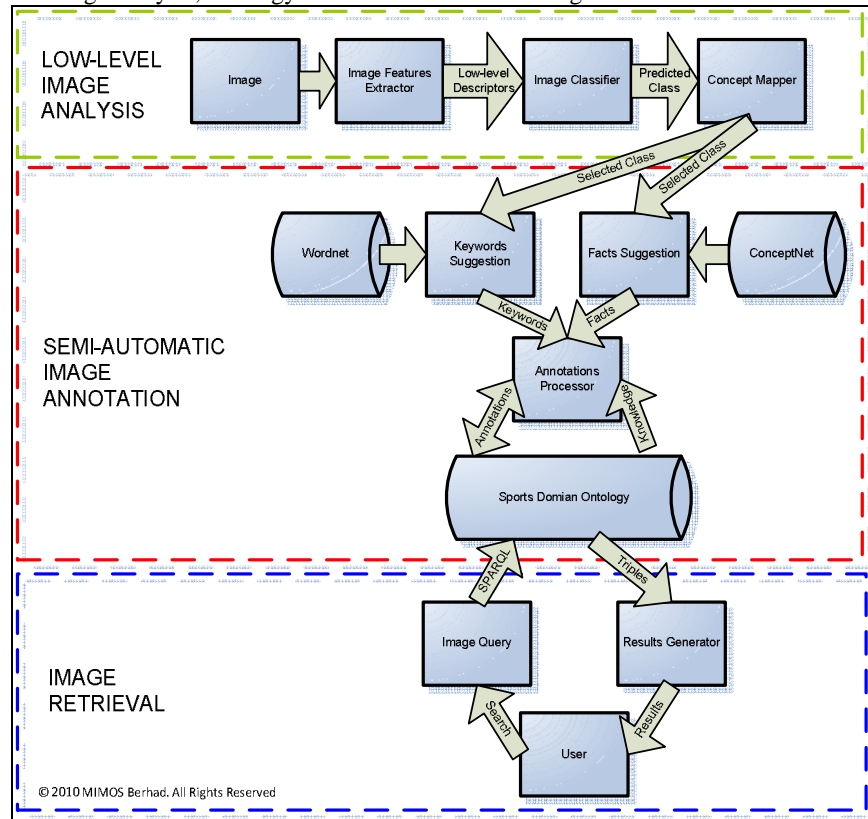


Fig. 2. High-level Architecture of the system

3.1 Low-level Image Analysis

Our approach relies on an automatic image classification component to identify the general genre of the sports input images. The underlying classification method employed is based on the work of Maree et al [16]. Due to the intention to handle images on Internet that come with diverse content variation, we needed a classifier that is robust enough to accommodate translation, rotation and skewness. Maree's randomized sub-windows technique seems to perform well under these conditions.

The result of the classification is a prediction confidence for each class and with the highest confidence normally being chosen as the genre, which in turn maps to a sport concept for further processing. This result is subsequently used in the next section.

3.2 Semi-Automatic Image Annotation

In this section we will explain the key part of the system. This part involves providing keyword suggestion and fact suggestion for annotating images and storing the annotation in the knowledge base.

3.2.1 Keyword Suggestion Using Wordnet

WordNet is a large lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept [13]. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts has been used in many semantic applications to measure and rank similarity of concepts [9].

We use Wordnet to provide keyword suggestions once the main concept has been identified for an image as discussed in Section 3.1. We retrieve the synsets based on the main concept's label from Wordnet. The returning synonyms serve as the loosely-structured controlled vocabularies to provide keyword suggestions in the annotation process. The images can then be assigned to one or more keywords and verified by a user.

3.2.2 Facts Suggestion Using Conceptnet

ConceptNet [17] is repository of common-sense knowledge, the kind of information that ordinary people know but usually leave unstated. The data in ConceptNet is being collected from Internet communities who contributed it on sites like Open Mind Common Sense [19]. It represents this data in the form of a semantic network (as shown in Fig. 2), and makes it available to be used in natural language processing and intelligent user interfaces.

ConceptNet contains some dynamic and ever-changing knowledge of the sports that we are interested in (such as Badminton, Basketball, Football, Golf and Tennis) which may be useful to annotate the images.

We use the main concept identified earlier to retrieve related common sense assertions, structured knowledge about concepts contributed by communities, from ConceptNet database [17] and these assertions, which are shown as 'facts' in the system and serves as additional annotation suggestions for the user.

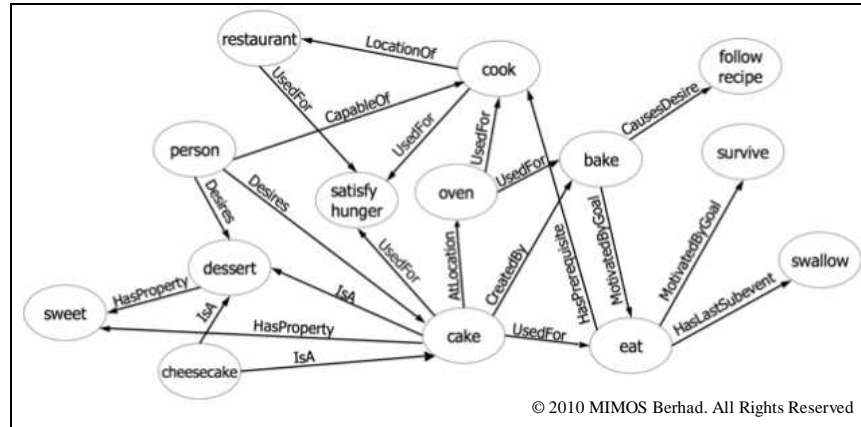


Fig. 3. Some concepts and relations in ConceptNet⁶

As shown in Fig. 3, the concept “Badminton” in ConceptNet is connected to other concepts such as the kind of equipment used and the venue in which it is normally played. There are 24 unique relation types in ConceptNet of which only seven are being considered in our work, because the other carry less useful information for image annotations. By selecting these annotations, extra relationships can be created between the images and other concepts in the ontology discussed in next section. ConceptNet has over 1 million assertions and 24 relations, the selected seven relations are shown in Table 1.

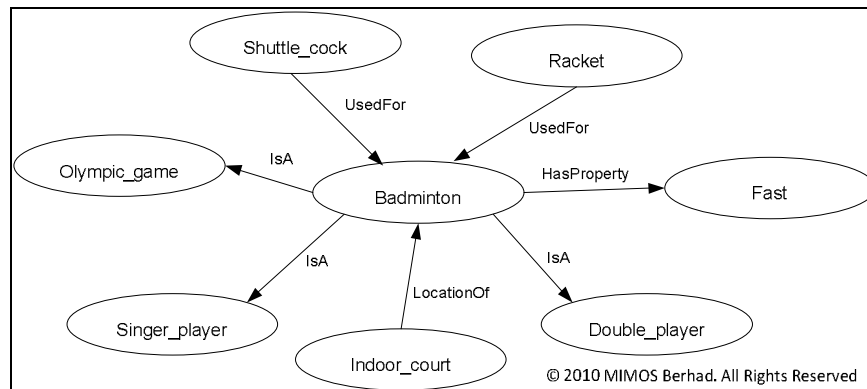


Fig. 4. ConceptNet of Badminton

⁶ Source: Common Sense Computing Initiative (<http://csc.media.mit.edu/node/49>)

Table 6. ConceptNet Relations

Relation	Description
IsA	What kind of thing is it?
HasA	What does it possess?
PartOf	What is it part of?
UsedFor	What do you use it for?
AtLocation	Where would you find it?
MadeOf	What is it made of?
HasProperty	What properties does it have?

3.3 The Domain Ontology

In general, ontology provides a formalism to define well structured concepts and their relationships. The ontology has been widely used in many areas of Artificial Intelligence (AI) and knowledge engineering [20] to develop intelligent applications and knowledge modeling. Ontology forms the integral part of any system of knowledge representation for a particular domain of interest. Ontology has been developed in respective of annotating the sport images. The sport ontology has several levels of class hierarchy as shown in Fig. 4.

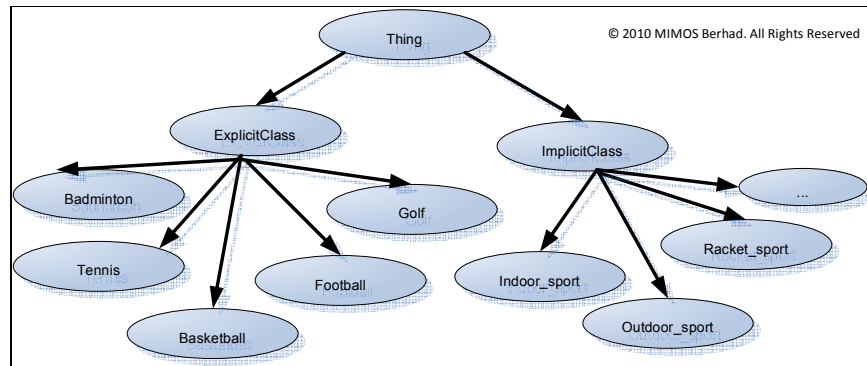


Fig. 5. Sport Ontology - Class hierarchy

The sport ontology consists of two main classes, they are known as the ExplicitClass and ImplicitClass. In the ExplicitClass, there are five sub-classes namely Badminton, Golf, Football, Basketball and Tennis which are of the interest of the system. These classes will hold the concepts that are basically mapped to the image classifier's classes as discussed in Section 3.1.

The ExplicitClass contains the main concepts that are mapped via the classification process discussed earlier, namely Badminton, Golf, Football, Basketball, Tennis. The ImplicitClass contains the ontological definition of further classification of sports based on other properties, for example RacketSport may be defined as any sports

concept which has the concept “racket” for their HasEquipment property. The implicit classes play a vital role in identifying the image similarity in the system.

After all annotations (keywords and facts) are finalized by users, the RDF triples of those annotations are constructed and stored in the triple store. The RDF triples annotation for an image is illustrated in Fig. 5 below.

The yellow triples represent the data properties such as filename, url, etc. The red triples denote the keyword selected and the green triples are generated from the facts selected for the image.

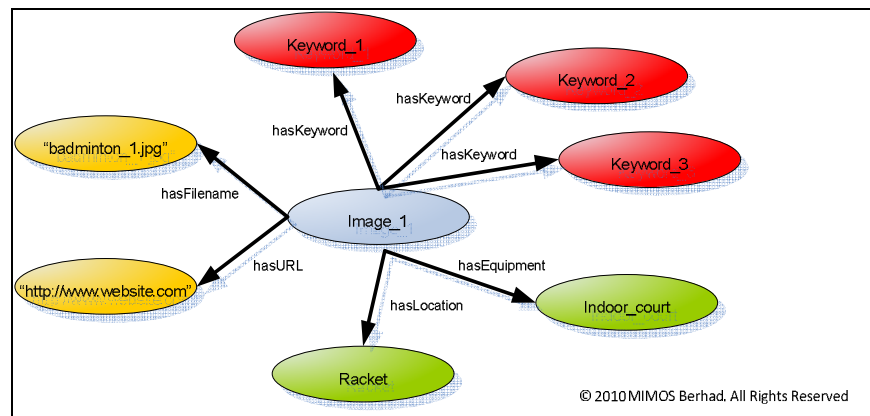


Fig. 6. Annotation triples

3.4 Ontology-based Image Retrieval

In CBIR, retrieval of similar images tend to involve measuring distances between low-level descriptors between two images, for example photos displaying similar colors profile are ranked more similar than photos having different colors. The outcome of distance measurement algorithm is without considering the semantic contents of the images.

