

**Workshop and Poster
Proceedings of the 4th Joint
International Semantic
Technology Conference
(JIST 2014)**

November 9-11, 2014

Chiang Mai, Thailand

Workshops of JIST 2014

Preface

The workshops are held in conjunction with the conference: 4th Joint International Semantic Technology Conference (JIST 2014), which takes place during November 9-11, 2014, Chiang Mai, Thailand. The aim of the workshops is to provide an opportunity for participants for in-depth discussion of current and emerging topics of the semantic technologies.

This year, we accepted two workshop proposals with the goal of exploring focused issues across various themes. All submissions to the workshops were reviewed by the program committees of the workshops. In total, 12 papers are included in the proceedings. We believe that the workshops would foster interactions between different communities within the scope of JIST.

We would like to thank the organizers of the two workshops and the authors of the papers, the workshop program committee members, and JIST 2014 local organizers for their effort to make this happen.

Thatsanee Charoenporn, *Burapha University, Thailand*
Sasiporn Usanavasin, *SIIT, Thailand*

JIST 2014 Workshop Co-chairs

Poster and Demonstration Session of JIST 2014

Preface

The Poster and Demonstration session is held in conjunction with the conference: 4th Joint International Semantic Technology Conference (JIST 2014), which takes place during November 9-11, 2014, Chiang Mai, Thailand. The aim of the Poster and Demonstration session is to provide an opportunity for poster presenters to present late-breaking research results, ongoing research projects, and speculative or innovative work in progress.

All submissions for the Poster and Demonstration session were reviewed by the program committees and four papers were accepted. In addition, six papers submitted to the regular technical sessions were recommended by the program committees to the Poster and Demonstration session. In total, ten papers are included in the proceedings. We believe that this Poster and Demonstration session provides a great opportunity for researchers to present and to receive direct feedback from audience about the ongoing significant work.

We are grateful to the authors of the papers, program committee members, and JIST 2014 local organizers for their effort to make this happen.

Hanmin Jung, *KISTI, Korea*

Ekawit Nantajeewarawat, *SIIT, Thailand*

Kultida Tuamsuk, *Khon Kaen University, Thailand*

JIST 2014 Poster and Demo Co-chairs

**The 3rd International
Workshop on Semantic
Web-based Computer
Intelligence with Big-data
(SWCIB 2014)**

The 3rd International Workshop on Semantic Web-based Computer Intelligence with Big-data (SWCIB 2014)

Workshop Organizers

- Hanmin Jung (KISTI, Korea)
- Seungwoo Lee (KISTI, Korea)

Program Committee

- Didier.El Baz (LAAS-CNRS, France)
- Ing-Xiang Chen (Ericsson Taiwan Ltd., Taiwan)
- Sung-Kwon Choi (ETRI, Korea)
- Sung-Pil Choi (Kyonggi University, Korea)
- Shengyin Fan (Ricoh Software Research Center, China)
- Michaela Geierhos (University of Paderborn, Germany)
- Hyunchul Jang (KIOM, Korea)
- Dongwon Jeong (Kunsan National University, Korea)
- In-Su Kang (University of Kyungsung, Korea)
- Haklae Kim (Samsung Electronics, Korea)
- Haksoo Kim (Kangwon Univ., Korea)
- Changki Lee (Kangwon Univ., Korea)
- Seungwoo Lee (KISTI, Korea)
- Yeong-Su Lee (Cylex, Germany)
- Qing Li (Southwestern University of Finance and Economics (SWUFE), China)
- Lijun Zhu (ISTIC, China)
- Fuyuko Matsumura (NII, Japan)
- Masaharu Munetomo (Hokkaido University, Japan)
- Seung-Hoon Na (ETRI, Korea)
- Jiandong Qi (Beijing Forestry University, China)
- Brahmananda Sapkota (University of Twente, Netherlands)
- Sa-Kwang Song (KISTI, Korea)
- Kazunari Sugiyama (National University of Singapore, Singapore)
- Shuo Xu (ISTIC, China)
- Yongwook Yoon (KT, Korea)
- Nobukazu Yoshioka (NII, Japan)
- Chengzhi Zhang (Nanjing University of Science and Technology, China)
- Honggang Zhang (Beijing University of Post and Telecommunications, China)
- Yunliang Zhang (ISTIC, China)
- Zhixiong Zhang (National Science Library, Chinese Academy of Sciences, China)
- Zhangbing Zhou (Institute TELECOM & Management SudParis, France)
- Hoon Ko (University of J.E. Purkinje, the Czech Republic)
- Folker Caroli (University of Hildesheim, Germany)
- Young-guk Ha (Konkuk University, Korea)

SWCIB 2014

Preface

SWCIB 2014 is the 3rd International Workshop on Semantic Web-based Computer Intelligence with Big-data, which is held in conjunction with the 4th Joint International Semantic Technology Conference (JIST2014) at Chiang Mai, Thailand, in 2014. The previous two editions of this workshop had been collocated with the 2nd Joint International Semantic Technology Conference (JIST2012) and the 3rd Joint International Semantic Technology Conference (JIST2013).

The characteristic of intelligence is usually attributed to humans but, recently, many products and systems are also regarded to be intelligent. The recent prevalence of smartphones and sensors is accelerating the generation of big-data to be utilized for computer-based intelligence and Apple's Siri and IBM's Watson are brightening the future prospects of the research on intelligence and knowledge constructed and implemented from big data. In this big wave of big data and intelligence, this workshop will provide seven accepted papers and be quite interesting and beneficial to researchers and practitioners from academy to industry. We hope that this workshop provides an opportunity for participants to discuss and share key issues and practices of Semantic Web and big data technologies for computational intelligence. Lastly, we express our thanks to the authors for their valuable contributions and also wish to express our profound gratitude to JIST 2014 Organizers including Dr. Marut Buranarach and Dr. Sasiporn Usanavasin for allowing and helping this workshop.

November 2014

Hanmin Jung and Seungwoo Lee, *KISTI, Korea*

SWCIB 2014 Workshop Organizers

Re-ranking ASR Outputs for Spoken Sentence Retrieval

Yeongkil Song, Hyeokju Ahn, and Harksoo Kim

Program of Computer and Communications Engineering, College of IT,
Kangwon National University, Republic of Korea
{nlpyksong, zingiskan12, nlpdrkim}@kangwon.ac.kr

Abstract. In spoken information retrieval, users' spoken queries are converted into text queries by using ASR engines. If top- I results of the ASR engines are incorrect, the errors are propagated to information retrieval systems. If a document collection is a small set of short texts, the errors will more affect the performances of information retrieval systems. To improve the top- I accuracies of the ASR engines, we propose a post-processing model to rearrange top- n outputs of ASR engines by using Ranking SVM. To improve the re-ranking performances, the proposed model uses various features such as ASR ranking information, morphological information, and domain-specific lexical information. In the experiments, the proposed model showed the higher precision of 4.4% and the higher recall rate of 6.4% than the baseline model without any post-processing. Based on this experimental result, the proposed model showed that it can be used as a post-processor for improving the performance of a spoken information retrieval system if a document collection is a restricted amount of sentences.

Keywords: Re-ranking, ASR outputs, spoken sentence retrieval

1 Introduction

With the rapid evolution of smart phones, the needs of information retrieval based on spoken queries are increasing. Many information retrieval systems use automatic speech recognition (ASR) systems in order to convert users' spoken queries to text queries. In the process of query conversion, ASR systems often make recognition errors and these errors make irrelevant documents returned. If retrieval target documents (so called a document collection) are a small set of short texts such as frequently asked questions (FAQs) and restricted chatting sentences (*i.e.*, chatting corpus for implementing an intelligent personal assistant such as Siri, S-Voice, and Q-Voice), information retrieval systems will not perform well because a few keywords that are incorrectly recognized critically affect the ranking of documents, as shown in Fig. 1 [1].

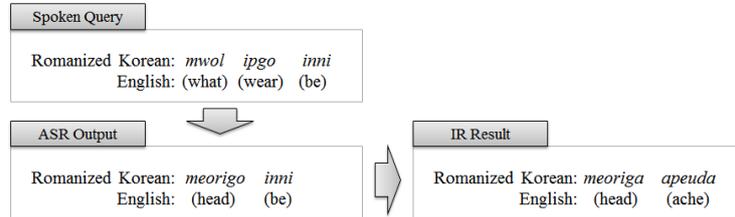


Fig. 1. Motivational example

To resolve this problem, many post processing methods for revising ASR errors have been proposed. Ringger and Allen [2] proposed a statistical model for detecting and correcting ASR error patterns. Brandow and Strzalkowski [3] proposed a rule based method to generate a set of correction rules from ASR results. Jung et al. [4] proposed a noisy channel model to detect error patterns in the ASR results. These previous models have a weak point that they need parallel corpus that includes ASR result texts and their correct transcriptions. To overcome this problem, Choi et al. [5] proposed a method of ASR engine independent error correction and showed the precision of about 72% in recognizing named entities in spoken sentences. Although the previous models showed reasonable performances, they have dealt with the first-ranked sentences among ASR results. The fact raised the result that low-ranked sentences are not considered although they are correct ASR outputs, as shown in the following Romanized Korean example.

Spoken query: *mwol ipgo inni* (What are you wearing?)
 Rank 1: *meorigo inni* (Is a head?)
 Rank 2: *mwol ipgo inni* (What are you wearing?)

To resolve this problem, we propose a machine learning model that re-ranks top- n outputs of an ASR system. In the above example, we expect that the proposed model changes *Rank 2* to *Rank 1*. If the volume of a document collection is big, it may be not easy to apply supervised machine learning models for re-ranking ASR outputs because the models need a large training data set that is annotated by human. However, if the document collection is a small set of short messages such as FAQs and chatting corpus, we think that the supervised machine learning models can be applied because the volume of the document collection is small enough to be annotated by human.

2 Re-ranking Model of ASR Outputs

2.1 Overview of the Proposed Model

The proposed model consists of two parts: a training part and a re-ranking part. Fig.1 shows the overall architecture of the proposed model.

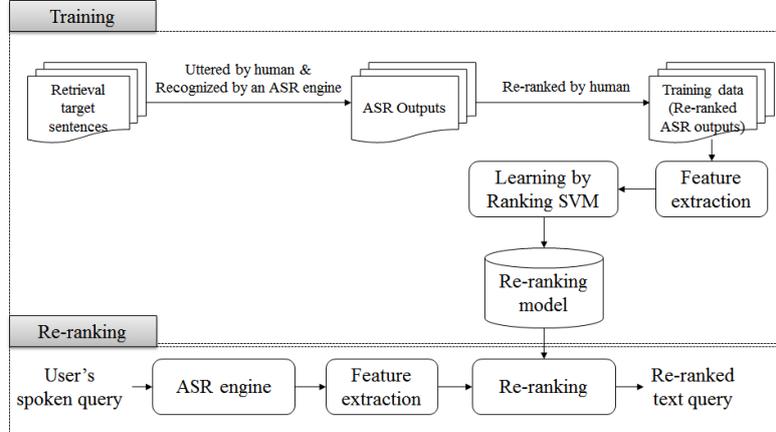


Fig. 2. Overall architecture of a re-ranking system

As shown in Fig. 1, we first collect top- n ASR¹ outputs of a document collection (a set of sentences in this paper) in which each sentence is uttered by 6 people. Then, we manually annotate the collected corpus with correct ranks. Next, the proposed system generates a training model based Ranking SVM (support vector machine) which is an application of SVM used for solving certain ranking problems [6]. When users input spoken queries, the proposed system re-ranks ASR outputs of the spoken queries based on the training model. Then, the system hands over the first ones among the re-ranked results to an information retrieval system.

2.2 Re-ranking ASR Outputs Using Ranking SVM

To rearrange top- n ASR outputs, we use a Ranking SVM which is a modification to the traditional SVM algorithm which allows it to rank instances instead of classifying them [7]. Given a small collection of ASR outputs ranked according to preference R^* with two ASR outputs $d_i, d_j \in R^*$, and a linear learning function f :

$$d_i \succ d_j \Rightarrow f(d_i) > f(d_j) \quad (1)$$

where the ASR outputs are represented as a set of features. The linear learning function f is defined as $f(d) = \mathbf{w} \cdot d$, as shown in Equation (2).

$$f(d_i) > f(d_j) \Leftrightarrow \mathbf{w} \cdot d_i > \mathbf{w} \cdot d_j \quad (2)$$

In Equation (2), the vector \mathbf{w} can be learned by the standard SVM learning method using slack variables, as shown in Equation (3).

¹ We use Google's ASR engine which returns top-5 outputs per utterance.

$$\begin{aligned}
& \text{minimize } \langle \mathbf{w} \cdot \mathbf{w} \rangle + C \sum_{i,j \in |R|} \xi_{ij} \\
& \text{subject to } \forall (d_i, d_j) \in R^* : \mathbf{w} \cdot d_i \geq \mathbf{w} \cdot d_j + 1 - \xi_{ij} \\
& \quad \forall (i, j) : \xi_{ij} \geq 0
\end{aligned} \tag{3}$$

To represent ASR outputs in the vector space of Ranking SVM, we should convert each ASR output into feature vectors. Table 1 show the defined feature set.

Table 1. Feature set of Ranking SVM

Feature Name	Explanation
ASR-Rank	Ranking of ASR outputs
ASR-Score	ASR score of the highest ranked ASR output
MOR-Bigram	Bigrams of morphemes
POS-Bigram	Bigrams of POS's
NUM-DUW	# of unknown content words that is not found in a domain dictionary
LEX-DUW	Unknown content words that is not found in a domain dictionary
NUM-GUW	# of unknown content words that is not found in a general dictionary
LEX-GUW	Unknown content words that is not found in a general dictionary

In Table 1, *ASR-Rank* has an integer number from 1 to 5 because Google's ASR engine returns five ASR outputs ranked by descending order. *ASR-Score* is represented by 10-point scale of ASR scores 0.1 through 1.0. In other words, if the ASR score is 0.35, the score in 10-point scale is mapped into 0.4. *MOR-bigram* and *POS-Bigram* are morpheme bigrams and POS bigrams that are obtained from a result of morphological analysis. For example, if a result of morphological analysis is "I/prop can/aux understand/verb you/prop", MOR-bigram is the set { ^;I I;can can;understand understand;you you;\$ }, and POS-bigram is the set { ^;prop prop;aux aux;verb verb;prop prop;\$ }. In the example, '^' and '\$' are the symbols that represent the beginning and the end of sentence, respectively. *NUM-DUW* and *LEX-DUW* are features associated with domain-specific lexicon knowledge. The domain dictionary used in *NUM-DUW* and *LEX-DUW* is a set of content words (so-called nouns and verbs) that is automatically extracted from a training data annotated with POS's by a morphological analyzer. *NUM-GUW* and *LEX-GUW* are features associated with general lexicon knowledge. The general dictionary used in *NUM-GUW* and *LEX-GUW* is a set of content words that is registered as entry words in a general purpose dictionary of a conventional morphological analyzer.

3 Experiments

3.1 Data Set and Experimental settings

We collected a chatting corpus which contains 1,000 sentences. Then, we asked six university students (three males and three females) for uttering the short sentences by

using a smartphone application that saves top-5 outputs of Google’s ASR engine. Next, we manually annotated with new rankings according to a lexical agreement rate between user’s input utterance and each ASR output. In other words, the more an ASR output lexically coincides with user’s input utterance, the higher the ASR output is ranked. Finally, we divided the annotated corpus into training data (800 sentences) and testing data (200 sentences). To evaluate the proposed model, we used precision at one (so-called P@1) and recall rate at one (so-called R@1) as performance measures, as shown in Equation (4). We performed 5-fold cross validation.

$$\begin{aligned}
 P@1 &= \frac{\text{\# of sentences corectly ranked in top-1 by the proposed model}}{\text{\# of sentences ranked in top-1 by the proposed model}} \\
 R@1 &= \frac{\text{\# of sentences corectly ranked in top-1 by the proposed model}}{\text{\# of sentences correctly ranked in top-1 by an ASR engine}}
 \end{aligned}
 \tag{4}$$

3.2 Experimental Results

We computed the performances of the proposed model for each user, as shown in Table 2.

Table 2. Performances per user

User	ASR-only		Proposed Model	
	P@1	R@1	P@1	R@1
1	0.487	0.647	0.545	0.726
2	0.408	0.634	0.461	0.717
3	0.437	0.665	0.469	0.715
4	0.466	0.669	0.499	0.717
5	0.440	0.604	0.494	0.678
6	0.460	0.608	0.493	0.654
Average	0.450	0.638	0.494	0.702

In Table 2, ASR-only is a baseline model that returns a top-1 output of an ASR engine without any re-ranking. The recall rate at five (so-called R@5) of Google’s ASR engine was 0.705. This fact reveals that Google’s ASR engine failed to correctly recognize 29.5% of the testing data. In other words, 29.5% of user’s utterances are not included in top-5 outputs of Google’s ASR engine. As shown in Table 2, the proposed model showed the higher precision of 4.4% and the higher recall rate of 6.4% than the baseline model. This fact reveals that the proposed model can contribute to improve the performance of a spoken sentence retrieval system if a document collection is a small set of short texts.

4 Conclusion

We proposed a re-ranking model to improve the top-1 performance of an ASR engine. The proposed model rearranges ASR outputs based on Ranking SVM. To improve the re-ranking performances, the proposed model uses various features such as ASR ranking information, morphological information, and domain-specific lexical information. In the experiments with a restricted amount of sentences, the proposed model outperformed the baseline model (the higher precision of 4.4% and the higher recall rate of 6.4%). Based on this experimental result, the proposed model showed that it can be used as a post-processor for improving the performance of a spoken sentence retrieval system.

Acknowledgements

This work was supported by the IT R&D program of MOTIE/MSIP/KEIT. [10041678, The Original Technology Development of Interactive Intelligent Personal Assistant Software for the Information Service on multiple domains]. This research was also supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074).

References

1. Kim, H., Seo, J.: Cluster-Based FAQ Retrieval Using Latent Term Weights. *IEEE Intelligent Systems*, 23(2), 58-65 (2008)
2. Ringger, E. K., Allen, J. F.: Error Correction via a Post-processor for Continuous Speech Recognition. In: *Proceedings of IEEE International Conference on the Acoustics, Speech and Signal Processing*, pp. 427-430 (1996)
3. Brandow, R. L., Strzalkowski, T.: Improving Speech Recognition through Text-Based Linguistic Post-processing. United States Patent 6064957 (2000)
4. Jeong, M., Jung, S., Lee, G. G.: Speech Recognition Error Correction Using Maximum Entropy Language Model. In: *Proceedings of the International Speech Communication Association*, pp.2137-2140 (2004)
5. Choi, J., Lee, D., Ryu, S., Lee, K., Lee, G. G.: Engine-Independent ASR Error Management for Dialog Systems. In: *Proceedings of the 5th International Workshop on Spoken Dialog System* (2014)
6. Joachims, T.: Optimizing Search Engines Using Clickthrough Data. In: *Proceedings of ACM SIGKDD*, pp. 133-142 (2002)
7. Arens, R. J.: Learning to Rank Documents with Support Vector Machines via Active Learning. Ph.D dissertation, University of Iowa (2009).

Relation Extraction Using TBL with Distant Supervision

Maengsik Choi and Harksoo Kim

Program of Computer and Communications Engineering, College of IT,
Kangwon National University, Republic of Korea
{nlpschoi, nlpdrkim}@kangwon.ac.kr

Abstract. Supervised machine learning methods have been widely used in relation extraction that finds the relation between two named entities in a sentence. However, their disadvantages are that constructing training data is a cost and time consuming job, and the machine learning system is dependent on the domain of the training data. To overcome these disadvantages, we construct a weakly labeled data set using distant supervision and propose a relation extraction system using a transformation-based learning (TBL) method. This model showed a high F1-measure (86.57%) for the test data collected using distant supervision but a low F1-measure (81.93%) for gold label, due to errors in the training data collected by the distant supervision method.

Keywords: Relation extraction, Transformation-based learning, Distant supervision

1 Introduction

In natural language documents, there are a huge number of relations between named entities. Automatic extraction of these relations from documents would be highly beneficial in the data analysis fields such as question answering and social network analysis. Previous studies on relation extraction [1,2,3,4] have been mainly conducted through supervised learning methods using Automatic Content Extraction Corpus (ACE Corpus) [5]. However, recent studies have investigated some methods based on simple rules rather than on complex algorithms because the supervised learning method using weakly labeled data generated by distant supervision has made it possible to use a large amount of data [6,7,8]. The distant supervision reduces the construction cost of training data by automatically generating a large amount of training data. In this paper, we propose a relation extraction system to generate weakly labeled data using DBpedia ontology [9] and classify the relations between two named entities by using transformation-based learning (TBL) [10].

2 Relation Extraction Method Based on TBL

As shown in Figure 1, the proposed system consists of three parts: (1) a distant supervision part that generates weakly labeled data from a relation knowledge base and

articles, (2) a training part that generates a TBL model by using the weakly labeled data as training data, and (3) an applying part that extracts relations from new articles using the TBL model.

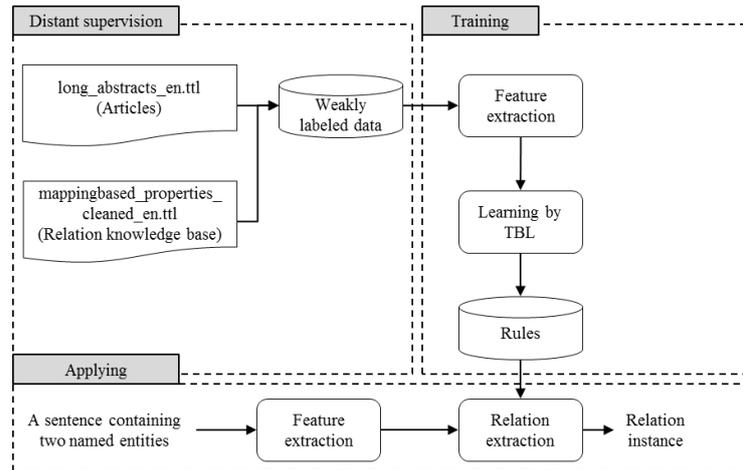


Fig. 1. Overall architecture of the proposed system

2.1 Construction of Weakly Labeled Data with Distant Supervision

For distant supervision, we use DBpedia ontology as a knowledge base. The DBpedia ontology is populated using a rule-based semi-automatic approach that relies on Wikipedia infoboxes, a set of subject-attribute-value triples that represents a summary of some unifying aspect that the Wikipedia articles share. As shown in Figure 2, the proposed system extracts sentences from Wikipedia articles by using the triple information of Wikipedia infobox.

Madonna (entertainer)

From Wikipedia, the free encyclopedia

	Madonna	
Born	Madonna Louise Ciccone August 16, 1958 (age 55) Bay City, Michigan, U.S.	<p>Madonna Louise Ciccone^[2] (/nɪˈkoʊnɛt/) (born August 16, 1958) is an American singer, songwriter, actress, and businesswoman. She achieved ...</p> <p style="text-align: center;"> Born Residence </p> <p>Born in Bay City, Michigan, Madonna moved to New York City to pursue a career in modern dance. After performing in the music groups <i>Breakfast Club</i> and <i>Emmy</i>, she signed with <i>Sire Records</i> (an affiliate of Warner Bros.</p>
Residence	New York City, U.S.	

Fig. 2. Example of an infobox and an article in Wikipedia

For example, the second sentence in Figure 2 is extracted from the weakly labeled data having a “Born” relation between “Madonna” and “Bay City, Michigan” and a “Residence” relation between “Madonna” and “New York City.”

2.2 Relation Extraction Using TBL

To extract features from the weakly labeled data, the proposed system performs natural language processing using Apache OpenNLP [10], as follows.

1. Sentences are separated using SentenceDetectorME.
2. Parts of speech (POS) are annotated using Tokenizer and POSTaggerME.
3. Parsing results are converted to dependency trees, and head words are extracted from the dependency trees.

For TBL, the proposed system uses two kinds of templates; a morpheme-level template and a syntax-level template. As shown in Figure 3, the morpheme-level template is constituted by the combination of words and POSs that are present in the left and right three words of the target named entities based on the POS tagging results.

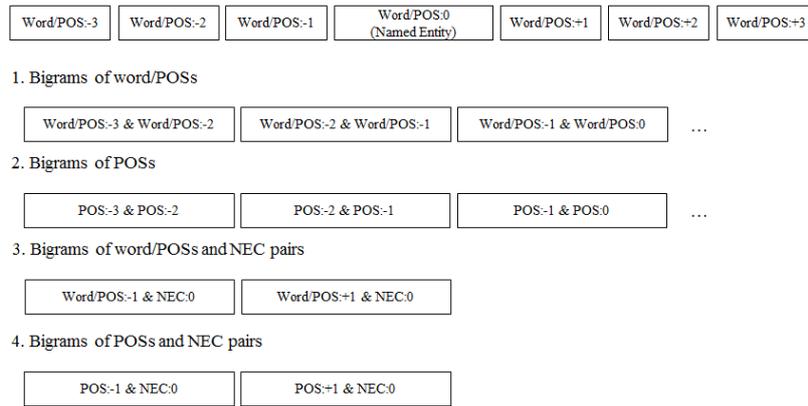


Fig. 3. Morpheme-level template

In Figure 3, the numbers described as “-n” and “+n” mean a left n -th word and right n -th word, respectively, from the target named entities. Then, “NEC” means the class name of the target named entity. As shown in Figure 4, the syntax-level template is constituted by the combination of two target named entities, common head word of the two named entities, head word (parent node in a dependency tree) of the two named entities, and dependent word (child node in a dependency tree) of the two named entities based on dependency parsing results.

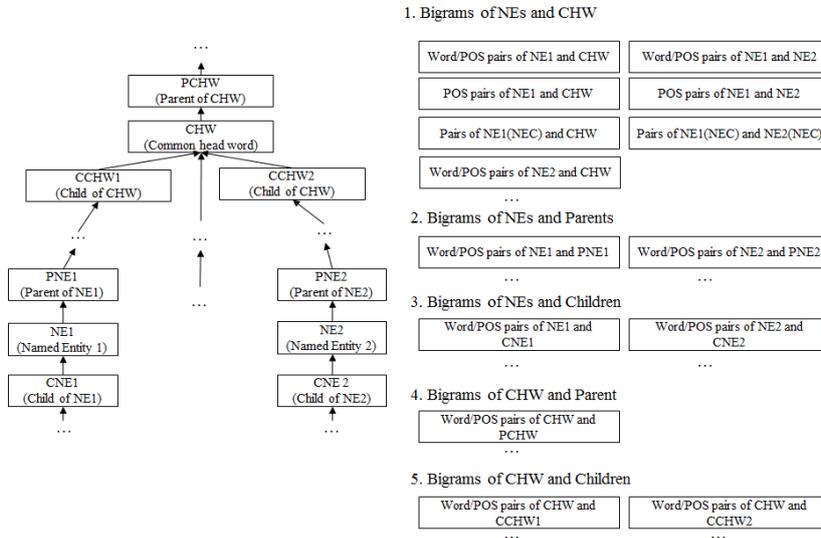


Fig. 4. Syntax-level template

3 Evaluation

3.1 Data Sets and Experimental Settings

We used DBpedia ontology 3.9 in order to generate weakly labeled data. Then, we selected the ‘person’ class as a relation extraction domain. Next, we selected five attributes highly occurred in the infobox templates of the ‘person’ class. Finally, we constructed a weakly labeled data set by using distant supervision. Table 1 shows the distribution of the weakly labeled data set.

Table 1. The number of instances in weakly labeled data

Attribute	All data	Weak-label test data	Gold-label test data
ActiveYearsStartYear	8,913	866	41
ActiveYearsEndYear	9,045	921	48
Award	1,627	140	4
BirthPlace	53,100	5,267	201
Nationality	8,754	869	133
Total	81,439	8,063	427

For the experiment, 90% of the weakly labeled data were used as training data, and the remaining 10% were used as test data (hereinafter “weak-label test data”). In addition, to measure the reliability of the weak-label test data, a gold-label test data set, which was manually annotated with correct answers after randomly selecting 822

items, was constructed. During manual annotation, 395 noise sentences out of 822 sentences were removed.

3.2 Experiment Results

As shown in Table 2, the proposed system showed high performance with regard to the weak-label test data but low performance with regard to the gold-label test data. This is because of noise sentences (*i.e.*, sentences that do not describe the relation between two named entities) that are included in the training data collected through the distant supervision method. For example, “Christopher Plummer” and “Academy Award” have an “Award” relation, and based on this, the extracted sentence (“The film also features an extensive supporting cast including Amanda Peet, Tim Blake Nelson, Alexander Siddig, Amr Waked and Christopher Plummer, as well as Academy Award winners Chris Cooper, William Hurt”) describes the “Award” relation between “Chris Cooper and William Hurt” and “Academy Award.” The results of analysis of the gold-label test data showed that there were 395 noise sentences out of 822 sentences.

Table 2. The performance of the proposed model

Data	Accuracy	Macro precision	Macro recall	F1-measure
Weak-label test data	0.9031	0.8735	0.8582	0.8657
Gold-label test data	0.7424	0.8275	0.8113	0.8193

The proposed system showed high performance for the weak-label test data. The fact reveals that the proposed system can show better performance for the gold-label test data only if good quality training data is ensured. In order to solve this problem, a method that can reduce the noise of the training data collected through distant supervision is required.

4 Conclusion

We proposed a relation extraction system using TBL based on distant supervision. In the proposed system, transformation rules were extracted based on a morpheme-level template that reflects the linguistic characteristics around named entities. They were also extracted based on a syntax-level template that reflects the syntactic dependency between two named entities. The experiment results indicated that higher performance could be expected only when the quality of the training data, which were extracted based on the distant supervision method, is improved. In the future, we will study on a method to increase the quality of training data collected by distant supervision.

5 Acknowledgments

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2013R1A1A4A01005074), and was supported by ATC(Advanced Technology Center) Program “Development of Conversational Q&A Search Framework Based On Linked Data: Project No. 10048448”. It was also supported by 2014 Research Grant from Kangwon National University(No. C1010876-01-01).

6 References

1. Culotta, Aron, and Jeffrey Sorensen: Dependency tree kernels for relation extraction. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, (2004)
2. Bunescu, Razvan C., and Raymond J. Mooney: A shortest path dependency kernel for relation extraction. In: Proceedings of HLT/EMNLP, pp. 724-731 (2005)
3. Zhang, Min, Jie Zhang, and Jian Su. Exploring syntactic features for relation extraction using a convolution tree kernel. In: Proceedings of HLT-NAACL, pp. 288-295 (2006)
4. Zhou, G., Zhang, M., Ji, D. H., & Zhu, Q.: Tree kernel-based relation extraction with context-sensitive structured parse tree information. In: Proceedings of EMNLP-CoNLL pp. 728-736 (2007)
5. NIST 2007. The NIST ACE evaluation website. <http://www.nist.gov/speech/tests/ace>
6. Mintz, Mike, et al: Distant supervision for relation extraction without labeled data. In: Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics, pp. 1003-1011 (2009)
7. Tseng, Yuen-Hsien, et al: Chinese Open Relation Extraction for Knowledge Acquisition. EACL 2014, pp. 12-16 (2014)
8. Chen, Yanping, Qinghua Zheng, and Wei Zhang.: Omni-word Feature and Soft Constraint for Chinese Relation Extraction. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 572-581 (2014)
9. <http://wiki.dbpedia.org/Downloads39>
10. Ngai, Grace, and Radu Florian: Transformation-based learning in the fast lane. In: Proceedings of NAACL, pp. 40-47 (2001)
11. <https://opennlp.apache.org/>
12. Aproso, Alessio Palmero, Claudio Giuliano, and Alberto Lavelli.: Extending the Coverage of DBpedia Properties using Distant Supervision over Wikipedia. In: NLP-DBPEDIA@ISWC. (2013)
13. Choi, Maengsik, and Harksoo Kim.: Social relation extraction from texts using a support-vector-machine-based dependency trigram kernel. In: Information Processing & Management 49.1 pp. 303-311 (2013)

Researcher Profiling for Researcher Analysis Service

Mikyoung Lee¹, Minhee Cho¹, Changhoo Jeong¹, Hanmin Jung¹

¹ Korea Institute of Science and Technology Information (KISTI),
245 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{jerryis, mini, chjeong, jhm}@kisti.re.kr

Abstract. This study examines a method of generating comprehensive profiling information for a researcher analysis service. In addition to basic and performance-based information about researchers necessary to generate profiling information, we introduce researcher performance index models for researcher analysis service. The models can that measure qualitative and quantitative performance, researcher influence, and growth potential, which are necessary to analyze the skills of a researcher from multiple perspectives. We measure the qualitative performance index of researchers by using the citation index of the papers they have published. The quantitative performance index can be measured based on a researcher's published papers. The Influence index measures the social influence of researchers according to their academic work. The growth potential index measures the speed at which a researcher improves research performance. We expect to develop a researcher analysis service that can evaluate a researcher's performance and enhance his or her research abilities.

Keywords: Researcher Profiling, Research Performance analysis, Researcher Analysis service

1 Introduction

Many services exist that provide researchers with a wealth of academic information such as scholarly literature, including Google Scholar, MS Academic Search and Elsevier SciVal. However, services are lacking that help improve the research abilities of researchers by analyzing their research skills. In Informetrics, various studies have represented researcher research skills in the form of objective numbers by examining various bibliometric indicators for research policy [1]. Currently, the h-index, which uses a citation index, is widely used to measure researcher skills. However, the h-index fails to produce objective assessments of the abilities of individual researchers. This is because it reflects only the combined achievements of co-researchers and is limited in the manner in which it comparatively analyzes them with the achievements of researchers in other fields. In addition, the h-index does not reflect the assessments of competent new researchers. Numerous studies have been conducted to compensate for the existence of a index, which is insufficient for analyzing the comprehensive research skills of researchers.

¹ Corresponding author

This study describes researcher profiling information used in a researcher analysis service for improving research skills.

2 Related Work

InSciTe Advisory(<http://inscite-advisory.kisti.re.kr/search>) deals with various textual big data including papers, patents, Web, social data, and linked data. It analyzes researcher's competitiveness and recommends attainable strategy and plan based on prescriptive analytics as well as descriptive analytics. It consists of two main analytics: descriptive analytics and prescriptive analytics. The former includes both activity history analysis and research power analysis for a selected researcher [2][3].

- **Commerciality:** Ability to produce practical products and profits
- **Scholarity:** Ability to produce new knowledge and academic outputs
- **Influentiality:** Ability to spread leverage to other researchers
- **Diversity:** Ability to extend research scope and degree of variation of research field
- **Durability:** Ability to keep research consistently in some research field
- **Technology emergability:** Degree of emerging about research area performed by researcher
- **Partner Trend:** Status of research area changing about partner researchers
- **Market Trend:** Status of market size changing about researcher area



Fig. 1. Descriptive analytics of InSciTe Advisory

ArnetMiner(<http://www.arnetminer.org>) aims to provide comprehensive search and mining services for researcher social networks. In this system, we focus on: creating a semantic-based profile for each researcher by extracting information from the distributed Web; integrating academic data from multiple sources; accurately searching the heterogeneous network; analyzing and discovering interesting patterns from the built researcher social network [4].

ArnetMiner offers insights into the capabilities of researchers by using the seven indices of activity, papers, citation, h-index, g-index, sociability, and diversity. As Fig. 2 shows, users can judge a researcher's ability by examining the graph related to each index.

- **Citation:** The number of citations of all publications by an expert.
- **Papers:** The number of all publications by an expert.

- **H-index:** An expert has index h if h of his or her N papers have at least h citations each, and the other papers have at most h citations each.
- **Activity:** People's activity is simply defined based on one's papers published in the last years.
- **Diversity:** An expert's research may include several different research fields. Diversity is defined to quantitatively reflect the degree
- **Sociability:** The score of sociability is basically defined based on how many coauthors an expert has.

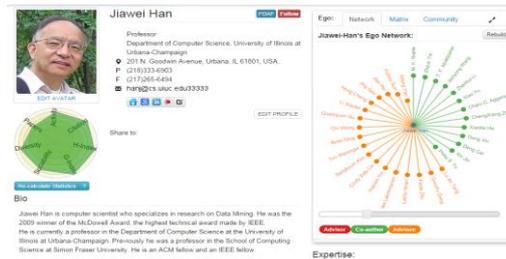


Fig. 2. Researcher Information on ArnetMiner

3 Researcher Performance Index Model

This chapter discusses the researcher performance index model used for researcher profiling. This index model collects the bibliography and citation information of researchers and compares them numerically to measure their research skills. We define four indices based on qualitative and quantitative performance, influence, and growth potential.

3.1 Qualitative Performance Index

We measure the qualitative performance index of researchers by using the citation index of the papers they have published. The qualitative performance index is a measurement of the influence of published papers and, thus, an indicator of their research quality. In general, indices such as citation index, h-index [5] and g-index [6] are used to measure researcher performance. A citation index can be used because the h-index and g-index for all authors in our collection are unavailable.

3.2 Quantitative Performance Index

The quantitative performance index can be measured based on a researcher's published papers. The more active the researcher has been producing recent studies, the more weight is given to the annual performance index.

The qualitative performance index is not proportional to the quantitative performance index because the more papers that are published, the higher is the

researcher's score on the quantitative index, thus indicating active research. However, the higher the quality of the papers, the higher is the researcher's quantitative performance index because the rank of the journals and conferences are reflected in the calculation of the quantitative performance index. If researchers being considered are the main or corresponding authors, they are more likely to score high on the quantitative index when they are actively involved in research. Using (1), the quantitative performance index can be calculated based on the rank of the journals and conferences, the weight of the author order.

$$QP = \sum_1^n Journal_{rank} \times author_{order} + \sum_1^m Proceeding_{rank} \times author_{order} \quad (1)$$

The rank of journals can be calculated by the journal's impact factor. The journals from each field are ranked according to the impact factor. Regarding conference proceedings, the rankings of the conferences in each field are used. With respect to authors, main and corresponding authors who write the majority of a paper can earn one point, whereas co-authors receive points equal to $1/n$ th, where n is the total number of authors.

3.3 Influence Index

Using a co-author network, the influence index measures the social influence of researchers according to their academic work. The influence index is an application of Google's PageRank [7] based on the co-author network and assumes that co-authors of influential research also have high influence.

$$II(A) = (1 - d)/n + d(\sum_1^n II(Tn)/C(Tn)) \quad (2)$$

In (2), $II(A)$ refers to the value of researcher A's influence, $II(TI)$ refers to the value of researcher TI 's influence, and $C(TI)$ represents the number of co-authors with researcher TI . In addition, d equals 0.85, as in the PageRank algorithm.

3.4 Growth Potential Index

The growth potential index measures the speed at which a researcher improves his or her research performance. We measure the growth potential index based on the amount of time a researcher spends reaching his or her current research performance, thus enabling to the system to predict how quickly the researcher can become a leader.

$$GP = QP/n \quad (3)$$

QP represents the quantitative performance index and n denotes the elapsed time from the beginning of the study to the present.

4 Researcher Profiling

4.1 Design

During the development of our researcher analysis service, we define research profiling, which includes all analytical information about the researcher. Researcher profiling consists of a list of basic information about researchers such as names and affiliations. In addition, it identifies essential information derived from their research such as main research skills (as shown in Fig. 3), research area (Fig. 3), and the co-author network. Moreover, it includes indices provided by research performance models.

Fig. 3 shows an example of researcher profiling designed in this study. The profiling extracts basic information from bibliographies and citations such as names, affiliations, beginning years of research, paper types, author order, number of citations, and major technology used. Information such as the level of technology and co-author network is obtained from analysis models. All acquired information is used to calculate the four indices of researcher performance index models. Most profiling preferences include time information, which facilitates the observation.

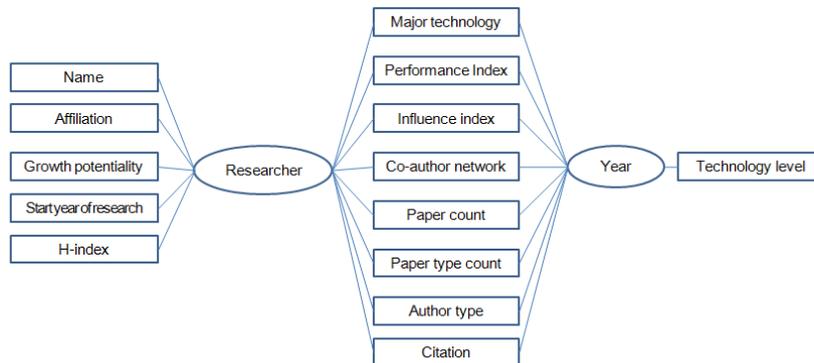


Fig. 3. Design of Researcher Profiling

4.2 Applications

Multiple aspects of research performance can be evaluated using the four indices of qualitative and quantitative performance, influence, and growth potential provided by research performance index models. Each index can be used to evaluate researcher tendencies and characteristics. For example, researchers can be divided into various types such as competent researchers who produce high quality research results, those who produce highly quantitative outcomes, and those who have immense research potential. In addition, yearly trends related to researcher profiling information can confirm changes in a researcher's performance based on flow patterns of falling, rising, and stagnancy. Moreover, the service can also be used for prescriptive analytics to reinforce researchers' strengths, correct research weaknesses, and

compare researchers of similar research capabilities and patterns. Finally, it can be used to analyze research performance patterns and to determine researcher's type.

5 Conclusion

This study described a researcher profiling method necessary to develop a researcher analysis service. Research profiling information consists of meta-information, performance information, and information of researchers acquired by analytical model. We evaluated researchers from multiple perspectives. We used a qualitative performance index, which indicates the quality of papers; a quantitative performance index, which indicates the number of papers published; an influence index, which refers to the amount of influence a researcher has among his or her peer researchers; and a growth potential performance index, which indicates degrees of change (i.e., levels of improvement) in research performance over time.

Our researcher profiling service represents an aggregation of all information that can be used to evaluate researcher and their research characteristics and patterns. This service is necessary to develop a service to assist researchers in improving their performance.

Acknowledgements

This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers].

References

1. Rodrigo, C., Maria, B.: The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. In: *Journal of Informetrics*, vol 1 issue 3, pp193-203, Elsevier (2007)
2. S. Song, D. Jeong, J. Kim, M. Hwang, J. Gim, H. Jung, :Research Advising System based on Prescriptive Analytics, In Proc. of International Workshop on Data-Intensive Knowledge and Intelligence in conjunction with the 9th FTRA International Conference on Future Information Technology. (2014)
3. Jinyung Kim, Myungwon Hwang, Janwon Gim, Sa-Kwang Song, Do-Heon Jeong, Seungwoo Lee, Hanmin Jung.: Intelligent Analysis and Prediction Model of Researcher's Capability for Prescriptive Analytics. In: AIM 2014 (2014)
4. Jie, T., Jing, Z., Limin, Y., Juanzi, L., Li, Z., Zhong, S.: ArnetMiner: extraction and mining of academic social networks. The 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 990-998, ACM, New York, NY, USA (2008)
5. H-index, <http://en.wikipedia.org/wiki/H-index>
6. G-index, <http://en.wikipedia.org/wiki/G-index>
7. Page, L., Brin, S., Motwani, R., Winograd, T.: The pageRank Citation Ranking: Bringing Order to the Web, Technical Report, Stanford InfoLab. (1999)

Ontology Schema-specific Rule Materialization

Seungwoo Lee, Chang-Hoo Jeong, Jung-Ho Um, Taehong Kim, Hanmin Jung

Dept of Computer Intelligence Research, KISTI
245 Daehak-ro, Yuseong-gu, Daejeon, 305-806, Korea
{swlee, chjeong, jhum, kimtaehong, jhm}@kisti.re.kr

Abstract. The reasoning should tackle big data issues like other domains as the size of ontology grows bigger and bigger. Especially, rule-based reasoning should overcome the following challenges: duplicate elimination and rule matching efficiency. To deal with these challenges, we introduce a new rule-based reasoning method which materializes each generic instance rule into several schema-specific instance rules and combines with Hadoop framework to deal with billions of triples. The experiment shows the materialization remarkably improves the efficiency of rule-based reasoning by reducing the amount of required memory and making it linear to the data size.

Keywords: rete reasoning; rule materialization; RDFS rule; OWL Horst rule

1 Introduction

Rule-based reasoning is a process that derives new knowledge – it is represented in triples composed of subject, predicate and object in ontology reasoning – from given set of knowledge by matching more than one rules. However, the reasoning process is also suffering from big data issues like other domains as the size of ontology has become bigger and bigger. To achieve efficient reasoning with overcoming big data issues, we have several challenges and two of them are follows: duplicate elimination and rule matching efficiency.

First, separate input sets of triples may derive same – i.e., duplicate – triples by one rule. Even different rules may derive duplicate triples. Urbani et al.[1] pointed out that reasoning might derive 50 times more duplicates than unique derived triples in their preliminary simulation. So, we need an efficient mechanism for eliminating duplicates and this challenge should be overcome by all means to achieve the scalability of reasoning process. Second, some parts of reasoning rules are often so generic to cause too many matches of triples. Rules are generally defined from the semantics of vocabularies of ontology description languages such as RDF (Resource Description Framework), RDFS (RDF Schema) and OWL (Web Ontology Language). Therefore, these rules are always valid independently of any specific ontology and have generic triple patterns in their condition. These rules often cause inefficiency in matching them to given set of triples because the generic patterns could be matched to too many given triples, most of which are eventually filtered out when joining with triples

matched to remaining patterns. So, we need an efficient mechanism for reducing such unnecessary triple matches.

In this paper, we present a method that removes unnecessary pattern matches, eventually reduces join operations in rule-based reasoning selectively fetches input triples, and efficiently eliminates duplicates derived by reasoning rules.

The remaining part of this paper is as follows: in section 2, some related works are explored and in section 3, our main approaches are described. Some experiments justifying our approaches are given in section 4, which is followed by conclusion in section 5.

2 Related Work

Rule-based reasoning is implemented widely based on rete algorithm[2][3] due to its efficiency in pattern matching. This algorithm achieves efficiency of pattern matching by enabling more than one rule to share triples matched to their common triple patterns. Most time-consuming operation in rete occurs when joining triples from more than one pattern because join operation causes repeated search and comparison of corresponding values to common variables. To perform this efficiently, indexing mechanisms such as hashing or Pyramid technique are usually adopted[3][4]. Rete has a big advantage in efficiency but also has a severe disadvantage in scalability. It requires quite large memory spaces because it maintains all triples matched to each pattern and joined by more than one pattern in main memory. This makes it impossible for rete-based reasoning to process billions of triples.

To achieve scalability of rule-based reasoning, recent research such as WebPIE[1] utilized Hadoop, a distributed and parallelized computing framework. It showed that the performance of Hadoop-based reasoning is highly dependent on how to design mappers and reducers for each rule. So, it designed rule-specific mappers and reducers and succeeded to achieve web-scale reasoning. To eliminate duplicate derivation in early stage, it tried to get mappers to group input triples by considering the output of the rule, not joining key. It, in addition, maintained schema triples in main memory to improve load balances between parallelized computing nodes.

In this paper, we describe a new approach that first removes unnecessary pattern matches in rete framework by materializing rules based on given ontology schema, next fetches input triples selectively by implementing rule-specific input formats, and finally eliminates duplicate derivations by grouping rules having same output forms.

3 Rule Materialization

In our previous work[5], we applied dynamic materialization on some rules having generic patterns in RDFS and OWL semantics[6][7] and in this paper, we extend the materialization to all rules having schema triple patterns in RDFS and OWL semantics. Triples can be divided into two types: schema and instance triples. Schema indicates triples defining classes, relationships between them, and attributes related to them while instance means triples describing individuals, relationship between them,

Table 1. RDF and RDFS rules[6]

	id	entailment rules
rdf	1	$(u p v) \rightarrow (p \text{ rdf:type rdf:Property})$
	2	$(u p v)$ (if v is a XML literal and $_ :n$ is a bland node allocated to v) $\rightarrow (_ :n \text{ rdf:type rdf:XMLLiteral})$
rdfs	1	$(u p v)$ (if v is a plain literal and $_ :n$ is a bland node allocated to v) $\rightarrow (_ :n \text{ rdf:type rdfs:Literal})$
	2	$(p \text{ rdfs:domain } c) (u p v) \rightarrow (u \text{ rdf:type } c)$
	3	$(p \text{ rdfs:range } c) (u p v) \rightarrow (v \text{ rdf:type } c)$
	4a	$(u p v) \rightarrow (u \text{ rdf:type rdfs:Resource})$
	4b	$(u p v) \rightarrow (v \text{ rdf:type rdfs:Resource})$
	5	$(p \text{ rdfs:subPropertyOf } q) (q \text{ rdfs:subPropertyOf } r) \rightarrow (p \text{ rdfs:subPropertyOf } r)$
	6	$(p \text{ rdf:type rdf:Property}) \rightarrow (p \text{ rdfs:subPropertyOf } p)$
	7	$(p \text{ rdfs:subPropertyOf } q) (u p v) \rightarrow (u q v)$
	8	$(c \text{ rdf:type rdfs:Class}) \rightarrow (c \text{ rdfs:subClassOf rdfs:Resource})$
	9	$(c \text{ rdfs:subClassOf } d) (u \text{ rdf:type } c) \rightarrow (u \text{ rdf:type } d)$
	10	$(c \text{ rdf:type rdfs:Class}) \rightarrow (c \text{ rdfs:subClassOf } c)$
	11	$(c \text{ rdfs:subClassOf } d) (d \text{ rdfs:subClassOf } e) \rightarrow (c \text{ rdfs:subClassOf } e)$
	12	$(p \text{ rdf:type rdfs:ContainerMembershipProperty}) \rightarrow (p \text{ rdfs:subPropertyOf rdfs:member})$
13	$(c \text{ rdf:type rdfs:Datatype}) \rightarrow (c \text{ rdfs:subClassOf rdfs:Literal})$	

and attributes related to them. Similarly, triple patterns forming rules can also be divided into two types: schema and instance triple patterns. Schema triples are generally small and static to a given ontology so as to be maintained in main memory while instance triples may continue to grow as much as not to be maintained in main memory. So, schema-only rules could be processed sufficiently on rete framework, but rules having instance triple patterns could not be processed on rete when the patterns are too generic.

For example, the generic triple pattern $(u p v)$ of rdfs2 in Table 1 could be matched to all given triples, but only small part of them could be joined with specific triples matched to the remaining triple pattern $(p \text{ rdfs:domain } c)$ due to the common variable ' p '. Indexing mechanisms such as hashing are usually applied to check such consistency efficiently, but they also require large memory spaces. More badly, as the target ontology grows, such indexing size also grows and may not be maintained in main memory. To solve this issue, we take following approaches according to types of rules:

- Schema-only rules (i.e., rdfs 5, 6, 8, 10, 11, 12, and 13, and owl-horst 9, 10, 12a, 12b, 12c, 13a, 13b, and 13c) are processed fully on rete framework.
- Generic-only rules (i.e., rdf 1 and 2, rdfs 1, 4a, and 4b, owl-horst 5a and 5b) are replaced and processed fully using dictionary which encodes all unique terms of input triples.

Table 2. OWL Horst rules[1][8]

id	entailment rules
1	$(p \text{ rdf:type owl:FunctionalProperty}) (u p v) (u p w) \rightarrow (v \text{ owl:sameAs } w)$
2	$(p \text{ rdf:type owl:InverseFunctionalProperty}) (u p w) (v p w) \rightarrow (u \text{ owl:sameAs } v)$
3	$(p \text{ rdf:type owl:SymmetricProperty}) (u p v) \rightarrow (v p u)$
4	$(p \text{ rdf:type owl:TransitiveProperty}) (u p v) (v p w) \rightarrow (u p w)$
5a	$(u p v) \rightarrow (u \text{ owl:sameAs } u)$
5b	$(u p v) \rightarrow (v \text{ owl:sameAs } v)$
6	$(u \text{ owl:sameAs } v) \rightarrow (v \text{ owl:sameAs } u)$
7	$(u \text{ owl:sameAs } v) (v \text{ owl:sameAs } w) \rightarrow (u \text{ owl:sameAs } w)$
8a	$(p \text{ owl:inverseOf } q) (u p v) \rightarrow (v q u)$
8b	$(p \text{ owl:inverseOf } q) (u q v) \rightarrow (v p u)$
9	$(c \text{ rdf:type owl:Class}) (c \text{ owl:sameAs } d) \rightarrow (c \text{ rdfs:subClassOf } d)$
10	$(p \text{ rdf:type rdf:Property}) (p \text{ owl:sameAs } q) \rightarrow (p \text{ rdfs:subPropertyOf } q)$
11	$(u p v) (u \text{ owl:sameAs } x) (v \text{ owl:sameAs } y) \rightarrow (x p y)$
12a	$(c \text{ owl:equivalentClass } d) \rightarrow (c \text{ rdfs:subClassOf } d)$
12b	$(c \text{ owl:equivalentClass } d) \rightarrow (d \text{ rdfs:subClassOf } c)$
12c	$(c \text{ rdfs:subClassOf } d) (d \text{ rdfs:subClassOf } c) \rightarrow (c \text{ owl:equivalentClass } d)$
13a	$(p \text{ owl:equivalentProperty } q) \rightarrow (p \text{ rdfs:subPropertyOf } q)$
13b	$(p \text{ owl:equivalentProperty } q) \rightarrow (q \text{ rdfs:subPropertyOf } p)$
13c	$(p \text{ rdfs:subPropertyOf } q) (q \text{ rdfs:subPropertyOf } p) \rightarrow (p \text{ owl:equivalentProperty } q)$
14a	$(c \text{ owl:hasValue } v) (c \text{ owl:onProperty } p) (u p v) \rightarrow (u \text{ rdf:type } c)$
14b	$(c \text{ owl:hasValue } v) (c \text{ owl:onProperty } p) (u \text{ rdf:type } c) \rightarrow (u p v)$
15	$(c \text{ owl:someValuesFrom } d) (c \text{ owl:onProperty } p) (u p v) (v \text{ rdf:type } d) \rightarrow (u \text{ rdf:type } c)$
16	$(c \text{ owl:allValuesFrom } d) (c \text{ owl:onProperty } p) (u p v) (u \text{ rdf:type } c) \rightarrow (v \text{ rdf:type } d)$

- Rules related to ‘owl:sameAs’ (i.e., owl-horst 6, 7, and 11) are replaced with *sameAs* table storing all same terms, defining a canonical term among them, and replacing all same term occurrences with their canonical ones.
- Remaining rules having combination of schema and instance triple patterns (i.e., rdfs 2, 3, 7, and 9, owl-horst 1, 2, 3, 4, 8a, 8b, 14a, 14b, 15, and 16) are processed first on rete to be materialized into schema-specific rules and then each materialized rule is processed on distributed and parallelized Hadoop framework.

The first and second ones are straightforward and the third one is similar to the approach of WebPIE[1]. So, the detailed explanation of them is omitted here. For the last one, our previous work[5] introduced rete-based framework that materializes some of the rules (i.e., rdfs 2, 3, 7, and 9, owl-horst 4, and 8a) into schema-specific rules and this paper extends the work into other rules in OWL Horst and incorporates Hadoop framework[9] additionally to deal with billions of triples.

For example, when a given ontology defines n functional properties, p_1, \dots, p_n , our rete-based reasoning framework will materialize the rule owl-horst1 into following n rules: $(u p_i v) (u p_i w) \rightarrow (v \text{ owl:sameAs } w)$ (here, $i = 1, \dots, n$). These rules can be implemented in one pair of a mapper and a reducer as in WebPIE[1] but having p_i as parameters. In addition, when input triples are stored and indexed with six possible

combinations of subject, predicate and object using Hbase[10], one of column-based, no-SQL databases, the input format of the mapper can be implemented to selectively fetch triples only matched to the corresponding pattern. Finally, we can combine rules having a common conclusion (e.g., rdfs 2, 3, and 9) and implement one pair of a mapper and a reducer to efficiently eliminate duplicates derived from different rules.

4 Experiments

To demonstrate the feasibility of the proposed materialization approach, we first checked memory usages according to materialization. Fig. 1 shows that the memory without materialization is exhausted quickly even though the size of data is quite small, while the memory with materialization is consumed smoothly. We also compared the elapsed time in reasoning with and without materialization using LUBM[11]. The result in Fig. 2 shows that materialization improves the reasoning remarkably and even makes it linear to the size of data.

Especially, rete-reasoning without materialization consumed and exhausted memories very quickly even for small size of data. However, materialization solved this issue effectively by maintaining only schema triples in memory and leaving reasoning of instance rules to Hadoop framework.

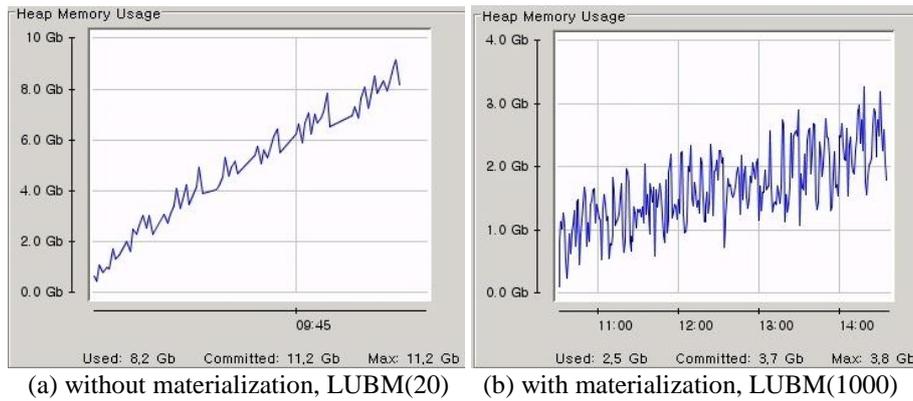


Fig. 1. Memory usages with and without materialization

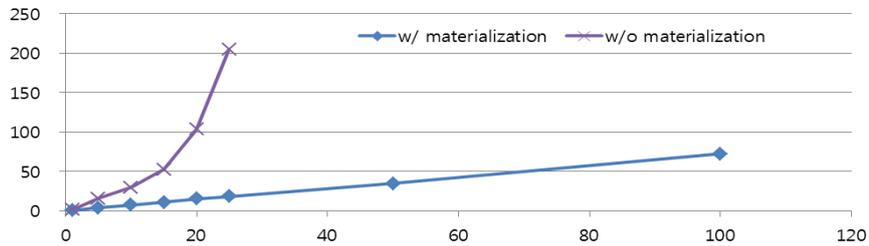


Fig. 2. The elapsed time in reasoning with and without materialization according to the size of data (LUBM)

5 Conclusion

This paper explained a rete-based reasoning method that materializes RDFS and OWL-Horst rules when an ontology schema is given and then can combine with Hadoop framework to deal with billions of triples. Each generic instance rule is materialized into several schema-specific rules, which can be implemented in a pair of a mapper and a reducer. Each mapper can selectively fetch input triples matched to its pattern using Hbase and rules having a common conclusion can be combined into a reducer to efficiently eliminate duplicate derivations from different rules.

The combination with Hadoop framework is being implemented and will be tested to check how much our method could improve reasoning performance, comparing to WebPIE[1] in near future.

Acknowledgement

This work was supported by the IT R&D program of MSIP/KEIT. [2014044034002, High performance database solution development for integrated big data monitoring and analysis]

References

1. Urbani, J., Kotoulas, S., Maassen, J., Harmelen, F., Bal, H.: WebPIE: A Web-scale Parallel Inference Engine using MapReduce. *Journal of Web Semantics: Science, Services and Agents on the World Wide Web* 10, 59--75 (2012)
2. Forgy, C.L.: Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. *Artificial Intelligence* 19(1), 17--37 (1982)
3. Doorenbos, R.B.: Production Matching for Large Learning Systems. Ph.D Thesis, Carnegie Mellon University, Pittsburgh, PA (1995)
4. Özacar, T., Öztürk, Ö., Ünalir, M.O.: Optimizing a Rete-based Inference Engine using a Hybrid Heuristic and Pyramid based Indexes on Ontological Data. *Journal of Computers* 2(4), 41--48 (2007)
5. Lee, S., Jung H., Kim, P., You, B.-J.: Dynamically Materializing Wild Pattern Rules Referring to Ontology Schema in Rete Framework. In *Proceedings of the 1st Asian Workshop on Scalable Semantic Data Processing (AS2DP)* (2009)
6. RDF Semantics, available at: <http://www.w3.org/TR/rdf-mt>
7. OWL Web Ontology Language Semantics and Abstract Syntax, available at: <http://www.w3.org/TR/owl-semantics>
8. Horst, H.J.: Completeness, Decidability and Complexity of Entailment for RDF Schema and a Semantic Extension Involving the OWL Vocabulary. *Journal of Web Semantics* 3(2-3) 79--115 (2005)
9. Apache Hadoop, available at <http://wiki.apache.org/hadoop>.
10. Um, J.-H., Lee, S., Kim, T.-H., Jeong, C.-H., Seo, K., Park, J., Jung, H.: MapReduce-based Bulk-Loading Algorithm for Fast Search for Billions of Triples, In *Proceedings of the 9th KIPS International Conference on Ubiquitous Information Technologies and Applications (CUTE 2014)*, (2014)
11. Guo, Y., Pan, Z., Heflin, J.: LUBM: A Benchmark for OWL Knowledge Base Systems. *Journal of Web Semantics* 3(2), 158--182 (2005)

Analyzing Email Patterns with Timelines on Researcher Data

Jangwon Gim¹, Yunji Jang¹, Do-Heon Jeong^{1,*}, Hanmin Jung¹

¹ Korea Institute of Science and Technology Information (KISTI)
245 Daehak-ro, Yuseong-gu, Daejeon (305-806), South Korea
{jangwon, yunji, heon, jhm}@kisti.re.kr

Abstract. This paper proposes a procedure that easily extracts a feature that helps differentiate between similar researcher names in articles. We examined email patterns and their timelines to identify researchers. Our statistical analysis results show multiple email address usage patterns are found in the case of approximately 43% researchers, and 5% of the patterns are overlapped. Base on the statistics, we conclude that the identification of researchers is still required to enhance performance of the researcher-centric analytics systems and applications.

Keywords: researcher name disambiguation, feature selection, researcher data set, timeline

1 Introduction

With ever-increasing amounts of research data and advancements in technology in big-data environments, a paradigm shift is required. Accordingly, studies on new business intelligence services are being conducted and forecasting and analysis methods are being developed. Prescriptive analytics first appeared in 2013 among several analytical methods and offers diverse strategies for achieving the objectives of and improving business competence. The 2014 Gartner Hype Cycle Special Report predicted that prescriptive analytics will advance rapidly and reach a technology maturity stage within the next ten years¹⁾. InSciTe Advisory is a service developed in 2013 for strengthening researcher research skills by using the 5W1H method with prescriptive analytics [1,2]. The service analyzes a researcher's skill set and provides analytical results by means of the 5W1H method in order to assist a researcher in attaining a role model group. However, exact diagnosis and analysis of researchers is required to provide them with an optimum strategy for reaching their research goals. To achieve this objective, a researcher's basic information as well their research data must be collected completely in order to examine research results and identify fields of study. This ensures that the researcher is properly identified. For example, a

* Co-responding author

¹⁾ <http://www.comworld.co.kr/news/articleView.html?idxno=48181>

researcher's research information is often confused with that of other researchers and thus retrieved together. This happens because of similar full or abbreviated names of researchers. Accurately identifying a researcher is thus difficult. If research results are integrated without accurate identification of the researcher in question, analysis of this researcher and his or her studies can be either overestimated or underestimated. In this study, we propose an accurate data acquisition procedure to properly identify researchers. Our proposed researcher identification method extracts researchers' email usage and timeline patterns. The structure of this paper is as follows. We discuss related studies in Section 2 and describe the feature selection procedure in Section 3. In Section 4, we present and analyze test data and results. Section 5 concludes our study and states avenues for our future research.

2 Related Work

The amount of academic literature published on the World Wide Web is ever-growing [3]. In this environment, researchers spend much time analyzing the following: research fields that are growing rapidly, well-known academic literature in specific research fields, and authors and their work that are most pertinent to their own research. Accordingly, researcher competence strengthening services that provide researchers with the most relevant and desirable information are being widely developed. These services help to ensure accuracy of researcher data and thus a researcher's credibility. Proper identification of researcher data is critical and many studies are being conducted in this area. Such studies on the accurate identification of research data have been published in databases such as DBLP and PubMed, which are popular sites for reviewing and collecting high quantities of researcher data [4,5].

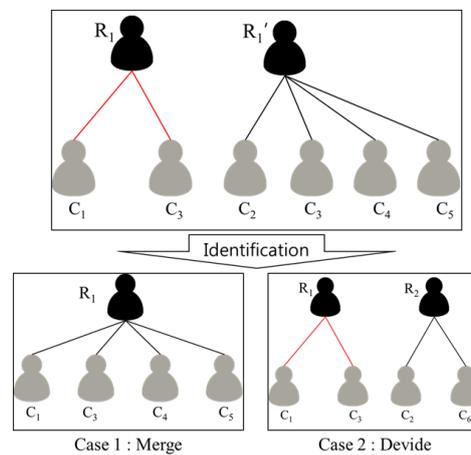


Fig. 1 An example of researchers who has the same name

Figure 1 shows an example of researchers who can have the same name. The authors might be divided into two different researchers or might be merged as a researcher.

Ensuring the accuracy of classification is difficult when the network automatically classifies Researcher 1 as the same person associated with data collected on specific research papers. Therefore, methods for automatically identifying researchers are necessary when largescale literature data is considered. In addition, a correct answer set with high accuracy and an experimental data set are required when researchers conduct studies based on researcher data. To achieve this, accurate identification of a researcher is required for certain works which are part of researcher data. Therefore, currently operating authentication services such as Elsevier SciVal Expert and ORCID are designed so that researchers can provide relevant information directly and manage it by themselves [6,7]. The accuracy of researcher information is improved through these services. However, researcher identification remains a problem when we try to integrate the data provided by these services with previously published data.

Therefore, studies on researcher identification based on researcher meta information extracted from papers and scientific literature data have been proposed. Studies exist that examine the use of researcher email and affiliation information. The study in [8] examines the email content of specific researchers and extracts names for identification purposes. The similarity of names is identified by examining the extracted name and a sentence containing that name. The researchers in [8] performed identification based on email contents, but they did not consider characteristics of email addresses themselves such as character strings. The study in [9] tried to solve disambiguation problems related to author names on the basis of researcher affiliation information. To accomplish this, it proposed the pairwise factor graph (PFG) method. This method generates pairs by randomly combining two papers a researcher has published and attempts to identify the researcher based on similarity information. In addition, it examines the distribution of atomic clusters by using the pairs to compare co-authors with the researcher, affiliation names, and titles of papers. However, identifying the exact author is difficult when another or several researchers exist who have the same name. Labeling Oriented Author Disambiguation (LOAD) method using a machine learning algorithm [10]. In LOAD, data was clustered with Precision Clusters (HPCs) and High Recall Clusters (HRCs). It clustered meta information, which can be extracted from each paper, including email and affiliation, and distinguished a different person with the same name by clustering papers by each author based on HPC. Comparing it to the existing automatic homonymy algorithm, LOAD improved the accuracy of disambiguation issues and can save a time for a human to label a specific cluster. However, it does not consider the timeline information of the features. One of the most important factors for identifying a researcher in researcher data is timeline information [11]. Email address and affiliation information of a researcher can be changed, added, or deleted. Therefore, a researcher's activity history can be tracked if timeline information is used for identifying a researcher [12].

In this paper, we introduce the extracting procedure for certain features, such as email address and affiliation name, which can play an important role in identifying a researcher. Further, we propose an analysis method based on the timeline, and state our experimental results.

3 Feature Selection

3.1 A procedure for selecting email patterns

This chapter explains the feature extraction procedure from researcher data.

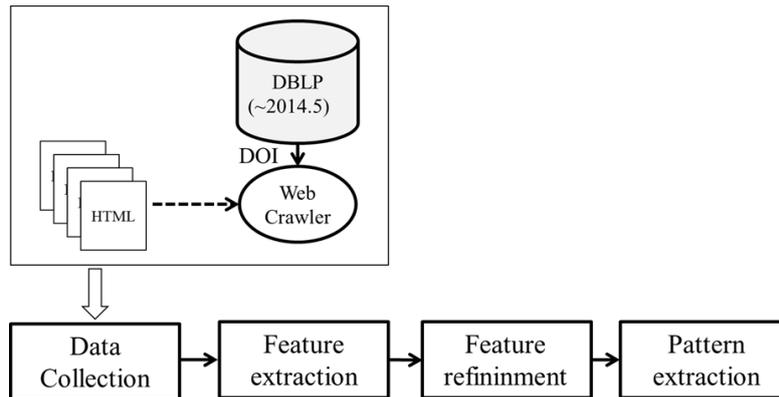


Fig. 2 A feature selection procedure

The feature extraction procedure involves four stages as shown Figure 2. The first stage involves the collection of researcher data; we collect the meta information of the published papers that are on the web to identify disambiguation of a researcher's name. To do so, we collect the Digital Object Identifiers (DOIs) of the papers; using these DOIs, we collect the published information on the pertinent sites. Since the websites where a paper is published are structured in different forms, we develop customized crawlers to collect data, taking into consideration the structure of each web page. The second stage is the feature extracting stage; email addresses of researchers are extracted. During this stage, the year when an email address was generated is extracted together with the email address to obtain the timeline information for the email addresses extracted. The third stage is the refining stage; during this stage, we remove the unnecessary data that may exist in the pertinent feature. For example, in addition to an affiliation name, the address and postal code of that organization are mentioned in some scientific papers; the same organization may be mistaken as different organizations owing to a different address or postal code mentioned in the papers, and therefore, such unnecessary information is removed. The last stage is the pattern extraction stage; during this stage, the unique pattern of a researcher is derived from the extracted pattern information, and this unique pattern can be used to accurately identify a researcher.

3.2 Email patterns of researchers

Figure 3 shows the general pattern of a researcher's email address based on a timeline. If a researcher's email address is considered with the timeline information,

the usage period of the email address can be defined. To estimate this period, the start and end time of a particular email address in use are defined as the time of first usage and the time of the latest usage (the last appearance of the pertinent email address), respectively. For example, more than 2 email addresses appeared for a particular researcher, and the periods during which each email address was used constitute a coprime relationship, i.e., the usage period for both the email addresses do not overlap like Case 1 as shown in Figure 3. The case 2 in Figure 3 shows that email addresses of researchers with the same name are different from each other but the appearance periods are overlapped; in the case, it is necessary to identify the researchers as same or different because it is possible that they are actually different researchers although their names are the same.

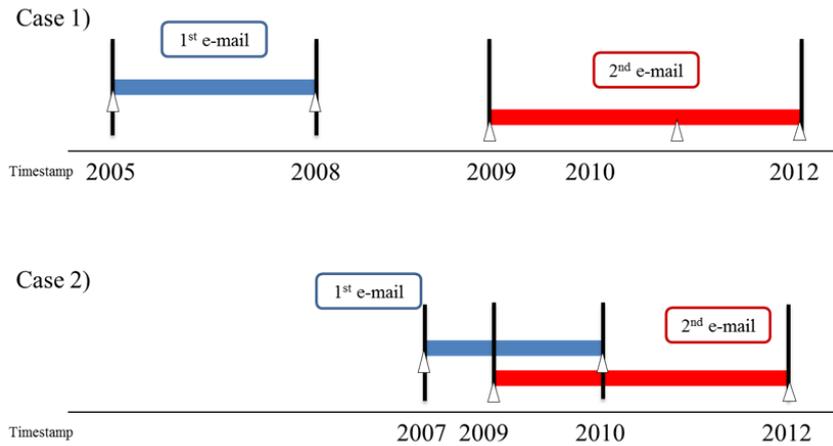


Fig. 3 Two cases of Email address patterns with timelines

4 Statistical Analysis

4.1 Data set

Metadata (title of the paper, the publishing year, coauthors, DOI, etc.) of a paper, as well as the researcher's name, are included in the researcher data extracted by DBLP. However, it does not include the email address and affiliation information of a researcher. Therefore, we implemented an experimental data set according to the procedure explained in Section 3.1. The number of researchers present on DBLP as of September, 2014 is about 1,465,700, and the number of papers is about 4,122,000. To obtain an experiment data set from the pertinent data, we collected website contents using the DOI of papers. The email information is collected automatically by a crawler, so the email addresses of all authors are collected. Therefore, it is not necessary that the n -th email address is the n -th author's email address. Thus, we considered only those email addresses for our experimental data, which had the number of authors equal to the number of email addresses. As a result, 64,802

researchers were extracted for the experimental data. To extract the first author, we compared the pertinent researcher's name with the co-author list, located it at the first instance, and deleted the overlapped name; finally an experimental data set including 18,867 researchers was implemented. We found that these researchers whose names were extracted using the aforementioned process, published 3,790 papers, which is 46.64% of the entire experimental data set.

4.2 Discussions

Through the statistical analysis of the results, we found that the number of overlapped email addresses whose appearance frequency are more than twice is 5.28%. In the data set, 3,162 researchers published at least two papers and a total of 8,126 papers were published by them. We set the minimum number of paper publications at two as a condition because the number of email addresses can be considered as one when the number of paper published is one. Further, 1,371 authors' emails appeared at least twice in the data set. Among the researchers, a total of 167 researchers showed the pattern depicted in the right side of Figure 4, and they published 574 papers (7.06% of the entire experimental data set). The result about the overlapped email address is lower than 2% of the result about the overlapped papers.

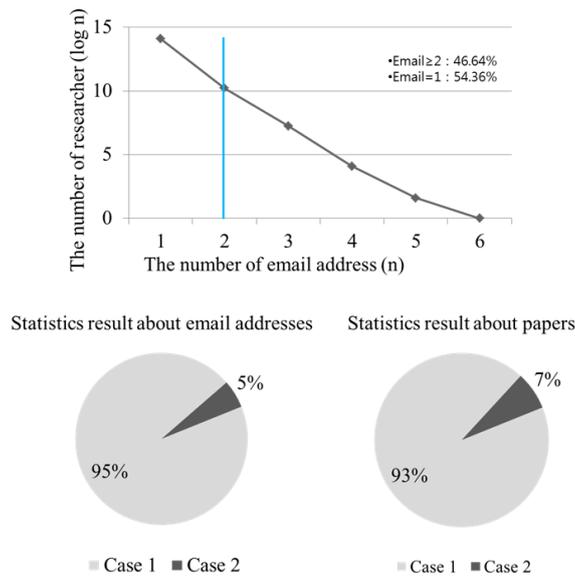


Fig. 4 Statistics results about two cases in terms of email and papers

It means that the productivity of researchers who have overlapped email patterns become high, and we are aware that additional methods are needed to classify them because 43% of the researchers who have more than 2 email addresses.

5 Conclusion and future studies

In the big data environment, academic data are generated at a very fast rate. In this vein, researchers need to quickly obtain and accurately grasp the information presented in studies related to their research, the possible co-author network, and the trend in a specific research field to strengthen researchers' research competence. Accordingly, prescriptive analytics are required for accurate analysis and establishing strategies. However, these services have to be implemented based on accurate data to establish customized strategies for researchers. To this end, the accurate identification of a researcher's name and improvement in the information credibility with regard to researcher data becomes important.

This paper proposed an extraction procedure for important features from researcher data to identify ambiguous researcher's names. To improve researcher identification, we defined the email address usage pattern by considering the timeline characteristic of the researcher's email information and carried out experiments based on the DBLP data set; we verified that our identification method based on email addresses and that considers timeline characteristic is effective, and can be used as an important factor for identifying a researcher. As a future study, we will find a unique pattern representing a researcher by collecting and extracting the affiliation information considering the timeline characteristic from researcher data, and will research on an automated researcher identification system method by applying the obtained pattern to identify researchers; to verify its effectiveness, we will implement an accurately refined data set and compare its performance with the experimental data set.