The expectation is that the similar image should belong to the same class and having similar properties. For example in Case 1, images in Badminton class demonstrate the similarity of images in Tennis class because both sharing the common equipment concept “Racket” and therefore it is classed as “RacketSport” in the knowledge base. In Case 2, Badminton images are not similar to Golf images because they do not have any common properties. A similarity ranking approach is used to rank distance by merely considering how many common properties and class membership they have in the knowledge base of the system.

An OWL-DL compatible reasoning engine is used to infer memberships of the implicit classes during the SPARQL query processing.

4 Implementation & System Walkthrough

4.1 The System Implementation

This system is implemented on Java platform using JDK 1.6.0 Update 12. The backend components run on Apache Tomcat while the frontend is developed with Google Web Toolkit 1.7.1 (GWT). The backend components in particular the classification module is deployed as a web service using JAX-WS 2.1 XML 1.0, SOAP 1.2, and WSDL 1.1.

The image classification module is pre-trained with Sports dataset which is collected over the Internet and it is exposed as a web service component for the use of the system which is called Semantic Image Organizer. The input of the system can be either an image or a URL of an online image. The output of the system is a set of RDF triples associated to the input image's annotations and they are stored in the triples store.

4.2 System Walkthrough

The screenshot of the system is shown in Fig. 6. Images are classified using the image classifier and a class is assigned to the image by the user. The system automatically suggests relevant keywords and facts for selection by the user again. The selected keywords and facts are then used to generate RDF annotation triples for the input image.

The system is capable to support several methods of image search and retrieval, keyword search, hierarchical browsing, related search and similarity search as depicted in Fig. 7.

Keyword search – It is a search process which is performed based on textual keyword matching.

Class Hierarchy search – The user selects the explicit class of images from the ontology for browsing.

Related search – This search starts with an image in the system as the input. The system queries the knowledge base to returns all images that share one or more common keywords with the image.

Similarity search – An existing image is used as the input to the search. The system returns all images that are “semantically similar” to the input image. This means that the system returns all images from the same ontological classes of the input image, and this includes both the explicit classes and the inferred members of the implicit classes. For example, “Badminton” and “Tennis” images maybe conceptually similar because they both maybe inferred to be member of “Racket Sport” because they have the same kind of sports equipment. Likewise “Football” and “Tennis” maybe be similar because they both belong to “Outdoor Sport” class via the common “hasVenue” property value of “Outdoor”.

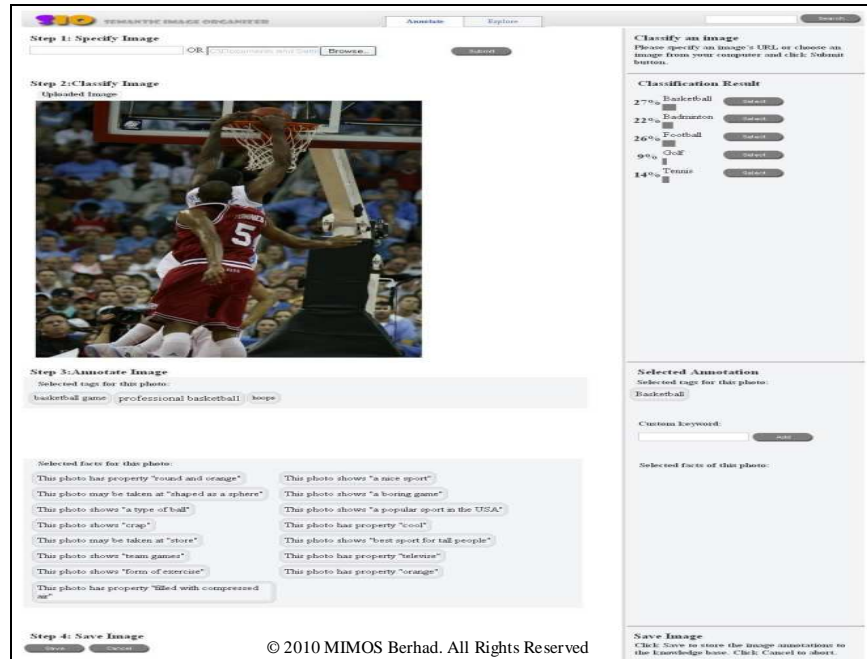


Fig. 6. Image annotation interface



Fig. 7. Search functionalities of the system

5 Results & Discussion

The image classifier was trained using 1302 manually classified images consisting five classes (Badminton, Basketball, Football, Golf and Tennis) of sport images collected from the Internet sources. The classification results show an overall mean of 85% accuracy as tabulated in Table 2.

Among the sports, the Basketball class has the lowest accuracy rate, we attribute that to the high level of noise in the training data as these data were collected and manually processed by human collectors.

Table 7. Image Classifier Performance

Class	Training Images	Accuracy (%)
Football	303	85
Golf	171	90
Basketball	245	75
Tennis	313	90
Badminton	270	85

Maree et al [16] classification method is adopted here because it can handle high dimensionality of input images and it is also robust to illumination changes, scale changes and orientation changes. Since the myriad of images from the web are dealing with high degree of such variants, this robustness is very essential for the classification accuracy.

In the semi-automatic annotation process, the users get to manually involve with the keywords and facts selection. The Wordnet and ConceptNet are useful in providing controlled vocabularies for the annotation process, but we believe the approach can be further improved by using the full richness of Wordnet's conceptual hierarchies and network of related concepts. For example, Wordnet's structure can be used to compute semantic distances between two concepts and the distance can be used to prioritise keyword suggestions.

The quality of ConceptNet's assertions need deeper evaluation, for instance there is a finding of conflicting assertions causing inconsistencies, such as Badminton is a fast_sport and at the same time, Badminton is a slow_sport. There is also duplication of concepts and this is causing ambiguity problems. It will be interesting to measure the impact of these problems in the future.

As shown in Fig. 8, the system has demonstrated that with an ontology-based method, it is possible to search for conceptually similar images by just considering the semantics annotations captured in the knowledge base as well as the class membership inferred by a reasoning engine, at the same time its physical attributes such as colors, shapes, etc. are not considered. In this case, a search for similar images of Badminton has returned Tennis images because these images are inferred to belong to the RacketSport class due to their semantic properties.



Fig. 8. Similarity based search for a Badminton image

6 Conclusions

In this paper, a low-level image analysis component is used to classify images into a certain class of sports. It then provides annotation suggestions to the user and stores the annotations in a form of semantic network which is governed by the domain ontology. It supports ontology-based retrieval of similar images.

The system has helped in the image annotation process by suggesting viable keywords from Wordnet and plausible facts about the image from the ConceptNet. It has provided a controlled environment mimicking controlled-vocabularies in many conventional annotation systems. These annotation triples coupled with the domain ontology allow further inferences to be made on the images. Thus, enriching the annotations of images by uncovering inherent semantic relationships (inferred triples) between the images. This capability is essential to the notion of semantic similarity-based retrieval in the system. The system has demonstrated that it is possible to retrieve conceptually similar images without considering the physical attributes of the images such as colors, shapes, textures, etc.

References

1. Exactly How Many Images Are Available Online?, <http://rising.blackstar.com/exactly-how-many-images-are-available-online.html>
2. Datta, R., Joshi, D., Li, J., Wang, J. Z.: Image Retrieval: Ideas, Influences, and Trends of the New Age. In: ACM Computing Surveys (CSUR), vol. 40, issue 2. ACM (2008)
3. Hyvonen, E., Harjula, P., Viljanen, K.: Representing metadata about web resources. In: E. Hyvonen (eds.) Semantic Web Kick-Off in Finland, number 2002-01 in HIIT Publications. Helsinki Institute for Information Technology (2002)
4. Harea, J. S., Lewisa, P. H., Enserb, P. G. B., Sandomb, C. J.: Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In: Chang, E.Y., Hanjalic, A., Sebe, N. (eds.) Proceedings of SPIE. vol. SPIE-6073, pp. 75–86. (2006)
5. Smuelders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-Based Image Retrieval at the End of the Early Years. In: IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 22, no. 12. (2000)
6. Bailer, W., Schallauer, P.: Metadata in the Audiovisual Media Production Process. In: Multimedia Semantics - The Role of Metadata, vol. 101/2008, pp. 65–84. Springer Berlin / Heidelberg (2008)
7. Mathes, A.: Folksonomies - Cooperative Classification and Communication Through Shared Metadata. In: Computer Mediated Communication - LIS590CMC, Graduate School of Library and Information Science, University of Illinois Urbana-Champaign (2004)
8. Hyvonen, E., Styman, A., Saarela, S.: Ontology-Based Image Retrieval. In: University of Helsinki, Department of Computer Science and Helsinki Institute for Information Technology (HIIT) (2002)
9. Popescu, A., Moellic P.A., Millet, C.: SemRetriev – An Ontology Driven Image Retrieval System. In: Proceedings of the 6th ACM international conference on Image and video retrieval, pp. 113–116. ACM USA (2007)
10. Chai, Y., Zhu, X., Zhou, S., Bian, Y., Bu, F., Li, W., Zhu, J.: Ontology-based Digital Photo Annotation using Multi-source Information. In: International Conference on Computational Intelligence for Measurement Systems and Applications, IEEE (2009)
11. Brickley, D., Guha, R. V.: RDF Vocabulary Description Language 1.0: RDF Schema. In: W3C Recommendation (2004)
12. Beckett, D. (edis): RDF/XML Syntax Specification (Revised). In: W3C Recommendation (2004)
13. Wordnet – A Lexical Database for English, <http://wordnet.princeton.edu/>
14. Heymann, P., Ramage, D., Garcia-Molina, H.: Social Tag Prediction. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, pp. 531–538. ACM USA (2008)
15. Golub, K., Moon, J., Tudhope, D., Jones, C., Matthews, B., Puzod B., Nielsen, M.L.: EnTag: Enhancing Social Tagging for Discovery. In: Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries, pp. 163–172. ACM USA (2009)
16. Maree, R., Geurts, P., Piater, J., Wehenkel, L.: Random Subwindows for Robust Image Classification. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 34–40, IEEE USA (2005)
17. Common Sense Computing Initiative, <http://csc.media.mit.edu/conceptnet>
18. Kesorn, K., Poslad, S.: Enhanced Sports Image Annotation and Retrieval Based Upon Semantic Analysis of Multimodal Cues. In: Advances in Image and Video Technology, vol. 5414/2009, pp. 817–828. Springer Berlin / Heidelberg. (2009)
19. Open Mind Common Sense, <http://openmind.media.mit.edu/>
20. Chandrasekaran, B., Josephson, J. R. V., Benjamins, R.: What Are Ontologies, and Why Do We Need Them? In: IEEE Intelligent Systems, pp. 20–26. IEEE (1999)

INFORMant™: Semantic Knowledge Base Application for 3D Visualized Digital Human

Pramod G Bagali and Susan Ong

INFOVALLEY® Group of Companies, Unit 1.1, Level 1, Block B, MINES
Waterfront Business Park, No.3, Jalan Tasik, MINES Resort City,
43300 Selangor, Malaysia
{pramod_bagali, susan}@infovalley.net.my

Abstract. Anatomy is the foundation of medicine; this knowledge supports patient physical examination, diagnosis, prognosis as well as treatments. Traditionally, anatomy teaching is mainly based on cadaver dissection; however, this dissection based anatomy has been reduced tremendously over the years due to non reproducibility and scarcity of cadavers. One of the innovative and advanced approaches of teaching and learning anatomy is through semantic knowledge base enabled 3D digital human. INFORMant™ is a novel approach to serve as a tool for teaching the intricate details of human anatomy and system physiology in future medical schools. It has applications in three (3) areas; namely 3D visualization of complex structures and anatomy; interactive and efficient access to delicate structures; and integration of knowledge bases to virtual human. INFORMant™ is designed and developed from actual MSCT DICOM data of human bodies. It captures the greatest definition of human anatomy, and is complete with easy to use navigational tools for effective learning. It is a neat interactive tool developed on a Semantic platform that allows users to explore the human anatomy and all its various systems either from the anatomy ontology to 3D anatomical volumes or vice versa.

Keywords: Semantic Knowledge Base enabled 3D Digital Human.

1 Introduction

Anatomy is the foundation of medicine; the knowledge that supports patient physical examination, diagnosis, prognosis as well as surgical treatments. For the past 30 years, there have been multiple studies which have reported the decline in undergraduates' knowledge of anatomy, based on evidences of reduction in allocated time, teaching staff and dissection in anatomy course [1]. Traditionally, anatomy teaching is mainly based on textbook with two dimensional drawing and graphical illustration (Gray's Anatomy), plastics models and cadaver dissection. However, the anatomy drawings and models tend to over simplify the complexity of the anatomy, as the depth, layering, thickness of the anatomy cannot be appreciated. On the other

hand, the dissection based anatomy has been reduced tremendously over the years. This is unavoidable as there is always the issue of scarcity of cadaver and its non reproducibility. Once the cadaver is dissected, it is damaged inadvertently, for instance, nerves and vessels are cut when the digestive system is examined; superficial muscles are often irreparably damaged when viewing the deep ones. Unusual morphologies or interesting conditions are not easily preserved for future classes due to the destruction and degradation of the specimens. Apart from that, there is also concern of insufficient dissection infrastructure, inadequate qualified teaching personnel and lack of sufficient bodies of different clinical cases in order to appreciate the anatomical variation and pathologies. As such, the traditional way of learning anatomy is insufficient to fulfill the needs of the modern medical student.

Beyond cadaver dissection, one of the innovative and advanced approaches of teaching and learning anatomy is through the 3D digital human. In fact, some of the world renowned IT giants such as IBM have envisioned that such a teaching tool will be the next disruptive technology for teaching and learning medicine in the future. These 3D digital humans are mostly based on Virtual Reality (VR) programs, stereoscopic 3D visual anatomy systems, and computer generated 3D models [2, 3]. Examples of the highly acclaimed works within this space are VOXEL MAN [4, 5], a computer program which visualizes 3D human body derived from Computer Tomography, Magnetic Resonance Tomography and Photography; and Body Voyage [6], a volume-rendering application for the visualization of large volumetric DICOM data sets.

Although these 3D applications offer incomparable dynamism and interactivity in anatomy teaching, they are still lacking the depth of the anatomical knowledge which is traditionally represented in text form. Anatomy is the scientific discipline devoted not only to the study of anatomical entities, but also to its correlation to physiological and pathological entities. As such, there is a need for a generalized, computable representation of anatomy which captures the nature of the diverse entities that make up the human structure together with the relations among these entities. One way to achieve this is through an anatomy ontology, a semantic knowledge base that is dynamic and relational in its nature⁶. The anatomy ontology can be further integrated and substantiated with the physical specification of the body i.e. 3D digital human, which outlines the morphology, topographical and geometrical relationships among the body parts in accordance to the ontology. Such combination of anatomy ontology and 3D visualisation will give a holistic and complete representation of a coherent body of knowledge on medical anatomy.

It is worthy to note that INFOVALLEY[®], a Malaysian company which is focused in Advance Medical Informatics, has designed and developed a Semantic knowledge base for 3D visualized digital human, INFORmant[™], for medical anatomy teaching and learning. The software system is comprised of medical knowledge bases built on the MIMOS Semantics Technology Platform; bi-directionally linked to 3D reconstructed volumes, MPR (multi planar reformatted) and 2D images from actual MSCT DICOM images taken from living human as well as the deceased.

2 System Overview

INFORmant™ is a novel approach to teach the intricate details of human anatomy and system physiology in future medical schools. It is an interactive tool for medical students, practitioners and teachers to explore the human anatomy in relation to various physiological systems either from the knowledge base to 3D voxel data or vice versa. It is a fresh approach in teaching intricate details of human anatomy and system physiology using advanced computer-imaging and visualization technologies to create a “Digital Cadaver” which can be repeatedly ‘dissected’ and explored without the fear of destroying the body. INFORmant™ has applications in three (3) areas:

- Visualization of complex structures and anatomy correlated to medical knowledgebase;
- More efficient access and navigation to delicate structures of a real human body; and
- Bi-directional synchronized visual integration of diverse information within a knowledge base to 3D reconstructed image with multiplanar reformatted images.

3 System Features

INFORmant™ is designed and developed to use actual MSCT DICOM images taken from human bodies that had natural death. The software renders a digital human from multislice computed tomography (MSCT) DICOM data with its high definition visualization technology and captures the greatest definition of human anatomy, in voxel, the elementary cuboids component of a digital representation of a 3D object. Since the 3D digital human is built from actual human DICOM data, the characteristics and dimensions of visualized human body and organs are more accurate than drawings in textbooks, animations, graphical illustrations, 2D photography or moulded models (see Fig.1). It is envisioned that in future versions, we will not only feature the gross anatomy DICOM images acquired from MRI and CT, but also features the physiological and histological aspects in relation to human anatomy through ultrasonography, echograms, microscopy or endoscopy for greater visual appreciation. For instance, the cellular related information such as histological 2D images (traumatic or non-traumatic, infectious or non-infectious) will further enrich the knowledge base with patho-physiological elements.



Fig. 4. Visualization of 3D rendered human body parts and organs.

The INFORMant™ anatomy ontology is a spatial structural ontology of entities and relations that form the phenotypic structure of human. It is a representation of classes and types of relationships in human anatomy which is parseable and interpretable by computing system [7]. In other words, the ontology is designed and structured to be understandable by human and navigable by computers. INFORMant™ has an extensive ontology with regards to gross human anatomy of four (4) main systems of human body, namely: osteology, respiratory, cardiovascular/circulatory and excretory (kidney urethra bladder) system. The medical ontology will be further enhanced to adapt the region based ontology, encompassing the pathology (diagnosis and disease), histology, development/embryology, pharmacology intervention and applied anatomy (with comparison studies) of entire human body systems.

The medical ontology and 3D visualized digital human are integrated and displayed on 2 separate browsers on parallel window (see. Fig. 2). The 3D Browser renders real-time 3D volumes and creates an intelligently annotated digital human. The browser is complete with easy to use navigational tools for effective learning which allow the user to interact with the visuals with various tools at the user's disposal. The digital human can be zoomed in/out, rotated in any axis, sliced, annotated with text and snapshot or video recorded for later reference.

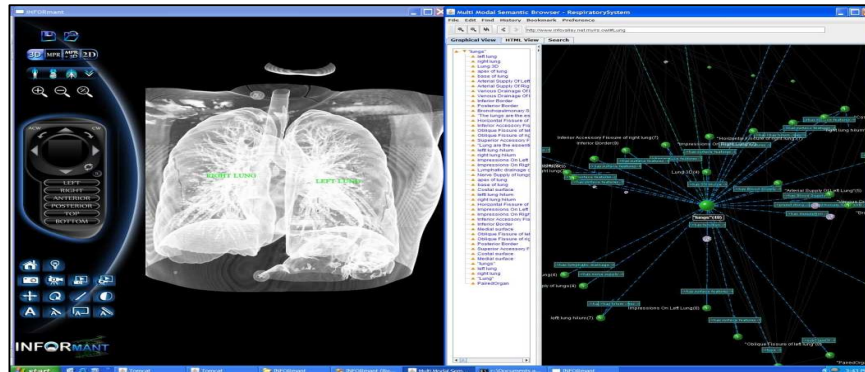


Fig. 2. Screenshot of INFORmant™ with the Semantic Browser window on the right shows the Lung concept and the 3D application on the left showing a rendered image of the Lung along with annotations.

The 3D browser has an in-built list of highly customized pathology specific presets that are based on differences in the densities (Hounsfield Unit) of the different anatomical parts (see Fig. 3). The preset enables the visualization of different layers of the human body at a mouse click, from the skin down to the bones, viscera and cavity. The 3D voxel data provide the medical students with the spatial depth impression that is important in understanding the anatomy, resembling the cadaver dissection [8]. On top of that, the system can also accommodate expanding 3D volume and MultiPlanar Reformatted images of cadaver with age, sex and pathological variation; which enriches and strengthens the medical knowledge bases.



Fig. 3. Different presets of 3D browser which allows visualization of different layers of the digital human.

The Semantic Browser displays the anatomy ontology in a graphical and interactive way which enables the user to explore and navigate easily through the knowledge base. The ontology can be displayed in 3 main views; namely:

- Graphical - concepts and relations are shown as nodes and links between nodes.
- HTML - concepts are represented as a dynamically generated HTML page with relations and target concept listed.
- Type Hierarchy - the medical knowledge base is shown as a tree. The current selected concept is shown along with its child concepts (if any) and a path through its ancestors to the root of the tree.

Advanced semantic querying will also be made possible whereby the user will be able to perform search/query within the knowledge base for effective learning. Hence, the relational medical or physiological information in relation to anatomy can be inferred or related for greater representation of knowledge and ease of understanding.

4 Synchronized Navigation

The 3D browser and the Semantic Browser are integrated and synchronized through socket communication, it allows user to explore the human anatomy and its various systems either from the medical ontology to images or vice versa. In this process, similar TAGS are used for the tagging of knowledge base and 3D voxel data in

INFORMant™ repository. When a user manipulates the ontology via the Semantic Browser, the TAGS of the central concept on display in the Semantic Browser will be transmitted to the Graphic Engine, which will in turn use the TAGS to fetch the corresponding image from 3D repository and display it on the 3D Browser (see Fig. 4). The same goes for the reverse approach. The bi-directional approach of associating domain knowledge to a specific anatomical structure of interest facilitates deeper understanding and relational observation in connecting anatomical structures to medical knowledge to appreciate significance, functionalities, abnormalities and relevance.