6 Acknowledgments

This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers].

7 References

1. Sa-kwang Song, Jinhyung Kim, Myunggwon Hwang, Jangwon Kim, Do-Heon Jeong, Seungwoo Lee, Hanmin Jung, Wonkyung Sung, "Prescriptive Analytics System for Improving Research Power," Proceedings of the 16th International Conference on Computational Science and Engineering (CSE), pp. 1144-1145, 2013.
2. Jinhyung Kim, Myunggwon Hwang, Jangwon Gim, Sa-Kwang Song, Do-Heon Jeong, Seungwoo Lee, Hanmin Jung, "Researcher Performance Analysis and Role Model Recommendation Model for Prescriptive Analytics," Proceedings of the Korea Computer Congress 2013 (KCC2013), Vol. 40, No. 2, pp. 241-243, 2013.
3. Madian Khabsa, Clyde Lee Giles, "The Number of Scholarly Documents on the Public Web," PLoS One, Vol.9, No.5, 2014 (DOI: DOI: 10.1371/journal.pone.0093949).

4. Ley, Michael. "The DBLP computer science bibliography: Evolution, research issues, perspectives," *String Processing and Information Retrieval*. Springer Berlin Heidelberg, pp. 1-10, 2002.
5. PUBMED, <http://www.ncbi.nlm.nih.gov/pubmed>
6. Emily Vardell, Tanya Feddern-Bekcan, Mary Moore, "SciVal Experts: a Collaborative Tool," *Medical Reference Services Quarterly*, Vol. 30, No. 3, pp. 283-294, 2011.
7. ORCID, <http://orcid.org/>
8. Einat Minkov, William Weston Cohen, Andrew Ng, "Contextual Search and Name Disambiguation in Email using Graphs," *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, Vol. 29, pp. 27 - 34 , 2006.
9. Xuezhi Wang, Jie Tang, Hong Cheng, Philip S. Yu, "Adana: Active name disambiguation," *Proceedings of the 11th International Conference on Data Mining (ICDM)*, pp. 794 - 803, 2011.
10. Yanan Qian, Yunhua Hu, Jianling Cui, Qinghua Zheng, Zaiqing Nie, "Combining machine learning and human judgment in author disambiguation," *Proceedings of the 20th ACM international conference on Information and knowledge management*, pp. 1241-1246, 2011.
11. Pei Li, Xin Luna Dong, Andrea Maurino, Divesh Srivastava, "Linking temporal records," *Proceedings of the VLDB Endowment*, Vol. 4, No. 11, 2011.
12. Jangwon Gim, Myungwon Hwang, Sa-Kwang Song, Jinhyung Kim, Do-Heon Jeong, Hanmin Jung, "Researcher History Tracking Service for Prescriptive Analytics based on Researcher Activities," In *Journal of KIISE : Computing Practices and Letters*, Vol. 20, No. 6, pp. 0359-0363, 2014.

Development of Framework System for Managing the Big Data from Scientific and Technological Text Archives

Mi-Nyeong Hwang¹, Myunggwon Hwang¹, Ha-Neul Yeom^{1,4},
Kwang-Young Kim², Su-Mi Shin³, Taehong Kim¹, and Hanmin Jung^{1,4,*}

¹Dept. of Computer Intelligence Research, ²Dept. of Overseas Information,
³Dept. of NDSL Service, Korea Institute of Science and Technology Information, Korea
⁴Korea University of Science and Technology, Korea
{mnhwang, mgh, lucetesky, glorykim, sumi, kimtaehong, jhm}@kisti.re.kr

Abstract. In today's era of big data, increasing attention is being paid to the relationships among different types of data, and not just to those within one type of massive data, while processing and analyzing these data. To analyze and predict the trends of technologies from literatures on the basis of the conventional form of textual documents, such as academic papers or patents, the objects of analysis should include recent information collected from news websites and social media sites, which indicate the user preferences. It is necessary to systematically collect multiple texts to integrate and analyze different types of data. This study introduces practical ways to implement a database on the basis of the global standard using the unstructured information management architecture (UIMA).

Keywords: Big data, Text Big data, Scientific and Technological Text, Text Crawling System, Web Crawler, SNS Crawler, UIMA

1. Introduction

At the 2012 World Economic Forum Annual Meeting in Davos, Switzerland, the big data processing technology was highlighted as the “most important scientific technology of the year” [1]. According to IBM, 80% of big data are scattered and unstructured and hence, cannot be managed as structured data. To process these scattered data, we need to develop a new method of collecting and analyzing data [2]. Businesses were the first to understand the value of the analysis of unstructured data. This analysis has now been expanded from conventional data such as those from academic papers, patents, and magazines to data extracted from news websites and social network services such as Twitter and Facebook in order to build business intelligence [3]. Such an expanded application of data analysis is utilized to analyze and predict the trends of the advances of scientific technology [4].

A big data processing platform consists of the four steps of collection, storage, analysis, and visualization [5]. To maximize the application of analysis and visualization, a thorough collection and an appropriate storage of big data are needed.

* Corresponding Author.

This paper explains the collection of big data related to scientific technology from different sources; this process is necessary to analyze the trends of scientific technology. It is expected to help researchers to improve their research and to better implement a database.

2 Related Work

The introduction and penetration of the World Wide Web has led to an increasing effort to crawl data from documents on the Web [6]. As the size of the Web increases, more studies are being conducted on the collection of documents on certain topics from the Web than on their storage in one place [7,8]. There have been efforts to develop a crawling process of scattered data to collect large documents. In this case, there are certain disadvantages related to the sorting of overlapped data and to data storage and management [9]. Furthermore, there has been research on the geographical partition of the server storing the original documents, focusing on the speed of the scattered crawlers of mass data [10]. However, a more macroscopic approach is needed in this era of big data where multiple sources of unstructured data are scattered because these crawlers focus on optimizing the crawling performance for one type of data. This paper suggests a framework for collecting multiple texts on scientific technology.

3 Scientific Technology Text Crawling System

Figure 1 shows the process of implementing the data collection system for texts related to scientific technology. Unstructured data related to scientific technology from academic papers, patents, Wikipedia, news websites, and social media were first collected as raw HTML, XML, and text data.

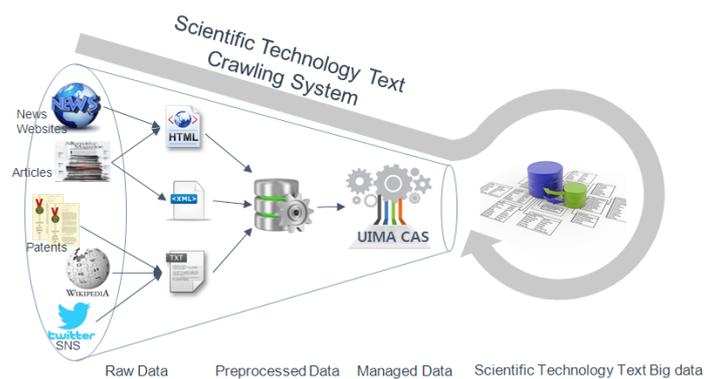


Fig. 1. Flow of scientific technology text crawling system

Metadata, which are needed to analyze papers related to scientific technology, were extracted from the collected data through this preprocessing procedure. The metadata were transformed into the common analysis structure (CAS) format of the unstructured information management architecture (UIMA) and were processed for implementing the data of big data texts on the basis of the global standard.

3.1 News Websites

News websites are an important source for collecting the latest news on scientific technology considering the fact that there is a gap between the time of research and the publication of academic papers and patents. 154 websites, whose services include news, magazines, and forums, such as Scientific Computing¹, ScienceNews², and Bioscience³, were selected to collect the latest news on scientific technology. Data from news websites were collected using three types of crawlers, namely Google crawlers, RSS crawlers, and direct crawler, as shown in Figure 2, because they contain news published since 2001.

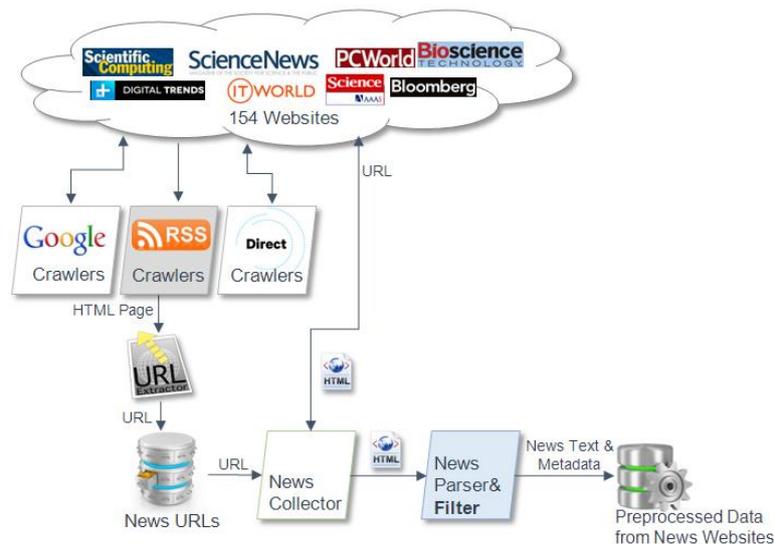


Fig. 2. Process of crawling news websites

News articles, which were already published in 2001, were collected by using both Google crawlers and direct crawlers that collect data from the websites. In the case of a website that is run by a keyword-based search engine, the direct extraction process showed the search results when keywords related to scientific technology were used. From a website that shows lists of news articles, the crawlers extracted the links of the news. From websites that provide an RSS service, the crawlers collected links of real-

¹ <http://www.scientificcomputing.com/>

² <http://www.sciencenews.org/>

³ <http://www.biosciencetechnology.com/>

time news in a parallel manner. These collected URLs were stored in the URL database to avoid overlapping data. Then, the News Collector collected the HTML code of the news items through the URLs, and the News Parser extracted the title, author, date, category, and the content from the HTML code. The News Filter eliminated the overlapping and irrelevant news.

3.2 Articles

In the case of collecting foreign papers from websites such as IEEE and NCBI PubMed, data were directly collected from the websites by using the news website crawler. Meta information, which includes the title, author information, keywords, and abstract of a paper, was collected in this manner. The objects of the real-time data collection included data published between 2001 and 2014.

3.3 Patents

Patent data are also needed to analyze the levels of originality and scientific progress. Patent data released internationally, particularly in the US and Europe, and registered in the US between 2001 and 2013 were collected in bulk. The metadata and abstracts from these data were used in this study.

3.4 Wikipedia

Wikipedia¹ is an Internet encyclopedia, whose contents are created directly by the users. Here, a uniform resource identifier (URI) is assigned to every piece of information and DBpedia², which provides the related meta information, is downloaded to implement the database.

3.5 Social Network Service Data

Along with data from papers, patents, news websites, and Wikipedia, data from social network services were collected. Among the social media contents, we collected tweets. Tweets that included 213 keywords on scientific technology, such as web, computer, and smartphone, were collected real-time using OpenAPI released by Twitter³. Data from 2014 were collected, and on average, about 700,000 tweets were extracted daily. Punctuation marks were not eliminated in the preprocessing step, and the entire contents were stored intact as tweets in general express user emotions.

¹ http://en.wikipedia.org/wiki/Main_Page

² <http://wiki.dbpedia.org/Downloads2014>

³ <https://about.twitter.com/what-is-twitter/>

4 Implementation of Database Based on the Global Standard Using UIMA

Documents collected by the Scientific Technology Text Crawling System were stored in the CAS format of UIMA after the preprocessing procedure. UIMA is an open Apache source project that defines the common systematic structures of software that analyzes large volumes of unstructured information in order to discover knowledge that is relevant to an end user. CAS is the defined form of structures that express the feature and annotation used in UIMA. CAS is redefined and used for meeting the characteristics of the collected metadata information. As information, which is collected in the CAS format of UIMA, is expressed as the structure of the global standard, it can be used as the input data for the engine that extracts unstructured information using UIMA. The text documents collected and preprocessed in the CAS format through this study were transmitted to the Hadoop-based information extraction system [12].

Table 1. Status of archiving big data from text related to scientific technology

Type of scientific technology text	Number of documents
Web News	5,656,465
Article(Korean)	962,984
Article(English)	13,744,480
Patent	9,427,117
Wikipedia	4,004,000
Tweet	204,445,587

Table 1 shows the current status of the database data that have been collected and implemented thus far. 5,656,465 documents were collected from news websites, including articles published between 2001 and 2014. Patent data published between 2001 and 2013 were collected in bulk. About 200 million tweets posted since January 2014 were collected. Articles from news websites, foreign papers, and tweets were collected real-time.

5 Conclusion

This study analyzed a practical method of collecting multiple types of big data texts related to scientific technology and of implementing a database. The system collected various types of information, such as patents, data from news websites, Wikipedia content, and social media content, and systematically implemented a text database and distributed it in the CAS format of UIMA, the global standard format.

In the future, we intend to study ways to apply the characteristics of the real-time data collection system, which is currently being applied only to the crawlers of Web news, foreign papers, and social media, to other types of information. Text information will also be used for analyzing and predicting the trends of scientific technology, which is necessary to help researchers to improve their research.

Acknowledgments

This work utilized scientific and technical contents constructed as part of the “Establishment of the Sharing System for Electronic Information with Core Science and Technology” project (K-14-L02-C01-S02) and the “S&T Contents Construction and Service” project (K-14-L02-C01-S04). We would like to thank the Department of Overseas Information and Department of NDSL Service for providing certain contents used in this study.

References

1. Big Data, Big Impact: New Possibilities for International Development, http://www3.weforum.org/docs/WEF_TC_MFS_BigDataBigImpact_Briefing_2012.pdf
2. IBM, Innovate with new analytics tools and technologies, <http://www.ibm.com/analytics/hk/en/what-is-smarter-analytics/innovate-with-analytics-tools.html>
3. Chen, H., Chiang, R., Storey, V.: Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, vol. 36, no. 4, pp.1165–1188 (2012)
4. Hwang, M., Cho, M., Hwang, M., Lee, M., Jeong, D.: Technical Terms Trends Analysis Method for Technology Opportunity Discovery. *INFORMATION-AN INTERNATIONAL INTERDISCIPLINARY JOURNAL*, vol. 17, no. 3, pp. 877–883 (2014)
5. Ferguson, M.: Architecting A Big Data Platform for Analytics. A Whitepaper Prepared for IBM (2012)
6. Burner, M.: Crawling towards eternity: Building an archive of the World Wide Web. <http://www.webtechniques.com/archives/1997/05/burner/> (1997)
7. Soumen, C., Berg, M., Dom, B.: Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, vol. 31, pp.1623–1640 (1999)
8. Aggarwal, C.C., Al-Garawi, F., Yu, P.S.: Intelligent crawling on the World Wide Web with arbitrary predicates. *Proceedings of the 10th international conference on World Wide Web*, pp.96–105 (2001)
9. Hafri, Y., Djeraba, C.: High performance crawling system. *Proceedings of the 6th ACM SIGMM international workshop on Multimedia information retrieval*, pp. 299–306 (2004)
10. Exposto, J., Macedo, J., Pina, A., Alves, A., Rufino, J.: Geographical partition for distributed web crawling. *Proceedings of the 2005 workshop on Geographic information retrieval*, pp. 55–60 (2005)
11. Apache UIMA, <https://uima.apache.org/>
12. Um, J., Jeong, C., Choi, S., Lee, S., Jung, H.: Fast Big Textual Data Parsing in Distributed and Parallel Computing Environment. *Mobile, Ubiquitous, and Intelligent Computing*, pp.267–271 (2014)

Business Event Extraction System Based on SSVM

Sungho Shin, Young-Min Kim, Choong-Nyoung Seon,
Seunggyun Hong, Sa-kwang Song*, Hanmin Jung

Dept. of Computer Intelligence Research, KISTI
245 Daehak-ro, Yuseong-gu, Daejeon, 305-806, South Korea
{maximus74, ymkim, wilowisp, xo, esmallj, jhm}@kisti.re.kr

Abstract. Information extraction from unstructured text data has been used essentially to provide new insights by collecting, storing, and analyzing text data in textual analysis. The research on event extraction has been recently getting more attention in information extraction area since lots of events happens and significantly affect our societies and countries. Related studies on event extraction use rules for identifying and extracting events from texts so far. However, rule based approaches have a limitation in terms of accuracy and rule construction. In this paper, we present an event extraction system which takes advantage of the machine learning method, especially SSVM. The system extracts event triggers and predefined event arguments, while existing rule based systems extract unknown event arguments. Ours provides 60.23 F1 score, which is higher than that of previous researches, in which rule based event extractions were performed. Even though rule-based and machine learning-based approaches cannot be compared against each other completely fairly, what is clear is that for the task in which event arguments are defined in advance, applying machine learning method can make better results.

Keywords: Event Extraction, Trigger, Argument, Temporal Information, SSVM, Machine Learning

1 Introduction

Recently, research on event extraction has been conducted to rapidly discover meaningful events with regard to business that are found in a massive quantity of data. In terms of meaningful information existing in text, an event might be viewed as a type of group that qualifies for information to be extracted. Data obtained from named entity recognition, relationship extraction, and event extraction can be employed as fundamental data for more detailed data analysis and intelligent services, such as natural language questioning and answering (NLQA) or predictive analytics (PA) which are emerging technologies presented in Gartner Hype Cycle [1].

Lots of researches have been conducted on named entity recognition and relation extraction, and the results have reached over 90% of human cognitive capacity [2]. In addition to the extensive works that has been performed in these areas, research on event extraction has attracted much attention recently. In particular, the prediction of

* Corresponding author

influenza spread and the assessment of the moving direction or damage status of a typhoon or tsunami, which are natural disasters, as performed by Google are the results of analyzing the extracted events. Similar to the early stages of named entity recognition and relation extraction, the research on event extraction, being still in the initial stage, has been mainly realized by rule-based methods. Known from researches on named entity recognition or relation extraction, rule-based methods have advantages of cost and time efficiency, although it considerably lacks in precision when being compared to machine learning methods. Accordingly, event extraction based on machine learning might hold greater benefits to services provided to users in actual business settings through NLQA or PA.

This study relates to event extraction as source data required to expand the knowledge database used for NLQA or PA. We aim to accurately extract a trigger and arguments of an event by using a machine learning method with applying a structural support vector machine (SSVM) algorithm. Furthermore, temporal information, another event argument, can be extracted more effectively by applying rules from a text or metadata rather than by applying the machine-learning method because of its various forms of expression. We designed and implemented a module for extracting temporal information.

2 Related Work

An event is real temporal data, whose temporal sequence is important and might be infinitely long [3]. An event can be defined based on the structure of “who, when, what, where, how (5W1H),” but could consist of only some of these. In addition, occurrences among compounds in the field of biology and data transmission/reception among computer devices are also referred to as events. In particular, researchers define social events as events that occur in a relationship between people and in a society. Social events include a wide range of accidents or incidents that affect society at a small or large scale [4]. Earthquakes, typhoons, traffic congestion, dialogues, accidents, international academic conferences, etc., are considered social events. Business events such as company mergers, product launches, and company bankruptcy are included as well. A trend of event extraction research is more inclined to comprehensive event extraction research than to business event-focused research.

A variety of techniques, such as language processing, text mining, data mining, and machine learning, are widely adopted to extract events from a large quantity of text. Extraction methods are largely divided into a machine-learning method and a rule-based method. The machine-learning method sets the structure of an extraction event in advance, constructs a learning group based on the determined structure, and extracts the event. This method is advantageous for extracting events with a predetermined structure. On the other hand, although the rule-based method also defines the structure of an event in advance, the structure of the event can subsequently be changed easily, thus permitting freely reflecting feedback from the extraction result. However, there exists a limitation to the rule-based method because of a significant

number of exceptions. Accordingly, the machine-learning method might be more efficient in extracting events with a predetermined structure.

3 Event Expression

This study focuses on business event extraction reported in the literature. A business event indicates an event arising from corporate activities, and the major concepts pertaining to this type of event are as follows [5]:

- **Event mention:** an event type. Common designation of events that have an identical meaning.
- **Event trigger:** the most exact term for expressing an event mention.
- **Event argument:** an event attribute associated with an event trigger, such as an entity mention, a temporal expression, or value.
- **Event instance:** an instance comprised of event arguments associated with an event trigger that is referring to an event mention presented in an identical sentence.

According to the above definition, the event in this study includes an event trigger, event subject, event object, and time. The event subject, event object, and time correspond to an event argument. An event is expressed as shown in (1).

Event mention <event trigger, subject (arg.1), object (arg.2), time (arg.3) > (1)

In this study, business events are limited to ‘Announce’ and ‘Launch’. The subjects of two events are corporation names and the objects are product names which are highly related to the companies.

4 Machine Learning based Event Extraction System

In this study, an event trigger and arguments, with the exception of temporal information, are extracted by the machine learning method. Temporal information is extracted by the rule-based method rather than by the machine learning method because such information can be expressed in various ways.

The machine learning algorithm used is the structural support vector machine (SVM) algorithm. The structural SVM algorithm is a machine learning algorithm extended from the conventional SVM algorithm. A structural SVM supports more general structural issues, for example, sequence labeling, parsing, etc., whereas a conventional SVM supports binary classification and multi-class classification, among other classification methods.

In addition, for the learning of the structural SVM algorithm, this study uses the Primal Estimated sub-GrAdient Solver for SVM (PEGASOS) algorithm among the Stochastic Gradient Descent (SGD) method as an extension of the structural SVM algorithm because the PEGASOS algorithm has shown high performance and rapid learning rate when applied to the SVM.

To extract an event by applying the extended structural SVM algorithm, training data is required. Training data can be directly constructed manually by domain experts, but because of limited time and labor, it is preferred to make it automatically. We also built it semi-automatic method [6]. The initial training data is constructed by using a simplified distant supervision method which is an automated method. For the method, seed data lists for named entity and for relation are respectively established first; Seed data is then used for key word searching to collect training sentences; sentences including the corresponding key words are extracted from the web; and a Silver standard training data is established.

Later, domain experts are hired to establish the gold standard through manual verification in order to enhance accuracy. The training data was prepared for two event types and two relation types which are just for the system evaluation. The number of training data of each type is presented in table 1.

Table 1. Types of relation & event and # of training sentences

Category	Type	Object*	# sentences
Relation	ElementOfTechnology	Technology name	1,031
	CompeteProduct	Product name	537
Event	Announce	Product name	2,448
	Launch	Product name	1,986
Total			6,002

With the machine learning, an event trigger, an event subject, and an event object are extracted; temporal information is extracted through an additional designed tool. Fig. 1 illustrates the process of extracting the event temporal information. Each module that comprises the tool is described as follows.

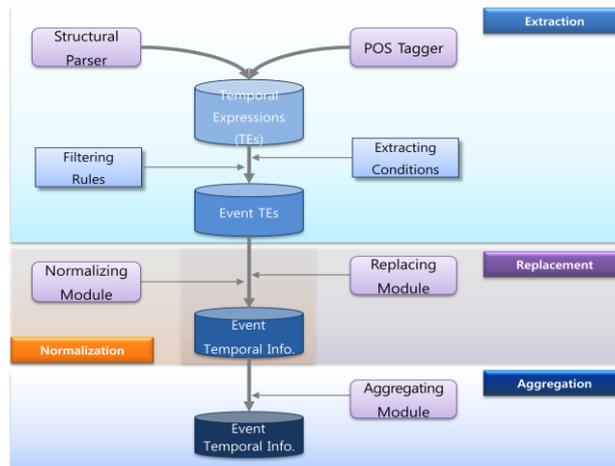


Fig. 1 Process of Capturing Temporal Information

<Extraction>

- Uses the results of parsing and POS tagging on each sentence
- Extracts temporal information from all time expression adverbs within a sentence by applying temporal information extraction conditions
- Filters or refines extracted time expression candidates according to purpose
- Excludes useless time expressions by filtering
- Deletes tokens not comprising time, such as prepositions, articles, and more, from the time information

<Replacement>

- Replaces or complements an accurate point of time by using metadata with respect to temporal information that is inferable by the metadata
- Applies temporal information replacement rules to the temporal information to be replaced

<Normalization>

- Classifies temporal information based on year-month-day combination
- Changes unstructured temporal information to a normalized form
 - * Normalized form: YYYY-MM-DD
- Expresses '0000' or '00' when lacking information in the normalized form

<Aggregation>

- Assumes that an identical triple (S-P-O) has identical temporal information
- Compares each temporal information by collecting identical triples
- Aggregates temporal information lacking in the triples

5 Experiments

The evaluation of the system which is used for event extraction in this study is performed with F1 score for two business relations. The result of the evaluation is presented in table 2. Our system is well performed in information extraction job, especially relation extraction which is similar with event extraction.

Table 2. Performance of our system for relation extraction

Types	Precision	Recall	F1 score
ElementOfTechnology	87.50	89.74	88.61
CompeteProduct	90.91	58.82	71.43
Total	89.20	74.28	80.02

The test data is 10% of the entire training data (table 1), and it is randomly sampled. The result of business event extraction is shown in table 3. Even though the F1 score of each event type is different, they are all over 60.0. The F1 score of our system exceeds that of other systems [5, 7] which are based on rules (table 3). This is not absolutely true because there is a limitation that the experiment environment is not same. However, what is clear is that for the task in which trigger and arguments are fixed in advance, applying the machine learning method can make better results.

Table 3. Result of Business Event Extraction

Types	Precision	Recall	F1 score
Announce	63.18	58.53	60.77
Launch	58.76	60.45	59.68
Total	60.97	59.59	60.23

Table 4. Comparison of our system and others

Research	Trigger F1	Arg F1	Average	Method
Ji and Grishman (2008)	67.30	42.60	54.95	Rule-based
Qi et al (2013)	67.50	52.70	60.10	Rule-based
Ours	-		60.23	ML

6 Conclusion

This study aims to extract business events by using machine learning based information extraction system. Different with events defined in other researches, arguments of an event is predefined. Thus, our system is only capable of extracting designated arguments in advance. For extracting the temporal information, one of event arguments, we use rules. The result of system evaluation says that our system accomplishes better performance than other rule based system in terms of F1 score. This means that for the event extraction in which the arguments are defined ahead, machine learning based method makes better results.

Acknowledgement. This work was supported by the IT R&D program of MSIP/KEIT. [2014-044-024-002, Developing On-line Open Platform to Provide Local-business Strategy Analysis and User-targeting Visual Advertisement Materials for Micro-enterprise Managers]

References

1. Gartner Inc., "Hype Cycle for Emerging Technologies, 2014," <http://www.gartner.com/technology/research/hype-cycles/>, 2014.
2. Bach, N., Badaskar, S., "A Review of Relation Extraction," 2007.
3. Complex Event Processing, http://en.wikipedia.org/wiki/Complex_event_processing.
4. Sakaki, T., Okazaki, M., Matsuo, Y., "Earthquake shakes Twitter users: Real-time event detection by social sensors," In Proceedings of the 19th International Conference on World Wide Web, pp. 851–860, 2010.
5. Li, Q., Ji, H., Huang, L., "Joint Event Extraction via Structured Prediction with Global Features," In Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics, 2013.
6. Shin, S., Choi, Y. S., Song, S. K., Choi, S. P., Jung, H., "Construction of Test Collection for Automatically Extracting Technological Knowledge," Journal of Korea Content Society, vol.12, no.7, 2012. (in Korean)
7. Ji, H., Grishman, R., "Refining event extraction through cross-document inference," In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, pp. 254-262, 2008.

**The 2nd International
Workshop on Linked Data
and Ontology in Practice
(LDOP 2014)**

The 2nd International Workshop on Linked Data and Ontology in Practice (LDOP 2014)

Workshop Organizers

- Hideaki Takeda (National Institute of Informatics, Japan)
- Jason J. Jung (Chung-Ang University, Korea)
- Kouji Kozaki (Osaka University, Japan)
- Thepchai Supnithi (NECTEC, Thailand)
- Marut Buranarach (NECTEC, Thailand)

Program Committee

- Ryutaro Ichise (National Institute of Informatics, Japan)
- Shinichi Nagano (Toshiba Corporation, Japan)
- Naoki Fukuta (Shizuoka University, Japan)
- Takahira Yamaguchi (Keio University, Japan)
- Kanda Runapongsa Saikaew (Khon Kaen University, Thailand)
- Nopphadol Chalortham (Chiang Mai University, Thailand)
- Athitaya Nitchot (Prince of Songkla University, Thailand)

Invited Talk

- Dumitru Roman (SINTEF, Norway)

LDOP 2014

Preface

The 2nd International Workshop on Linked Data and Ontology in Practice (LDOP 2014) is the second version of the Joint International Workshop: 2013 Linked Data in Practice Workshop (LDPW2013) and the First Workshop on Practical Application of Ontology for Semantic Data Engineering (PAOS2013) that took place at the 3rd Joint International Semantic Technology Conference (JIST2013). LDOP 2014 is co-located with the 4th Joint International Semantic Technology Conference (JIST2014). The workshop focuses on three related themes: Linked Data and Ontology, Linked Data in Practice and Ontology in Practice. Five submissions were selected and are included in the proceedings. Each submission was reviewed by at least two program committee members. The organizers would like to thank the program committee members for their contribution in carefully reviewing the submissions. We also would like to thank Dr. Dumitru Roman for his invited talk for the workshop. Lastly, we would like to thank JIST 2014 organizers who helped making this event happens.

November 2014

Hideaki Takeda

Jason J. Jung

Kouji Kozaki

Thepchai Supnithi

Marut Buranarach

LDOP Workshop Organizers

Invited Talk

The DaPaaS Platform: Data-as-a-Service Solution for Open Data

Dumitru Roman

SINTEF, Forskningsveien 1a, Oslo, Norway

Abstract: Data-as-a-Service (DaaS) is emerging as a new paradigm for cost-effective and agile data provisioning, promising to simplify data management for organizations with limited expertise in the field, and to reduce the costs for data integration, publishing and consumption. This presentation will provide a brief overview of existing relevant DaaS solutions, with a focus on Open Data and the DaPaaS Platform (<http://dapaas.eu/>) --- an emerging open DaaS solution for Open Data. The goal of this presentation is to familiarize the audience with the DaaS concept, the emerging solutions in this domain, and to provide an overview of the DaPaaS platform.

Bio: Dr. Dumitru Roman works as a Senior Research Scientist at SINTEF ICT, Oslo, Norway. He is currently involved in research around simplification of data access, integration, and data publication and consumption at large scale. He currently acts as the DaPaaS project coordinator (<http://dapaas.eu>) – a project developing a Data- and Process-as-a-Service for simplifying access to data and data-intensive applications. He holds an adjunct associate professorship at the University of Oslo, Norway.

A Framework of Automatic Alignment of Concept in Ontology with Confidence Score based on Inner Concept Information

Panadda Jaiboonlue¹, Supot Nitsuwat¹, Wasan Na Chai²,
Prasert Luekhong³, Taneth Ruangrajitpakorn² and Thepchai Supnithi²

¹Faculty of Information Technology

King Mongkut's University of Technology North Bangkok, Thailand

pjaiboonlue@gmail.com, sna@kmutnb.ac.th

²Language and Semantic Technology Laboratory

National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand

{wasan.na_chai, taneth.rua, thepchai.supnithi}@nectec.or.th

³College of Integrated Science and Technology

Rajamangala University of Technology Lanna, Chiang Mai, Thailand

prasert@rmutl.ac.th

Abstract. This paper presents a framework to align synonymous concepts of multiple ontologies. It applies the information attached to the concept including label, description and property relations. Label is a feature to consider for likeness of concept's name which can be the same, partial alike, or totally different. Description is an optional feature in case the given definition of the concepts is similar. Properties of the concept are the major feature to indicate the equivalent relation of the concepts to another concepts and their datatype. After the equivalent concepts are assigned, confidence score is calculated to provide a confidence value of the alignments. From the result, the system gains the impressive result as it can align synonymous concepts as same as the manual mapping concept list.

Keywords: ontology alignment, ontology matching, sericulture

1 Introduction

Ontological products become more popular nowadays due to ontology advantage [1] such as re-usability, interoperability of human and machine, etc. and several supported tools for ontology development such as ontology editor [2][3], inference engine [4], etc. Hence, there are many implemented ontologies using in active researches at the moment. From the observation, several ontologies in the same topic were developed and published freely for re-using and knowledge-sharing. However, ontology and ontology-based system developers often ignore the existing ontologies and decide to design and develop their own ontology since the scope of ontology of the same topic is slightly or subjectively different. This leads to the increasing number of

several new ontologies in the same topic and the reusability and extendable benefit of ontology cannot practically be explicit as claimed.

Creating a new ontology is not the hardest part in the development process, but to include the well-designed class and properties of the existing one. Normally, ontology developers review the existing relevant ontologies in the topic as a reference to overcome the weakness or fill out their own interesting scope. However, the reference ontologies can contain a large number of non-relevant concepts and their relations, and as aforementioned, the number of ontologies in such topic can be numerous. Hence, the assisting tools to help on finding out the classes in ontologies can be useful to indicate the interesting concepts. Moreover, to review many ontologies in the same topic can be helpful on comparing the coverage and missing applicable concepts, but the number of concepts to examine can be greatly burden to reviewers or developers who want to extend existing ontologies.

For comparing several ontologies, it is simple to acknowledge the equivalent class with the same or similar label. However, there are the cases which are 1) ontological classes are equivalent in different label, and 2) classes refer to different concepts with the same label. These issues require much knowledge and understanding in the field from the readers.

Since the ontological concepts include the essential attributes such as concept label, properties of concepts and hierarchical structures, the mentioned information is the hint to inform the likeness of ontological concept. The more similar the information is, the more likely those concepts are synonym to each other. Thus, we use the information as a clue to identify the likeness of concepts from between ontologies to develop the assisting tools to align synonym concepts.

In this paper, the rest is organised as following. Section 2 gives information on related work on existing concept alignment systems. Section 3 provides the methodology of the proposed framework. Experiment setting and results are given in Section 4. Section 5 is filled with discussion over the results and methodology. Section 6 concludes the paper and lists the plan for future development.

2 Related work

This section shows the existing matching ontology approaches. Several matching application were proposed such as SAMBO[5], Falcon[6], DSsim[7], RiMOM[8], ASMOV[9] and Anchor Flood [10]. The efficiency of those approaches were described in Table 1.

Table 1. Analytical comparison of the recent matching systems from[11]

System	Input	Output	Terminological	Structural
SAMBO	OWL	1:1 Alignments	n-gram, Edit distance, UML, WordNet	Iterative structural similarity base on <i>is-a, part-of</i> hierarchies
Falcon	RDFS,OWL	1:1 Alignments	I-SUB, Virtual Documents	Structural proximities, Clustering, GMO
DSsim	OWL,SKOS	1:1 Alignments	Tokenization, WordNet Monger- Elkan, Jaccard	Graph similarity base on leaves
RiMOM	OWL	1:1 Alignments	Edit distance, WordNet Vector distance	Similarity Propagation
ASMOV	OWL	n:m Alignments	Tokenization, WordNet String Equality, UML Levenstein distance	Iterative fix point computation, hierarchical, Restriction similarities
Anchor Flood	RDFS,OWL	1:1 Alignments	Tokenization, WordNet String Equality, Winkler-base sim.	Internal/External similarities, Iterative anchor-based similarity propagation
Agreement Maker	XML,RDFS, OWL,N3	n:m Alignments	TF IDF, Edit distance, Substring, WordNet	Descendant, sibling similarities
Propose System	OWL	1:1 Alignments	String similarity, WordNet, Description	Structural Properties

Those approaches are considered the fine systems as they were publically used and tested in several ways. They have their own advantages and disadvantages as shown in Table 1. However, they can decide the matching of the concepts with their criteria, but none of them can give a confident reason to endorse their decision. To solve such problem, we propose a concept matching system which provides the automatic matching of synonym concepts with confidence score in this paper.

3 Methodology

In this work, the new alignment process for matching synonym concepts between the multi-ontologies is proposed. The system employs the alignment function to identify the synonymy concepts between two ontologies based on the information within a concept. The features to identify synonymy concepts consist of four parts: label of the concept, description of a concept, object and data properties of a concept, and the hierarchical structure of a concept. All features together are used as a measurement to determine the semantic relation that holds between two concepts that express the same meaning. Each feature alone, such as label, cannot conclude the synonymy result since a label is an apparent concept name which can ambiguously be polysemy. These features are a certain hint to scope the synonym and similarity to each other concept among several ontologies. As aforementioned features, the framework is designed into five modules to handle each feature separately and to sum up the simi-

larity score. The expected result is the list of concepts which are synonymy in the different ontologies. The overview of the proposed framework is illustrated in Fig. 1.

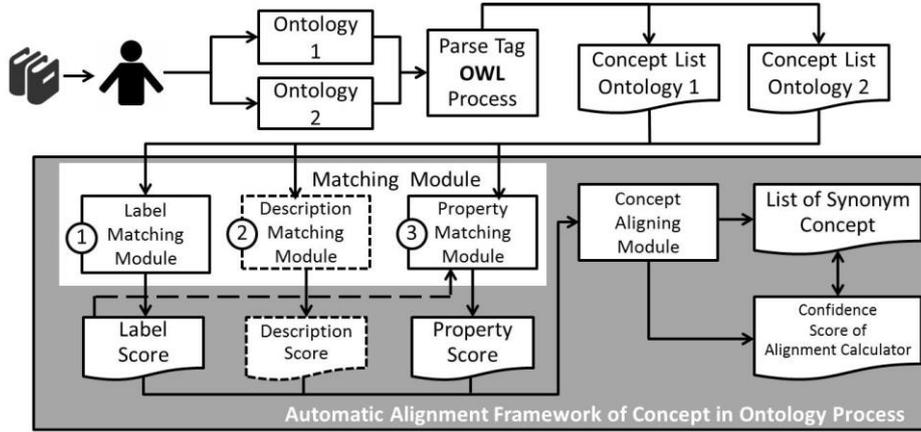


Fig. 1. Overview of the proposed framework

An input for the framework is a list of concepts in two or more ontologies with their corresponding information such as description, object property, data property and its hierarchy. These details are extracted by the exploited of owl parsing. The notations using in this paper are assigned as following:

Let $O1$ and $O2$ are ontologies that are considered. Ontology can be defined as follows:

$$O1 = \{C11, C12, C13, \dots, C1p\} \text{ and } O2 = \{C21, C22, C23, \dots, C2q\}$$

where Cij is a concept in ontology i^{th} and order at j^{th} . p, q are the number of concept in each ontology. For any concepts in ontology, it composes of properties which are categorised into two types; 1) part-of property or object property (PP) and 2) attribute-of property or data property (PA). Those are defined as $Cij = \{PPij, PAij\}$.