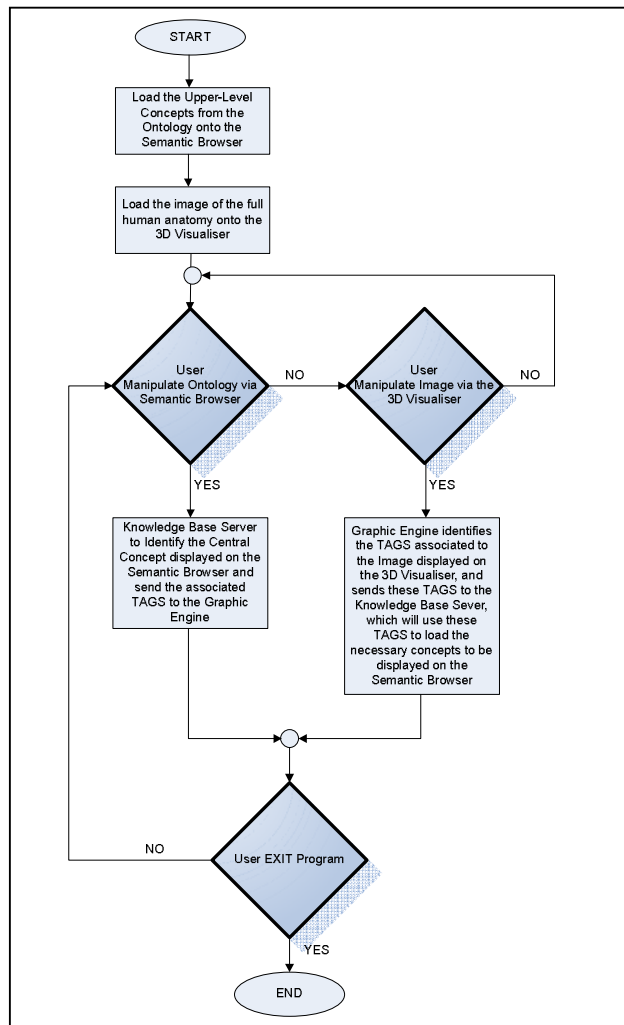


Fig. 4. The TAGGING system in INFORMant™ socket communication.

5 Conclusion

Advancements in 3D visualization have bridged the gap between the classroom learning and real clinical practice [9,10]. The fact that students can now study anatomy at their leisure at their own choice, on their own computer screens outside of the anatomy laboratory, has revolutionized the way of medical anatomy teaching and learning. There is no doubt a semantic knowledge base for digital Human will bring about a major change in the way medical anatomy is taught in medical schools. It is an integrated and holistic approach of anatomy learning as the application offers physical, structural and functional clarity comparable to the textbook. The availability of an application like this will make the students and teachers more inquisitive and proactive, which in turn will lead to a paradigm shift in medical teaching from “passive” to “active” learning and teaching. In addition, it also solves the issues of cadaver scarcity, non reproducibility of the cadaver dissection, as well as shortage of qualified educators. [9]

References

1. Turney, B.W.: Anatomy in a Modern Medical Curriculum. *Ann R Coll Surg Engl.* 89, 2 (2007) pp. 104—107
2. Ribas, G., R. Bento, and A. Rodrigues. Anaglyphic three-dimensional stereoscopic printing: Revival of an old method for anatomical and surgical teaching and reporting. *Journal of Neurosurgery* 95 (2001) 6
3. Frisby, A.G. Part I: Advances in educational technology: IVD, CD-I and journeys into Virtual Reality. *Journal of Allied Health* 22 (1993) 131–38
4. Höhne, K. H., Petersik, A., Pflesser, B., Pommert, A., Priesmeyer, K., Riemer, M., Schiemann, T., Schubert, R., Tiede, U., Urban, M., Frederking, H., Lowndes, M., Morris, J.: *VOXEL-MAN 3D-Navigator: Brain and Skull. Regional, Functional, and Radiological Anatomy.* 2nd edn. Heidelberg: Springer-Verlag Electronic Media (2003)
5. Höhne, K. H., Pflesser, B., Pommert, A., Riemer, M., Schiemann, T., Schubert, R., Tiede, U. A new representation of knowledge concerning human anatomy and function. *Nat. Med.* 1 (1995) pp. 506—511
6. Tsiaras, A.: *Body Voyage: A Three-Dimensional Tour of a Real Human Body.* Warner Books, New York (1997)
7. Rosse, C., Mejino Jr, J. L. V.: *The Foundational Model of Anatomy Ontology, Anatomy Ontologies for Bioinformatics: Principles and Practice,* Springer, New York
8. Perry, J. Kuehn, D. and Langlois, R.: Teaching Anatomy and Physiology Using Computer-Based, Stereoscopic Images. *Journal of College Sciences Teaching.* 36 (2007) pp. 18—23
9. Henn, J., M. Lemole, M. Ferreira, F. Gonzlez, M. Schornak, M. Preul, and R. Spetzler. Interactive stereoscopic virtual reality: A new tool for neurosurgical education. *Journal of Neurosurgery* 96 (2002) 1
10. Shaffer K. Teaching anatomy in the digital world. *N Engl J Med.* 351 (2004) pp. 1279—1282.

An Integrated Health Ontology System

Yip Chi Kiong¹, Sellappan Palaniappan², Nor Adnan Yahaya³

Department of Information Technology,
Malaysia University of Science and Technology,
Kelana Square, Kelana Jaya, 47301 Petaling Jaya, Selangor, Malaysia
¹yipzqiang@yahoo.com, ²sell@must.edu.my, ³noradnan@must.edu.my

Abstract. Most healthcare institutions such as hospitals and clinics store their data in the form of databases of various formats. The Health Ontology System that we have developed provides a means to integrate these data with concepts and semantics in the form of a shared cumulative ontology for enabling machines to interpret them. This involves two major software tools, namely, Ontology Generator and Ontology Distiller. The Ontology Generator is used to create ontology from a selected database using metadata provided by its database management system. In a reverse process, the Ontology Distiller enables a subset of data from an ontology to be distilled into a database for further analysis. This Integrated Health Ontology System will pave the way for integrating existing data with ontologies that will be useful for developing semantic agents for healthcare domain.

Keywords:

Ontology encoding and generation, database schema, ontology viewing, ontology information extraction and integration, extracting ontology into database, knowledge sharing.

1 Introduction

Most institutions in the healthcare industry in this country store their data in various forms of database management systems. While database management systems are efficient in the storage and retrieval of data, they are mostly proprietary and locked into applications that access them. Some of these data may be confidential; especially the health records of certain patients who wish to remain private. However, some of the information should best be shared within the healthcare community. Ontologies provide a standard method for sharing this information, with the added advantage of including semantics and relationships that enable software agents to interpret the information stored in these repositories. The ability to create and evolve an ontology is vital towards developing a system for sharing healthcare knowledge.

An ontology can be viewed as a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and relationships between them. It is typically represented as a knowledge base which then becomes available to applications that need to use and/or share the knowledge of a domain.

Within health informatics, an ontology is a formal description of a health-related domain [1]. In recent years, ontologies have been adopted in many business and scientific communities as a way to share, reuse and process domain knowledge [2]. Ontologies play a major role in the development of the Semantic Web [6]. In the context of our research, an ontology is a shared conceptualization that describes the terminology used in a particular domain, which in this case is the healthcare domain. This ontology is expressed using the Web Ontology Language or OWL in short [3]. OWL is based upon the Resource Description Framework or RDF in short [4] [5].

The Integrated Health Ontology System consists of a set of tools designed to manage the creation, evolution, merging of ontologies and extraction of their subsets for various applications. The first of these tools is the Ontology Generator [11], which generates an ontology using metadata from a database management system. The generator is used as the first stage in building a healthcare ontology data store. This ontology can then be integrated into a cumulative ontology. The cumulative ontology contains concepts integrated from various kinds of ontologies. There is little or no tools currently available in the market for data mining on ontologies directly. Hence, we have also developed an Ontology Distiller [12], which can extract a subset of an ontology and produce a database which can be the source data for existing data mining tools. This process is the reverse of the Ontology Generator. Other tools which will be described later are still in the process of development.

2 The Overall Design

Central to our Health Ontology System is the Cumulative Ontology and the Target Database. The Cumulative Ontology integrates some ontologies that are relevant to the domain, which in our current research, covers patient records, doctors, and diseases. The Target Database stores the data which is referenced by the Cumulative Ontology. The initial setup of the system involves the use of the Ontology Generator, which generates an ontology from an initial database that supplies the primary source of information. The instances from the record data are stored in the Target Database. There are few or no tools available currently to perform data mining on the Cumulative Ontology. Hence, some data may be extracted from the Cumulative Ontology to form a subset database from which it may be possible to perform data-mining using existing tools. The whole design is shown in Figure 1.

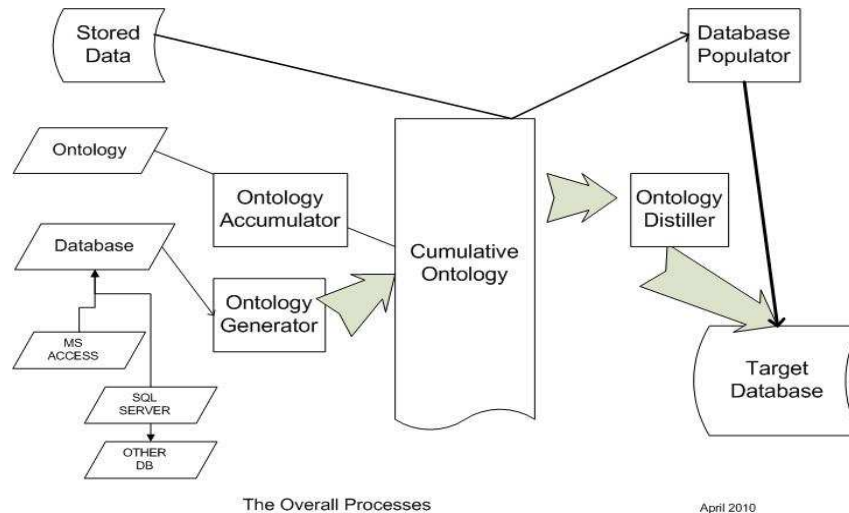


Figure 1: The Overall Design of the Health Ontology System.

The Ontology Accumulator, which is currently in the design stage, is used to integrate additional information from other ontologies.

3 The Ontology Generator

The Ontology Generator is a tool designed to create an ontology from a selected database. The design is based upon the DataMaster Plugin for Protégé 3.4 [7]. Protégé is written in Java. However, we use a different algorithm that uses C# as the programming language, and the Microsoft Visual Studio as the programming platform. This enables us to use the newer features in C# and rapid prototyping in Visual Studio. We have named it as Health Ontology Generator (HOG) for reference.

3.1 Algorithm to Extract Schema from Database

The first step involves selecting the type of database. We started with Microsoft Access and SQL Server. The connection string is created to connect to the database. For Microsoft Access databases, we use the Microsoft.Jet.OLEDB.4.0 provider, for SQL Server, we use the OLEDB provider. Currently the system supports only Microsoft Access and SQL Server. Support for MySQL will be added later.

HOG uses the schemaTable method to query the tables in a database and returns the result as a DataSet. After the table names are obtained, it extracts the column names and their data types and puts them into another DataSet. Finally, if the user chooses, the row data is extracted into a third DataSet.

3.2 Method for Encoding Ontology

The ontology generation is an automatic process which encodes and stores the ontology physically as an RDF file that includes declarations of classes, properties and instances. In addition, the ontology also includes the semantics that describe the meaning of the data included in it. Typically, the file is given the file extension owl. The first part of the encoding process of ontology is the generation of the header. The body of the ontology includes the classes, the properties and the instances. The final part is the trailer. Figure 2 shows these stages diagrammatically.

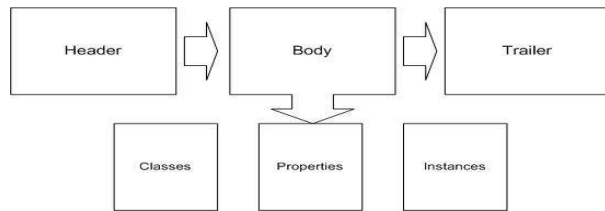


Figure 2: Stages in encoding the ontology.

3.3 Encoding the Header

The header specifies the RDF start tag (with namespace attributes) and the ontology element. It starts with the version information of the XML encoding. This is followed by some standard namespaces, which includes XML schema (for data types), RDF, RDFS and OWL. Each of these standard namespaces is declared using their usual URIs. For example, XMLS is declared as

```
xmlns:xsd="http://www.w3.org/2001/XMLSchema#" . The
ontology's own namespace is declared as xmlns:
db="http://zhiq.tripod.com/db_table_classes?DSNtype=Acces
s:dbHealth_1#", which is a reference to the database to link to the ontology.
Finally, the ontology element is declared simply as <owl:Ontology rdf:about=""/>.
```

3.4 Encoding the Body

In the ontology, tables are converted to classes. This is done by constructing the RDF statement as an OWL class. Field names (or column names) in the table are converted to functional attributes. In addition, other functional attributes are added, which describe the semantics of the ontology. The rows of each table are converted into instances, beginning the first instance in the form of instance_1. The annotation properties, such as #hasForeignKeys are then added.

3.5 Encoding the Trailer

The trailer consists of the closing RDF tag and information about the creator of the ontology.

```
</rdf:RDF><!--creator -->
```

3.6 The User Interface of HOG

The User Interface is shown in Figure 3. The user first selects Data Source Type, whether it is MS Access or SQL Server database. Clicking the Connect button will display the Tables, Fields and Records. The user then selects the destination filename to Output the owl file. Clicking the Generate button will generate the owl file in the selected location. The generated owl file is compatible with Protégé 3.4.

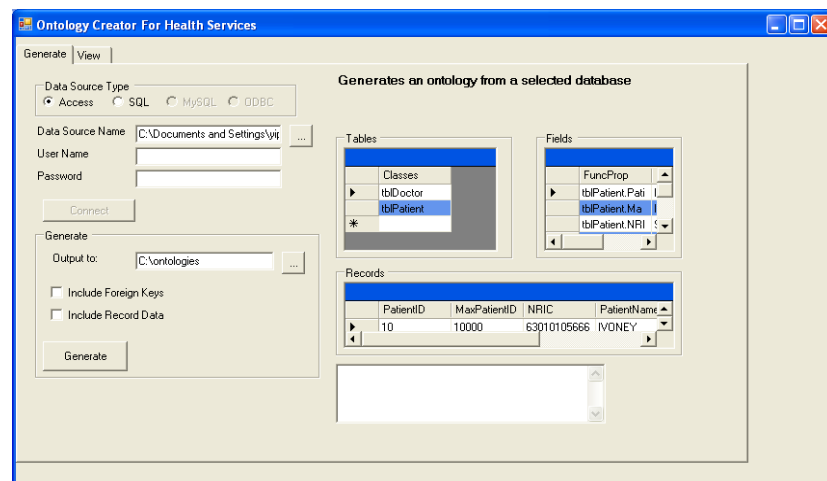


Figure 3. The User Interface of HOG

4 The Ontology Distiller

The Ontology Distiller is the reverse process of the Ontology Generator. However, it uses a totally different algorithm. It extracts concepts from an ontology owl file and creates a database. The output database can be a MS Access or a SQL Server database. The current programming requires the database to be created first by MS Access or SQL Server. However, we are working on a direct creation of the database. This database can be operated on by existing data mining tools. We have not been able to access any tool from the Web which can perform a similar task, as we were able to do so with Protégé when we built the Health Ontology Generator.

There is very little work published on the process of creating databases from ontologies. After we have developed the Ontology Distiller, we have found a US

patent and a paper describing Knowledge Bus [9], which can create databases from an ontology using the Java platform. This paper uses application program interfaces (APIs) to reflect the entity types and relations (classes and methods) that are represented by the database.

4.1 The Distiller Algorithm

The algorithm used in programming the Distiller involves the following steps:

1. Count the number of classes in the ontology file
2. Get Class properties
3. Build the DataSets
4. Get the Instances
5. Store the data in the database.

Counting the number of classes is necessary to declare the amount of storage space to be allocated for the array of datasets. This is the first pass of the entire encoded owl file before storing any data. The second pass involves allocating the actual classes and their properties and building the datasets. The third pass would read in the instances, their data types and their values. The resulting data would be stored. It may be displayed if necessary.

The process involves some rules that form the basis of the programming. Classes in the ontology are stored as tables in the database. Similarly, functional properties are stored as attributes for each table. Instances from the ontology form the records in the database.

4.2 The Distiller User Interface

The user interface is shown in Figure 4. First, the user searches for the location of the Ontology Source then clicks the Extract button. The Distiller will extract the classes and the class properties. Then the user can select the type of database to be output. Next is the name of the database.

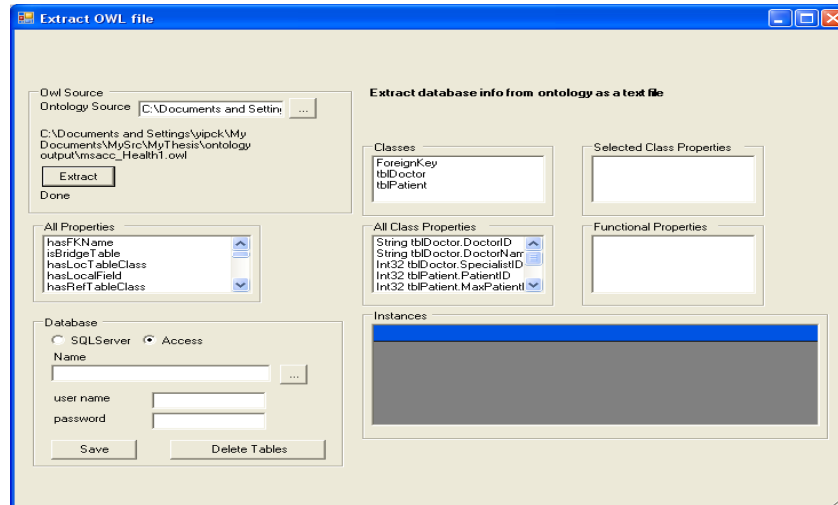


Figure 4. User interface of the Ontology Distiller

In the case of MS Access, it is necessary to first create the Access database using Access, and then search for the database. In the case of SQL Server, it is only necessary to name the database. Clicking the Save button will create the tables from the classes, together with the corresponding attributes and records.

5 Conclusion

We have described a Health Ontology System consisting of an integrated set of tools for creating and processing ontologies. The combination of the Health Ontology Generator and the Ontology Distiller that works on the Cumulative Ontology offers the following side benefit. You can start with a MS Access database, use the Ontology Generator to create an ontology, then use the Ontology Distiller to convert that database into an SQL Server database. These are two of the tools of the Integrated Health Ontology System that we have created so far. The next step is the design and programming of the Ontology Accumulator, which will integrate similar ontologies into the Cumulative Ontology.

References

1. Open Clinical Knowledge for medical care Web Site <http://www.openclinical.org/ontologies.html>
2. Protege Overview, <http://protege.stanford.edu/overview/>
3. McGuinness, D. L. and Harmelen, F. V. (Eds), OWL Web Ontology Language Overview, W3C Recommendation, 10 February 2004. Available at :<http://www.w3.org/TR/owl-features/>

4. Manola, F. and Miller, E. (Eds), RDF Primer, W3C Recommendation, 10 February 2004. Available at : <http://www.w3.org/TR/rdf-primer/>
5. Brickley, D. and Guha, R. V. (Eds), RDF Vocabulary Description Language 1.0 : RDF Schema, W3C Recommendation, 10 February 2004. Available at <http://www.w3.org/TR/rdf-schema/>
6. Lee, T.B., Hendler, J., and Lasilla, O, "The Semantic Web," *Scientific American*, May 2001.
7. Nyulas, C., O'Connor, M., Samson Tu, *DataMaster – a Plug-in for Importing Schemas and Data from Relational Databases into Protégé*, Stanford University School of Medicine, Stanford, CA 94305
8. Wikipaedia, Choosing between versions of Protégé
9. Andersen, William A., Brinkley, Paul M., Engel, Joshua F., Peterson, Brian J, (2003) *Ontology for database design and application development*, United States Patent 6640231
10. Horridge, M., Knublauch,H., Rector,A., Stevens,R., Wroe, C. (2004) , *Protégé OWL Tutorial*, The University Of Manchester, Stanford University, United Kingdom, pp.11-14.
11. Yip Chi Kiong, Sellappan Palaniappan, Nor Adnan Yahaya, (2009), Health Ontology Generator: Design And Implementation, Malaysia University of Science and Technology. IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.2, February 2009, pp. 104-112.
12. Yip Chi Kiong, Sellappan Palaniappan, Nor Adnan Yahaya, (2009), Ontology Distiller: Extracting Databases From Health Ontologies, Malaysia University of Science and Technology. IJISCE International Journal of Information Sciences and Computer Engineering, to be published.

Using the Spiral Process Model to Develop a Medical Knowledge Base

Saniah Mohamed, Supiah Mustafa, Dickson Lukose

Knowledge Engineering Center, MIMOS Berhad
Technology Park Malaysia, Kuala Lumpur, Malaysia 57000
{saniah | supiah | dickson.lukose}@mimos.my

Abstract. The power of ontology lies in its ability to explicitly describe semantic data in a common way, independently of data source characteristics, and providing a schema that allows data interchange among heterogeneous information systems and users. Ontology Engineering is seen as a challenge in the application of Semantic Technology for problem solving and application development. This paper presents our experience in using the Spiral Process Model to develop medical ontology, which was intended to implement a Medical Advisor and a Medical Diagnosis System to be used by medical interns. It is discovered that the use of a spiral process resulted in an adequate ontology, and also helped to reduce mistakes by making it possible to identify and solve problems faced in each “cycle” of the process. Verification from domain experts was required in each stage, in order to ensure that the knowledge encapsulated in the ontology was presented in the correct way.