In case of part-of properties, it will be linked to a concept within its ontology to define a constraint on the range of properties and thus we assign the linked class of each part-of property as LC . We define a pair of property label and linked class ($ppij, LCij$) in each property below:

$$PPij = \{(pp11, LC11), (pp12, LC12), \dots, (ppPPk, LCPPk)\}$$

For the attribute-of properties, they link the properties to a defined data-type symbol (s) such as, integer, float, string and boolean. Hence, we define each property as a paired list of property label and symbol ($paij, sij$) as below:

$$PAij = \{(pa11, s11), (pa12, s12), \dots, (paPAk, sPAk)\}$$

3.1 Label Matching Module

This process is designed to find a similarity of the labels which are a given surface word of the concept. To compare likeness of the label, three types of a comparable concept are identified as 1) exact sameness, 2) partial sameness and 3) none sameness. Though the labels of two concepts are completely different in terms of characters, they can mean to the same concept as a synonym. Hence, the label matching of concepts is invented to two separated calculating functions.

String based Similarity Matching.

To consider the sameness of the apparent concept names which are exactly the same and partially alike, string similarity calculation proposed by [12] is exploited to calculate the score.

A merge of normalised longest common subsequence (*NLCS*), maximal consecutive longest common subsequence starting at character 1 (*NMLCS₁*) and maximal consecutive longest common subsequence starting at any character *n* (*NMCLCS_n*) are applied in this module. Where *label-c_{1i}* and *label-c_{2j}* are a label of concept in ontology 1 and a concept in ontology 2 respectively, The formulae are obtained as:

$$v_1 = NLCS(label - c_{1i}, label - c_{2j}) = \frac{length(LCS(label - c_{1i}, label - c_{2j}))^2}{length(label - c_{1i}) \times length(label - c_{2j})} \quad (1)$$

$$v_2 = NMLCS_1(label - c_{1i}, label - c_{2j}) = \frac{length(NMLCS_1(label - c_{1i}, label - c_{2j}))^2}{length(label - c_{1i}) \times length(label - c_{2j})} \quad (2)$$

$$v_3 = NMLCS_n(label - c_{1i}, label - c_{2j}) = \frac{length(NMLCS_n(label - c_{1i}, label - c_{2j}))^2}{length(label - c_{1i}) \times length(label - c_{2j})} \quad (3)$$

The weighted sum of these individual values v_1 , v_2 and v_3 is used to determine string similarity score, where w_i is weights with the sum of $w_i = 1$. w value is set by using an EM algorithm to find a significance of each parameter by v . Therefore, the string similarity of the two concepts is:

$$Sim(label - c_{1i}, label - c_{2j}) = w_1 v_1 + w_2 v_2 + w_3 v_3 \quad (4)$$

From the abovementioned formulae, the score of the string similarity calculation is at maximum as 1.0 in case of exact sameness whilst the partial sameness will gain the decreasing score based upon the apparent difference. With string similarity calculation, the labels with little different writing style such as plurality form, gerund form, capitalisation and localising form (American - British English) can be handled systematically. For example given in Fig. 2, the exact sameness example is the class "Method" in both ontology#1 and 2 which is equivalent in label therefore the score is calculated as 1.0. Furthermore, the

class “Rope” from ontology#1 and the class “Ropes” from ontology#2 are partially different so the calculation returns the score as 0.8 based on the equation (4). However, string similarity calculation cannot determine the completely different surface of the concept name as exemplified in a line with X mark in Fig. 2.

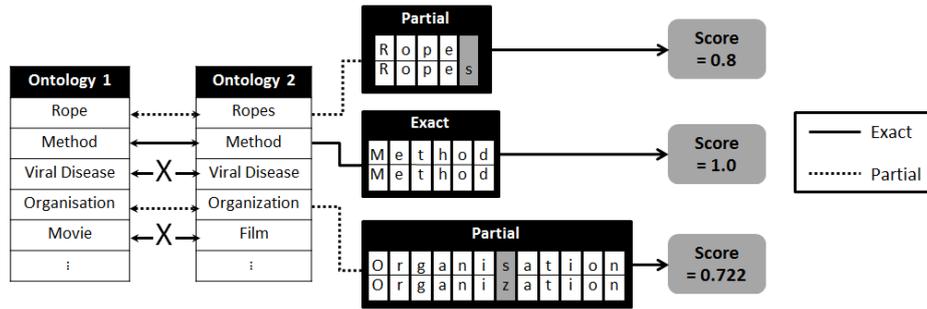


Fig. 2. An example of String Based Matching Result

Sense Based Similarity Matching

This process is designed to deal with the completely different surface of the semantically equivalent concept. WordNet [13], [14] is chosen as a source for lexical relations. The relations include synonym within the given entry and the relation across POS type. Normally, a label of a concept in an ontological product is a phrasal expression. To employ WordNet, those phrases should be split into words. Each word is searched with the headword in WordNet entry and is examined the related information given in WordNet as a medium to another label in another ontology. For more detail, please see examples in Fig. 3.

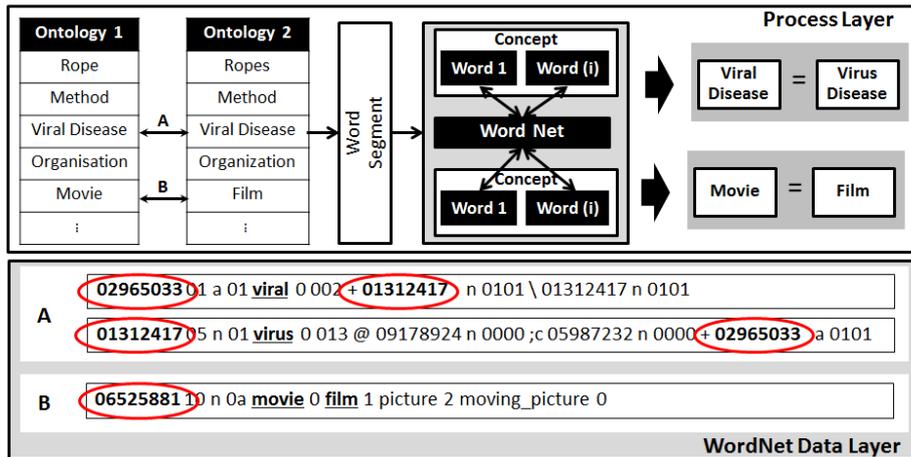


Fig. 3. An example of Sense Based Matching Result

In the WordNet, relations of lexicon are assigned by the pattern which links two or more senses with sense ID. As shown in Fig. 3 A, the lexicon “virus” in noun file is assigned with sense ID “01312417”, and there is a link (signal as the circle in Fig. 3) to another sense ID as “02965033” in adjective file which is where the word “viral” is located. The link informs that the word sense relation has the related meaning but different part-of-speech. Thus, this information provides us that the word “virus” and “viral” have the relevantly equivalent meaning. In Fig. 3 B, the word “movie” from ontology#1, once is searched through WordNet is found that it is in the sense ID “06535881” entry which has several another words such as “film” in synonym set (SynSet). This information hence can be concluded that the label “movie” and “film” are semantically equivalent.

Cases of relation of the sense between concepts are exact, partial and non-matching. Each concept label was segmented to list of string. Then calculate the sense based similarity matching (Sense) for two concepts as follows:

$$\text{Sense}(\text{label} - c1i, \text{label} - c2j) = \frac{\text{amount of equivalent senses}}{\text{maximum length of string}} \quad (5)$$

3.2 Description Matching Module

This function is designed as an optional score in case there is a description (Des) attached to the concept. Naturally, the description of the concept is according to ontology developer to select the description from well-known reference, and it can be chosen freely. Therefore, there will be less chance to capture synonymous meanings to identify the equivalent concepts. However, in case that the descriptions of two concepts are exactly the same as using the same reference of meaning, they can be assured that those two concepts are the synonymy to each other. Thus, the matching description is consider as a positive information as a hint for informing the equivalent concept.

The string matching is applied to capture the sameness of description in this work. The Des value can solely be 1.0 if the descriptions of both concepts are the same. The exact sameness will only be counted as a plus score towards the total score while other cases will be ignored by the system. For the case of the Des value is not 1.0, the non-matched description will not be calculated in the total score.

3.3 Property Matching Module

This module is to calculate the likeness of properties related to concepts between ontologies. We assume that the concepts which contain equivalent properties in terms of property label, cardinality, and range of class are likely to be synonym to each other. Moreover, the inherited properties from mother concepts are also considered as attached properties. Please see the exemplified illustration in Fig. 4.

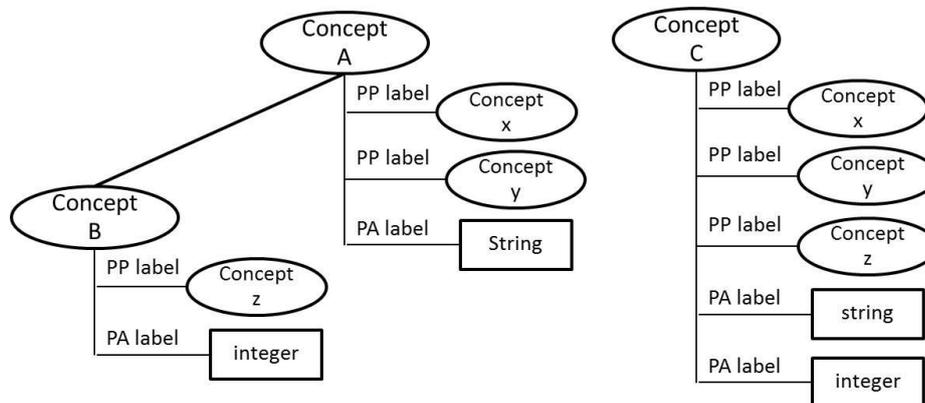


Fig. 4. An example of Property Based Matching

From

Fig. 4, Concept A and B are the concepts from ontology#1 while Concept C is from ontology#2. Each concept has its own property as shown, but please be reminded that Concept B also gets the inherited properties from Concepts A thus Concept B has five properties in total. To identify equivalence of the concepts, properties of the concepts should be the same or mostly similar. In this work, ontological property [15][16] is categorised into two types.

1. *part-of* property or object property (*PP*) – containing a constraint and number on the range of properties as an object of the relation
2. *attribute-of* property or data property (*PA*) – containing a constraint on the instantiated data by data type, i.e. string, integer, float, Boolean, etc.

From

Fig. 4, Concept B from ontolog#1 with inherited properties from Concept A and Concept C from ontology#2 contain the same PP properties in terms of linked concepts and PA properties in terms of data-type. To draw the matching method, we applied the best first search in our methodology as given in pseudo code in Fig. 5.

```

Property Similarity (O1,O2)
1. Let O1[i] and O2[j] is a list of all concepts in ontology O1 and O2 respectively
2. Let Initial C[m] is an initial list of concept which contains only attribute-of properties from O1
3. Let Diff O1 = O1 - Initial C
3. While concept o1[i] in Initial C is not null
4.   For each concept o2[j] in O2
5.     Sim [i][j] = Sim[j][i] = Similarity_Calculation (o1[i], o2[j])
6.     Calculated_C ← o1[i]
7.   End For
8. End While
9. While concept in DiffO1 is not null
10.  SimCalcCandidate = Update (DiffO1 , Calculated_C)
11.  If SimCalcCandidate is not null
12.    DiffO1 ← DiffO1 - SimCalcCandidate
13.    While concept o1[i] in SimCalcCandidate is not null
14.      For each concept o2[j] in O2
15.        Sim [i][j] = Sim[j][i] = Similarity_Calculation (o1[i], o2[j])
16.        Calculated_C ← o1[i]
17.      End For
18.    End While
19.  Else
20.    UnabletoCal ← DiffO1
21. End While

Function Update (DiffO1 , Calculated_C)
1. For each concept O1[i] in DiffO1
2.   For each PP[i,j] in O1[i]
3.     Unless LinkClass PP[i,j] is in Calculated_C
4.       Break
5.     End For
6.   SimCalcCandidate ← O1[i]
7. End For

```

Fig. 5. Psuedo Code of Property Matching Module

From Fig. 5, the pseudo code is designed to handle PA and PP of concept from two ontologies. The PA of the concept will be handled first hand and compare with the candidate concepts in another ontology. After PAs are collected, PPs of the concept are focused for similarity calculating. Each property will be compared and once the

calculation is done, the set of PA and PP will be sent to compare with another concept until all possible concepts are scored.

To score the property similarity, the following equations are obtained.

$$S_{PM} = \frac{\sum_{PP}^{ij} S_{LM}(PP_{kC_{1i}}, PP_{mC_{2j}}) + \sum_{PA}^{ij} (PA_{kC_{1i}}, PA_{mC_{2j}})}{n_{pp} + n_{pa}} \quad (6)$$

Where

- $PA_{kC_{1i}}$ is an *attribute-of* property k^{th} of concept i^{th} in ontology#1
- $PA_{mC_{2j}}$ is an *attribute-of* property m^{th} of concept j^{th} in ontology#2
- $PP_{kC_{1i}}$ is a *part-of* property k^{th} of concept i^{th} in ontology#1
- $PP_{mC_{2j}}$ is a *part-of* property m^{th} of concept j^{th} in ontology#2
- S_{LM} is score of label matching selected from higher score between *Sim* from (4) and *Sense* from (5)
- n_{pp} is amount of *part-of*
- n_{pa} is amount of *attribute-of*

3.4 Alignment

To decide which pair of ontologies is a synonym, the scores from all the features are employed. The result of this module is a list of possibly equivalent classes based on alignment score and the confident score to inform the degree of confidence which system makes. Since several features are used in this work, the alignment score can be above alignment criterion from the calculation though the pair is not guaranteed from any features. Confident score (ConfScore) is applied to distinguish the trustable pair from another.

To get alignment score, we apply equation (7) while equation (8) is designed to generate ConfScore.

$$\text{AlignmentScore}(C_{1i}, C_{2j}) = \frac{S_{LM} + S_{DM} + S_{PM}}{F} \quad (7)$$

$$\text{ConfScore} = (\text{intial} - ((1 - S_{LM}) + (1 - S_{DM}) + (1 - S_{PM})) * 100) + \text{bonus} \quad (8)$$

Where

- AlignmentScore is a similarity score of a concept pair
- ConfScore is a confidence score
- C_{1i} is a concept i^{th} in ontology 1
- C_{2j} is a concept j^{th} in ontology 2
- S_{LM} is score of label matching selected from higher score between *Sim* from (4) and *Sense* from (5)
- S_{DM} is score of description matching (in case it exists)
- S_{PM} is score of properties matching from (6)
- F is amount of feature apply in used

The criterion to assign equivalent concepts is the AlignmentScore is over 0.5.

ConfScore is given based on the strong score from each feature. The more strong score from features, the higher of the ConfScore will be. The initial ConfScore is 50. Once the score of the feature is found at maximum, the bonus of 50 ConfScore will be added. Otherwise, the score will be decreased from the missing point from the matching feature score.

Table 2. An Example Alignment Result and ConfScore Calculation

		O2								
		A			B			C		
		Slm,Spm	AlignScore	ConfScore	Slm,Spm	AlignScore	ConfScore	Slm,Spm,Sdm*	AlignScore	ConfScore
O1	X	1.0,1.0	1	150	0.22,0.1	0.61	78	0,0	0	not aligned
	Y	0.1,1.0	0.55	50	0.91,0.75	0.83	16	1.0,0.87,1.0*	0.96	137
	Z	0.72,0.3	0.51	0	1.0,0.7	0.85	70	0.2,0	0.1	not aligned

For example from Table 2, focusing on concept “X”, concept “A” and “B” are chosen as a equivalent concept since concept “C” do not meet the criteria from alignment score which is below 0.5. The pair of X-A obtains 150 ConfScore from two maximum feature scores which will give two of 50 bonus scores. For concept “Y”, all of concept A, B and C are aligned as Y’s synonym since they all give 0.55, 0.83 and 0.956, respectively. As shown, pair of Y-C gets another a plus score (marked with asterisk symbol) from description exact matching. In details, the pair of Y-A obtains ConfScore as 50 according to $((50^1 - (90 - 0)^2) = 0^3) + 50^4$. For the pair of Y-B, the ConfScore is 16. The pair of Y-C obtains ConfScore as 137. Hence, the pair from concept Y shows that the pair Y-C has the highest ConfScore.

4 Experiment

To test an ability of the framework, three related ontologies were selected. The ontologies are mulberry ontology (O1), silk worm ontology (O2) and trade statistic of silk-mulberry products ontology (O3). Those ontologies share several synonymous concepts since they are in the same agriculture topic. For the statistic, O1, O2 and O3 contain 303, 372, and 96 concepts respectively. For a test result, ontology developers were asked to manually align the equivalent concepts as a gold standard.

The gold standard shows that there are 59 equivalent concepts comparing O1-O2, and O1-O3 pair has 27 equivalent concepts while O2-O3 gives 12 equivalent pairs.

¹ Initial score of ConfScore is 50

² This is minus score from missing score of missing feature score. 90 is obtained from missing 0.9 point from feature#1 multiples with 100 while 0 is from non-missing score from 1.0.

³ The score is set as an absolute integer which does not allow negative value.

⁴ 50 is the bonus score from existing of a maximum feature score.

The sum of all equivalent concepts from all three ontologies is 97 concepts. An example of the equivalent concept list is given in Table 3.

Table 3. An example of the equivalent concept list

Concept in O1	Concept in O2	Concept in O3
X	Acre	Acer
Viral Disease	Virus Disease	X
X	Cocoon	Cocoon
Fungal Disease	Fungal Disease	X
X	Raw Silk	Raw Silk
Mulberry tree	X	Mulberry Plant
Count Unit	Count Unit	Count Unit

From comparing to gold standard, we measured the result in terms of precision, recall and f-measure. The results from the system comparing with gold standard are given in Table 4.

Table 4. Precision, Recall and F-measure of the Matching Result

Feature	Gold Standard	System Found	Precision	Recall	F-measure
label	60	60	1	1	1
properties	7	9	0.78	1	0.88
description	0	0	0	0	0
label + description	4	4	1	1	1
label + properties	24	25	0.96	1	0.98
properties + description	0	1	-	-	-
label + properties + description	2	2	1	1	1
All matches	97	101	0.96	1	0.98

5 Discussion

From the result, we found that the proposed framework gave a good accuracy result. The system can capture all 97 equivalent concepts assigned in gold standard list. However, there are four concepts that the system returned as synonymous concept pair but not in the list. Those concepts were examined in details and found that they are the synonymous concepts which experts overlooked from manual mapping since the labels are ambiguous.

We found that label matching plays the main role for capturing 91 concepts of the result while properties matching can capture 35 concepts. In the given ontologies, there are some descriptions attached to the concept, and it helped on matching 7 con-

cepts which were already considered as a pair by label matching or property matching. However, the description matching gave an extra confidence score to those pairs to assure the reliable aligning.

From 91 concept pairs by label matching, 69 concepts were found by the string based criteria while 22 concept pairs were recognised by sense based matching module. The examples of found pairs with the score gained from system calculation are shown in Table 5.

Table 5. An Example of found pairs by label with the score

Concept	Matched Concept	Score	Criterion
Method	Method	1.0	String similarity
Viral Disease	Virus Disease	1.0	Sense - WordNet
Rope	Ropes	0.8	String similarity
Mulberry Plant	Mulberry Tree	1.0	Sense - WordNet
Food	Nutrition	1.0	Sense - WordNet
Bacterial Disease	Bacteria infection	0.5	Sense - WordNet

Based on property matching module, we found that properties can be a great method to capture phrasal terms. All of the nine concepts that property matching can solely capture are a phrasal label with domain-specific terms as exemplified in Table 6. From examples in Table 6, the first row is the concepts with exactly same range of concepts and data-types while the second row shows the concepts that required sense based criteria to map the range concept.

Table 6. An Example of found pairs with the score and matching type

Concept	Matched Concept	Score
Product (<i>in silk ontology</i>) <ul style="list-style-type: none"> • PP – range_class: Material, label: made_of • PP – range_class: Country, label: import_to • PA – datatype: integer, label: has_retail_price 	Silk Goods (<i>in Trading Stat ontology</i>) <ul style="list-style-type: none"> • PP – range_class: Material, label: made_of • PA – datatype: integer, label: has_retail_price • PP – range_class: Country, label: import_to 	1.0
Harvesting (<i>in Trading Stat ontology</i>) <ul style="list-style-type: none"> • PP – range_class: Season, label: has_season • PP – range_class: Cocoon_Product, label: has_output 	Product (<i>in silk ontology</i>) <ul style="list-style-type: none"> • PP – range_class: Time, label: has_time • PP – range_class: Silk_Product, label: has_output 	0.747

As for description matching, seven concepts are matched. Those concepts were also aligned with other matching, thus it can be additional plus score to weight up the

confidence score. By focusing on confidence score, we found that 87 concepts from the 101 matched equivalent concepts are assigned with over 100 confidence score points especially the concepts with description matched. Unfortunately, the confidence score cannot be measured systemically, but ontology developers has no complain against the score and satisfy with the given confidence score.

6 Conclusion

In this paper, we present a new method to capture synonym concepts from several ontologies. The framework exploits information within ontological concepts including a label of a concept, properties of a concept and a concept's description. The aforementioned information is treated as features for considering similarity. Once the score of each feature is calculated, those scores are used for making decision to align a pair of concepts. Not only alignment of equivalent concepts is implemented in this work, but the confidence score is also calculated to distinguish the guaranteed pair from ambiguous pairs. From testing the framework against manual pair alignment, the system shows the potential to work equivalently to human selection. Moreover, there are some captured concepts which can be considered similar concepts that were overlooked by manual selection.

To improve the performance, we plan to add more features from concept's information such as hierarchical structure of the concept and other relevant ontological details. We also plan to test the system with large scale ontologies to approve its speed and robustness.

7 References

- [1] D. Fensel, F. van Harmelen, I. Horrocks, D. L. McGuinness, and P. F. Patel-Schneider, "OIL: an ontology infrastructure for the Semantic Web," *IEEE Intell. Syst.*, vol. 16, no. 2, pp. 38–45, Mar. 2001.
- [2] S. Aitken, R. Korf, B. Webber, and J. Bard, "COBrA: a bio-ontology editor," *Bioinformatics*, 2005.
- [3] A. Bernstein and E. Kaufmann, "GINO—a guided input natural language ontology editor," *Semant. Web-ISWC 2006*, 2006.
- [4] V. Haarslev and R. Möller, "Racer: A Core Inference Engine for the Semantic Web.," *EON*, 2003.
- [5] H. Tan and P. Lambrix, "SAMBO Results for the Ontology Alignment Evaluation Initiative 2007.," *OM*, pp. 2–9, 2007.
- [6] W. Hu and Y. Qu, "Falcon-AO: A practical ontology matching system," *Web Semant. Sci. Serv. Agents World Wide Web*, vol. 6, no. 3, pp. 237–239, Sep. 2008.
- [7] and E. M. Nagy, Miklos, Maria Vargas-Vera, "DSSim-ontology mapping with uncertainty Conference Item," 2006.
- [8] J. Li, J. Tang, Y. Li, and Q. Luo, "Rimom: A dynamic multistrategy ontology alignment framework," *Knowl. Data Eng. ...*, vol. 21, no. 8, pp. 1218–1232, 2009.

- [9] Y. Jean-Mary, "Asmov: Results for oaei 2010," *Ontol. Matching* 126, 2010.
- [10] M. Seddiqui and M. Aono, "Anchor-flood: results for OAEI 2009," in *Proceedings of the ISWC 2009 Workshop on Ontology Matching*, 2009.
- [11] P. Shvaiko and J. Euzenat, "Ontology matching: state of the art and future challenges," *Knowl. Data Eng. IEEE Trans.* 25, vol. 1, no. X, pp. 158–176, 2013.
- [12] P. Luekhong, T. Ruangrajitpakorn, T. Supnithi, and R. Sukhahuta, "Pooja : Similarity-based Bilingual Word Alignment Framework for SMT," in *Proceedings of the 10th International Symposium on Natural Language Processing, Phuket, Thailand*, 2013.
- [13] G. Miller, "WordNet: a lexical database for English," *Commun. ACM*, 1995.
- [14] "WordNet Search - 3.1." [Online]. Available: <http://wordnetweb.princeton.edu/perl/webwn>. [Accessed: 28-Jul-2014].
- [15] K. Kozaki, Y. Kitamura, M. Ikeda, and R. Mizoguchi, "Hozo: an environment for building/using ontologies based on a fundamental consideration of 'Role' and 'Relationship,'" ... *Manag. Ontol. ...*, 2002.
- [16] N. Noy, M. Crubézy, and R. Fergerson, "Protege-2000: an open-source ontology-development and knowledge-acquisition environment," in *AMIA Annu Symp Proc*, 2003, p. 2003.

Advancing Underutilized Crops Knowledge using SWRL-enabled Ontologies - A survey and early experiment

Abba Lawan¹, Abdur Rakib¹, Natasha Alechina², and Asha Karunaratne³

¹ School of Computer Science
The University of Nottingham, Malaysia Campus
{khyx3alw, Abdur.Rakib}@nottingham.edu.my

² School of Computer Science
The University of Nottingham, UK
natasha.alechina@nottingham.ac.uk

³ Crops For the Future Research Centre (CFFRC), Malaysia
asha.karunaratne@cffresearch.org

Abstract. Due to their powerful knowledge representation formalism and associated inference mechanisms, ontology-based approaches have been increasingly adopted to formally represent domain knowledge. We propose the use of ontologies to advance knowledge-sharing on underutilized crops and propose how to integrate those ontologies with rules for added expressiveness. We first present a survey on the use of ontologies in the field of life-sciences with emphasis on crop-related ontologies, and justify why we need a new formalism. We then present the UC-ONTO (an Underutilized Crops Ontology) as a case study showing the integration of OWL (Web Ontology Language) ontologies with Semantic Web Rule Language (SWRL) rules for added expressiveness.

Keywords: Crop Ontology, OWL Ontology, Semantic Web Rule Language, Underutilized Crops, Knowledge representation, Reasoning

1 Introduction

A shared concern among knowledge engineers and domain experts is the formalization of knowledge domains with minimum ambiguities. One possible solution is the use of ontologies, which serve as an explicit specification of terms that formally define and structure the concepts of a shared domain and the relationships that exist between them [6]. In essence, ontologies help to provide a common understanding of a domain while enabling knowledge-sharing among experts and software tools.

In the field of Life Sciences, ontologies have proven to be increasingly valuable by providing the semantic framework for defining domain concepts and their relationships coupled with automated reasoning and analysis tools that support knowledge organization and sharing [45, 51]. Thus, breathing new life into biological/agricultural classifications by providing common understanding

of terms among researchers and bridging the gaps in semantic and organizational differences between tools and databases. In the Crops domain, various ontologies do exist, such as the Crop Ontology [7], the Plant Ontology [12], the Gene Ontology [5] and the popular AGROVOC [10, 11], among others. However, information on specific crops (categorized as Underutilized) hardly exist in these crops vocabularies and ontologies (see figure 1). *Underutilized Crops* [37], are those that are currently neglected though previously grown and consumed with considerable nutritional and/or market value [15].

Moreover, existing crop-related ontologies such as those listed above are usually available in the Open Biomedical Ontology (OBO) format, being developed using an open-source OBO-Edit environment [12]. Though, OWL versions of these ontologies are also provided in most cases, and since tools for converting OBO to OWL ontologies do exist, such as OboInOwl⁴, it is always easier to adapt OBO ontologies to OWL-based development environments and semantic web applications [19].

The high-expressive power of OWL - owed to its rich collection of constructs and its support for rule languages such as SWRL, offer developers greater flexibility in domain modeling and expressing declarative knowledge (using rules) over ontologies. Moreover, the integration of OWL ontologies with rules help in expressing implicit domain knowledge by utilizing existing rule-based reasoning supports. With OWL being the standard ontology language approved by the World Wide Web Consortium (W3C), efforts to provide Crop Ontologies in RDF and OWL format will undoubtedly boost crop knowledge-sharing and allows interoperability between participants of the platform and beyond. Also Semantic Web applications can be developed to utilize such ontologies.

The focus of this paper is to practically explore how rules can be used to increase the expressive powers of ontologies focusing on the SWRL rules. By so doing, we develop OWL ontology for the Underutilized Crops domain and further integrate the ontology with basic SWRL rules. One significant role of ontologies is that they facilitate knowledge reuse. As such, we utilize some domain-independent as well as crop ontologies in the underutilized crops ontology. FAOs geopolitical ontology and OWL-time ontology are some of the ontologies imported. We hope in the future, to see more complex representation of general crops knowledge other than concept hierarchies and the simple is-a and part-of relationships currently offered by the popular crop ontologies (section 3.2). In a similar gesture, authors of Crop Ontology: vocabulary for crop-related concepts in [32], have suggested the use of OWL-DL in the future works of for added expressiveness and complex domain modeling.

The remainder of the paper goes as follows: We present our motivation and the scope of the review in the next section and section 2 discusses the expressive powers of OWL and the need for integrating OWL ontologies with rules. This is followed by a brief introduction of the SWRL formalism. Section 3 presents the relevant works on using ontologies to model a knowledge domain with emphasis on the crops domain. Section 4, which introduces the UC-ONTO, describing the

⁴ <http://www.bioontology.org/wiki/index.php/OboInOwl> : *MainPage*

problem background, approaches, and development methodology. In section 5, we present an implementation of the SWRL rules extension for the UC-ONTO case study. We evaluate the ontology and SWRL rule assertions in section 6 and finally conclude in 7.

1.1 Motivation and Scope

Inspired by [32, 28], the work presented in this paper is part of a PhD project which among others, aims at using ontologies (and related formalisms) to standardize knowledge representation in the field of Underutilized Crops. The review part of our work focuses on extending ontologies with rules and is restricted to the literature that discusses the use of SWRL rules and its expressive extensions. However, evaluation of computational and reasoning capabilities of OWL + SWRL combination is not provided in this paper.

Our work can serve as an introduction to the rule-based formalisms and a guide to new researchers and non-logic experts that plan to utilize these formalisms for their problem domain. The complete ontology can be found online⁵.

2 Background

In this section, we discuss the expressive powers of OWL ontologies and the importance of integrating such ontologies with SWRL rules. The SWRL formalism is then briefly discussed highlighting its condition for decidability, the DL-safeness.

2.1 OWL Expressiveness and the need for rules

Description Logic (DL)-based OWL is the standard ontology language approved by W3C for modeling domain knowledge in the Semantic Web [55]. In the quest for a more expressive web ontology language, the OWL family [33, 50], evolves from OWL 1 that consists of three sub-languages namely: OWL-Lite, OWL-DL and OWL-Full, to the more recent OWL 2, which is also partitioned into OWL2EL, OWL2QL and OWL2RL [34]. These languages offer different expressiveness and computational desirability with the current version, OWL 2, able to provide a wider range of constructs such as transitive and inverse properties, cardinality restrictions, as well as inheritance among others.

However, despite its success in achieving hierarchical definition and efficient classification of domain concepts when compared to Resource Description Framework (RDF) its predecessor, OWL suffers from other expressive limitations, such as its lack of support for composite role definition between concepts. Hence, there is the need for a more expressive domain modeling language than OWL as established by various researchers citing both theoretical and practical example

⁵ <https://www.dropbox.com/s/4l4bbcdus0bv7zm/BG1BG2MergeFinalOWL2RL.owl?dl=0>

[24, 13, 35, 29, 17, 26]. Rule formalisms were consequently adopted to provide the needed support for more expressive power to the OWL language both being fragments of the classical logic.

The expressive limitations of OWL and the choice for Rules are not just mere coincidences. While OWL-DL ontologies provides simple, reusable and easy to understand knowledge models, they lack the expressiveness offered by rules. Furthermore, the rule formalisms apart from being in common practice, provides an efficient reasoning support to ontologies with the added expressiveness.

The integration of OWL-DL and SWRL provides many advantages that cannot be achieved using either OWL DL or Horn rules alone. Moreover, extending ontologies with rules is favored due to the wide acceptance of rules in knowledge modeling and the success of Rule-based formalisms in commercial applications among others.

2.2 SWRL Formalism

In the literature, various formalisms exist to extend DL ontologies with rules and they are often classified into Hybrid (loosely coupled) and Homogenous (tightly coupled) approaches [1, 40]. Among the homogenous formalisms, SWRL has received a considerable attention from the Semantic Web community over the last few years [24, 23, 8, 43] and forms the basis of our survey. The classification of the rule languages into hybrid or homogenous can be in terms of syntax, semantics or both. we refer the interested reader to [28] for a more detailed list of the popular formalisms.

SWRL is a direct extension of OWL-DL that exploits its model theoretic semantics while combining the syntaxes of OWL-DL with that of Rule-ML. SWRL, originally called ORL (OWL Rule Language) [25], is a horn-like rule formalism having antecedent (body) as well as consequent (head) with both having conjunctions of rule atoms. Usually in the form:

$$atom_1, atom_2, atom_3, \dots, atom_n \rightarrow atom_1, atom_2, atom_3, \dots, atom_i$$

As initially defined in [24] and further discussed in [28, 8], SWRL extensions are bindings that provides a mapping between variables used in the rules to elements of a given domain. Ontology elements in SWRL are identified using their URI⁶ references. For technical details on the syntax and semantics of SWRL, we refer the reader to [25] and for background theory and implementations of Description Logic, see [2].

Decidability of SWRL Formalism: DL safeness. SWRL rules added to OWL ontologies need to be DL-safe to retain the decidability offered by OWL and ensure sound and complete reasoning over their ontologies. A DL-safe SWRL rule [35], ensures that only named concepts are used in the rules to avoid generating anonymous individuals during inference. In other words, only those

⁶ Uniform Resource Identifiers, strings similar to URLs, used to identify all objects on the semantic web

variables (or named individuals) already declared in the antecedent may be used in the inference no new concepts may be introduced.

3 The Context: Ontologies in the Crop domain

This section discusses the relevant works of using ontologies to model a knowledge domain with emphasis on the crops domain. Starting with classifying ontologies, we present the popular crop-related ontologies showing their inadequacy in representing underutilized crops knowledge. Finally, we point the benefits of ontologies in life-sciences.

3.1 Ontologies as Knowledge Repositories - Classification

Ranging from generic taxonomies to specific application-based knowledge models, ontologies have commonly been categorized into three levels [48, 27] namely: (i) The foundational ontologies, (ii) Domain ontologies and (iii) Application-level ontologies.

Foundational Ontologies also called top-level or reference ontologies, provide general taxonomies with multi-domain knowledge. The Unified Foundational Ontology (UFO) [20], Basic Formal Ontology (BFO) [47], General Formal Ontology (GFO) [21], and the GFO-Bio [22] among others, are common examples of foundational ontologies. Foundational ontology being a repository of general knowledge provides a means for semantic evaluation of lower ontologies such as the domain ontologies.

Domain ontologies on their part provide conceptual and more descriptive definition of terms within scoped domain boundaries, usually for an organization or knowledge community comprising of concepts, their relationships and individual instances. They offer a common vocabulary for sharing, reuse and standardizing knowledge of a specific community or domain of discourse. Larger domain ontologies are sometimes referred as upper-domain, such as BIOTOP [4], which is an upper-domain ontology for molecular biology linking smaller domain ontologies with the BFO, FAOs AGROVOC [10, 11, 39], which has in the past thirty years grown from simple multilingual agricultural index to a Linked-Open-Data (LOD) set. Other examples of domain ontologies include the Crop Ontology [46], Plant ontology [12], Gene Ontology [5], and the Underutilized Crops Ontology (UC-ONTO), which is currently under development by Crops For the Future Research Center (CFFRC)⁷.

Application ontologies are developed to be used for specific applications and usually utilize the domain ontologies by restricting conceptualizations to model a specified application domain. For example, the Food Ontologies for nutritional applications in [42, 9, 30] and sensor ontologies for manufacturing application reviewed in [44].

⁷ <http://www.cropsforthefuture.org/>

3.2 Domain Ontologies in Life Sciences

In this section, we review some of the popular crop-based domain ontologies with emphasis on the expressiveness provided by their development languages.

Gene Ontology. The Gene Ontology [5] is a popular biological upper-domain ontology developed by the Gene Ontology Consortium to establish standards in the representation of gene-related knowledge for various species of organisms. It is designed as a collaborative community-based ontology development effort providing gene ontologies with three components: molecular functions, biological processes and cellular components, their annotations as well as tools to access and process the ontologies [51]. Like many existing biological ontologies, the Gene Ontology is available mostly in the OBO format. Though, OWL versions of these ontologies are provided in some cases. However, OBO ontologies even when converted to OWL formats are less expressive. This is due to the well-defined semantics, interoperability with other ontologies and the various tools and services that facilitate development, maintenance and reuse of OWL ontologies,

Plant Ontology. Considering it as a comparative tool for plant anatomy and genomic analysis [12], the Plant Ontology is developed to provide formal specification of terms that describe plant anatomy, morphology and growth stages with the first and later developed as components of the whole ontology. Plant Ontology utilizes the data model available in the Gene Ontology (GO) [5], for annotating the plant anatomy and growth stage ontologies with gene expressions and phenotype data from the GO. Similar to the Gene Ontology, the Plant ontology is also guided by the OBO Foundry ontology for seamless collaboration with other biological ontologies [12] and most of the ontology is available in the OBO format. However, some parts of the ontology are available in the OWL format. For efficient comparison of disparate data with similar terms, such as that of genomics, the use of ontologies is necessary for data curation and analysis as it helps to provide common structured vocabulary that permits automated reasoning.

Crop Ontology. Citing data management, accessibility and retrieval challenges as the main motivation, Generation Challenge Program (GCP) ⁸ developed the Crop Ontology to facilitate community sharing of crop-related information by semantically characterizing and annotating historic generic crop data sets (traits, phenotype, germplasm, breeding, etc.) [7, 46]. With a simple web-based interface and the help of semantic experts as moderators of the ontologies, the Crop Ontology platform allows community-based collaborative ontology development, where users can create and add their own ontologies to the pool. Originally in

⁸ <http://www.pantheon.generationcp.org>

Open Biomedical Ontology (OBO) formats, the Crop ontology has evolved to utilize more terminological standards such as RDF and OWL [32].

With OWL being the most widely used standard for developing ontologies, effort to provide crop ontologies in RDF and OWL format will no doubt improve knowledge-sharing among researchers. This is basically due to the high expressiveness, efficient reasoning support, and the added advantage integrating OWL ontologies with declarative rule languages such as SWRL. Moreover, Semantic Web applications can be developed to easily utilize OWL ontologies.

From the foregoing exploration, the ontologies are able to provide an efficient and comprehensive hierarchical representation of their domains with common roles between concepts being of the form is-a and part-of relationships, which simply put, denotes that a concept is either a subtype of the connecting concept or that of the root/ancestral concept (see Fig. 1 on the right panel). However, they seem to lack complex representation of roles or relationships between concepts, which is one of the major differences between ontologies and hierarchical taxonomies such as thesauri. In a similar gesture, authors of Crop Ontology: vocabulary for crop-related concepts in [32], have suggested the use of OWL-DL in their future work for added expressiveness and complex domain modeling.

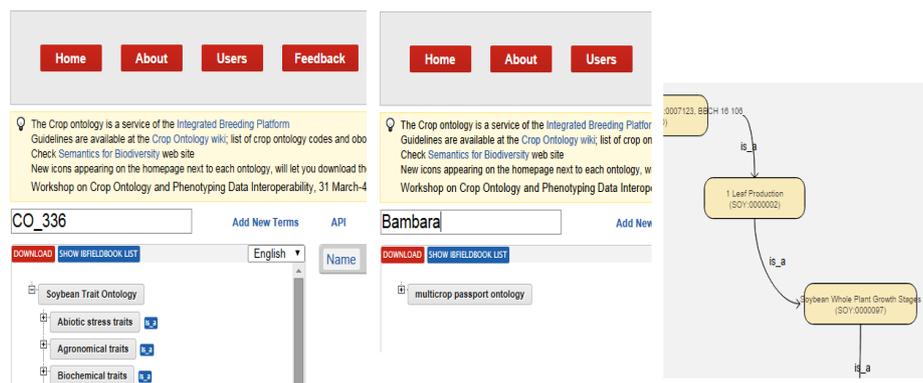


Fig. 1. Crop Ontology Curation Tool showing: general crops info. (left), lack of similar info. on underutilized crop (center) and the simple is-a relationship (right). Image source: <http://www.cropontology.org/>

Contribution of ontology to the crops domain (not exclusive though) can be summarized in the following points: i) For organization and sharing crop information ii) As integrative comparative tools iii) For standardization of domain knowledge and also iv) Useful for developing semantic web applications.

It should be noted however, that these contributions are not exclusive to the crops domain as they are simply benefits brought about by the use of ontologies. Though, the comparability advantage is more pronounced in the field of life sciences. Moreover, as ontologies are designed to be the knowledge modeling

formalisms for an open-web [16, 6], their advantages may not be restricted to a particular domain. A review on the recent trends and applications of ontologies citing examples from various domain ontologies is presented in [14] and text book detail on uses of ontologies in bio-informatics is given in [49].

4 Case Study: Underutilized Crops Ontology (UC-ONTO)

This section introduces the motivation as a case study on the use of SWRL rules for integrating ontologies in the crops domain. The approach and specific design issues as related to our case, underutilized crops knowledge modeling are discussed.

Problem Background. With the United Nation’s decade long efforts on biodiversity and food security, there is an awakening on the need to revitalize the cropping of neglected or underutilized crop species, many of which have the potential of providing food security as well as nutritional sustainability [15, 37, 53]. The Crops for the Future Research Center (CFFRC) , is one of the research bodies dedicated for research and development on Underutilized Crops. With many researchers working on different underutilized-crops related projects, there is a need for domain-level ontology to provide explicit specification of terms, the relationships between those terms and how they are related across the various research fields and outside partners.

Reusability Approach. As stated earlier, one of the benefits of developing ontology for a domain is knowledge reuse. Considering the available crop-domain collaborative ontologies (see section 3.2) and in line with the principle of ontology reuse, we ought not develop a new ontology but simply tailor these ontologies to present relevant information on underutilized crops species. Similar approach has been proposed in [3], where AGROVOC is used as a base vocabulary to develop the *CropOnt* a framework for relevant knowledge on crop production life cycle for individual farmers.

However, despite their nutritional, dietary-diversity, and economic importance [15], basic concepts definition on underutilized crop species are very rare and in some cases non-existent [53]. Consequently, most of the general crop ontologies do not have information on the underutilized or neglected crops due to the lack of available information on underutilized crops in general. In their book *Global research on Underutilized Crops* [37], the authors cited the lack of technical knowledge as one of the constraints to research and development on Underutilized crops.

Knowledge Gathering Approach. Two approaches have been considered in the early stages of our project: either to develop a complete Underutilized-crops ontology from scratch, or to utilize existing crop ontologies by importing relevant

and shared concepts. The latter, which support knowledge-reuse and favored in the field of ontology engineering, was thus accepted.

To do this however, there is a need to analyze some of these general crop ontologies and critically evaluate them for possible integration, while considering compatibility issues. Specifically, how the Underutilized Crops ontologies, which share much of the concepts of the general crops, can fit together with proper source formats and linkages with the imported ontologies. Furthermore, this will not only support the reusability spirit of ontologies but will also save a great amount of development time on the part of CFFRC knowledge engineers. The choice will also ensure conformity of our ontology to the existing standards in crop-domain modeling.

4.1 UC-ONTO development methodology.

We employ the collaborative ontology development methodology, which is necessary to enable knowledge engineers work closely with the domain experts (underutilized-crops researchers in our context).

To achieve a comprehensive modeling, the general guidelines advised in the work of Noy and McGuinness [36], the METHONTOLOGY [18], DILIGENT [41], and the Onto-Knowledge methodology, were utilized. These guidelines help to structure the ontology engineering process by identifying important but non-obvious aspects, such as the target users of the ontology, supporting tools, and specifying what values can be allowed for properties. Other aspects that are apparent and also common to all methodologies - such as defining domain terms and roles, asserting their hierarchy, and filling the concept slots with individual instances - are performed iteratively for each source of data to populate the underutilized crops ontology. Similarly, our user-defined SWRL rules are added iteratively while ensuring the consistency of the ontology by invoking the *Pellet* reasoner. The major steps for UC-ONTO development can be summarized as follows: i) Ontology requirement specification ii) Domain knowledge gathering and conceptualization iii) Model implementation and iv) Evaluation of the model.

These steps were performed repeatedly for each component version of the UC-ONTO, leading to the final complete version. The two final stages were termed *versioning* and *assembly*. In 'versioning', we assign a label to represent each ontology fragment, specifying where it fits to the larger ontology. While in the 'assembly' stage, smaller ontology modules are put together and a reasoner is invoked to assert the overall classification and check for consistency. A common problem with the assembly stage however, is that for each module added to the main ontology, inconsistencies are bound to arise. As such, to minimize such inconsistencies, the assembly is carried out with the ontology Reasoner in active mode. Moreover, for each smallest ontology module assembled, the reasoner need to be invoked to check for the consistency. This way, it is easier to keep track of what causes the inconsistencies and where to correct them.

Other common issues in ontology development include, the failure to reuse existing ontologies in the beginning of development and also the failure to familiarize with basic domain concepts (by ontology engineers) - leading to the

problems of modeling roles as classes and vice versa. Advisably, a domain expert should be available at all times to continuously check the progress of ontology modeling. This is because, while a Reasoner can cross-check inconsistencies arising from hierarchical representation and incorrect assertions, it is incapable of highlighting domain-related inconsistencies, among others.

4.2 The UC-ONTO

We have developed the first version of the underutilized crops ontology (UC-ONTO) using the Protégé 4.2 ontology editor. The ontology currently consists of SWRL built-ins, OWL-time ontology, and FAO geopolitical ontology as direct imports. This is because these ontologies being domain-independent and available in the OWL format can stand-alone without posing compatibility problems and inconsistencies.

While details on the development methodology and the aspects of the UC-ONTO (such as agronomic, physiological traits) are beyond the scope of this paper, we give a brief account of the composition of the ontology with a glimpse on the naming convention and structure. In the ontology, all crops related concepts are grouped together under the *DomainConcepts* as super class and all other concepts such as *DaysOfWeek*, *TimeZone*, etc. offered by imported ontologies, are composed as siblings. The *UnderutilizedCrops* class contains four sub classes with *Taro*, *Tef*, *Millet* and *BambaraGroundnut* class, which dominates most of the object properties such data-type property modeling in this version.

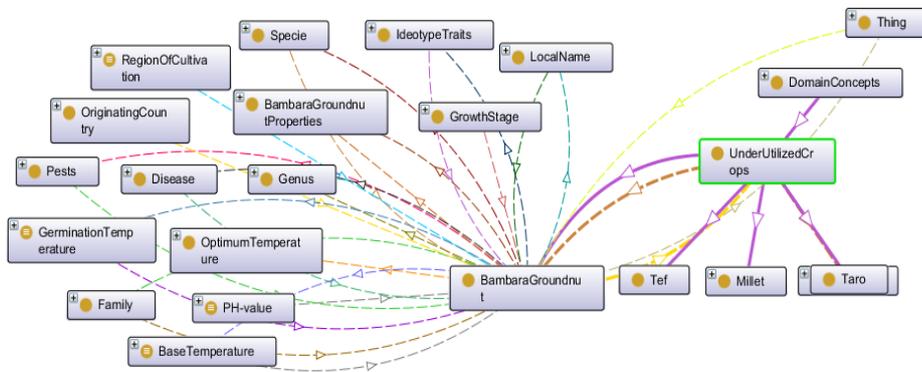


Fig. 2. Graphic view of concepts in UC-ONTO

The SWRL rules are written using the SWRL tab to specify more relationships between concepts on top of our ontology giving more flexibility to declarative property assertions in the UC-ONTO. Fig. 2 gives a partial graphic overview of the concepts and roles specified in the UC-ONTO.

5 Extending UC-ONTO with SWRL rules

This section presents an implementation of the SWRL rules extension for the UC-ONTO case study. The rules are intended to allow modeling declarative knowledge and for expressing complex roles (such as composite relations between concepts) that are not easily expressible with OWL alone.

The addition of our user-defined, DL-Safe, SWRL rules was delayed until the final version of the ontology was checked for consistency using *Pellet* reasoner and found to be consistent. Moreover, considering the main reason of using SWRL rules in our ontology, which is to express complex relations between domain concepts and utilize the SWRL built-ins to define and assert domain-specific concepts, it will still do no harm to our ontology if we express the OWL axioms using SWRL.

In the rules interface depicted in Fig. 3, we begin with a simple assertion in rule 8 that asserts a relationship between members of *BambaraGroundnut* and those of *BambaraGroundnutProperties* class using *hasProperty* relation. We then continue to assert more roles that are easily expressed with declaration, thereby extending the expressive power of the ontology. For example, the fourth rule:

$$\begin{aligned} & BambaraGroundnut(?y), Leaf(?z), isFeatureOf(?z, ?y) \rightarrow \\ & hasLeafType(?y, "Trifoliolate") \end{aligned}$$

States in simple terms, that if *BambaraGroundnut* class has a feature leaf, then it will be asserted that the leaf type is 'trifoliolate'. However, since features such as leaf are not exclusive to *BambaraGroundnut* class, then unless the leaf individual is related to *BambaraGroundnut*, the leaf type trifoliolate, cannot be asserted. Rules of these types that are based on certain conditions being true or otherwise, are hard to be expressed with OWL syntax alone.

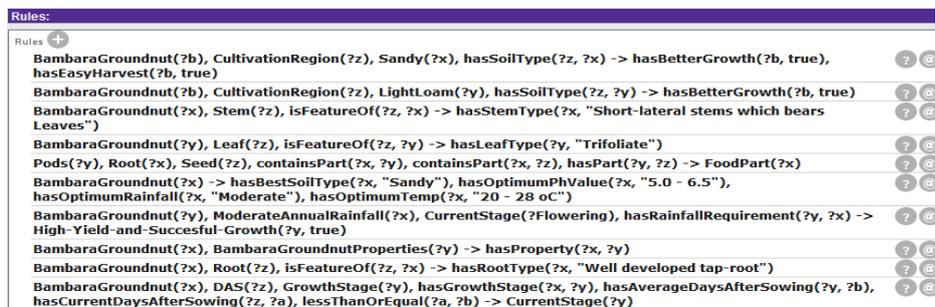


Fig. 3. Rules interface showing some user-defined SWRL rules

6 Evaluation

In this section, we evaluate the ontology and SWRL rules assertions by invoking the Pellet reasoner to classify and check for the consistency of the ontology.

Additional knowledge implicit in the crop ontology can then be inferred by this reasoner. Also to verify the conceptual facts and individual assertions, DL queries are used to probe the ontologies. We evaluate 2 to 3 queries for each SWRL rule, making a total of 46 DL queries. Results of frequent queries are saved and added as part of the ontology thereby evaluated automatically once the reasoner is invoked.

6.1 Reasoning and query processing

Using ontologies allow measuring performance at the design as well as run-time via a reasoner to compute the ontology classification and ensure consistency. As such a reasoner needs to be active and the ontology classified before writing any DL Queries. Our user-defined SWRL rules are validated by writing DL queries to check their inference or otherwise by the reasoner. For example, the query result of the sixth rule, determines the current 'growth stage' of a *BambaraGroundnut*. The rule uses a SWRL built-in '*swrlb:lessThanOrEqual*', to compare the days an individual *BambaraGroundnut*(*BG*) is planted with the number of days asserted for the different growth stages (e.g. the flowering stage *hasAverageDaysAfterSowing* = 50). If there is a match, the reasoner will then assert this growth stage as the current stage of the individual *BG*. Results for some of the rules, which assert Datatype properties to *BambaraGroundnut* individual, can be seen from the Inference provided by the Pellet reasoner in Fig. 4 (right).

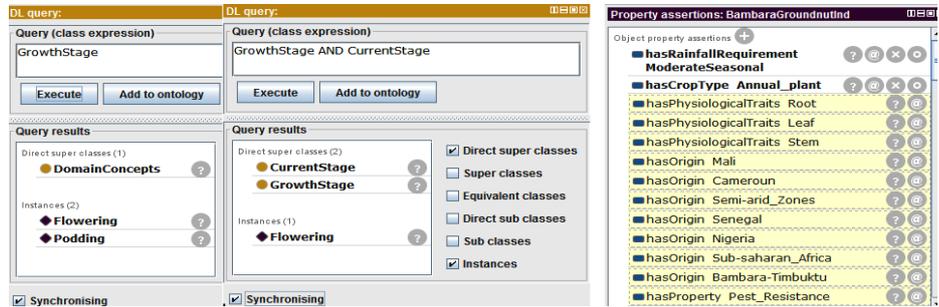


Fig. 4. Interface showing DL-Query Results (left) and Reasoner Inferences for Rules (right).

We would like to mention that the Underutilized crops ontology presented in this paper has: 24701 axioms, 111 classes, 397 individuals, with 94 object properties and 133 data properties. However, size and functionality of the ontology is expected to be continuously growing as more underutilized-crops data becomes available. The expressiveness of our ontology borders on SHOIN(D) algorithm and all SWRL rules added are DL safe, thereby decidable. The queries considered in the experiment were originated from the competency questions generated in

our ontology engineering stage; due to space constraints we are unable to present those in details.

7 Conclusions and future work

In this paper, we propose a framework for representing knowledge using OWL ontologies and SWRL rules. Using the crops domain as a case study, we review and justify the need for integrating ontologies with rules. We present the SWRL-extended underutilized-crop ontology (UC-ONTO), highlighting our motivation, approach and development methodology. This is followed by an evaluation, which involves validation of the knowledge represented in the UC-ONTO through Reasoner inferences and writing appropriate DL queries. In the future, we aim to populate the ontology with more standard crop-related data from relevant Foundational Ontologies. Also we plan to publish the ontology and present the domain-knowledge to the public through a web-based, social-networking styled decision support system for underutilized crops.

For added expressiveness to our ontology, we intend to study and utilize the available SWRL extensions such as the first-order logic extension SWRL-FOL [40], the non-monotonic extensions for dealing with negation, exclusion and rule priority as in [8], and the X-SWRL [31], which allows for dealing with existential quantification of new individuals. Others extensions considered important includes the Fuzzy-SWRL [38], vague-SWRL [52], and SWRL-F [54] for modeling imprecise knowledge - a situation commonly encountered when dealing with domain experts, especially in the field of crops where informal and undocumented practices still hold sway.

Acknowledgement. This work is partially supported by a CFFRCPLUS CropBase PhD scholarship scheme.

References

1. Antoniou, G., Damásio, C.V., Grosz, B., Horrocks, I., Kifer, M., Maluszynski, J., Patel-Schneider, P.F.: Combining Rules and Ontologies. A survey. Tech. rep. (2005)
2. Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.: The Description Logic Handbook: Theory, implementation, and applications. Cambridge University Press (2007)
3. Bansal, N., Malik, S.K.: A framework for agriculture ontology development in semantic web. In: Proceedings - 2011 International Conference on Communication Systems and Network Technologies, CSNT 2011. pp. 283–286 (2011)
4. Beisswanger, E., Schulz, S., Stenzhorn, H., Hahn, U.: BIOTOP : An Upper Domain Ontology for the Life Sciences. Applied Ontology 3(4), 205–212 (2008)
5. Berardini, T.Z.: The Gene Ontology in 2010: Extensions and refinements. Nucleic Acids Research 38 (2009)
6. Berners-Lee, T., Hendler, J., Lassila, O.: The semantic web. Scientific American pp. 29–37 (May 2001)

7. Bruskiwich, R., Davenport, G., Hazekamp, T., Metz, T., Ruiz, M., Simon, R., Takeya, M., Lee, J., Senger, M., McLaren, G., Van Hintum, T.: Generation Challenge Programme (GCP): standards for crop data. *Omics : a journal of integrative biology* 10(2), 215–9 (2006)
8. Calero, J.M.A., Ortega, A.M., Perez, G.M., Blaya, J.A.B., Skarmeta, A.F.G.: A non-monotonic expressiveness extension on the semantic web rule language. *J. Web Eng.* 11(2), 93–118 (2011), <http://dl.acm.org/citation.cfm?id=2230896.2230897>
9. Cantais, J., Dominguez, D., Gigante, V., Laera, L., Tamma, V.: An example of food ontology for diabetes control. In: Working notes of the ISWC 2005 Workshop on Ontology Patterns for the Semantic Web. p. 9 (2005)
10. Caracciolo, C., Morshed, A., Stellato, A., Johannsen, G., Jaques, Y., Keizer, J.: Thesaurus maintenance, alignment and publication as linked data: The AGROOVOC use case. In: Communications in Computer and Information Science. vol. 240 CCIS, pp. 489–499 (2011)
11. Caracciolo, C., Stellato, A., Morsheda, A., Johannsen, G., Rajbhandaria, S., Jaques, Y., Keizer, J.: The agrovoc linked dataset. *Semantic Web* 4(3), 341–348 (2013)
12. Cooper, L., Walls, R.L., Elser, J., Gandolfo, M.a., Stevenson, D.W., Smith, B., Preece, J., Athreya, B., Mungall, C.J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T.Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y., Jaiswal, P.: The plant ontology as a tool for comparative plant anatomy and genomic analyses. *Plant & cell physiology* 54(2), e1 (Feb 2013)
13. Cregan, A., Mochol, M., Vrandecic, D., Bechhofer, S.: Pushing the limits of owl, rules and protg – a simple example
14. Deshpande, N.J., Kumbhar, R.: Construction and applications of ontology: Recent trends. *DESIDOC Journal of Library Information Technology* 31, 84–89 (2011)
15. Ebert, A.: Potential of Underutilized Traditional Vegetables and Legume Crops to Contribute to Food and Nutritional Security, Income and More Sustainable Production Systems. *Sustainability* 6(1), 319–335 (Jan 2014)
16. Eiter, T., Ianni, G., Krennwallner, T., Polleres, A.: Reasoning web. chap. Rules and Ontologies for the Semantic Web, pp. 1–53. Springer-Verlag, Berlin, Heidelberg (2008)
17. Eiter, T., Lukasiewicz, T., Schindlauer, R., Tompits, H.: Well-founded semantics for description logic programs in the semantic web. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 3323 LNCS, pp. 81–97 (2004)
18. Fernández-López, M., Gómez-Pérez, A., Juristo, N.: Methontology: from ontological art towards ontological engineering. In: Proc. Symposium on Ontological Engineering of AAAI (1997)
19. Golbreich, C., Horridge, M., Horrocks, I., Motik, B., Shearer, R.: OBO and OWL: Leveraging semantic Web technologies for the life sciences. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). vol. 4825 LNCS, pp. 169–182 (2007)
20. Guizzardi, G., Wagner, G.: Using the Unified Foundational Ontology (UFO) as a foundation for general conceptual modeling languages. In: Theory and Applications of Ontology: Computer Applications, pp. 175–196. Springer Netherlands (2010)
21. Herre, H.: General Formal Ontology (GFO): A foundational ontology for conceptual modelling. In: Theory and Applications of Ontology: Computer Applications, pp. 297–345. Springer Netherlands (2010)

22. Hoehndorf, R., Loebe, F., Poli, R., Kelso, J., Herre, H.: GFO-Bio: A biomedical core ontology. *Applied Ontology* 3(4), 219–227 (2008)
23. Holford, M.E., Khurana, E., Cheung, K.H., Gerstein, M.: Using semantic web rules to reason on an ontology of pseudogenes. *Bioinformatics (Oxford, England)* 26(12) (Jun 2010)
24. HORROCKS, I., PATELSCHNEIDER, P., BECHHOFFER, S., TSARKOV, D.: OWL rules: A proposal and prototype implementation. *Web Semantics: Science, Services and Agents on the World Wide Web* 3(1), 23–40 (Jul 2005)
25. Horrocks, I., Patel-schneider, P.F., Boley, H., Tabet, S., Grosz, B., Dean, M.: SWRL : A Semantic Web Rule Language Combining OWL and RuleML (2004), <http://www.w3.org/Submission/SWRL/>
26. Hustadt, U., Motik, B., Sattler, U.: Data Complexity of Reasoning in Very Expressive Description Logics. In: Kaelbling, L.P., Saffiotti, A. (eds.) *Proc. of the 19th Int. Joint Conference on Artificial Intelligence (IJCAI 2005)*. pp. 466–471. Morgan Kaufmann Publishers (2005)
27. Keßler, C., Raubal, M., Wosniok, C.: Semantic rules for context-aware geographical information retrieval. *Smart Sensing and Context* 5741 LNCS, 77–92 (2009)
28. Krisnadhi, A., Maier, F., Hitzler, P.: Owl and rules. In: *Proceedings of the 7th International Conference on Reasoning Web: Semantic Technologies for the Web of Data*. pp. 382–415. RW’11, Springer-Verlag (2011), <http://dl.acm.org/citation.cfm?id=2033313.2033320>
29. Krötzsch, M., Maier, F., Krisnadhi, A.A., Hitzler, P.: A Better Uncle for OWL: Nominal Schemas for Integrating Rules and Ontologies. In: *20th International World Wide Web Conference (WWW2011)*. p. 645 (2011)
30. Li, H.C., Ko, W.M.: Automated food ontology construction mechanism for diabetes diet care. In: *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, ICMLC 2007*. vol. 5, pp. 2953–2958 (2007)
31. Li, W., Tian, S.: XSWRL, an Extended Semantic Web Rule Language and prototype implementation. *Expert Systems with Applications* 38(3), 2040–2045 (Mar 2011)
32. Matteis, L., Chibon, P., Espinosa, H., Skofic, M., Finkers, R., Bruskiwich, R., Hyman, G., Arnaud, E.: Crop Ontology: Vocabulary For Crop-related Concepts. In: Larmande, P., Arnaud, E., Mougenot, I., Jonquet, C., Libourel, T., Ruiz, M. (eds.) *Proceedings of the first international Workshop on Semantics for Biodiversity*. vol. 979 (2013)
33. McGuinness, D.L., Van Harmelen, F.: OWL Web Ontology Language Overview. W3C recommendation 10, 1–22 (2004), <http://www.w3.org/TR/owl-features/>
34. Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C.: OWL 2 Web Ontology Language Profiles (2009), <http://www.w3.org/TR/owl2-profiles/>
35. Motik, B., Sattler, U., Studer, R.: Query answering for owl-dl with rules. *Web Semant.* 3, 41–60 (Jul 2005)
36. Noy, N.F., McGuinness, D.L.: *Ontology development 101: A guide to creating your first ontology*. Tech. rep., Stanford (2001)
37. Padulosi, S., Hodgkin, T., Williams, J.T. and Haq, N.: Underutilized Crops: Trends, Challenges and Opportunities in the 21st Century. In: Jackson, J.E., Rao, V., M. (eds.) *Managing plant genetic diversity*, pp. 323–338. CAB International (2002)
38. Pan, J.Z., Stoilos, G., Stamou, G., Tzouvaras, V., Horrocks, I.: f-swrl: A fuzzy extension of swrl. *Journal on Data Semantics, special issue on Emergent Semantics* (2006)

39. Paziienza, M.T., Stellato, A., Tudorache, A.G., Turbati, A., Vagnoni, F.: An architecture for data and knowledge acquisition for the semantic web: The AGROVOC use case. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. vol. 7567 LNCS, pp. 426–433 (2012)
40. Peter F. Patel-Schneider, (Bell Labs Research, L.T.: A Proposal for a SWRL Extension towards First-Order Logic (2005), <http://www.w3.org/Submission/SWRL-FOL/>
41. Pinto, H.S., Tempich, C., Staab, S.: Diligent: Towards a fine-grained methodology for distributed, loosely-controlled and evolving engineering of ontologies. In: *Proceedings of the 16th European Conference on Artificial Intelligence*. pp. 393–397. IOS Press (2004)
42. Pizzuti, T., Mirabelli, G., Sanz-Bobi, M.A., Gómez-González, F.: Food Track & Trace ontology for helping the food traceability control. *Journal of Food Engineering* 120(1), 17–30 (2014)
43. ROSATI, R.: On the decidability and complexity of integrating ontologies and rules. *Web Semantics: Science, Services and Agents on the World Wide Web* 3(1), 61–73 (2005)
44. Schlenoff, C., Hong, T., Liu, C., Eastman, R., Foufou, S.: A literature review of sensor ontologies for manufacturing applications. *2013 IEEE International Symposium on Robotic and Sensors Environments (ROSE)* pp. 96–101 (2013)
45. Schuurman, N., Leszczynski, A.: Ontologies for bioinformatics. *Bioinformatics and biology insights* 2, 187–200 (2008)
46. Shrestha, R., Mauleon, R., Simon, R., Balaji, J., Channelière, S., Alercia, A., Senger, M., Manansala, K., Metz, T., Davenport, G., Bruskiwich, R., McLaren, G., Arnaud, E.: Development of GCP Ontology for Sharing Crop Information. *Nature Precedings* p. 6 (Apr 2009)
47. Smith, B., Grenon, P.: Basic formal ontology (bfo). *INFOMIS Reports* (2006)
48. Staab, S., Studer, R.: *Handbook on Ontologies*. Springer Science & Business Media (2010)
49. Stevens, R., Lord, P.: Application of Ontologies in Bioinformatics. In: *Handbook on Ontologies*, pp. 735–756 (2009)
50. W3C OWL Working Group: OWL 2 Web Ontology Language Document Overview (2012), <http://www.w3.org/TR/owl2-overview/>
51. Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.a., Jaiswal, P., Mungall, C.J., Preece, J., Rensing, S., Smith, B., Stevenson, D.W.: Ontologies as integrative tools for plant science. *American journal of botany* 99(8), 1263–75 (Aug 2012)
52. Wang, X., Ma, Z.M., Yan, L., Meng, X.: Vague-swrl: A fuzzy extension of swrl. In: *Proceedings of the 2Nd International Conference on Web Reasoning and Rule Systems*. pp. 232–233. RR '08, Springer-Verlag, Berlin, Heidelberg (2008)
53. Williams, J., Haq, N.: Global research on underutilized crops: An assessment of current activities and proposals for enhanced cooperation. *Biodiversity International* (2002)
54. Włodarczyk, T.W., Rong, C., O'Connor, M., Musen, M.: Swrl-f: A fuzzy logic extension of the semantic web rule language. In: *Proceedings of the International Conference on Web Intelligence, Mining and Semantics. WIMS '11, ACM, New York, NY, USA* (2011)
55. World Wide Web Consortium: Semantic Web "Layer Cake", =<http://www.w3.org/2004/Talks/0412-RDF-functions/slide4-0.html>

Publishing Linked Open Data from Semantic Relation Extraction for Thai Cultural Archive

Watchira Buranasing, Marut Buranarach
National Electronics and Computer Technology Center,
National Science and Technology Development Agency
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
{watchira.buranasing, marut.buranarach@nectec.or.th}

Abstract. Culture is a key dimension of the information society, that refers to the cumulative knowledge, experience, beliefs, attitudes, meanings, hierarchies, religion, spatial relations and material objects and possessions acquired by a group of people in each generations. This research proposed a method to create linked open data from semantic relation extraction for Thai cultural archive. It describes in detail a methodology for creating a linked data resource and developing a useful application with resource description framework (rdf) format.

Keywords: Linked Open Data, Thai Cultural Archive, Relation Extraction, Resource Description Framework (RDF)

1. Introduction

Culture is a key dimension of the information society, that refers to the cumulative knowledge, experience, beliefs, attitudes, meanings, hierarchies, religion, spatial relations and material objects and possessions acquired by a group of people in each generations. A cultural archive derived from Thai Cultural Information Center Website (<http://www.m-culture.in.th>), which is one of an important database for education, economy and society. The content database associates with person, organization, place and artifact. A size of database has been increasing in terms of volume of data from cultural specialist in 76 provinces of Thailand. There are more than 100,000 records uploaded in 3 years since November, 2010 to September, 2014 .

Relation extraction is one of the interesting topics in natural language processing. It is the task of extracting related pairs of entities from text. The goal is to discover the relationships between pairs of entities in texts. Previous research developed by Watchira Buranasing et al. is “Semantic Relation Extraction for Extensive Service of a Cultural Database” [1]. It is the method to discover semantics relation among a set of entities in a cultural archive. The approach is based on a set of relation templates which are determined by relation type and their arguments.

Linked Open Data is a method for publishing of data structure, that allows metadata to be connected. Therefore the different representations of the same content can be

found, and links made between related resources. One of a research developed by Je Edelstein et al. , “Linked Open Data for Cultural Heritage: Evolution of an Information Technology”. [2] It surveys the landscape of linked open data projects in cultural heritage, examining the work of groups from around the world. Traditionally, linked open data has been ranked using the five star method proposed by Tim Berners-Lee. This research developed a six-stage life cycle based on the five-star method, describing both dataset development and dataset usage. It uses this framework to describe and evaluate fifteen linked open data projects in the realm of cultural heritage. In addition, “Linked Open Data and its Implications for Artistic and Cultural Resources” developed by Allana Mayer [3] , “Amsterdam Museum Linked Open Data” developed by Victor de Boer et al. [4], all of the above researches proposed a method to create linked open data from database, but this research proposed a method to create linked open data from semantic relation extraction.

The remainder of the paper is organized as follows. Section II gives an overview of the methodology for creating linked data from semantic relation extraction for extensive service of a cultural database. Section III shows an application. Section IV concludes and discusses some directions.

2. Overview of the methodology for creating linked open data from semantic relation extraction

2.1 Semantic Relation Extraction

The content of each document from a cultural database including four components, there are images, title , description and category .There is one main subject of relations in each documents and the main subject belongs to one cultural domain. This research focusing on three cultural domains. There are place, person and artifact. Based on these domains, the possible subject of the relations is a place, a human and a man-made object. Therefore, the set of relations corresponding to the subject, such as the subject is a place, consequently, the related information has to be *where* it is, *when* it was built and *who* built it. The formal expressions for these relations are IsLocatedAt, IsBuiltIn and IsBuiltBy. The surface forms of the relations used for searching the relation texts. Named entity types, associated with the main subject domain and their relations.

This research controls semantic drift of the target arguments using named entities. The named entity recognizer has been built from an annotated corpus. [5] According to the relation templates, this method trained the model with four named entity tags. The list of named entity tags are location (LOC), person name (PER), organization name (ORG) and date (DAT). The samples of relation instances produced by the approach is shown in table 1.

Table 1. The samples of relation instances

Record ID	Subject	Relation	Object	Argument	Image	Date	Source
48081	วัดท่าเจดีย์ WatThaJedi	ตั้งอยู่ที่ IsLocatedAt	ตำบลบางเลน Tambon Bangsan	LOC	http://m-culture.in.th/media/big/201945.jpg	2013-01-01 00:00:01	http://m-culture.in.th
99722	นายชاکาเรีย หะมะ Mr. Sakareeya Hama	บิดาชื่อ HasFatherName	สะมะแอ หะมะ SamaAir Hama	PER	http://m-culture.in.th/media/big/152560.jpg	2013-01-01 00:00:01	http://m-culture.in.th
68704	พิพิธภัณฑ์หนัง ระโมทัย NangPraMoThaiMuseum	สร้างขึ้นโดย IsBuiltBy	องค์การบริหาร ส่วนตำบลโพนทัน Pon Tan Subdistrict Administrative Organization	ORG	http://m-culture.in.th/media/big/100604.jpg	2013-01-01 00:00:01	http://m-culture.in.th

2.2 Conversion and Modeling

In this sub section briefly explain the process of conversion from the result of relation extraction to linked open data.

The results of relation extraction have been presented as a set of statements. Each statements giving a value. For example, a description of the record information about Title, Relation and Object. The attribute pairs are reformatted into subject predicate object statements by using the record identifier as the subject of each statement. A URI represents a Uniform Resource Identifier (URI) reference, that identifies the name and location of a file or resource in a uniform format. The subject of record must be a URI and globally unique. A cultural archive derived from Thai Cultural Information Center Website (<http://m-culture.in.th/>) by Thailand's Ministry of Culture can create a unique URI for a resource in it's collections as “MOC” plus “record id”. The record ID can be replaced with the resource URI in the set of statements, as shown in Figure 1.

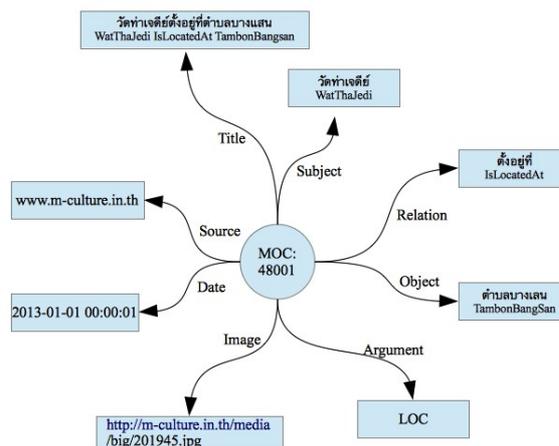


Fig. 1. The resource URI in the set of statements

Each statement of the resource is represented by an RDF property. Next step is to find a URI for a property matching the attribute in each statement. This research taken the set of namespace from Dublin Core terms[6], ISBD[7], RDA[8] and Resource Description Framework[9], as shown in Table 2.

Table 2. Statements with the attribute replaced by a URI

Statement	dct	isbd	rda	rdf
Title	title			
Subject				subject
Relation				predicate
Object				object
Argument		hasNoteOnTitleProper		
Image	image			
Date			dateOfPublication	
Source	source			

The attributes in the statements derived from the example record can be replaced by the matching property URIs, as shown in Table 3.

Table 3. The attributes in the statements replaced by the matching property URIs

Subject URI	Attribute property URI	Value
MOC:48081	dct:title	วัดท่าเจตีย์ตั้งอยู่ที่ตำบลบางเลน
MOC:48081	rdf:subject	วัดท่าเจตีย์
MOC:48081	rdf:predicate	ตั้งอยู่ที่
MOC:48081	rdf:object	ตำบลบางเลน
MOC:48081	isbd:P1068	LOC
MOC:48081	dct:image	http://m-culture.in.th/media/big/201945.jpeg
MOC:48081	Rda:dateOfPublication	2013-01-01 00:00:01
MOC:48081	dct:source	http://www.m-culture.in.th

The final step is to publish the set of RDF triples derived from the example record as shown in Figure 2. and the rdf graph is shown is Figure 3.

```

@prefix dct: <http://purl.org/dc/terms/>
@prefix isbd: <http://iflastandards.info/ns/isbd/elements/>
@prefix MOC: <http://m-culture.in.th/>
@prefix rda: <http://rdvocab.info/elements>
@prefix ref: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
MOC:48081 isbd:P1014 “Notes on an electrical experiment”
MOC:48081 dct:title “วัดท่าเจดีย์ตั้งอยู่ที่ตำบลบางเลน”
MOC:48081 rdf:subject “วัดท่าเจดีย์”
MOC:48081 rdf:predicate “ตั้งอยู่ที่”
MOC:48081 rdf:object “ตำบลบางเลน”
MOC:48081 isbd:p1068 “LOC”
MOC:48081 dct:image “http://m-culture.in.th/media/big/ 201945.jpeg”
MOC:4801 rda:dateOfPublication “2013-01-01 00:00:01”

```

Fig. 2. a set of triples

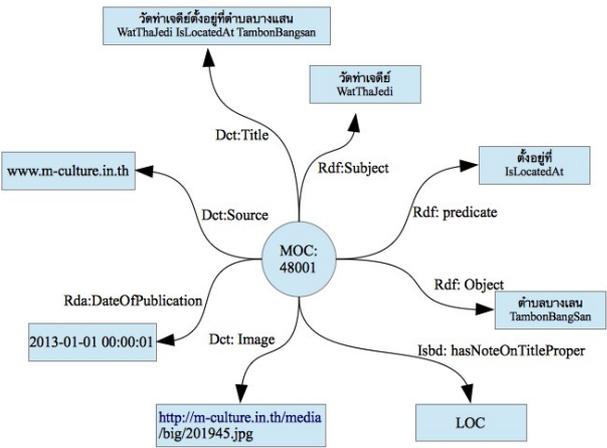


Fig. 3. a set of triples

3. *Linked Open Data Applications*

Linked Open data from semantic relation extraction can build the interesting and useful applications upon them. For example, creating a knowledge map as shown in Figure 4.