Keywords: Semantic Technology, Ontology Engineering, Spiral Process Model, Medical Advisor, Medical Diagnosis

1 Introduction

Medical Knowledge base (MKB) plays an important role in Patient Diagnosis and Management System [1]. It also fosters sharing and reuse of knowledge, facilitates collaboration between medical experts and medical interns, and enables nurses in rural areas to better manage difficult cases. In this paper, we report our experience in using the Spiral Process Model to develop the medical knowledge base for cardiovascular, paediatric and occupational health.

A knowledge base (KB) is a machine-readable resource for the dissemination of information [2]. Engineering ontology for the domain is the first stage of developing the KB. They are a means for people to specify the meanings of terms used in knowledge that they might generate, share, or consume [3]. Ontology defines a common vocabulary for researchers and practitioners who need to share information in a domain [4]. The description of concepts, as well as of their interrelationships and their classification into taxonomy, gives rise to a framework in which information resources (text, images, video, etc.) associated with the domain can be organized.

Ontology together with a set of individual instances of classes constitutes a knowledge base [4].

There are many ontology engineering tools available in the market. Among them include DogmaModeller [5], KAON [6], OntoClean [7], HOZO [8], Protégé [9] and TopBraid Composer [10]. Most of these tools claim to enable no IT experts to model ontologies, but as yet, there are no standardized methodologies for building ontologies [11]. Such a methodology would have to include a set of stages to be executed when building ontologies, a set of guidelines and principles for each of the stages, and an ontology life-cycle that would describe the relationships among the stages [11]. The best-known ontology construction guidelines were developed by Gruber [3] to encourage the development of more re-usable ontologies. Noy and McGuinness [4] also provided a good guideline on developing ontology. Recently, increased efforts have been devoted to the task of trying to develop a comprehensive ontology-building methodology [3]. Some methodologies referred to established approaches for building ontology include Methontology [12] and DILIGENT [13]. The most widely used of these alternatives, Methontology, adopts some ideas from Software Engineering in order to improve its applicability. These activities include Ontology Management, Ontology Development and Ontology Support. DILIGENT on the other hand is an ontology engineering methodology for distributed, loosely controlled and evolving engineering of shared ontologies.

As yet, none of these methodologies enable us to realize a comprehensive enterprise level knowledge base development that will address consistency and completeness of the knowledge base with respect to the knowledge of the domain expert. In view of these challenges, we proceeded by adopting an existing Software Engineering methodology to knowledge engineering. This paper will outline what forms of adoptions was carried out, and our experience in utilizing it to engineer the MKB. The remaining sections of this paper will describe the following: Section 2 will outline reasons why we were unable to utilize the existing Unified Medical Language System (UMLS) Metathesaurus [14]. Section 3 will outline the Knowledge Engineering Methodology based on the Spiral Process Model and demonstrate our experience in applying it. Section 4 will discuss the pros and cons of this methodology, and finally Section 5 will conclude this paper.

2 Non-Suitability of Existing Generic Medical Ontology

In the first instance of developing the Medical Ontology, we attempted to take advantage of the existing UMLS Metathesaurus [14] knowledge sources as a source of medical terminology. The Metathesaurus is a very large, multi-purpose, and multi-lingual vocabulary database that contains information about biomedical and health-related concepts and the relationships among them. It is built from the electronic versions of many different thesauri, classifications, code sets, and lists of standard terms used in a range of biomedical contexts including patient care, health services billing, public health statistics, indexing and cataloguing biomedical literature, and basic, clinical, and health services research. [11]

The Metathesaurus was created to facilitate the development of computer systems related to biomedicine and health, but is not optimized for any particular application. It can be applied in systems that perform a range of functions involving one or more types of knowledge. The Metathesaurus is organized around concepts. In essence, its purpose is to link alternative names and views of the same concept together and to identify useful relationships among different concepts.

Unfortunately, the UMLS scope was too broad and the size was too large to support our purpose. In addition, other information such as patient management, which is essential for our application in Medical Diagnosis and Medical Advisor, was not specific enough for our purposes. Furthermore, many of the terms used in the UMLS differ from normal usage in Malaysia because the UMLS was created for the United States National Library of Medicine while Malaysia usually follows UK standards of terminology.

Medical experts⁷ suggested that it would be better to focus on a limited body of information supplied directly by them, rather than trying to manipulate the entire contents of a large database such as the UMLS. In light of this advice, we decided to develop our MKB from scratch, and to restrict its scope on the three main areas of cardiovascular, paediatric and occupational health. The MKB was based on the domain experts' own knowledge of these topics.

Based on the above challenges, the decision was made to build the medical ontology from ground up by modelling the knowledge of local medical experts. To this end, we had to utilise a suitable and easy to implement knowledge engineering methodology to successfully build the overall MKB. We adopted the Spiral Process Model, typically used in Software Engineering, by customizing it for our purpose of Knowledge Engineering. The following Section will outline how we applied the Spiral Process Model for engineering the MKB.

3 Knowledge Engineering Methodology

3.1 Adaptation of the Spiral Process Model

The knowledge engineering methodology that we utilized is based on the Spiral Process Model. The Spiral Process Model as depicted in Figure 1, proceeds by repeated iterations of a cycle that passes through four quadrants. The first quadrant identifies purpose and scope of knowledge base development. The second quadrant of the cycle consists of actually acquiring the needed information from the domain experts. In the third quadrant, the information is conceptualized through knowledge modeling and encoding. The fourth quadrant involves validation of expanded knowledge based and planning for the next iteration of the cycle. Each cycle of the spiral model iterates through these four quadrants. Each cycle also builds on the previous one by expanding more detailed knowledge. The number of cycle that will be required to complete the development of a knowledge base is very much dependent

⁷ The medical experts involved in this exercise include Prof Dr Abdul Rashid Abdul Rahman, Dr Nor Mahani Harun and Dr Azrul Rozaiman from Medical Interest Group Sdn Bhd.

on the complexity of the domain, the ability of the domain experts to articulate their domain expertise, and the ability of the knowledge engineer to model and encode the knowledge base. This knowledge engineering methodology was adopted to develop the MKB. The MKB focused specifically on the areas of cardiovascular, pediatric and occupational health.

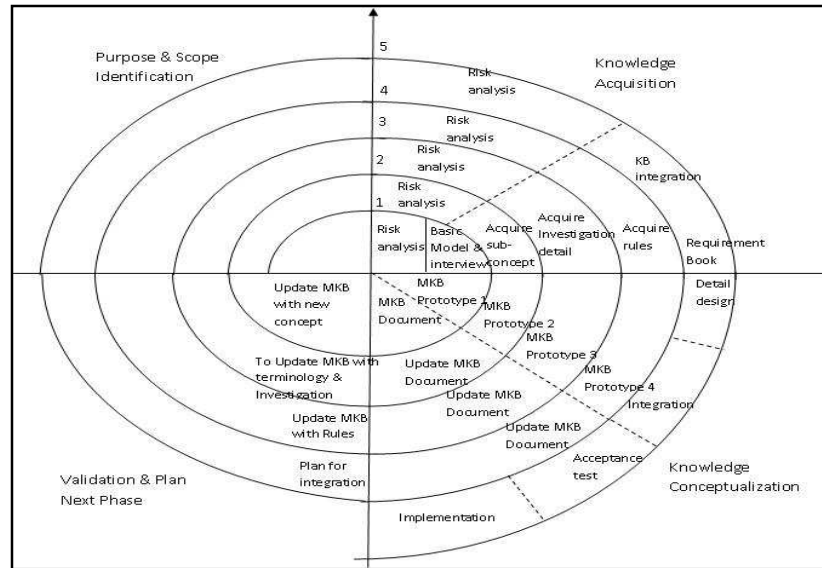


Fig. 5. Spiral Process Model of Medical Knowledge Base Development
 ©2008-2010 MIMOS Berhad. All Rights Reserved.

3.2 Challenges in Knowledge Engineering

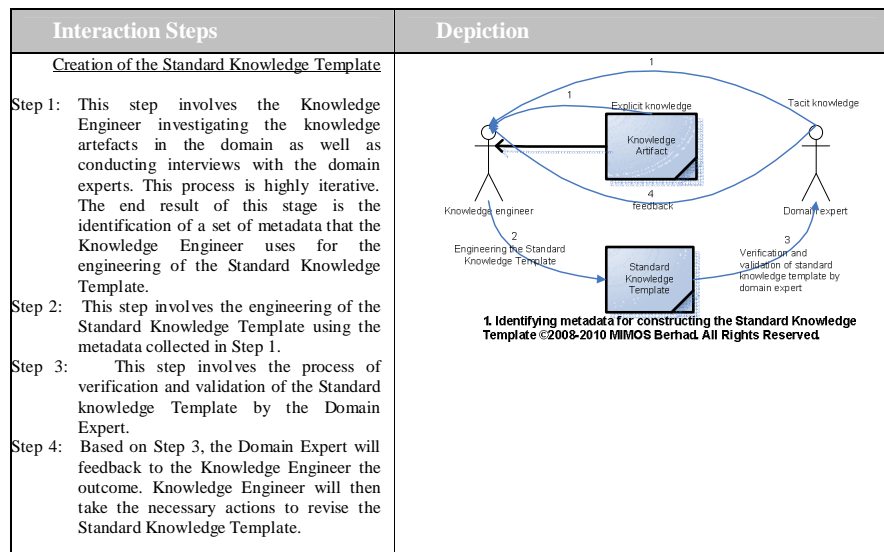
To successfully apply the knowledge engineering methodology described in Section 3, we had to overcome a number of challenges. We identified 5 inter-related challenges. A description of each of these challenges is listed in Table 1, while Table 2 depicts and lists the interactions between the knowledge engineers and the domain experts.