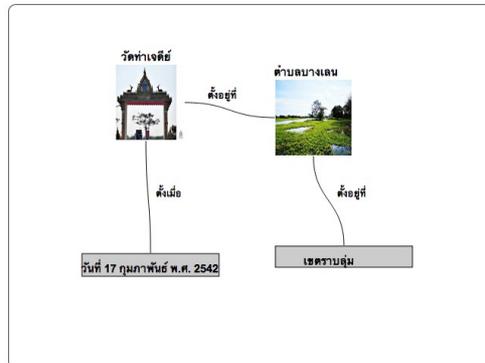


Fig. 4. Linked Open Data Applications

4. Conclusion

The semantic relation extraction for extensive service of a cultural archive is a collection of relation among two entities include place, organization, personal and date. This research approach to creating and publishing linked open data by a simple model; moreover, developing an application for showing relationships between two entities using resource description framework (rdf) in knowledge graph. Possible future work will include more repository of cultural domain documents containing a wider variety of contents is also in progress.

References

1. Watchira Buranasing, Virach Sornlertlamvanich, Thatsanee Charoenporn :Semantic Relation Extraction for Extensive Service of a Cultural Database : The tenth symposium on natural language processing. 2013.
2. Jeff Edelstein, Lola Galla, Carolyn Li-Madeo, Julia Marden, Alison Rhonemus, Noreen Whysel: Linked Open Data for Cultural Heritage: Evolution of an Information Technology: The 31st ACM international conference on Design of communication.2013.
3. Allana Mayer:Linked Open Data and its Implications for Artistic and Cultural Resources:2013.
4. Victor de Boer ,Jan Wielemaker, Judith van Gent, Marijke Oosterbroek, Michiel Hildebrand , Antoine Isaac, Jacco van Ossenbruggen, Guus Schreiber: Amsterdam Museum Linked Open Data: 2012
5. Theeramunkong, T., Boriboon, M., Haruechaiya- sak, C., Kittiphattanabawon, N., Kosawat, K., On- suwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. "Thai-nest: A Framework for Thai Named entity Tagging Specification and Tools". In *Proceedings of CILC.* , 2010.

6. Expressing Qualified Dublin Core in RDF / XML: <http://dublincore.org/documents/dcq-rdf-xml/>
7. INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS: ISBD(M):International Standard Bibliographic Description for Monographic Publications, 2002.
8. Joint Steering Committee for the Revision of Anglo-American Cataloguing Rules: *RDA – Resource Description and Access: Scope and Structure*. Available: <http://www.collectionscanada.ca/jsc/docs/5rda-scope.pdf>.:2006.
9. Eric Miller:An Introduction to the Resource Description Framework: <http://www.dlib.org/dlib/may98/miller/05miller.html>,1998

Ontology Refinement Using Implicit User Preferences: A case study in cultural tourism domain

Krich Nasingkun^{1,2,3}, Mitsuru Ikeda¹, Boontawee Suntisrivaraporn²,
and Thepchai Supnithi³

¹ Japan Advance Institute of Science and Technology
{krich, ikeda}@jaist.ac.jp

² Sirinthorn International Institute of Technology, Thammasart University, Thailand
sun@siit.tu.ac.th

³ Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand
thepchai@nectec.or.th

Abstract. Recommender systems employ static knowledge elicited from experts, causing high cost of maintenance for making the knowledge up-to-date. The contribution of this paper is the proposed method to collect potential concepts from users, in order to assist experts or development of automatic approaches to refining an ontology. Implicit knowledge induced from the users, which is much less expensive to maintain ontology. Ultimately, it offers finer-grained, more effective recommendations that match expectation of the users.

Keywords: Ontology refinement, implicit knowledge, cultural tourism

1 Introduction

Cultural tourism (or culture tourism) is a subset of tourism concerned with a country or region's culture, specifically the lifestyle of the people in those geographical areas, the history of those people, their art, architecture, religion(s), and other elements that helped shape their way of life. The web site of Thai Cultural Knowledge Center [1] is a cultural archive project, implemented through close cooperation between National Electronics and Computer Technology Center and Ministry of Culture under the 2011 Memorandum of Understanding (MOU). The first phase of the project was to develop a technology platform for acquisition, digitization, documentation, preservation, security, and management of complex data in the cultural domain. The second phase focused on integrating data from different sources using different storage technologies, and providing a unified view of the collected data. From November 2010 to June 2013, the database contains more than 100,000 records, linking relevant persons, organizations, places, and artifacts.

It is quite difficult to find recommendations for tourists based on the cultural aspect, since there is abundant knowledge and data. Fig.1 shows an overview of the

recommender system framework for cultural tourism. The cultural portal is the central database storing cultural data obtained by data collection module which is done by officer in Ministry of Culture. To utilize the cultural database, an expert may constructs an ontology based on his/her expertise. Relation extraction is a key process for eliciting knowledge in terms of ontology's instances, concepts, and relations from cultural database. Relation templates which are done in the ontology construction process enable us to extract semantic relation among a focused set of entities in cultural archive [2], which is constrained by relation types and their arguments. In this paper, we focus on the ontology refinement process, to improve and clarify existing knowledge. The better understanding provide the better alternatives for recommendation. User constrains (from user profile) and selection algorithm are deployed to create the final recommendation output.

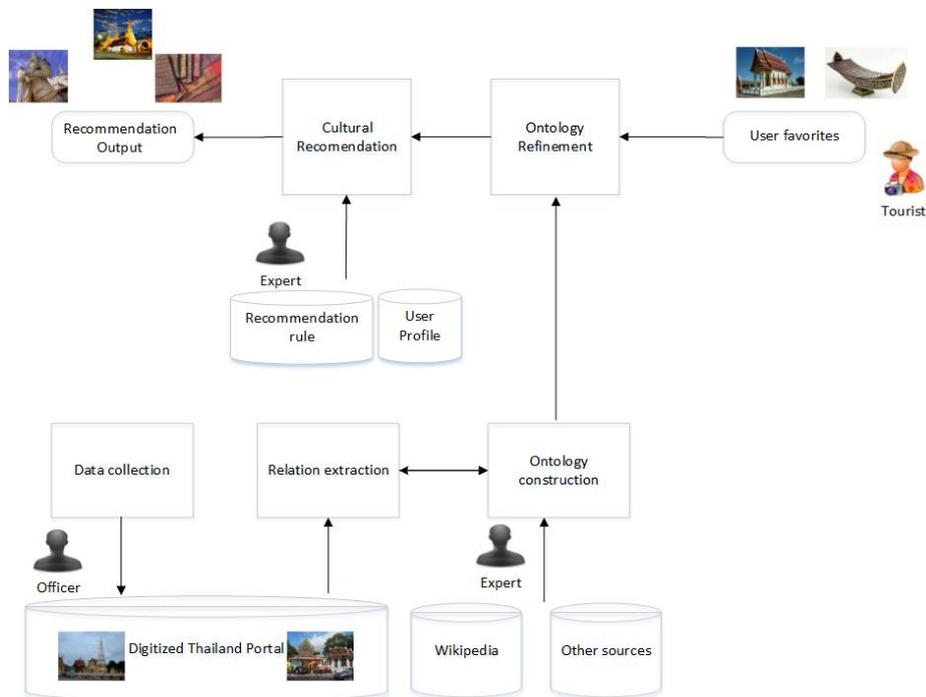


Fig. 1. Cultural tourism recommendation framework

The rest of paper is organized as follows. Section 2 explains our related work. Section 3 presents the proposed method and details of our algorithm. Section 4 illustrates usage scenarios of recommendation system that employ the proposed method. Section 5 shows the discussion of this work. Section 6 provides conclusion and some future development directions.

2 Related Work

Ontology refinement can be categorized into two approaches: semi-automatic and automatic approaches. In the semi-automatic approach, the refinement algorithm aims to help the knowledge engineer find relevant information. This can be done by nominating the terms to reduce the effort of looking for new relevant pieces of information. An example of a technique could be the exploration of statistically significant terms. Term co-occurrences are exploited to identify related terms based on statistical means [3–4]. The automatic approach, on the other hand, does not require a knowledge engineer during ontology refinement but require some principled way to drive the integration of new knowledge in the ontology. These automatic methods rely either on heuristics (like some quality measure), or on information extraction from unstructured source [5]; for example, the expansion of WordNet to the tourism domain [6]. In the biomedical domain, an automated method to refine the Gene Ontology is proposed [7]. The idea is to extract rules based on term variations for automatic term expansion and validate them with the literature. By using IR techniques, the ontology query model identifies missing knowledge in the ontology relevant to IR tasks. An automatic method to revise the ontology accordingly is proposed for generating better queries [8]. Many applied NLP techniques to this approach, but, to the best of our knowledge, none of them concentrate on interests from system users. In our work, we focus on semi-automatic technique to collect a potential concepts using evident from user interest, in order to assist ontology engineer in culture domain.

3 Ontology Refinement Framework

Based on the definition, ontology refinement is a method to improve existing knowledge to more clarify in specific domain. In our work, we proposed the ontology refinement process based-on user interest, to collect the potential concepts which may use to refine ontology in the future. Cultural tourism domain is used to demonstrate an idea of our approach.

3.1 Resource Description Framework

The Resource Description Framework (RDF) [9] is a framework for expressing information about resources. Resources can be anything, including documents, people, physical objects, and abstract concepts. RDF is intended for situations in which information on the web needs to be processed by applications, rather than being only displayed to people. RDF provides a common framework for expressing this information so it can be exchange between applications without loss of meaning. Since it is a common framework, application designers can leverage the availability of common RDF parsers and processing tools. The ability to exchange information between different applications means that the information may be made available to applications other than those for which it was originally created. RDF allows us to make state-

ments about resources. The format of these statements is simple. A statement always has the following structure:

<subject> <predicate> <object>

An RDF statement expresses a relationship between two resources. The subject and the object represent the two resources being related; the predicate represents the nature of their relationship. The relationship is phrased in a directional way (from subject to object) and is called in RDF a property. Because RDF statements consist of three elements they are called triples. Fig.2 show examples of RDF triples (informally expressed in pseudo code).

```
<Bob> <is a> <person>.
<Bob> <is a friend of> <Alice>.
<Bob> <is born on> <the 4th of July 1990>.
<Bob> <is interested in> <the Mona Lisa>.
<the Mona Lisa> <was created by> <Leonardo da Vinci>.
<the video 'La Joconde à Washington'> <is about> <the Mona Lisa>
```

Fig. 2. Example RDF statement (Source: RDF 1.1 Primer N.d.)

In the example above, Bob is the subject of four triples, and the Mona Lisa is the subject of one and the object of two triples. This ability to have the same resource be in the subject position of one triple and the object position of another makes it possible to find connections between triples, which is an important part of RDF's power

3.2 Cultural Tourism Ontology

Cultural tourist has their specific character, their interest not only limit to target destinations/activities itself. But they may gain knowledge in some more aspect around cultural resources. Existing ontology-based recommendation approach has a deep investigate on “*is a*” and “*part of*” relations, meanwhile the similarity measurement among instances of the same or similar concepts are well investigated. As shown in Fig.3, tourist make interest in “*Wat Chong Kham*” and “*Grand Palace*”. Using “*is a*” and “*part of*” from existing ontology approaches, recommender system may recommend another temple or palaces that related to user favorites. Limitation of existing approached cannot capture interest that may related to resources in other aspect. For example, “*King*”, “*Art*”, “*Minority*”, “*Religious*” or “*Traditional and Ritual*” will never been concern.

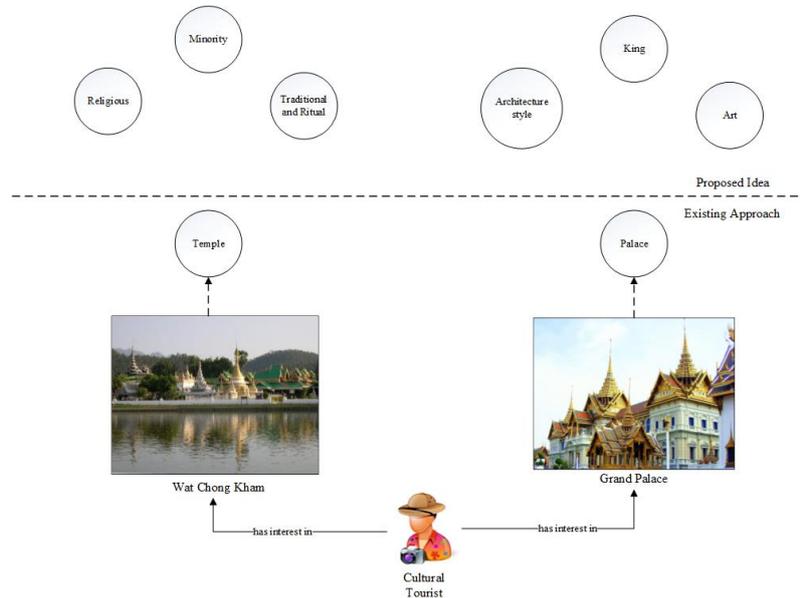


Fig. 3. Conceptual Idea of Cultural Tourist Recommendation

Table.1 show an example relations that we used to model cultural ontology in our approached. Relations are defined to capture cultural aspects of cultural resources. By this approached, cultural aspect of user interest will be analyses in order to recommend the most related on some aspect.

Table1. Example Relations in Cultural Ontology

Relations	Domains	Ranges	Meaning
<i>has_periods</i> (X, Y)	Festival	Date/Time, Date, Periods	X held on specific date or period Y
<i>has_location</i> (X, Y)	Festival, Tradition	Location	X held on specific location Y
<i>has_activities</i> (X, Y ₁ , Y ₂ , .. Y _n)	Festival, Tradition	Activities	X which includes the activities Y ₁ , Y ₂ , .. Y _n
<i>to_celebrate</i> (X, Y)	Festival, Tradition	Religious_event, Seasoning_event, Living_activity	X has a purpose to celebrate an Event Y
<i>to_respect</i> (X, Y)	Festival, Tradition	Animal, God, Water, Rice, etc...	X has a purpose to show respect/thanks to Y
<i>sign_of</i> (X, Y)	Event	Religious	Event X is a sign of religious Y
<i>place_of_religious</i> (X, Y)	Religious_attraction	Religious	X is a religious place belong to religious Y
<i>founded_by</i> (X, Y)	Attraction	Person, Organization	X is an important person who support to create Y
<i>lived_by</i> (X, Y)	Attraction	Person, Organization	X is live in Y

- 1: <Temple> <is a> <Attraction>.
- 2: <Temple> <founded by> <Person>.
- 3: <King> <is a> <Person>.
- 4: <Wat Arun> <is a> <Temple>.
- 5: <Rama II> <is a> <King>.
- 6: <Wat Arun> <founded by> <Rama II>.

Fig. 4. Example RDF Statements in Cultural Ontology

Fig.4 show a partial of cultural ontology that we will use to explain algorithm in next section. Relation “*founded by*” and “*is a*” are a key relations that we will use to identify potential concepts related to user interested. The output of our approached can collected to assist ontology engineer, to refine ontology according to real interested of users.

3.3 Ontology Refinement Process

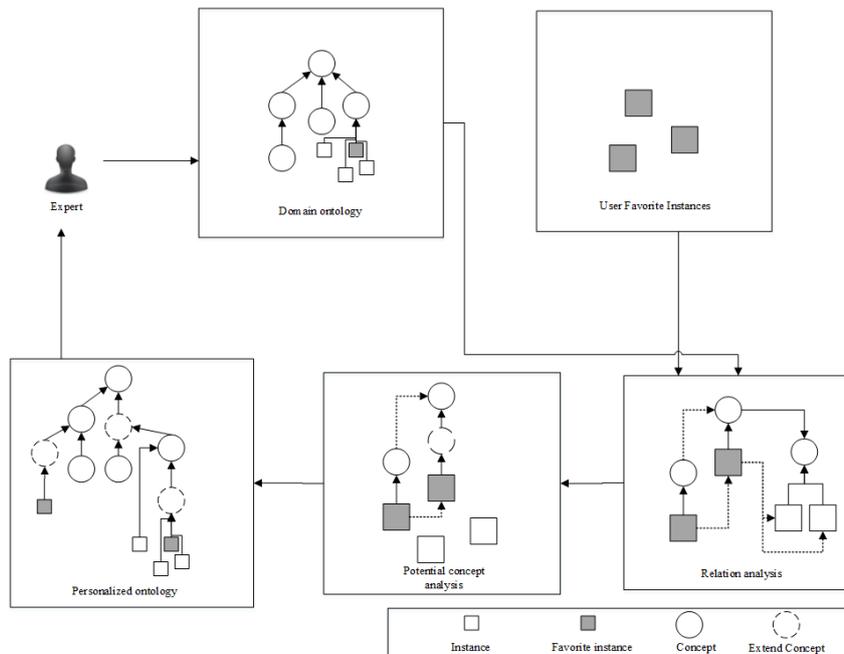


Fig. 5. Ontology Refinement Process

As showed in Fig.5, Domain ontology and user favorite resources are used as an input of Relation Analysis Process. By using evident from user favorite, the related concepts and instances of user favorites are analyzed. Potential concepts will be nominate by Potential Concept Analysis Process, all relations of related concepts are take into account. Possible concepts that may clarify user interest are formulated. However, only the concepts that share common relation are nominated as a potential new knowledge. In Personalized Ontology Process, new knowledge are collect in order to assist expert to update existing domain ontology in future.

3.4 Ontology Refinement Algorithm

In this section, we explain the pseudo code for ontology refinement. The input of our framework is a set of users favorite's resources that input directly from users. Let O be an ontology that modeling by RDF statements, $F = \{f_1, f_2, f_3 \dots f_n\}$ be a set of users' favorite instances. First, we look at each instance to find identify domain concept of relation DC and the related instances RI . Next, we create a temporary concept that contain only a related instances in each relations, then list of concept that related instances belong to set of relation concept RC . the intersection operation is used to identify the unique relation that share common between the same domain concept and range concept. Finally the set of relation $Result$ which store the list of relation of concepts are return as an output. The output of our approach is a personalized extended ontology.

Algorithm 1: Refinement Algorithm

```

Input:  $O = \{\text{rdf triples}\}$ ;  $F$  set of  $n$  user's favorite individuals from  $O$ 
Output:  $\{\text{extended rdf triples}\}$  w.r.t  $O$  and  $F$  (to be added to the personalized ontology)
1:  $ADC \leftarrow \emptyset$ ;  $Result \leftarrow \emptyset$ ;
2: for each  $f \in F$  do
3:    $DC \leftarrow \emptyset$ ;  $RI \leftarrow \emptyset$ ;  $TempC \leftarrow \emptyset$ ;  $RC \leftarrow \emptyset$ ;  $Relation \leftarrow \emptyset$ ;
4:    $DC \leftarrow \{c \mid f \text{ is an instance of concept } c\}$ 
5:    $RI \leftarrow \{g \mid r(f, g) \text{ for some role } r\}$ 
6:    $TempC \leftarrow \{G \mid G \text{ is a temporary concept that has only } RI \text{ as a member}\}$ 
7:    $RC \leftarrow \{G \mid r(f, g) \text{ for some role } r \text{ that } g \text{ is an instance of concept } G\}$ 
8:    $Relation \leftarrow \{r(DC, RC) \text{ and } r(DC, TempC) \mid r(f, g) \text{ for some role } r\}$ 
9:   if not exist  $(DC \text{ in } ADC)$  then
10:     $ADC \leftarrow ADC \cup DC$ 
11:     $Result \leftarrow Result \cup Relation$ 
12:   else
13:     $Result \leftarrow Result \cap Relation$ 
14:   end if
15: end for
16: return  $Result$ 

```

The output from our technique capture the real interest concepts based-on existing ontology structure in the user point of view. Ontology engineer can use this technique to collect the realistic concepts to assist the ontology refinement process.

4 Recommendation Scenarios

This section shows the usage scenarios recommendation system in Cultural recommendation framework. Existing ontology has the structure as shown in Fig.4, tourist identify "Wat Arun" as his favorite place. By our approach, all related instances

and concepts will be investigate. The unique character of user interest will be identify. Finally, the potential knowledge are nominate as an extended of personalized ontology.

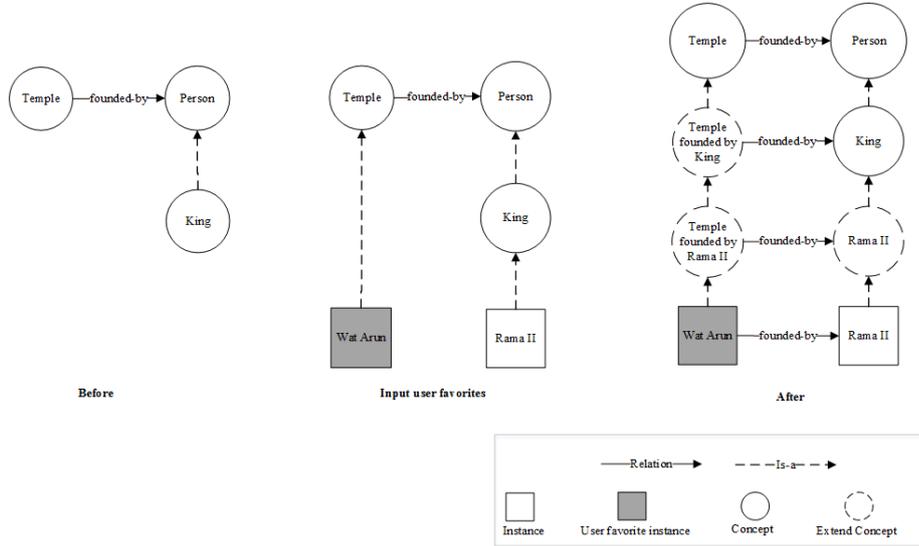


Fig. 6. Example Ontology Extension Generated by the Proposed Technique

Fig.6 show an example user favorite instances and existing ontology structure. The output of our approached is an extended knowledge, Concept “*Temple founded by King*”, “*Temple founded by Rama II*” and “*Rama II*” are nominated to be the part of the personalized ontology.

In the recommendation, instead of interpret the interest to temple according to initial ontology. The emerging of new concept ‘*Temple founded by King*’ lead the interpretation of user interest into a specific group of temple. Instead of 40,717 temples in Thailand, we may scope the number of recommendation using our approach into 217 items. When we collect the list of concepts from many users, ontology engineer will use it to analyze and refine existing domain ontology in the future work.

5 Discussion

In our approach, induced related concepts require instances as supporting evidence. For example, if we have instance of temples as a member of user favorites, it is possible to discover sub concepts of the temples. The shared commons among different types of concepts (ex. Festival, Palace, Temple) without instances, cannot infer the new knowledge. Although we can identify the links among items, the support evidence (instant of concept) still required to prove the intention from users. For example, our technique cannot infer concept ‘*King founded Temple*’ from favorite instance of *Palace* or *Buddhism_Related_Festival* concept (even we have some relation between this two concepts).

Without this approach, existing taxonomy of ontology can produce the similar outputs (for example *Temple founded by King*). However, that concept may not be interested (never be used) by real users. In addition, it will make the over size of ontology problem. In contrast, expert will decide to approve/ignore the inferred knowledge in our approach.

6 Conclusion and Future work

We have presented an ontology refinement approach in cultural tourism domain using implicit knowledge induction from users' favorite resources. New potential concepts based-on user interest are discovered to improve and clarify the existing knowledge.

Some future work includes an implementation of a recommendation framework for cultural tourism and evaluation of the recommendation result. In addition, the ontology refining approaches can be applied to other specific user-oriented domains.

Acknowledgment

This research was conducted under a grant in the SIIT-JAIST-NECTEC Dual Doctoral Degree Program.

References

1. Cultural Knowledge Center, Ministry of Culture, Thailand. <http://www.m-culture.in.th>, accessed September 15, 2014.
2. Buranasing, W., Sornlertlamvanich, V., Charoenporn, T.: Semantic Relation Extraction for Extensive Service of a Cultural Database. Proceedings of the SNLP2013. pp. 241-247 (2013).
3. Faatz, A., Steinmetz, R.: Ontology enrichment with texts from the WWW. Semantic Web Mining. 20 (2002).
4. Cimiano, P., Handschuh, S., Staab, S.: Towards the self-annotating web. Proceedings of the 13th international conference on World Wide Web. pp. 462-47 (2004).
5. Hahn, U., Schnattinger, K.: Towards text knowledge engineering. Hypothesis. 1, (1998).
6. Navigli, R., Velardi, P.: Automatic Adaptation of WordNet to Domains. Proceeding of the LREC2002. pp. 1023-1027 (2002).

7. Lee, J.-B., Kim, J. -j., Park, J.C.: Automatic extension of Gene Ontology with flexible identification of candidate terms. *Bioinformatics*. 22, 665–670 (2006).
8. Jimeno-Yepes, A., Berlanga-Llavori, R., Rebholz-Schuhmann, D.: Ontology refinement for improved information retrieval. *Information Processing & Management*. 46, 426–435 (2010).
9. RDF 1.1 Primer N.d. <http://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>, accessed October 5, 2014.

Community-Driven Approach to Large-Scale Ontology Development based on OAM Framework: a Case Study on Life Cycle Assessment

Akkharawoot Takhom¹, Prachya Boonkwan², Mitsuru Ikeda³,
Boontawee Suntisrivaraporn¹, and Thepchai Supnithi²

¹School of Information, Computer and Communication Technology
Sirindhorn International Institute of Technology (SIIT), Thammasat University, Thailand
akkharawoot.t@gmail.com, sun@siit.tu.ac.th

²Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand
{prachya.boonkwan, thepchai.supnithi}@nectec.or.th

³School of Knowledge Science
Japan Advanced Institute of Science and Technology (JAIST), Nomi, Japan
ikeda@jaist.ac.jp

Abstract. This paper presents a community-driven development framework for large-scale ontology that offers intra- and inter-community communication, voting and endorsement system, and version control. Our system design addresses three lacks in the traditional ontology development: (1) constructive communication among the relevant stakeholders, (2) consensual endorsement and voting system for concept and structural augmentation, and (3) bookkeeping and version control. To cope with these shortcomings, we introduce three more tiers: user management with technical profiles, knowledge augmentation via a web-board, and community collaboration via the vote and endorsement system, on top of the existing Ontology-based Application Management Framework (OAM) [1]. When applying the design to the task of Life Cycle Assessment, we discover additional needs to partition a large ontology to modularize the users' responsibilities and to provide thread-based conversation for keeping track of the topics.

Keywords: ontology development framework, community-driven approach, large-scale ontology, knowledge management, version control, Life Cycle Assessment

1 Introduction

Life Cycle Assessment (LCA) is a kind of measures of environmental impact which is used in environmental impact assessment (EIA). Offering opportunities to environmental decontamination, it quantifies energy and materials that are consumed in the

process and released to the environment, and assesses the environmental impacts of human activities and industries. The comprehensive knowledge of LCA is large-scale, voluminous, and industry-specific. It has to be acquired from the published standard documentations (ISO 14040 [2] , 14044 [3], and 14048 [4]), and numerous relevant case studies, resulting in a very steep learning curve for the novice users. Development and management of such knowledge are laborious and require a close collaboration among experts. Therefore, knowledge transfer in the relevant stakeholder community becomes non-trivial in this task.

For ease of knowledge transfer, parts of LCA knowledge have been elicited in terms of ontologies such as CASCADE [5], LCAO [6], Semantic Oil [7], and O-LCA [8, 9], and transcribed into several representations, such as Semantic Web [10] encoded as OWL [11], Predicate Logic, and Description Logic [12]. The notion of ‘ontology’ promotes common understanding among the experts, makes the knowledge scalable and reusable, and improves the interoperability between the experts and the users and across relevant applications. Despite precise elicitation, the development of a large-scale ontology is still an immense challenge because it requires effective knowledge management when the entire community of experts are involved.

This issue can be alleviated by the notion of community-driven development approach. In this approach, it is assumed that every stakeholder is willing to participate in the development and maintenance and contribute to the community with his expertise. We focus on multidisciplinary knowledge integration and cross-checking among domain experts and relevant stakeholders. A large-scale ontology is constructed and maintained on a common platform on which the community of experts, knowledge engineers, and users controls the development and maintenance processes, where each decision making is done based on the community’s consensus. The platform provides a means to communicate, transfer tacit knowledge, and discuss the changes of the worked ontology with respect to immediate needs. Community-driven development has been used in several ways, such as ontology matching [13, 14] and knowledge curation [15]. There have been attempts to large-scale ontology development [16–20].

The community-driven approach is suitable for the development of LCA ontology because of the following reasons. First, the domain experts specialize in their particular subfields of LCA. Since these fields sometimes share common knowledge, cross-checking becomes necessary in a large-scale development project. Second, the development of the LCA ontology is operated by a group of experts in parallel. In practice, they usually branch (or *fork*) the current version of the ontology to work on their own. This causes problems in updating when the finished ontologies are to be merged back to the main ontology; thus, the need of version control. Third and last, some parts of the ontology have to be cross-checked by specialists from other relevant fields; for example, some parts of the ontology regarding earth and water can also be validated by geophysicists, chemists, and environmentalists.

We propose the use of the Ontology-based Application Management Framework (OAM) [1] as a community-driven development platform of the LCA Ontology. The current framework offers the following facilities:

1. an ontology management tool that allows the community to get involved in the evolution of the ontology,

2. a sandbox toolkit with application templates for a semantic search engine and a recommender system, where no programming skills are required, and
3. APIs and web service deployment for practical application development.

However, it lacks collaborative features that enables constructive communication among the stakeholders. The contribution of this paper is: we introduce to OAM the notion of modified webboard, where any concept and structural augmentations are proposed by experts and they have to be endorsed by the community. This allows the system and the knowledge to grow along with the users' expertise.

In this paper, we will show that the OAM Framework does not fully support community-driven development of a large-scale ontology where constructive communication is needed (section 2). Then we will show that our modified webboard enables such communication oriented by concept and structural augmentations and consensual endorsement (section 3). We will also explain how to apply this technique to the field of LCA (section 4).

2 Problem Statement

OAM Framework [1] is a software platform that aims to simplify the development and maintenance of a semantic web and an ontology as well as to automate the implementation of a semantic search and a web service. The architecture of OAM is illustrated in Fig. 1.

Ontology development via OAM entails three fundamental steps and three user roles: domain experts, a knowledge engineer, and application developers. First, a domain expert designs his own ontology according to the task of interest and export it in the OWL format. On OAM, the knowledge engineer maps the ontology designed by the experts to the database schema and the vocabulary and imports it into the system. At this step, the knowledge engineer also maintains the current ontology according to the experts' requests via personal communication.

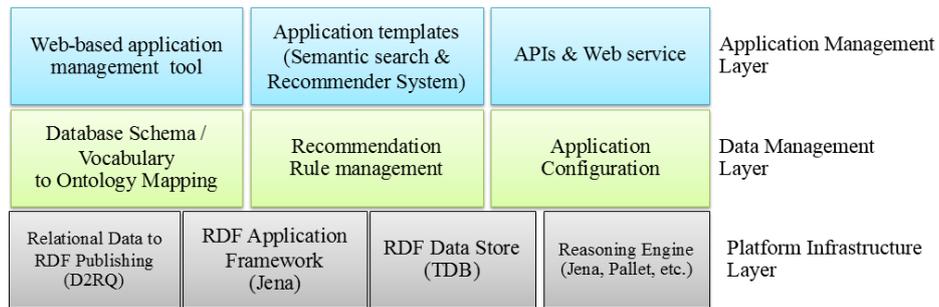


Fig. 1 System architecture of OAM Framework [1]

Second, the domain experts design recommendation rules for the current ontology in terms of Prolog-like first-order logic. The knowledge engineer then transcribes

these rules into JENA Language via the Recommendation Rule Management Module. Finally, the knowledge users implement their own knowledge-enhanced applications with the Application Configuration Module. At this stage they can deploy a semantic search engine and a web service using the ontology developed by the domain experts.

The key struggle in this paradigm is the ontology development entirely relies on personal communication. This method is prone to the loss of communication; i.e. it is very hard to keep track of conversations and consensus as time goes by. The history of development evolution, or version control, plays a crucial role in community-based development, especially for a large-scale ontology, in which a group of domain experts, knowledge engineers, and knowledge users are involved. These lacks necessitate the use of a tractable communication means where conversations and consensus are structurally organized for ease of bookkeeping, knowledge transcription, versioning, and deployment.

We are proposing an extension of the OAM Framework that incorporates the notion of thread-based webboard, version control, and status notification to solve the aforementioned problems. These features allow the community with the three user roles to co-create a large-scale ontology and maintain it by means of community endorsement. By doing so, the system and the knowledge grow along with the users' expertise.

3 System Design

In this paper, we design a community-driven ontology-based application management framework (CD-OAM). We extend the canonical OAM Framework (shown as Data Tier and Application Tier, all in dashed borders) with three more tiers: Collaboration Tier, Knowledge Tier, and User Tier (all in solid borders). A system overview of CD-OAM is illustrated in Fig. 2.

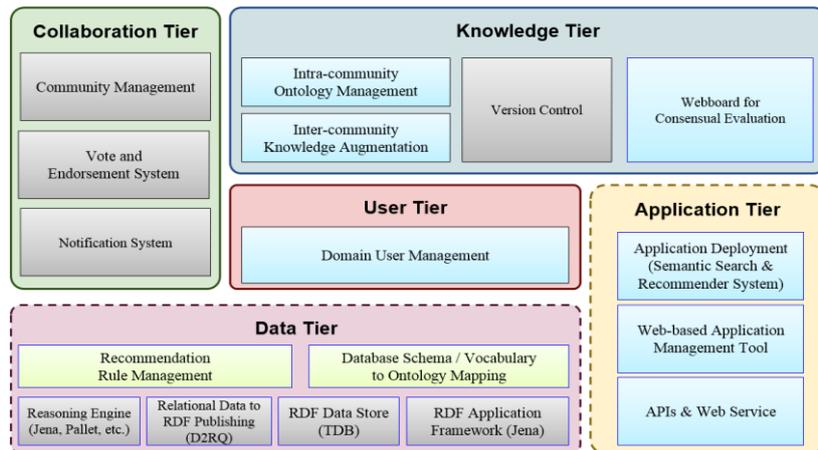


Fig. 2 System Overview

First, the User Tier facilitates the admin to manage domain expert members, including adding, deleting, and assigning to some of the communities. The users are required to create a profile annotated with their expertises so as to classify them with respect to their interests, technical backgrounds, and objectives. Our user policy is merit-based [14]: the more a user participates in votings or change proposals, the more merits he gets.

Next, The Collaboration Tier, the heart of this framework, encourages the community-driven development of a large-scale ontology. It contains three modules: community management, vote and endorsement system, and notification system. The Community Management module allows the community's administrator to manage their members and their user roles. The Vote and Endorsement System prepares a platform for voting and endorsement by the community when changes in the ontology take place. Finally, the Notification System disseminates the voting results and changes to the community.

Finally, the Knowledge Tier is composed of four modules that facilitates both intra- and inter-community communication in large-scale ontology development and keep track of the evolution. We assume that a large-scale ontology can be separated into parts and distributed to all involved communities. The Intra-community Ontology Management Module [15] allows the community to manipulate the part of the ontology the community is responsible for. The Inter-community Knowledge Augmentation Module facilitates the integration of each part of the ontology. Should modification is necessary, this can be voted and endorsed by all relevant communities. All proposed changes must be endorsed by the community. The Webboard Module organizes a vote for intra- and inter-community changes in the ontology and approves the consensus. For example, an expert may propose to augment some concepts or the hierarchy in the Webboard. If changes are endorsed and committed, the history is kept in the Version Control Module.

4 Case Study: LCA

Ontology development for LCA is a very challenging task. Standardized by the ISO into ISO 14040, 14044, and 14048, LCA is a family of best-practice procedures for information sharing and guidelines for environmental impact evaluation. It minimizes operations in the organization which affect the environment, to comply with laws, regulations, and other environmentally oriented requirements and continual improvement. Figure 3 shows the ISO guideline for LCA that consists of four main phases: (1) setting the goal and scopes (2) listing up life cycle inventory from a given supply chain (3) assessing the life cycle impacts and (4) data interpretation. The ontology for LCA is usually very large because there are numerous concepts and a complicated hierarchy and relations between them in the product's life cycle.

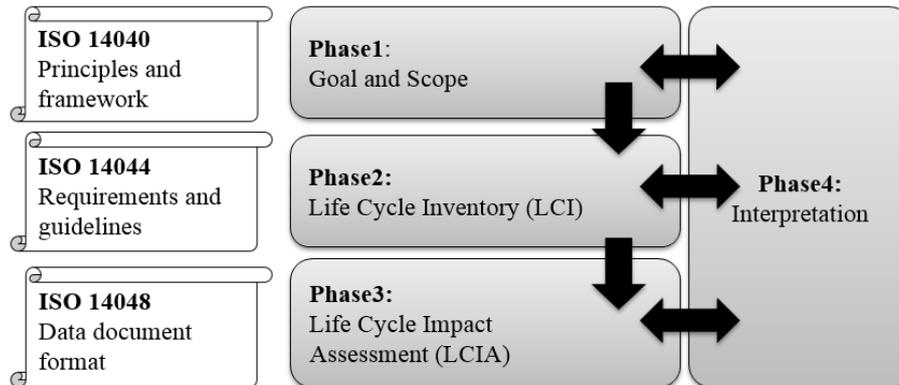


Fig. 3 ISO standards guideline of LCA

Due to its sheer scale, there are a number of stakeholders working side by side on the ontology. We classify them into five categories with respect to their objectives:

1. Domain experts: experts and specialists on LCA
2. Domain beginners: beginners who just started in the field and are learning the knowledge
3. Knowledge users: direct users of the knowledge via a recommender system
4. System developers: software developers who make use of the knowledge
5. Knowledge engineers

Complying with the community-based approach, these stakeholders have different roles in developing the LCA ontology and they all need a means to communicate to each other. Any requests of change in the ontology have to be endorsed by the entire community. Moreover, version control becomes necessary as the ontology grows by the community. Our system design fit all these requirements for the community-based LCA Ontology Development.

Our system design has to be specialized as per the roles and interests of the users. We envisage the following modifications in the following tiers when applied to the field of LCA.

- **User Tier:** Each user should specify his objectives, role, interest, and technical backgrounds regarding LCA. These users are then assigned to a different part of the ontology according to their request or technical backgrounds.
- **Collaboration Tier:** To modularize the user responsibilities, the community are able to partition (and repartition) the ontology via voting and endorsement.
- **Knowledge Tier:** There are at least two sources of knowledge in LCA: ISO standards and field data collection. Knowledge integration becomes a non-trivial issue because domain experts may have different perspectives on the received data. The webboard module has to offer a place for open discussion before leading to voting and endorsement. To keep track of the decision, a thread-based conversation is best suitable for this scenario.

5 Conclusion and Future Work

We have presented a design for a community-driven development framework for large-scale ontology, where all relevant stakeholders can participate in the evolution of the ontology. Three additional layers: user management, knowledge augmentation, and community collaboration, respectively, are put on top of the OAM Framework to facilitate communications among the stakeholders, voting and endorsement, and version control. We also find that, in practice, ontology partitioning and thread-based conversation are also needed for teamwork as found in Life Cycle Assessment.

Our future work remains as follows. First, we will implement the system closely following this design and conforming to User Experience (UX). Second, we will incorporate the notions of conflict detection and concept similarity into the knowledge augmentation module. This will help knowledge engineers and the community pre-determine holistic conflicts in the ontology before committing any changes. Third and finally, we will develop a data modeling module [21] which automatically maps input databases with a different data scheme to the ontology.

Acknowledgment

This research is financially supported by Japan Advanced Institute of Science and Technology (JAIST), Sirindhorn International Institute of Technology (SIIT), and Thammasat University (TU). Materials and information provided by National Metal and Materials Technology Center (MTEC, Life Cycle Assessment Laboratory, Thailand). The authors are grateful to Marut Buranarach for fruitful technical discussions.

References

1. Buranarach, M., Thein, Y., Supnithi, T.: A Community-Driven Approach to Development of an Ontology-Based Application Management Framework. In: Takeda, H., Qu, Y., Mizoguchi, R., and Kitamura, Y. (eds.) *Semantic Technology*. pp. 306–312. Springer Berlin Heidelberg (2013).
2. ISO 14040:2006 - Environmental management -- Life cycle assessment -- Principles and framework. ISO, Geneva, Switzerland (2006).
3. ISO 14044:2006 - Environmental management -- Life cycle assessment -- Requirements and Guidelines, (2006).
4. ISO/TS 14048:2002 - Environmental management -- Life cycle assessment -- Data documentation format, (2002).
5. Cappellaro, F., Masoni, P., Moreno, A., Scalbi, S.: CASCADE. In: Werner Pillmann, K.T. (ed.) *The 16th Internationale Conference: informatics for environment protection*. pp. 490–493. IGU/ISEP (2002).

6. Brascher, M., Monteiro, F., Silva, A.: Life cycle assessment ontology. the 8th Conference of the International Society for Knowledge Organization. pp. 169–177. Congreso ISKO-España, España, Spain (2007).
7. Bertin, B., Scuturici, V., Risler, E., Pinon, J.: A semantic approach to life cycle assessment applied on energy environmental impact data management. Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 87–94 (2012).
8. Takhom, A., Suntisrivaraporn, B., Supnithi, T., Theeramunkong, T., Manabu, O.: Ontology-enhanced life cycle assessment: toward formalizing the standard guidelines. The International Conference on Information and Communication Technology for Embedded Systems (ICICTES 2013). , Samutsongkhram, Thailand, (2013).
9. Takhom, A., Suntisrivaraporn, B., Supnithi, T.: Ontology-enhanced life cycle assessment: a case study of application in oil refinery. The Second Asian Conference on Information Systems, (ACIS 2013) (2013).
10. Horrocks, I.: Ontologies and the semantic web. *Commun. ACM.* 51, 58–67 (2008).
11. Schreiber, G., Dean, M.: OWL Web Ontology Language Reference. (2004).
12. Baader, F., Horrocks, I., Sattler, U.: Description Logics. *Stud. Health Technol. Inform.* 101, 137–41 (2004).
13. Zhdanova, A. V: Towards Community-Driven Ontology Matching. *Computer (Long. Beach. Calif.)*. 1–2 (2005).
14. Zhdanova, A. V, Shvaiko, P.: Community-driven Ontology Matching. Proceedings of the 3rd European Conference on The Semantic Web: Research and Applications. pp. 34–49. Springer-Verlag, Berlin, Heidelberg (2006).
15. Groza, T., Tudorache, T., Dumontier, M.: Commentary: State of the Art and Open Challenges in Community-driven Knowledge Curation. *J. Biomed. Informatics.* 46, 1–4 (2013).
16. Gendarmi, D., Lanubile, F.: Community-Driven Ontology Evolution Based on Folksonomies. *Move to Meaningful Internet Syst.* 2006 OTM 2006 Work. 181–188 (2006).
17. Hepp, M., Bachlechner, D., Siorpaes, K.: OntoWiki: Community-driven Ontology Engineering and Ontology Usage Based on Wikis. Proceedings of the 2006 International Symposium on Wikis. pp. 143–144. ACM, New York, NY, USA (2006).

18. Siorpaes, K.: Lightweight community-driven ontology evolution. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 951–955 (2007).
19. Aseeri, A., Wongthongtham, P.: Community-driven ontology evolution based on lightweight social networking in oil and gas domain. IEEE International Conference on Digital Ecosystems and Technologies. pp. 197–202 (2011).
20. Maleewong, K., Anutariya, C., Wuwongse, V.: A Semantic Argumentation Approach to Collaborative Ontology Engineering. Proceedings of the 11th International Conference on Information Integration and Web-based Applications and Services. pp. 56–63. ACM, New York, NY, USA (2009).
21. Bertin, B., Scuturici, V.-M., Pinon, J.-M., Emmanuel Risler: A semantic approach to life cycle assessment applied on energy environmental impact data management. Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 87–94. ACM, New York, NY, USA (2012).

Poster and Demonstration Papers of JIST 2014

Gathering Photos from Social Networks using Semantic Technologies

M.B. Alves^{1,2}, C. V. Damásio¹, N. Correia³

CENTRIA, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.
ESTG, Instituto Politécnico de Viana do Castelo 4900-348 Viana do Castelo.
CITI, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.
mba@estg.ipv.pt, cd@fct.unl.pt, nmc@fct.unl.pt

Abstract. We present a system based on a Semantic Web approach to retrieve photos from events, organizations or even related to semantic concepts using context and social information, without require extra annotation work by the publishers. We make use of the knowledge of the system, represented in ontologies and semantic rules, to allow users search using its own terminology.

1 Introduction

The motivation of our work is moving towards an easy searching and browsing of photos collections as is required by the huge amount of digital photo collections made available on the web by the photographers in different repositories, like Facebook, Picasa or Flickr. The gap between the photo information and the users desire is addressed by us through a Semantic Web approach, namely by a) associating photo metadata with precisely defined semantics, represented through ontologies [4], and b) through reasoning over this information. In this work, we describe a system that searches for photos on the web, that can be tailored to a specific domain like events, organisations, sports. We combine information that we have in ontologies with the information about the photos, either photo metadata or contextual information, to retrieve photos of a given event, from a given person, or in a given place. This system acts as a semantic mashup of photos collection, in other words, as a personal collection, but in reality these photos are spread in the web, social networks and photo databases. Our system requires a knowledge model of the domain, represented by a domain ontology and by rules. This information model gives to the system knowledge about the domain, allowing queries in a terminology recognised by the user. As a result, we are dealing with the semantic gap that exists in multimedia content [9], the lack of coincidence between the information that can be extracted from the photo and the meaning of that photo to the human.

This document is organised as follows. In section 2 we present our approach and we explain how a semantic web approach can meet our purposes in photo retrieval. In section 3 we detail our system architecture and we explain each component of the architecture and we finish with the conclusions where we analyze our work and we present some benchmarking in section 4.

2 Semantic Web approach for multimedia retrieval

In our Semantic Web approach we use ontologies to formally describe the domain. Upper-level ontologies, describing very general concepts that are the same across all domains, are used to modelling concepts such as Events, Photos, Time, etc. In this work we have used: Ontology for Media Resources [5], to describe media resources; LOD2 [7], ontology for Linking Open Descriptions of Events; Time ontology [1], FOAF [2] for describing persons. Each domain, for which we want to implement a photo retrieval engine with the approach described in this work, requires a particular domain ontology to represent the domain-specificities. We will model the domain ontology using OWL 2 [3]. To overcome some of the OWL limitations [6], we also use semantic rules to add expressivity and expertise to our model, giving support to some object properties, which link individuals to individuals.

Now, we will illustrate how our Semantic Web approach is used to support photos retrieval, giving meaning to the content. To simplify reading we will use the well known prefixes of the used ontologies and the prefix *don* is our domain ontology. In our domain ontology, we define an **objectProperty** *isPhotoOfEvent* having as domain *ma:Image* and as range *lode:Event* to make the relationship between photos and events instances. We define some rules to deduce, in our system, when a given *Photo* can belong to a given *Event*. These are general rules independent of the specific domain. **Rule 1** is one of these rules, present using the syntax Conclusion \leftarrow Premises (body atoms).

Rule 1: (?Photo don:isPhotoOfEvent ?Event) \leftarrow
(?Photo don:wasTakenAtTimeOf ?Event), (?Photo don:wasTakenInSamePlace ?Event),
(?Person don:isTaggedIn ?Photo), (?Person don:participates ?Event).

With **Rule 1**, we declare that a given *Photo* belongs to a given *Event* if: *s1*) the *Photo* was taken at the same time of the *Event*; *s2*) that *Photo* was taken in the same place of the *Event*; *s3*) there is at least one *Person* that was tagged in the *Photo*; *s4*) who participated in the *Event*. The first statement, *s1*, makes use of the **objectProperty** *wasTakenAtTimeOf* which is supported by rules to define if a given photo was taken while the event occurred. The *isTaggedIn* and *participates* are also object properties that define if a *Person* is tagged in a *Photo* and if a *Person* participates in a *Event*. Notice that **Rule 1** cannot be captured by OWL 2 role inclusion chain axioms.

In our system, we allow to give a confidence to a relation (real number between 0.0 and 1.0). The RDF reification vocabulary is extended to associate a confidence to each inferred triple used in search dimensions/values, according to the following pattern stating that a given *Relation* between a *Subject* and an *Object* has a given *Confidence*:

```
?bn rdf:subject Subject . ?bn rdf:predicate Relation .  
?bn rdf:object Object . ?bn don:confidence Confidence .
```

As we can define queries with confidence we can define weaker relations, for instance between photos and events, and give a confidence to them. This confidence will be used in ranking the photos.

3 System Architecture

In figure 1 is represented the architecture of our tool. Our system uses the APIs of Facebook, Flickr and Picasa to retrieve the photos, the information associated with the photo and the context information. We simply use the information that we can obtain with the photo and the normal operations that the users usually perform: tagging photos, with text tags or people, indicate the place where the photos were taken, or give a name to an album. The information is extracted from the multimedia databases and is kept in a knowledge base to support the inference engine. The photos are available to the users through a web interface that retrieve the photos that answer to the users queries.

ETL process - The Extract, Transform and Load process (ETL), acts like a web crawler. The metadata that we can get from these new photos is extracted and saved using the vocabulary specified by the Ontology for Media Resources. The transform task can make use of the semantic model and the rules to do inference and materialize some of the consequences in the pre-processing task due to performance issues.

Semantic Model and rules - Each domain has its own vocabulary, the terms that are recognised by the users. Therefore, a domain ontology must be developed to represent the application domain, providing the knowledge of the domain that allows us to achieve a higher precision in the retrieval process. Without a domain ontology, we couldn't use these terms as a search dimension/facet.

The query process and the interface - The users make queries through the Web interface. The system returns the links to the photos that are related to the queries. These queries are answered using the information kept in the knowledge base, making inferences from the semantic model and the rules. We have a semantic model to support the querying. This semantic model is a meta-model within which the classes of search dimensions are defined in the web interface. This meta-model, that must be tailored to each different domain, allows the system to know what must be searched. With this approach, we do not need any change in code in a new system. Everything is knowledge provided to the system and everything is kept outside in simple configuration files. As we can give a confidence to a relation, we can rank the images using the confidence value of inferred triples.

System details - Our implementation is based on Jena framework¹, which offers an "all-in-one" solution for Java, OWL reasoning, inference and rule engine. We use Pellet [8] together with Jena to make OWL 2 inference. The TDB component of Jena is used for RDF storage and query.

¹ <https://jena.apache.org>

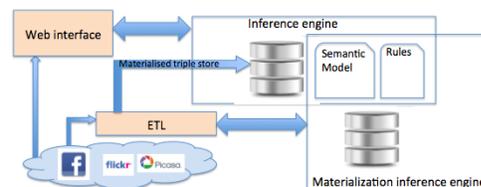


Fig. 1. System architecture of the photos retrieval engine

4 Conclusions

In this work, we presented a system to retrieve photos using a Semantic Web approach. Our system uses the context information of the photo or annotations done by the user and other metadata that can be retrieved from social networks to combine with the knowledge of the system to classify the photos. The user can search for photos using its own terminology, creating dynamically its own personal collections despite these photos are distributed along the web. . We can perform this because we have knowledge of the domain and, in this way, we try to overcome the problem of the semantic gap between the photos information and the means of that photo to the user.

The precision and the recall of our system it is an open question, and depends of each new implemented domain and the requirements. If we want to give relevance to the precision, we define rules that represent exactly one relationship or, at most, with a high probability of occurrence. If we want to give relevance to the recall, we define rules to represent how relationships may happen, even if those rules bring some fake positive results. As we can introduce a confidence factor in our rules, we can give relevance to the recall but improving the F1-Score, a measure that combines precision and recall. Even though precision and recall be an open question, we performed some benchmarking with a small example using a Facebook account of a swimming club with 3599 friends. The system focused on 49 albums and 1148 photos. It were retrieved 431 photos, 425 of which were distributed over 9 albums. We had 100% precision in the retrieved photos. We inspected the photos of the swimming club events that were published but that were not retrieved, achieving a recall of 95,5%. An implementation in a real situation is on-going work that can be tried at <http://www.estg.ipv.pt/~mba/SemanticPhotosSearch/>.

The system presented in this work was tailored to be enhanced at production time. The semantic model and the rules are kept in configuration files. Any change to these files only requires an re-initialisation of the system. Thus, we can readily improve our semantic model or our rules, adding new knowledge to our system or refining the existing one.

References

1. Time ontology in OWL. Electronic, September 2005.
2. D. Brickley and L. Miller. FOAF Vocab. Spec. 0.97. Namespace doc., 2010.
3. Cuenca et al. OWL2: The next step for OWL. *Web Semantics*, 6(4):309–322, 2008.
4. T. Hofweber. Logic and ontology. In *The Stanford Ency. of Phil.* Spring, 2009.
5. Lee et al. Ontology for Media Resources 1.0, Feb. 2012. W3C Recommendation, <http://www.w3.org/TR/mediaont-10/>.
6. B. Parsia et al. Cautiously approaching SWRL, 2005.
7. R. Shaw et al. Lode: Linking open descriptions of events. In *Proc. of the 4th Asian Conf. on The Semantic Web, ASWC '09*, pages 153–167, Berlin, Heidelberg, 2009.
8. E. Sirin et al. Pellet: A practical OWL-DL reasoner. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(2):51–53, jun 2007.
9. Smeulders et al. Content-based image retrieval at the end of the early years, 2000.

An Overview of the Linked Data AppStore

~ Demo/Poster Paper ~

Dumitru Roman, Claudia D. Pop, Roxana I. Roman, Bjørn M. Mathisen,
Leendert Wienhofen, Brian Elvesæter, and Arne J. Berre

SINTEF, Oslo, Norway

Contact: dumitru.roman@sintef.no

Abstract. This demo/poster paper provides an overview of a Software-as-a-Service platform prototype for data integration on the Web – The Linked Data AppStore (LD-AppStore). It builds upon Linked Data technologies, targets data scientists/engineers and data integration application developers, and aims to provide a solution for simplifying tasks such as data transformation, querying, entity extraction, data visualization, crawling, etc. This paper focuses on the overall architecture of the LD-AppStore, basic data operations supported by the current prototype, and outlines the demonstration of the prototype.

1 Motivation

In recent years a significant amount of data has been made available as Open and/or Linked Data, however applications utilizing such data have been rather few.¹ Reasons include, amongst others, the technical complexity and economical cost of integration, publishing, interlinking and providing reliable access to the data, and lack of simplified and unified solutions for data consumption, and lack of tools and infrastructures where datasets and 3rd party components can be made easily available to application developers to reuse, combine and develop novel data-driven applications. At present, Linked Data publishers and application developers need to rely on generic platforms (like the Amazon Web Services or Google App Engine cloud providers), and build, deploy and maintain complex Linked Data software and data stacks from scratch. Tools addressing various aspects of data integration process, though available in a Linked Data context, are difficult to use for more complex, interesting data integration tasks. This results in a high cost of data integration at large scale, a rather complicated and time consuming process. New innovative ways of simplifying data integration in a Linked Data context are needed.

¹ As of Sept 2014, for example, the official EU public open data portal (<http://publicdata.eu/>) contains more than 48,000 datasets but lists less than 80 applications using the data. The situation is not much different for other open data portals (see e.g. <http://www.datacatalogs.org/>).

To simplify the data integration process, and support data publishers and application developers, this paper provides an overview of a Software-as-a-Service platform—*The Linked Data AppStore (LD-AppStore)*—for data scientists/engineers aiming to enable them to use, in a rather simplified manner, tools/services for tasks such as data transformation, entity extraction, data visualization, crawling, etc. At the same time, data integration application developers have the possibility of exploiting the use of their tools/services by plugging them into the LD-AppStore.

2 The LD-AppStore Platform Overview

The LD-AppStore is meant to be a service where data engineers can get access to various types of data operations, such as data transformation, storage, querying, linking, visualization, etc., which they can apply on their data, and have access to various tool/service implementations of those data operations – implementations provided by developers. The LD-AppStore serves as a registry of data operations and their implementations.

Figure 1 provides a high level overview of the LD-AppStore architecture.

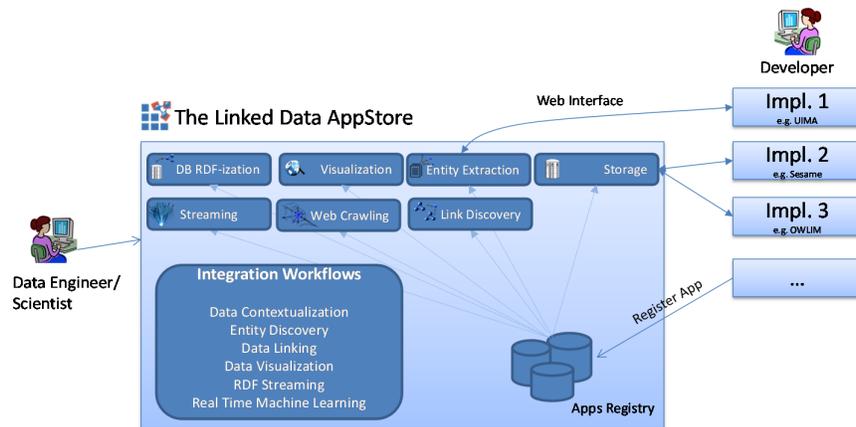


Figure 1. LD-AppStore Architecture Overview

The upper part of the picture depicts components for basic data operations, currently being considered: RDF-ization of relational databases (mapping relational tables to RDF graphs), data visualization (visualization of RDF graphs), entity extraction (extracting entities from various sources), data storage (storage of RDF data manipulated in the platform), link discovery (finding links between data in RDF graphs), crawling (searching through RDF graphs), and data streaming (querying streams of RDF data). A set of Web APIs have been designed for these data operations. The set of tools/services that implement these basic data operations are made available through the registry functionality of the platform (lower right part of the figure). When using a specific data operation, the data engineer may select which implementation of that operation he/she wants to use. The Linked Data tool/service developers have access to the platform for

registering their implementations, i.e., the implementations of the Web APIs corresponding to the data operations APIs. The lower left part depicts a set of data integration workflows meant to seamlessly combine the basic data operations in workflows (configurable by the data engineers) that can eventually provide further useful insights into the data on which they are applied.

In the current design, the platform offers six different types of basic data operations for which Web APIs have been designed: *DB-RDFization* (for mapping data from relational databases to RDF); *Entity Extraction* (for extracting entities from various sources); *Data Visualization*; *Storage* (for storing/querying data); *Streaming* (for querying streams of data); *Link Discovery* (for discovering relations between different datasets); and *Web Crawling* (for searching Linked Data).

3 The LD-AppStore Prototype and Demonstration

The current implementation of the LD-AppStore that will be demonstrated consists of the backend infrastructure for registering applications/tools implementing the APIs of data operations, the graphical frontend infrastructure through which data engineers can access the various data operations and the tools/services that implement them, as well as a set of tools that have been modified to implement the above mentioned APIs. Figure 2 provides a screenshot of the LD-AppStore homepage.

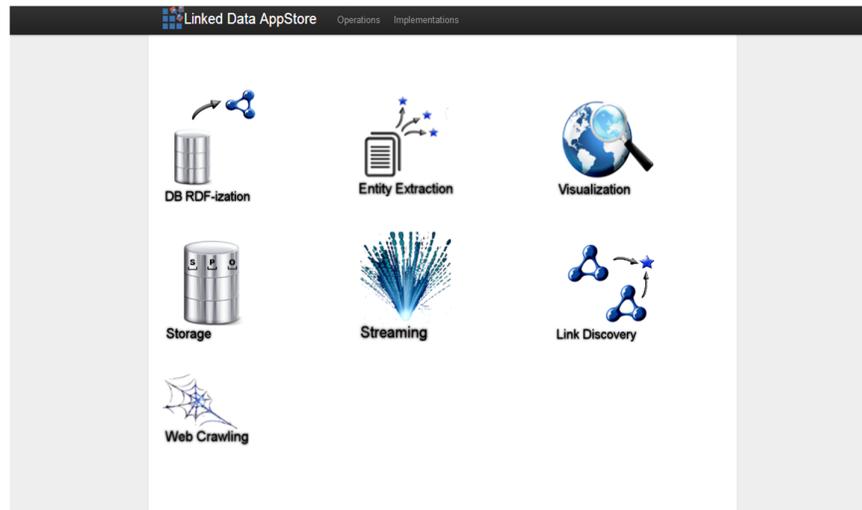


Figure 2. Screenshot of the LD-AppStore homepage.

The platform offers the possibility to register new tools/services as implementations for various operations. For each of the already registered tools a programmatic Web interface has been made which follows the one for its corresponding operation. In this way, implementation independence has been obtained, as long as each of the new added tools implement the operation's interface. The following tools have been integrated in

the current prototype: DB2Triples² for the DB-RDFization operation; The Unstructured Information Management Architecture (UIMA)³ for the entity extraction operation; LodLive⁴ for the visualization operation; OpenRDF Sesame⁵ for storage operations; Continuous SPARQL (C-SPARQL)⁶ for the streaming operation; The Silk framework⁷ for the link discovery operation; and LDSpider⁸ for the crawling operation.

The demonstration will show the current implementation focusing on overall the capabilities of the prototype and exemplify the registration and use of existing tools (e.g. DB2Triples) in the LD-AppStore.

Related Approaches. The LD-AppStore follows the research line of bundling well-established technologies and tools for publishing and consuming Linked Data in order to ease data integration on the Web. Notable approaches developed in this area include toolchains such as the Linked Data Stack⁹ and the LarkC platform¹⁰. Such approaches do not provide an as-a-service hosted solution where 3rd party tool developers can plug-in their implementations for different data operations and where data publishers can configure and execute workflows of data operations implementations on their data --- which is what LD-AppStore targets. DaPaaS¹¹, COMSODE¹², and LinDA¹³ are a number of recent EU funded research projects addressing the problem of simplifying access, integration, and usage of open data based on Linked Data technologies, primarily focusing on data publication and consumption aspects. The projects are in early stages of development with their approaches not entirely defined yet, however ideas from the LD-AppStore are finding traction in the DaPaaS project.

Acknowledgements. This work was partly funded by the following projects: BigFut (SINTEF internally funded project 102003299), DaPaaS (FP7 610988)¹⁴, SmartOpenData (FP7 603824)¹⁵, and InfraRisk (FP7 603960)¹⁶.

² <https://github.com/antidot/db2triples>

³ <https://uima.apache.org/>

⁴ <http://en.lodlive.it/>

⁵ <http://www.openrdf.org/>

⁶ <http://streamreasoning.org/download/>

⁷ <http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/>

⁸ <https://code.google.com/p/ldspider/>

⁹ <http://stack.linkeddata.org/>

¹⁰ <http://www.larkc.eu/>

¹¹ <http://dapaas.eu/>

¹² <http://www.comsode.eu/>

¹³ <http://linda-project.eu/>

¹⁴ <http://project.dapaas.eu/>

¹⁵ <http://www.smartopendata.eu/>

¹⁶ <https://www.infrarisk-fp7.eu/>

Implementing Tourism Service Based on Linked Data with Social Experiments

Takuya Makiyama¹, Yu Ono², Takahiro Sugiyama², Takeshi Morita³, Hiroaki Kogusuri⁴, Hideyuki Tejima⁴ and Takahira Yamaguchi¹

¹Keio University 3-14-1 Hiyoshi, Kohoku-ku, Yokohama-shi, 223-8522 Japan

²Shizuoka University 3-5-1 Johoku, Naka-ku, Hamamatsu-shi, 432-8011 Japan

³Aoyama Gakuin University 5-10-1 Fuchinobe, Chuo-ku, Sagamihara-shi, 252-5258 Japan

⁴Central Nippon Expressway Company Limited 2-18-19, Nishiki, Naka-ku, Nagoya-shi, 460-0003 Japan

¹yamaguti@ae.keio.ac.jp

Abstract. The purpose of this paper is to implement a web service with Linked Data and evaluating the service. These days, Japanese government sets Open Data as a new strategy and focuses on "Linked Open Data (LOD)". However, experiments to show the effect by consuming Linked Data have not been conducted yet. We implemented a tourism service with Linked Data. Moreover, we conducted social experiments on verification of our tourism service. As a result, the possibilities of Linked Data to respond to various queries easily and to apply for information services were explored.

Keywords: Linked Data, Tourism, Mobile application

1 Introduction

Recently, Japanese government sets Open Data as a new strategy and focuses on "Linked Open Data (LOD)". However, experiments to show the effect by consuming Linked Data have not been conducted yet. The purpose of this study is to develop a web service with Linked Data and to evaluate the service. Therefore we implemented a tourism service which makes users to drop in a tourist spot as a case study. Finally we evaluated the service through verification experiments.

A representative service using Linked Data is "DBpedia Mobile" [1], which is an application to display a map containing information about nearby locations based on the current GPS position. Users can explore background information about locations and can navigate into DBpedia and other interlinked datasets. Conventional applications including DBpedia Mobile, however, are not evaluated their usefulness by appropriate experiments. Okawara [2] implemented a mobility service based on Japa-

nese Linked Data and conducted a verification experiment. He gives details on creating and consuming Japanese Linked Data. However, his experiment does not really evaluate the application because the users are very few in his experiment.

2 Implemented System

Our system can be divided into two sections, creating Japanese Linked Data and Tourism service. Fig 1 shows an overview of system.

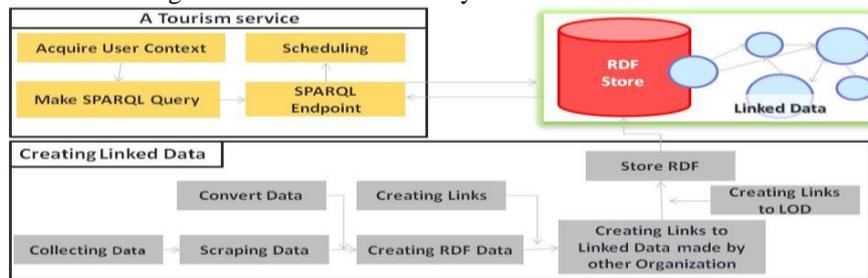


Fig. 1 System Overview

2.1 Creating Linked Data

In this step, we employ the method to create Japanese Linked Data from a conventional system "A Mobility Service based on Japanese Linked Data" [2] and also consume some Linked Data stated in the paper such as Expressway Linked Data and Traffic Regulations Linked Data. We created tourism information Linked Data, coupon information Linked Data and Linked Data of photos on Flickr newly. The links between datasets and an example of their model are shown in Fig2.

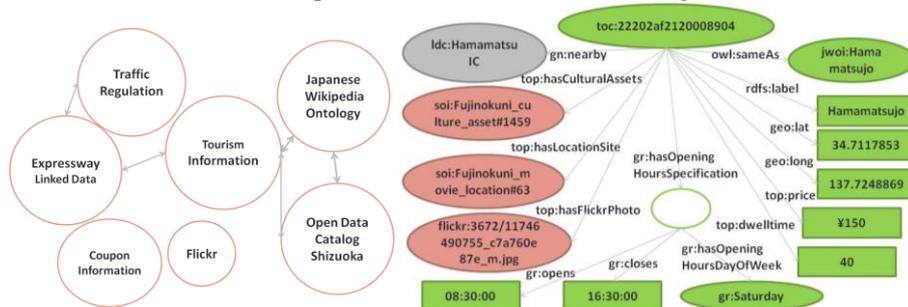


Fig. 2 Links between Datasets and an Example of its Model

Tourism information Linked Data is converted from CSV files created by students of Shizuoka University who are familiar with the area our service provided. It contains information of tourist spots and what tourist can enjoy there. Coupon information Linked Data is a dataset of coupons used at tourist spots. Linked Data of Flickr is dataset of photos taken at tourist spots. Additionally, we link our Linked Data with Open Data Catalog Shizuoka and Japanese Wikipedia Ontology [3]. Open Data Cata-

log Shizuoka is an open dataset composed of various information owned by Shizuoka prefecture. The total number of triples used in our service is 195665. We stored them in a database and set up a SPARQL endpoint.

2.2 A Tourism Service

We implemented our application targeting drivers on an expressway and based on the presupposition that they use our application while taking rest on SA/PA or driving. Therefore, our application is a web application for smart phones. For the sake of helping user to find appropriate tourist spots, our application provides 11 functions. These functions are divided into 2 merger functions, recommendation and search.

Our application recommends tourist spots to users by analyzing contexts of users. The application asks users some question to know the type of group, their interests, available time and money. The application makes a query based on the user context and sends query to get locations matching for the user. Fig3 presents an example of an user context and the SPARQL query based on it.

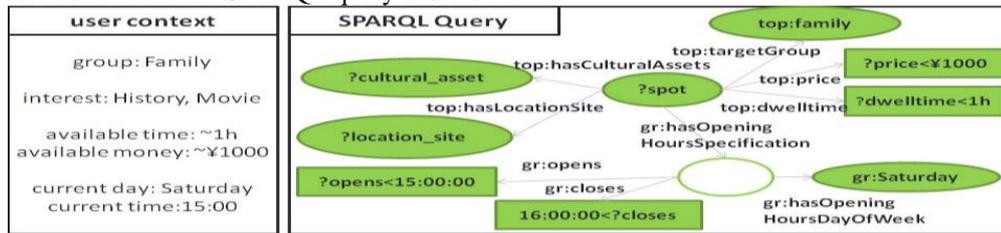


Fig. 3 An Example of User Context and SPARQL Query

Users can also search tourist spots by themselves. We prepare 10 search methods, for example, users can search location of a movie in “Location search”. In this method, application gets a title and description about a movie. Users can refer the casts and detail of the movie from Japanese Wikipedia Ontology.

You can access our application in the following URL. <http://tatiyori.jp>

3 Experiments and Results

If users decide to drop in a tourist spot, it costs much money and time. Therefore we would like to emphasize that it is difficult to make user drop in a tourist spot.

Table 1. Contents of Verification Experiment

	Preparing Experiment	Social Experiment
Period	January 25, 2014 (1day)	June 6, 2014~ July 7, 2014(1month)
SA/PA	Hamanako SA, Hamamatsu SA	Kamigo SA, Makinohara SA, Hamanako SA, Hamamatsu SA

We conducted a social experiment on verification of our tourism service. Before the experiment, we also conducted an experiment for preparing. Table1 shows the contents of these experiments.

3.1 Preparing Experiment

In this experiment, 10 students asked tourists to use our application actively and interviewed to fill questionnaires in the SAs. The number of tourists was 63. 4 tourists dropped in tourist spots and 12 tourists undecided to drop in or not. According to the questionnaires, tourists were prior recommendation to searching. Actually 74% tourists answered that they wanted to visit spots recommended.

3.2 Social Experiment

In this experiment, 10 students passed out leaflets about our experiments in the SAs on Saturday and Sunday. They didn't ask tourists to use our application actively in this experiment differently from the preparing experiment. Not only passing out leaflets, we spread our experiment by news papers, posters and web pages. The number of unique mobile accesses was 554 and 35 tourists showed possibility of dropping in. According to the activity logs, 17 tourists found the spots they dropped in because these spots were affordable compared with their available money. The target tourists have strong constraints because they have to arrive at their goal on time. Our service is helpful for them in looking for appropriate spots rapidly because it can respond their request more flexible than conventional tourism web services thanks to Linked Data and SPARQL.

4 Conclusion and Future Works

In this paper, we implemented a web application consuming Linked Data and conducted social experiments on verification of our tourism service. In these experiments, the possibilities of Linked Data to respond to various queries easily and to apply for information services were explored. For the future work, we improve our recommendation system by applying other semantic web technology such as OWL.

References

1. Christian Becker and Christian Bizer: "DBpedia Mobile: A Location-Enabled Linked Data Browser", *Web Semantics: Science, Services and Agents on the World Wide Web*, vol.7, Issue4, pp.278-286, 2008
2. Wataru Okawara, Takeshi Morita, Takahira Yamaguchi, "Implementing Mobility Service with Less Cognitive Load Based on Japanese Linked Data", 3rd Joint International Semantic Technology (JIST) conference, Springer, LNCS8388, pp. 51-66 (2013)
3. Susumu Tamagawa, Shinya Sakurai, Takuya Tejima, Takeshi Morita, Noriaki Izumi, and Takahira Yamaguchi: "Learning a Large Scale of Ontology from Japanese Wikipedia", 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, wi-iat, pp.279-286, 2010

Estimation of Character Diagram from Open Movie Database using Markov Logic Network

Yuta Ohwatari, Takahiro Kawamura,
Yuichi Sei, Yasuyuki Tahara, and Akihiko Ohsuga

Graduate School of Information Systems,
University of Electro-Communications, Tokyo, Japan
{y-ohwatari@ohsuga.is.uec.ac.jp, kawamura@ohsuga.is.uec.ac.jp,
sei@is.uec.ac.jp, tahara@is.uec.ac.jp, ohsuga@uec.ac.jp}

Abstract. In this paper, we propose the estimation method of interpersonal relationships of characters from movie script databases on the Web using Markov Logic Network. By using Markov Logic Network, we can infer while allowing the violation of rules. In experiments, we confirmed that our proposed method can estimate favors between the characters in a movie with a precision of 69.8%.

Keywords: Markov Logic Network, Semantic Analysis, Open Movie Database

1 Introduction

Every year, a large number of movies have been released. If a user want to quickly know about a movie, he/she will see the summary of the movie. Therefore, It is effective summarization of a movie is required in order for better understanding of the movie.

An overview of our proposed method is illustrated in Figure 1. Our method is separated into the estimation of interpersonal relationships and the generation of character diagrams.

First, we prepared script data for *learning* and *inferring* by extracting who speak what to whom from a movie database. Then, we estimate the sentiment

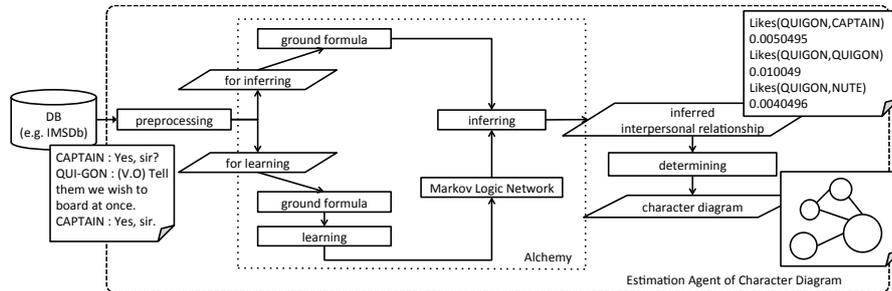


Fig. 1. Method workflow

polarity for lines in the script and the favorable impression between a speaker and a listener in a movie using Markov Logic Network. Finally, we generate the character diagram of a movie from the estimated interpersonal relationships.