Table 1. Knowledge Engineering Challenges

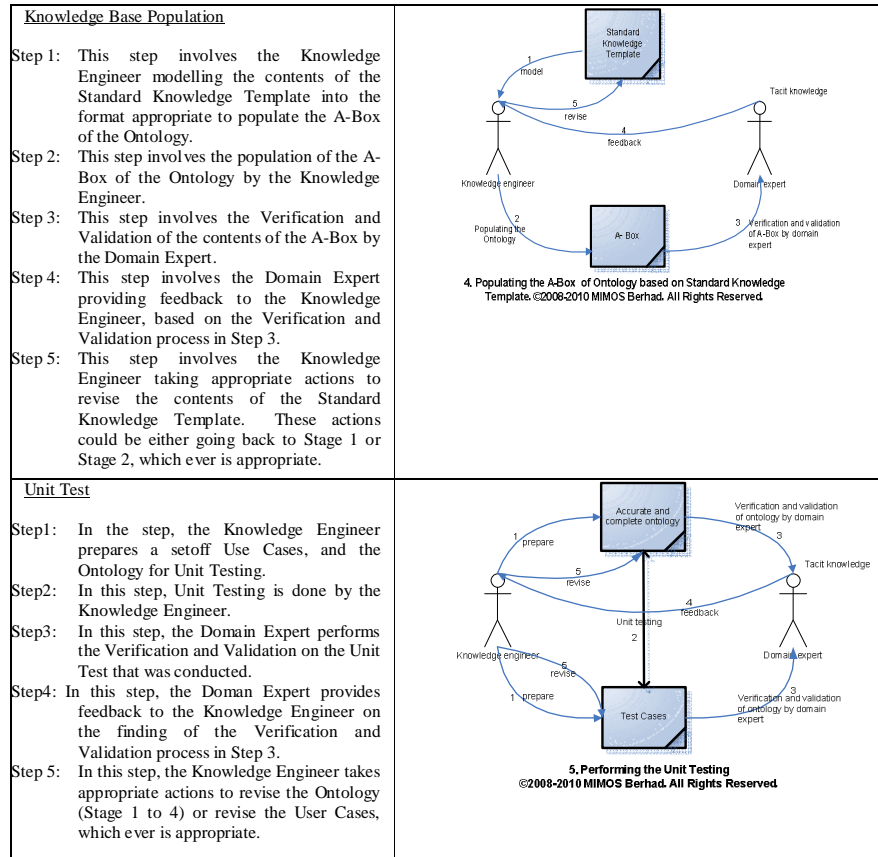
Challenges	Description
Creation of the Standard Knowledge	This is where the Knowledge Engineer interacts (via interview and electronic communication) with the Domain

Template	Expert to elicit the necessary meta-knowledge to put together the Standard Knowledge Template. This is an iterative process of elicitation, engineering, and verification and validation with the domain expert.
Elicitation and documentation of domain specific explicit and tacit knowledge	This is where the Knowledge Engineer documents (using the Standard Knowledge Template) the explicit knowledge from the domain as well as the tacit knowledge from the domain expert. This is an iterative process of elicitation, engineering, and verification and validation with the domain expert.
Ontology Engineering	This is where the Knowledge Engineer utilizes the documented knowledge from the Standard Knowledge Template and engineer the T-Box of the Ontology. This is also an iterative process of transformation, engineering, and verification and validation with the domain expert.
Knowledge Base Population	This is where the Knowledge Engineer utilizes the documented knowledge from the Standard Knowledge Template to populate the A-Box of the Ontology. This is also an iterative process of transformation, engineering, and verification and validation with the domain expert.
Unit Test	This is where the Knowledge Engineer will create a series of test cases and utilizes them to validate the ontology. The same test case will be used against the domain expert. If both provide consistent result then one could conclude the ontology to be sufficiently complete, otherwise, one could expect revision of the ontology based on steps (functionalities) outlined above.

Table 2. Interactions between the knowledge engineers and the domain experts



<p style="text-align: center;"><u>Elicitation and documentation of domain specific explicit and tacit knowledge</u></p> <p>Step 1: This step involves the Knowledge Engineer modelling and eliciting the necessary knowledge from the knowledge artefacts and the domain expert, respectively.</p> <p>Step 2: This step involves the Knowledge Engineer populating the Standard Knowledge Template with the elicited domain knowledge from Step 1.</p> <p>Step 3: This step involves the Domain Expert conducting the Verification and Validation on the knowledge that was populated into the Standard Knowledge Template.</p> <p>Step 4: This step involves the Domain Expert providing feedback to the Knowledge Engineer, based on the Verification and Validation process in Step 3.</p>	<p style="text-align: center;">2. Extracting the Tacit and Explicit Knowledge into Standard Knowledge Template. ©2008-2010 MIMOS Berhad. All Rights Reserved.</p>
<p style="text-align: center;"><u>Ontology Engineering</u></p> <p>Step 1: This step involves the Knowledge Engineer modelling the contents of Standard Knowledge Template into appropriate format to engineer the T-Box of the Ontology.</p> <p>Step 2: This step involves the engineering of the T-Box of the Ontology by the Knowledge Engineer.</p> <p>Step 3: This step involves the Domain Expert conducting the Verification and Validation process on the T-Box of the ontology that was create in Step 2.</p> <p>Step 4: This step involves the Domain Expert providing feedback to the Knowledge Engineer based on the Verification and Validation process in Step 3.</p> <p>Step 5: This step involves the Knowledge Engineer taking appropriate actions to revise the contents of the Standard Knowledge Template. These actions could be either going back to Stage 1 or Stage 2, which ever is appropriate.</p>	<p style="text-align: center;">3. Engineering the T-Box of Ontology based on Standard Knowledge Template. ©2008-2010 MIMOS Berhad. All Rights Reserved.</p>



4 Case Study

4.1 Spiral Cycle 1

During the first meeting with the domain experts, the KE explains the purpose of the MKB, its medical applications, and the objectives of the knowledge collection phase to the subject matter experts. This is to ensure that the domain experts associated to this project will have a clear picture of the scope and objectives of the MKB and Patient Diagnosis and Management System, despite being unfamiliar with the planned ontology itself. Because of this unfamiliarity, however, the elicitation rate in the initial stage was slow. The task associated to enabling the four quadrant of the spiral process model Spiral Cycle 1 listed in Table 3.

Table 3. Summary of the Tasks Involved in Spiral Cycle 1

Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
Purpose and scope identification	Knowledge acquisition	Knowledge conceptualization and formalization	Validation and planning of next phase
KE explains the purpose of Patient Diagnosis and Management System, and specific objectives of knowledge collection phase, to domain experts.	Risk analysis: Domain expert not familiar with KE's expectations. KE prepares the knowledge template (KT) of MKB based on literature study of medical domain and target application. Subject matter experts identify types of knowledge elements that KE is expected to collect.	KE identifies main concepts Prepare medical ontology T-Box (data structure incorporating concept and their interrelationships) Prepare documentation on ontology.	Internal review of ontology. Medical experts reviewed documentation of ontology. They identify knowledge refinement and gap.

In execution of this task, the Knowledge Engineer prepares a Knowledge Template (KT) shown in Table 4 and populated it. This template is then reviewed and refined by domain expert. In addition they will also identify knowledge gap. This the point of which the next cycle of the Spiral Cycle 2 process model will begin.

Table 4. Knowledge Template

Unstable Angina

Symptoms	Signs	Investigations
<ul style="list-style-type: none"> • Chest pain: squeezing, central, heavy. Crushing ; Commonly substernal; Epigastrium; Radiation to neck, left shoulder, left arm • New onset of chest pain • Chest pain at rest • Deterioration of pre-existing angina • Atypical presentation: Indigestion; Pleuritic chest pain; Dyspnoea 	<ul style="list-style-type: none"> • May be unremarkable • Diaphoresis (excessive sweating) • Pale cool skin • Sinus tachycardia (HR >100bpm) • 3rd and/or 4th heart sound • Basilar rales • Hypotension (BP < 100./60mmH) 	<ul style="list-style-type: none"> • ECG <ul style="list-style-type: none"> ◦ ST-segment depression ◦ Transient ST-segment elevation(in 30-50% of patients) ◦ and/or T-wave inversion • Cardiac biomarkers <ul style="list-style-type: none"> ◦ Elevated Troponin ◦ Elevated CK-MB

4.2 Spiral Cycle 2

Based on the opportunities identified cycle 1, the task list for Spiral Cycle 2 was established in each of the quadrants as shown in Table 5. The main activity in the second cycle is to refine and fill in the knowledge gaps.

Table 5. Summary of the Tasks Involved in Spiral Cycle 2

Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
Purpose and scope identification	Knowledge acquisition	Knowledge conceptualization/ formalization	Validation and plan next phase
Identified major concepts and sub-concepts for ontology. Acquire the related detailed knowledge and update the SKT model.	Risk Analysis: Availability of domain expert to meet schedule. Extra man-hours need to fulfill the gap. Knowledge related to four concepts and their sub concepts acquired.	Update T-Box according to new knowledge acquired, and populate A-Box. Update ontology documentation accordingly.	Internal review. Medical expert reviewed the SKT. Medical experts reviewed documentation of ontology. Identified gap on Terminology.

The Knowledge Template is updated based on the new knowledge acquired in this cycle, and the refined Knowledge Template is listed in Table 6. It is self evidence in this cycle; the Knowledge Engineers have identified additional information on disease management. In this case, the information is on how to manage Unstable Angina.

Table 6. Refined Knowledge Template based on Spiral Cycle 2

Unstable Angina

Symptoms	Signs	Investigations	Management
<ul style="list-style-type: none"> • Chest pain: squeezing, central, heavy. Crushing ; Commonly substernal; Epigastrium; Radiation to neck, left shoulder, left arm • New onset of chest pain • Chest pain at rest • Deterioration of pre-existing angina • Atypical presentation: Indigestion; Pleuritic chest pain; Dyspnoea 	<ul style="list-style-type: none"> • May be unremarkable • Diaphoresis (excessive sweating) • Pale cool skin • Sinus tachycardia (HR >100bpm) • 3rd and/or 4th heart sound • Basilar rales • Hypotension • (BP < 100./60mmH) 	<ul style="list-style-type: none"> • ECG <ul style="list-style-type: none"> ○ ST-segment depression ○ Transient ST-segment elevation(in 30-50% of patients) ○ and/or T-wave inversion • Cardiac biomarkers <ul style="list-style-type: none"> ○ Elevated Troponin ○ Elevated CK-MB 	<ul style="list-style-type: none"> ▪ Clopidogrel 300mg stat, 75mg daily ▪ Atovarstatin 80mg stat, (if troponin +ve) ▪ Enoxaprin 1mg/kg ▪ Beta Blocker <p>** if LVH present-ACE inhibitor</p>

4.3 Spiral Cycle 3

Here again, based on the validation in Spiral Cycle 2, new set of tasks were identified for the next cycle (Cycle 3), and each of the quadrants are populated as shown in Table 7. One of the main tasks is to identify the descriptions of the terms in layman's language. For this purpose a new Knowledge Template need to be designed and agreed upon between the Knowledge Engineers and the Domain Experts. This template is listed in Table 8.

Table 7. Summary of the Tasks Involved in Spiral Cycle 3

Quadrant 1	Quadrant 2	Quadrant 3	Quadrant 4
Purpose and scope identification	Knowledge acquisition	Knowledge conceptualization /formalization	Validation and plan next phase
Acquire medical terms, description of the terms used and layman's language. Update MKB ontology with this information on terminology.	Risk Analysis: Time allocation for expert to review and validation definition populated if the definition is not fit. KE prepare terminology list and acquire description of terms from online dictionary. Get validation of terminology from domain expert. Detail diagnosing process and next action identified by domain expert.	Update SKT with property that incorporates medical terms with definition. Update terminology list.	Review and validate changes made to ontology. Domain expert suggested adding more detail to "investigations" component of KB. KE identified need to develop rules to support decision making process.

Table 8: Knowledge Template for Terms List and its Description

Sign / Symptom	Layman's term	Also known as
1. 3 rd and/or 4 th heart sound	3 rd and 4 th sound is abnormal heart sound	
2. abdominal swelling	swelling seen from outside or inside by pressing the organ	
3. anorexia	loss of, or severe reduction in, appetite for food	
4. anxiety	anxiety abnormal and overwhelming sense of apprehension and fear	

Several more cycles were conducted (in total 5 cycles) before it was decided collectively by the Knowledge Engineers and the Domain Expert that the knowledge modelled was of sufficient details, and coverage. For this particular case study, the final Knowledge Template is listed in Table 9.

Table 9. Final Knowledge Template for Unstable Angina

Unstable Angina

Symptoms	Signs	Investigations	Management Heuristics	
<ul style="list-style-type: none"> • dyspnoea • indigestion • Epigastric Pain • Chest Pain <ul style="list-style-type: none"> ○ heavy ○ crushing ○ radiating (to left arm, to neck, to left shoulder) ○ squeezing 	<ul style="list-style-type: none"> • pulse (sinus tachycardia) • third heart sound • fourth heart sound • basilar rales • hypotension • pale cool skin • sweating 	<ul style="list-style-type: none"> • ECG <ul style="list-style-type: none"> ○ Ischaemic Change <ul style="list-style-type: none"> ▪ Tall T waves ▪ Pathological Q Waves ▪ ST elevation ▪ Inverted T waves ▪ ST depression ▪ No ST elevation • Cardiac biomarkers <ul style="list-style-type: none"> ○ Elevated Troponin ○ Elevated CK-MB 	<ol style="list-style-type: none"> 1. IF one of (symptoms) AND one of (signs) AND (investigations) THEN 2. IF (Investigations: Cardiac Biomarker:: Troponin:: Present) THEN 3. IF pain persist 	<p>Aspirin 300mg stat followed by 150mg daily AND Clopidogrel 300mg stat, 75mg daily AND beta blocker AND Atorvastatin 80mg stat</p> <p>Same as above PLUS Enoxaprin 1mg/kg</p> <p>Give morphine 50mg stat and stemetil 12.5mg stat and consider referral for angiography</p>

5 Results

Using the Knowledge Engineering Methodology based on the Spiral Process Model, we were able to conduct the necessary systematic knowledge acquisition, elicitation, and modelling activities to eventually develop a Medical Ontology and the Medical Knowledge Base that can be used by a Patient Diagnosis and Management System. Figure 2 depicts a very small portion of the knowledge base, specific to the case study discussed in this paper. The complete Cardiovascular Knowledge Base consists of 2240 triples. The concepts and properties modelled in this knowledge base are listed in Table 10.

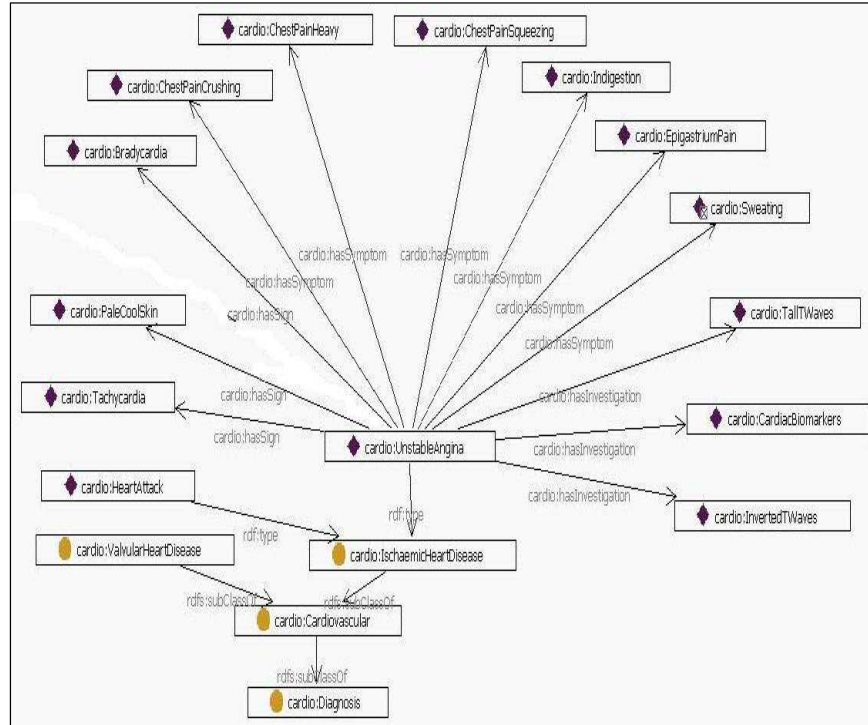


Fig 2. Portion of the Medical Knowledge Base
 ©2008-2010 MIMOS Berhad. All Rights Reserved.

Table 10. List of Concepts and Properties in the Cardiovascular Knowledge Base

List of Concepts	List of Properties
cardio:ECHO	cardio:hasApexBeatCharacteristic
cardio:EchoResultDetail	cardio:hasAtrialFibrillationResult
cardio:Echocardiogram	cardio:hasBiopsyResult
cardio:EmotionDisorder	cardio:hasBloodTestResult
cardio:GastrointestinalDisorder	cardio:hasCardiacBiomarkersResult
cardio:GastrointestinalDisorder	cardio:hasCharacteristic
cardio:GeneralSymptom	cardio:hasChestX-RayResult
cardio:HeartAscultation	cardio:hasDetail
cardio:HeartBeat	cardio:hasDisease
cardio:HeartBeat	cardio:hasECGResult
cardio:HeartCondition	cardio:hasExtraInvestigation
cardio:HeartCondition	cardio:hasHeartRhythm
cardio:HeartRhythm	cardio:hasInvestigationResult
cardio:HeartSound	cardio:hasIschaemicChangesResult
cardio:Investigation	cardio:hasIschaemicChangesResultDetail
cardio:InvestigationResult	cardio:hasLVHResult
cardio:IschaemicChanges	cardio:hasMurmurCharacteristic

cardio:IschaemicHeartDisease	cardio:hasMurmurSite
cardio:JugularVenousPulse	cardio:hasMurmurType
cardio:L VH	cardio:hasOccurance
cardio:LeftBundleBranchBlockLBBB	cardio:hasOtherInvestigation
cardio:LungRelatedSign	cardio:hasPhase
cardio:Management	cardio:hasPulseCharacteristic
cardio:Medicine	cardio:hasPulseRate
cardio:MurmurCharacteristic	cardio:hasPulseRhythm
cardio:MurmurSite	cardio:hasRadiologicalTestResult
cardio:OtherHeartSound	cardio:hasSign
cardio:OtherInvestigation	cardio:hasSignCharacteristic
cardio:Pain	cardio:hasSite
cardio:PhysicalAppearance	cardio:hasSoundBehaviour
cardio:PhysicalDeterioration	cardio:hasSymptom
cardio:Pulse	cardio:hasTremor
cardio:PulseCharacteristic	cardio:hasUrinalysisResult
cardio:RadiologicTest	cardio:hasVentricularEctopicsResult
cardio:RenalRelatedSign	cardio:hasPulseRhythm
cardio:Sign	cardio:hasRadiologicalTestResult
cardio:SignCharacteristic	cardio:hasSign
cardio:Site	
cardio:StomachDiscomfort	
cardio:Swelling	
cardio:Symptom	
cardio:Treatment	
cardio:Urinalysis	
cardio:ValvularHeartDisease	
cardio:VentricularEctopics	
cardio:MurmurCharacteristic	
cardio:MurmurSite	
cardio:OtherHeartSound	
cardio:OtherInvestigation	
cardio:Pain	

6 Conclusion and Future Work

In a nutshell, the experience gained in adopting and using the Spiral Process Model for engineering a Medical Knowledge Base has been in-valuable, both in terms of determining the usability of Spiral Process Model for Knowledge Engineering, as well as its ability to facilitate incremental knowledge acquisition and modelling of the knowledge base. This capability become extremely crucial when you have a situation where you are unable to acquire the complete knowledge first time around. It also allows the parties involved in the knowledge engineering activities to be focused on the tasks to be carried out and in a systematic matter. These features are extremely important when you are attempting to develop knowledge bases for commercial purpose.

This Knowledge Engineering Methodology was applied successfully for several knowledge bases including medical, agriculture, health and financial. This methodology is robust in its ability to handle the varied situations one encounters when performing the knowledge engineering activities. In future, we are looking at

making this methodology more formal, from the perspective of establishing a series of standard operating procedures (SOPs) in using the Spiral Process Model for Knowledge Engineering.

Acknowledgements. We would like to acknowledge and extend our heartfelt gratitude to the following persons who have made the completion of this paper possible: Prof Dr Abdul Rashid Abdul Rahman, Dr Nor Mahani Harun and Dr Azrul Rozaiman from Medical Interest Group Sdn Bhd, for their ideas, knowledge, supports and time.

References

1. Giuse, NB., Bankowitz, RA., Giuse, DA., Parker, RC., Miller, RA.: Medical Knowledge Base Acquisition: The Role of the Expert Review Process in Disease Profile Construction, American Medical Informatics Association (AMIA), November 8, 1989, pp.105-109.
2. Knowledge Base : Whats.com, http://searchcrm.techtarg.com/sDefinition/0,sid11_gci75339,00.html
3. Ontology of Folksonomy: A Mash-up of Apples and Oranges, Int'l Journal on Semantic Web & Information Systems, 3(2), 2007. <http://tomgruber.org/writing/ontology-of-folksonomy.htm>
4. Noy, N., McGuinness, D.: Ontology Development 101: A Guide to Creating Your First Ontology, Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, March 2001.
5. DogmaModeler, <http://starlab.vub.ac.be/website/node/47>
6. KAON, <http://kaon.semanticweb.org/frontpage>
7. Guarino, N., Welty, C.A.: An Overview of OntoClean, in S. Staab, R. Studer (eds.), Handbook on Ontologies, Springer Verlag 2004, pp. 151-172.
8. Mizoguchi, M., Sunagawa, E., Kozaki, K., Kitamura, Y.: The Model of Roles Within An Ontology Development Tool : Hozo, Journal of Applied Ontology, 2(2):159-179.
9. Protégé, <http://protege.stanford.edu/>
10. TopBraid Composer, http://www.topquadrant.com/products/TB_Composer.html
11. Uschold, M., Gruninger, M.: Ontologies: Principles, Methods and Applications, Knowledge Engineering Review, Vol. 11, No. 2. (1996), pp. 93-155.
12. Fernandez, M., Gomez-Perez, A., Juristo, N.: METHONTOLOGY : From Ontological Art Towards Ontological Engineering, AAAI Technical Report, Madrid, Spain (1997).
13. Pinto, H., Staab, S., Tempich, C.: DILIGENT : Towards a fine-grained methodology for Distributed, Loosely-controlled and Evolving Engineering of Ontologies, European Conference on Artificial Intelligence (ECAI) 2004.
14. UMLS Metathesaurus, <http://www.nlm.nih.gov/pubs/factsheets/umls.html>