A first-order knowledge base can be seen as a set of hard constraints on the set of possible worlds. However, the solution in the real world is often on the set of impossible worlds. In contrast, Markov Logic Network solves this problem by associating weight that reflects how strong a constraint is with each formulas. Also, it is laborious to construct Markov Networks. Markov Logic Network can be viewed as a template for constructing Markov Networks. Markov Logic Network (MLN) is a probabilistic extension of a finite first-order logic[4], which makes up the disadvantages of Markov Networks and a first-order logic.

Note that we used the learning and inference algorithms provided in the open-source Alchemy¹ as an implementation of the MLN in this paper.

2 Related Work

Tanaka et al[7] presents interpersonal relationships extracted from sentence structures as a summary of a story. We considered that it is effective to present the relationships of characters as a summary. On the other hand, analysis of e-mails[1] and estimation from co-occurrence of the name[2] are studies of estimating the relationships of persons in the *real world*. However, it has not been studied about the estimation of interpersonal relationships of *fictional* characters.

There are many studies using MLN, for example, entity resolution[5], information extraction[3]. These studies focus on global constraints, and built a model by using MLN. We also targets a text and extracts information on global constraints.

3 Defined Rules

We defined rules to estimate interpersonal relationships for MLN. These rules determine the sentiment polarity for lines in the script using sentiment polarity for the word and favor between characters using the sentiment polarity for lines. To use sentiment polarity of words, we incorporated as the Semantic Orientations of Words Dictionary that is built by Takamura et al[6]. This assigns a real value in the range from -1 to +1 to where the words assigned with values close to -1 are supposed to be negative, and the words assigned with values close to +1 are supposed to be positive. Vocabulary was extracted from WordNet².

In this paper, we limited to two-valued attribute of positive(+1) and negative(-1). A observed predicate is a predicate with all arguments given by inferring and training. A hidden predicate is a predicate with an argument not given by inferring but given by training. Observed predicates and hidden predicates in this paper are shown in Table 1.

¹ <http://alchemy.cs.washington.edu/>

² <http://wordnet.princeton.edu>

Table 1. Observed and hidden predicates

	Predicate	Description
Observed predicates	Line(text, speaker, listener)	<i>speaker</i> speak <i>text</i> to <i>listener</i>
	Word(text, position, word)	<i>word</i> in <i>text</i> and the position is <i>position</i>
	Wpol(word, pol)	The sentiment polarity of <i>word</i> is <i>pol</i>
Hidden predicates	Lpol(text, pol)	The sentiment polarity of <i>text</i> is <i>pol</i>
	Likes(person, person)	Favor

We describe some of the logical rules for each script line below. t and l is variable. A constant is enclosed in double quotes. Underscore means an arbitrary value. If (+) gets attached to the front of the variable, it is replaced by all the constants that is deployed from the actual data (grounding).

$$\begin{aligned}
 \text{Word}(t, l, +w) \wedge \text{Wpol}(+w, +p) &\Rightarrow \text{Lpol}(t, +p) \\
 \text{Line}(t, +sp, +li) \wedge \text{Lpol}(t, "P") &\Rightarrow \text{Likes}(+sp, +li) \\
 \text{Line}(t, +sp, +li) \wedge \text{Lpol}(t, "N") &\Rightarrow \neg \text{Likes}(+sp, +li) \\
 &\vdots
 \end{aligned}$$

4 Experiment on Relation Extraction

Datasets

In the experiment, we used movie script data from IMSDb: The Internet Movie Script Database³ on the Web. The title of movies used in the experiment are *Back to the Future (1985)*, *Good Will Hunting (1997)*, *Harry Potter And The Sorcerer’s Stone (2001)*, *The Lord of the Rings The Fellowship of the Ring (2001)*, and *Star Wars Episode I The Phantom Menace (1999)*. The average number of lines and characters are 704.6 and 42.6, respectively.

Setting

We used a movie for testing, and the remaining 4 movies as training data. We treated as true above the mean value of the probability, because estimation results are expressed in a probability. Note, we ask the person for a description of *Likes* predicates in the training data that is familiar with the movies and has seen actually.

Result

The experimental results are shown in Table 2. The training time was about 19 hours in total, and the inferring time was about 3 hours in total. As a result, recall is lower than precision. In addition, Figure 2 shows an example of the generated character diagram from the estimated interpersonal relationships. In this figure, a node represents a person, an edge represents a relationship. The

³ <http://www.imsdb.com/>

information with edge shows the estimated probability of the predicate *Likes()* and the mean of the probability (like or not like). A dashed edge means false estimation. This figure generally represents the interpersonal relationships of *Star Wars Episode I The Phantom Menace (1999)*.

Table 2. Estimated relationships and the number of grounded rules

	Precision	Recall	F-measure	ground clauses
mean	69.752	45.332	53.392	46,203

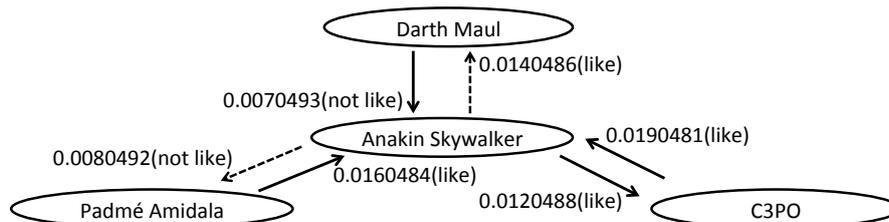


Fig. 2. Part of character diagram generated from the estimated relationship

5 Conclusion

In this paper, using MLN on the movie script database, we estimated the sentiment polarity of script lines and the interpersonal relationships of the characters in a movie. In the experiments, we confirmed that our proposed method estimated favors between the characters in a movie with a precision of 69.8%. In the future, we will improve the model to achieve the higher accuracy.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 24300005, 26330081, 26870201.

References

- Adamic, L.A., Adar, E.: Friends and neighbors on the web. *Social networks* 25(3), 211–230 (2003)
- Matsuo, Y., Tomobe, H., Hashida, K., Nakajima, H., Ishizuka, M.: Social network extraction from the web information. *Transactions of the Japanese Society for Artificial Intelligence* 20(1), 46–56 (2005)
- Poon, H., Domingos, P.: Joint inference in information extraction. In: *AAAI*. vol. 7, pp. 913–918 (2007)
- Richardson, M., Domingos, P.: Markov logic networks. *Machine learning* 62(1-2), 107–136 (2006)
- Singla, P., Domingos, P.: Entity resolution with markov logic. In: *Data Mining, 2006. ICDM'06. Sixth International Conference on*. pp. 572–582. IEEE (2006)
- Takamura, H., Inui, T., Okumura, M.: Extracting semantic orientations of words using spin model. In: *Proc. the 43rd Annual Meeting on Association for Computational Linguistics*. pp. 133–140. Association for Computational Linguistics (2005)
- Tanaka, S., Okabe, M., Onai, R.: Interactive narrative summarization. *Workshop on Interactive Systems and Software* pp. 06–01–06–6 (2011)

News Recommendation based on Semantic Relations between Events

Ryohei Yoko, Takahiro Kawamura, Yuichi Sei, and Yasuyuki Tahara,
and Akihiko Ohsuga

Graduate School of Information Systems University of Electro-Communications,
1-5-1 Chofugaoka, Chofu-shi, Tokyo, 182-8585 Japan
{r-yokoh, kawamura, sei, tahara, ohsuga}@ohsuga.is.uec.ac.jp

Abstract. In recent years, “News Curation Services” that recommend news articles on the internet to user have been popular. In this study, we propose a new “News Curation Service” that collects and recommends novel articles by using semantic relationships between events in the news articles that a user feels interest. The semantic relationships between events are represented by Linked Data. In order to recommend the news articles to the user, we create search queries by using the sentence structure features. Finally, we collect the news articles on the internet and recommend the articles to the user.

Keywords: Linked Data, News Recommendation System, Information Retrieval

1 Introduction

Recently, web services such as “paper.li”¹ and “The Tweeted Times”² that automatically gather news articles, and recommend to users has been popular. The users can easily get information with topicality and novelty by the web services. Such web services are called “News Curation Services”. In this study, we propose a “News Curation Service” that recommends news articles with novelty according to interest of a user. We focus on the semantic relationships between terms of news articles that the user shows interest. The semantic relationships between the terms are represented as Linked Data.

2 Related Work

Khrouf et al. [1] recommended event information to convert meta-information (place, time, tag, genre, etc.) represented by Linked Data. They have built a

¹ <http://paper.li>

² <http://tweetedtimes.com>

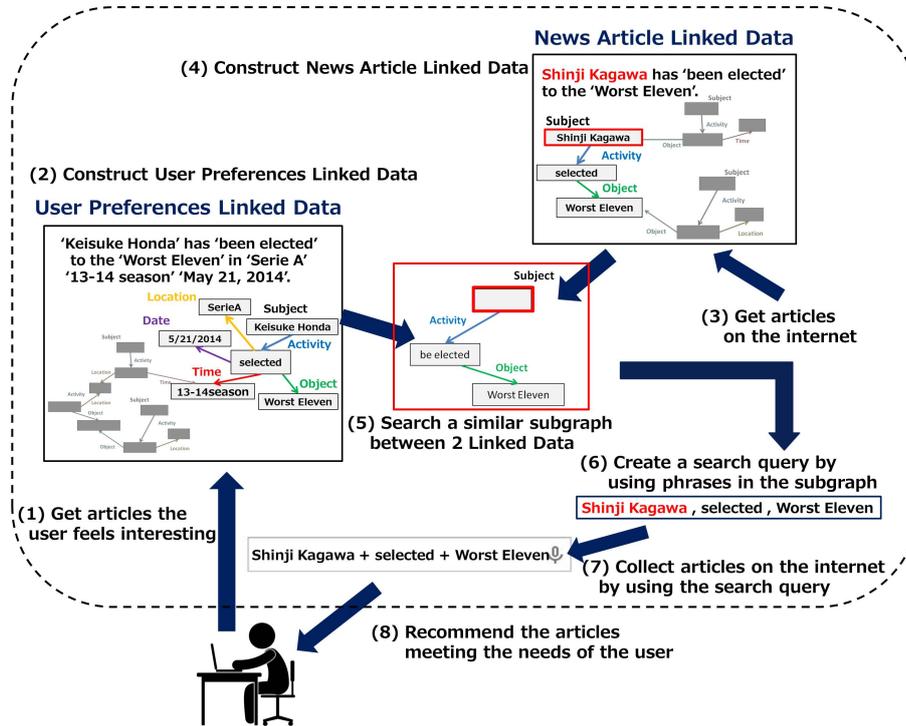


Fig. 1. Summary of Our Approach

hybrid recommendation system by a method using Linked Data and collaborative filtering. In other research, Ostuni et al.[2] has applied user’s implicit feedbacks in the service, and recommended top n items by using Linked Data. In recent years, many recommendation systems applying Linked Data have been proposed. However, there is no recommendation system that converts phrases and the semantic relationships within sentences to Linked Data. In addition, use of Bag-of-Words model is a typical method in many recommendation systems, but the model cannot represent the semantic relationships. Our approach can incorporate the interest of the user by the semantic relationships in comparison with the existing methods of using Bag-of-Words vectors.

3 Approach for Creating Linked Data

In this study, we recommend news articles which are collected on the internet according to users’ preference. Fig.1 indicates a summary of our method. (1)(2)-first, we collect news articles that a user feels interest and construct “User Preference Linked Data” by each sentence of the article. (3)(4)- next, we gather articles that are candidates of the search queries on the internet to construct

“News Articles Linked Data”. Both Linked Data are composed of phrases and semantic relations, that are extracted from sentences of the articles. (5)(6)- finally, if there is a similar subgraph between the two Linked Data, a search query is created as the set of the words of the subgraph.

In order to construct the Linked Data, we extract sentence structures from the articles. We define the sentence structures consists of Subject, Activity, Object, Date, Time, and Location. Each phrase of a sentence is annotated by one of these labels. We represent the sentence structures by using labeled phrases. As an example of Fig. 1, the sentence structure is represented such as “Kagawa→(Activity)→selected” and “selected→(Object)→Worst Eleven” in the Linked Data. We use Conditional Random Fields(CRF) in a similar way as the method of Nguyen et al.[3]. Nguyen et al. annotated the labels in web pages and tweets in Twitter. We recreated a training data set and constructed a learning model for CRF from 100 news articles about soccer written in Japanese in “Sponichi”³. “Sponichi” is a famous internet news media in Japan. The average F-measure of labeling with CRF by 10-folds cross-validations is Subject:63.19%, Activity:66.43%, Object:50.06%, Date:91.00%, Time:46.86%, and Location:48.06%.

4 Evaluation

In order to create “News Article Linked Data”, we used 14,904 articles from 6/30/2013 to 7/1/2014 in “Sponichi”. Likewise, in order to create “User Preferences Linked Data”, we used 10 articles that a test user feels interest from 152 articles from 8/1/2014 to 8/5/2014 in “Sponichi”.

We asked to the test user whether the recommended article is fun(feels fun by the information of the article), novelty(feels new) and serendipity(feels discovery) to each testers. There are eight test users who are our university students. The test users answered in 4 levels: “I thinks so”, “I think so a little”, “I don’t think so a little”, and “I don’t think so”. We recommended top five articles that are searched in Google by the search query created in the previous section. We choose five search queries at random for each test user, so that we recommend 25 articles to a test user. We also defined that a score of the first article is 5 point, the second is 4 point ..., and the fifth is 1 point. Thus, the perfect score is 15 for each query. In the case that the test users answers “I thinks so”, and “I think so little” to the recommended article, the article was regarded as OK, and we added the corresponding point.

We conducted comparison to the method using tf-idf method. We extracted frequent words from 10 articles that the test user feels interest as our proposed method. The search query is created by three of top 10 words calculated by tf-idf method. Fig. 2 indicates the average scores by our method and the tf-idf method. Our method was better than the user tf-idf method. Thus, our approach is effective as a method of news recommendation. In particular, we confirmed that it recommends news articles that are novel to the user.

³ <http://sponichi.co.jp>

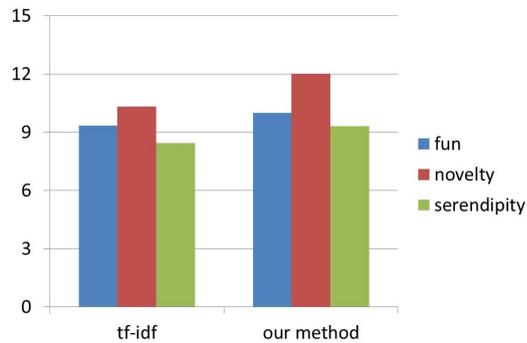


Fig. 2. Comparative Evaluation Results

5 Conclusion and Future Work

In this paper, we proposed a “News Curation Service” by using semantic relationships within sentences. Our approach incorporates interest of a user in comparison with the existing methods. In the future, we will improve accuracy of the labeling.

References

1. H. Khrouf, R. Troncy, Hybrid event recommendation using linked data and user diversity, Proceedings of the 7th ACM conference on Recommender systems, pp.185-192, 2013.
2. Ostuni, Vito Claudio and Di Noia, Tommaso and Di Sciascio, Eugenio and Mirizzi, Roberto: Top-N recommendations from implicit feedback leveraging linked open data, Proceedings of the 7th ACM conference on Recommender systems, pp.85-92, 2013.
3. T.M. Nguyen, T. Kawamura, Y. Tahara, and A. Ohsuga: Self-supervised capturing of users’ activities from weblogs, International Journal of Intelligent Information and Database Systems, Vol.6, No.1, pp.61-76, 2012.
4. Lafferty, John D. and McCallum, Andrew and Pereira, Fernando C. N. : Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data, ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning, pp.282-289, 2001.

A Multi-Aspect Comparison and Evaluation on Thai Word Segmentation Programs

Chaluemwut Noyunsan¹, Choochart Haruechaiyasak² Seksan Poltree³, and Kanda Runapongsa Saikeaw^{1**}

¹ Department of Computer Engineering, Faculty of Engineering, Khon Kaen University 123 Mittrapap, Mueang, Khon Kaen 40002, Thailand
`chaluemwut@kkumail.com`, `krunapon@kku.ac.th`

² Human Language Technology Laboratory (HLT) National Electronics and Computer Technology Center (NECTEC), Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
`choochart.haruechaiyasak@nectec.or.th`

³ Morange Solution Company Limited, 456/1 Klang Mueang, Mueang, Khon Kaen 40000, Thailand
`seksan@orange.co.th`

Abstract. Word segmentation is an important task in natural language processing, especially for languages without word boundaries, such as Thai language. Many Thai word segmentation programs have been developed. Researchers and developers in Thai documents usually spend a tremendous amount of time in studying and trying different Thai word segmentation programs. This paper presents the performance of six Thai word segmentation programs which include Libthai, Swath, Wordcut, CRF++, Thaisemantics, and Tlexs. Based on experimental results, we compare these programs in terms of usage, response time, time outs, and relevance.

Keywords: Word segmentation, term tokenization, software tools, natural language processing

1 Introduction

Natural Language Processing (NLP) enables computers to understand human languages. It consists of many processes such as word segmentation, part-of-speech tagging, automatic summarization, and speech synthesis. Most NLP applications require input text to be segmented into words before being processed

** Corresponding author

further. For example, in sentences similarity application, text must first be tokenized into a series of terms before being analyzed grammatically and semantically. Word segmentation is an essential part for Asian languages such as Thai, Chinese, Japanese, and Korean. This is because these languages are written without delimiter spaces for words in the same sentence.

2 Thai Word Segmentation Programs

Many researches and several programs have been developed for word segmentation. We chose six Thai word segmentation programs which were Libthai, Swath, Wordcut, CRF++, Thaisemantics, and Tlexs. They were selected because they were actively maintained and widely used. Libthai [1] is a set of C programming function to support Thai word segmentation. Swath [7] and Wordcut [6] are command-line programs. CRF++ is a tool for supporting Condition Random Field (CRF) [2]. Thaisemantics has been developed by using Restful web service [5]. Tlexs uses CRF to train models for segmentation [4].

3 Experimental Analysis

Fig. 1 shows the system overview. Our system used the Benchmark for Enhancing the Standard of Thai language processing (BEST) corpus which has been developed by NECTEC [3] and widely accepted for Thai language processing. Our system sent an origin message to the selected Thai word segmentation programs, and then received the results from each program. Then the results from each system were compared with manually tagged words in BEST.

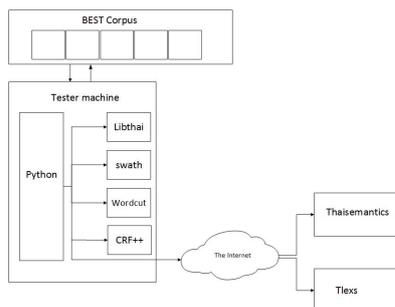


Fig. 1. System overview

Usage Table 1 summarizes the features about usage, offline support, and whether installation is needed.

Table 1. Comparison of word segmentation programs in terms of usage

Features	Libthai	Swath	Wordcut	Thaisemantics	Tlexs	CRF++
Usage	C function	Wrapper	Wrapper	REST API	SOAP API	Wrapper
Offline support	Yes	Yes	Yes	No	No	Yes
Installation needed	Yes	Yes	Yes	No	No	Yes

Response Time We ran the experiments and measured response times on the computer with Intel Dual 1.73GHz and 3 GB of RAM using Ubuntu 64bit as an operating system. The program execution times are shown in Table 2. Libthai performed the best with the response time 0.022 seconds while Swath and CRF++ had response times as 0.024 seconds. On the other hand, the responses times of Thaisemantics, Wordcut, and Tlexs were large. This is because Thaisemantics and Tlexs are programs that are used over the internet thus their response times depend on the internet bandwidth and traffic time. Wordcut is implemented using JavaScript language which may increase the response time of the program.

Table 2. Comparison of word segmentation program response times (in seconds)

Libthai	Swath	Wordcut	Thaisemantics	Tlexs	CRF++
0.022	0.024	0.203	1.520	0.144	0.024

Number of Time Outs Tlexs and Thaisemantics had several time outs because they are called over the internet. Tlexs had average 2,070 time outs while Thaisemantics had 5 time outs. Tlexs caused a large number of occurrences of time outs might be due to server settings or errors.

Relevance Messages from the BEST were sent to each program and the word segmentation output was kept in a list. After that, we compared this log list with a correct list from BEST by using the percentage of precision, recall and F-measure. We ran this test by using 5-fold cross-validation, and then computed

the average value as shown in Table 3. Both Tlexs and CRF++ have the best F-measure because they use CRF.

Table 3. F-measure of Thai word segmentation programs

Measurement	Libthai	Swath	Wordcut	CRF++	Thaisemantics	Tlexs
Precision	61.23	65.09	57.27	59.91	66.03	74.80
Recall	54.97	55.96	62.05	67.80	60.58	75.88
F-measure	57.61	59.60	59.30	63.14	63.03	75.26

4 Conclusions

This paper presents the comparison of six Thai word segment programs in terms of usage, response time, time outs, and relevance. Swath, Libthai, and CRF++ programs provide the smallest response times because they are native programs. Thaisemantics yields the largest response time because Thaisemantics is called over the internet and uses a dictionary. Although Tlexs is also called over the internet, it has better response time because it uses CRF. Both Tlexs and CRF++ give the highest F-measure because they employ CRF.

References

1. T. Karoonboonyanan, C. Silpa-Anan, P. Kiatisevi, P.Veerathanabutr and V. Ampornaramveth, *Libthai Library*, retrieved on Jul 1, 2014 from <http://linux.thai.net/projects/libthai>.
2. T. Kudo, *CRF++: Yet Another CFT toolkit* retrieved on July 2, 2014 from <http://crfpp.googlecode.com/svn/trunk/doc/index.html?source=navbar>.
3. National Electronics and Computer Technology Center (NECTEC), *BEST: Benchmark for Enhancing the Standard for Thai Language Processing* retrieved on Jul 5, 2014 from <http://thailang.nectec.or.th/best/>.
4. National Electronics and Computer Technology Center (NECTEC), *Tlexs: Thai Lexeme Analyser*, retrieved on Jul 3, 2014 from <http://sansarn.com/tlex/>.
5. S. Poltree, *Thaisemantics: Free Thai Language Resources and Services*, retrieved on Jul 2, 2014 from <http://www.thaisemantics.org/>.
6. V. Satayamas, *wordcut program* retrieved on Jul 3, 2014 from <https://github.com/veer66/wordcut>.
7. P. Charoenpornasawat, *SWATH - Thai Word Segmentation*, retrieved on Jul 1, 2014 from <http://www.cs.cmu.edu/~paisarn/software.html>.

An Online Learning-based Efficient Search System for Sufficient SPARQL Endpoints using Extended Multi-armed Bandit Algorithm

Yoshiaki Kadono and Naoki Fukuta

Graduate School of Informatics, Shizuoka University
{gs14014@s, fukuta@cs}.inf.shizuoka.ac.jp

Abstract. On searching and integrating various kinds of open data, one of crucial issues is to efficiently find sufficient SPARQL endpoints that store sufficient data to be retrieved in the query. In this demonstration, we will present our on-line learning-based efficient search algorithm and our prototype implementation based on an extended multi-armed bandit approach, which takes into account how the demanded data could be stored in each SPARQL endpoint by the results obtained from a series of test queries.

1 Introduction

At the beginning of 2014, we have over 500 of information sources known as endpoints of Linked Open Data(LOD)*. To utilize such endpoints effectively, we often make some federated queries to some of useful endpoints available in the world. It is not an easy task to accurately predict which endpoints could be helpful to be used for the queries, since there are so many endpoints and also those endpoints are dynamically updating their data but little information are given about how their stored data are useful for the queries[1].

To examine how an endpoint is useful and appropriate for a given query, we need to pre-check how each endpoint has useful data as well as how such data can be effective to retrieve further data stored in other endpoints. However, when we try to examine all the endpoints to be accessed at one time, it will cause serious network congestions due to its heavy network loads, as well as unwanted server loads in those endpoints. Therefore, we need to effectively predict the performance of possible endpoints to be used in a query. In this demonstration, we present a conceptual model of our online learning-based efficient search algorithm and our prototype implementation based on an extended multi-armed bandit approach, which takes into account how the demanded data could be stored in each SPARQL endpoint by the results obtained from a series of test queries.

2 Extending Budget-Limited Multi-Armed Bandits Model

The Multi-Armed Bandits(MAB) Problem classified in decision theory[2] assumes that there is a slot machine with K arms and each arm has its own revenue distribution. In

*<http://sparqls.okfn.org/>

the situation, the problem seeks how an algorithm can choose the best arm to obtain maximal revenues as well as the cost for estimating the revenue distributions of arms.

The Budget-Limited Multi-Armed Bandits(BLMAB) model[3][4] extends the MAB model to handle limited budgets and costs for pulling arms to estimate revenue distributions. Here, each arm i has its own cost c_i to pull the arm, and the payment will be withdrawn from the total budget B .

The Extended BLMAB model for the end-point search problem(ESP) is an extended BLMAB model, which introduces a two-dimension cost constraint, i.e., the cost for network load and the cost for query execution time. In our extended BLMAB model, the player can choose an arbitrary number of arms from the total K arms to pull in each step. Each arm i has its cost c_i to obtain the revenue based on an unknown distribution μ_i . The player has budget B so that the cost to be payed should be in the budget, and also the maximum steps is limited to be $T \in \mathbb{N}$. The objective of the player is maximizing its revenue by choosing arm i with the limited budget B within the limited T steps. Therefore, we extend the original BLMAB constraints to the following:

$$P\left(\sum_i^K N_i^A(B)c_i \leq B \cap t \leq T\right) = 1 \quad (1)$$

Here, as described in [4], we consider $N_i^A(B)$ be the random variable that represents the number of pulls of arm i by A , with respect to the budget limit B . Note that, as we mentioned, the goal is to discover the best arm to be used, rather than total revenues obtained by exploration queries.

3 Algorithm for Extended BLMAB

On the end-point search problem, we can assign each endpoint as each arm in the extended BLMAB model. Here, often each endpoint has a limited number of variations in its effective exploration queries. Therefore, the extended BLMAB model should satisfy that, let Q be the number of variations in its exploration queries, an arm i should satisfy $N_i^A(B) \leq Q$.

While the main objective of MAB is to maximize the total revenue, the main objective in end-point search problem is to find the best arm to be used for the final (but not exploration) query. Therefore, let $i(A)$ be the best arm obtained from A , $i(A^*)$ be the theoretically best arm, we re-define the regression function for our extended BLMAB model as follows:

$$R(A) = i(A^*) - i(A) \quad (2)$$

4 Implementation and Preliminary Evaluation

Figure 4 shows the results of both ϵ -greedy and an extended version of the KDE algorithm[4] on the prepared test environment.

In the standard MAB problem, a typical value of ϵ is about 0.1. However, from the shown result in Figure 4, around 0.4 seems optimal. This might be caused by the

difference of objectives, since our model is based on an extended BLMA Model. In Figure 4, the vertical axis shows the regression value defined in Equation 2. Notice that, less regression value is better in this context.

Figure 1 shows the overview of our prototype system, called *SPARQL-MAB*. We are implementing the two algorithms in the system*. Further evaluations and refinements about the models and algorithms are our future work.

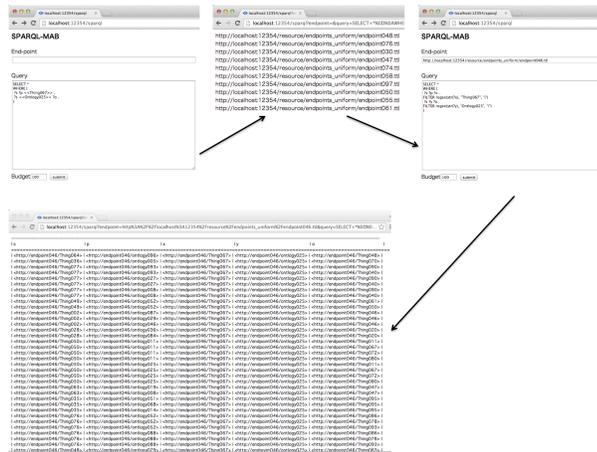


Fig. 1. An overview of our prototype system

For the preparation of the target query to be examined, an abstract query[5] is used which can include some *unknown* nodes, denoted by “<<” and “>>”. Figure 2 shows an example of abstract query.

```
SELECT *
WHERE {
  ?x ?y <<Object>>.
  <<Subject>> <<Predicate>> ?x.
}
```

Fig. 2. An example of user’s query

In Figure 3, two example queries are shown. Those queries were generated from the user’s query shown in Figure 2.

* A demonstration is available at http://whitebear.cs.inf.shizuoka.ac.jp/sparql_mab/

<pre> SELECT ?a WHERE { ?x ?y ?a. FILTER regex(str(?a), "Object", "i") } </pre>	<pre> SELECT ?a ?b WHERE { ?a ?b ?x. FILTER regex(str(?a), "Subject", "i") FILTER regex(str(?b), "Predicate", "i") } </pre>
---	---

Fig.3. Example of queries for endpoint evaluation

In this evaluation, we prepared a hundred of pseudo endpoints each of which stores 10 thousands of triples. Each triple is composed from a randomly selected concepts generated from a hundred kind of nouns or concepts in the given ontologies.

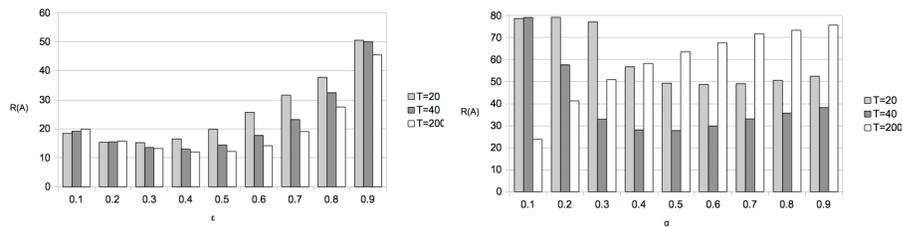


Fig.4. Performance comparison of ϵ -greedy and KDE algorithm

References

1. T. Fujino and N. Fukuta, *Utilizing Weighted Ontology Mappings on Federated SPARQL Querying*, In, Proc. the 3rd Joint International Semantic Technology Conference, pp.331–347, 2013.
2. H. Robbins, *Some aspects of the sequential design of experiments*, Bulletin of the AMS 55:527-535, 1952.
3. L. Tran-Thanh, A. Chapman, J. Munoz De Cote Flores Luna, A. Rogers and N. Jennings *Epsilon-First Policies for Budget-Limited Multi-Armed Bandits*, In, Proc. Twenty-Fourth AAI Conference on Artificial Intelligence(AAAI-10), pp.1211-1216, 2010.
4. L. Tran-Thanh, A. Chapman, A. Rogers and N. Jennings *Knapsack based optimal policies for budget-limited multi-armed bandits*, In, Twenty-Sixth AAI Conference on Artificial Intelligence (AAAI-12), pp.1134-1140, 2012.
5. H. Noguchi, T. Fujino, and N. Fukuta, *On Implementing SPARQLoid and its Query Coding Support Framework – Querying with Weighted Ontology Mappings*, In, Proc. the 3rd Joint International Semantic Technology Conference (JIST2013),2013.(demonstration)

LinkedCJ: A Knowledge Base of Chinese Academic Journals Based on Linked Data

Peng Xu¹, Xin Wang¹, and Haofen Wang²

¹ School of Computer Science and Technology, Tianjin University, Tianjin, China
{wxlyf, wangx}@tju.edu.cn

² Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai, China
whfcarter@ecust.edu.cn

Abstract. Nowadays the amount of academic articles published in Chinese is growing rapidly, however, existing methods of managing and querying Chinese academic journals and articles are not semantic-based. Our work consists of creating an ontology for represents and organizing bibliographic information of Chinese academic journals and articles. Moreover, we develop software applications based on Nutch, jsoup, and Drools for transforming millions of Web pages from website of Wanfang into approximately 15 million triples stored in a triple store. Finally, the knowledge base is evaluated using the Semantic Service Platform of Chinese Academic Journals and Articles (SSPCAJA). Results of the functional test show that information of Chinese academic journals and articles is effectively represented on the platform.

Keywords: linked data, knowledge base, Chinese academic journals, ontology

1 Introduction

The three major Chinese Web publishers, VIP, Wanfang, and CNKI have embodied over 38 million, 20 million, and 36 million in articles respectively, as well as 12000, 7000, and 8000 in journals, until July 2013. However, existing methods on managing and querying Chinese academic journals and articles are not semantic-based.

The Semantic Web, which stores all information in the form of Linked Data^[1] instead of hyperlinked Web pages, focuses on semantic interpretation of the data on the Web. At present, leading publishers, such as NPG, DBLP, and CrossRef, have been organizing semantic data from journals on the basis of linked data. But most of the vocabularies in NPG, DBLP, and CrossRef cannot adapt to Chinese academic journals because the structure of knowledge organization cannot be reused directly. However in China, semantic-based knowledge organizations for Chinese academic journals have not yet been constructed, while the popular alternatives are data centralizing platforms, such as C-DBLP and Not Old academic search.

In this paper, we seek to organize data from several Chinese academic journals on the basis of linked data. The main contributions of this paper are:

- we build the LinkedCJ ontology for representing information of Chinese academic journals;
- we develop a method for extracting RDF triples from the Web pages of Wanfang using a series of tools including Nutch plugins, jsoup, and Drools;
- we conduct a set of functional tests on the SSPCAJA using LinkedCJ knowledge base and compare SSPCAJA with linked data platforms of NPG and DBLP to evaluate the LinkedCJ knowledge base.

The rest of the paper is organized as follows. Section 2 describes the construction of the LinkedCJ ontology. Section 3 gives the method for extracting RDF triples from the crawled Web pages of Wanfang. The functional test and evaluation of the knowledge base are presented in Section 4. Finally, Section 5 concludes the paper and gives the future work.

2 LinkedCJ Ontology

In order to organize and construct the knowledge base of Chinese academic journals, it is necessary to create an ontology according to the characteristics of Chinese academic journals.

LinkedCJ, our new ontology for Chinese academic journals, inherits a list of classes, object properties and data properties from existing ontologies of semantic publishing, such as Functional Requirements for Bibliographic Records (FRBR)^[2], FRBR-aligned Bibliographic Ontology (FaBiO)^[3], Citation Typing Ontology (CiTO)^[4], Dublin Core Metadata (DC)^[5] and The Friend of a Friend (FOAF). LinkedCJ adds several new classes, object properties and data properties in order to indicate exclusively information of Chinese academic journals. Fig.1 shows the top level classes and object properties of LinkedCJ.

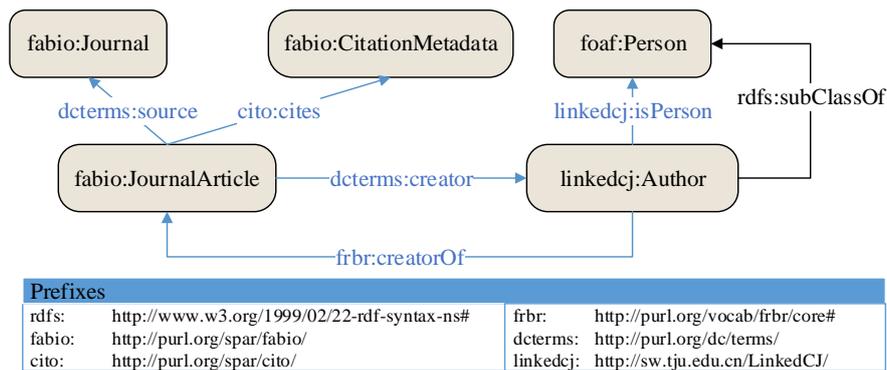


Fig. 1. Top level classes and object properties of LinkedCJ

LinkedCJ ontology contains 5 classes, 5 object properties, and 35 data properties. A list of original classes, object properties, and data properties, such as linkedcj:cn, linkedcj:hasProject, and linkedcj:hasSecondarySubjectTerm, take part in the ontology

for representing items with Chinese characteristics that does not include in existing ontologies, such as FRBR, FaBiO, CiTO, DC, and FOAF.

In summary, the vast majority of the information from Chinese academic journals and articles could be represented by LinkedCJ normatively.

3 Triple Extraction

In order to construct the knowledge base, we develop software applications based on Nutch, jsoup and Drools for transforming HTML pages from website of Wanfang into triples stored in a triple store. Fig. 2 shows the whole process of triple extraction.

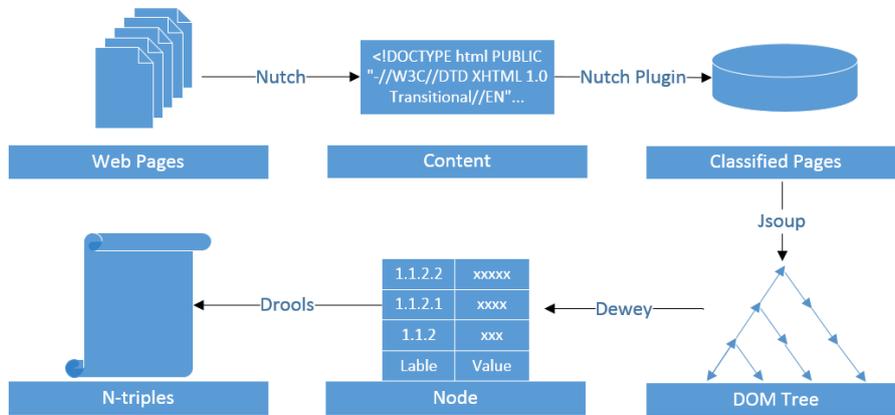


Fig. 2. The process of triple extraction

Wanfang is the data resource of our knowledge base. Triples have been gathered from the 13 journals of China Computer Federation (CCF), such as the Chinese Journal of Computers. After approximately 300000 pages fetched in this way, we have acquired a copy of HTML raw data including approximately 48000 journal articles, 590000 citations and 180000 authors. Finally, an N-triples (.nt) file containing about 15 million triples has been generated.

4 Evaluation and Comparison

In order to evaluation the LinkedCJ knowledge base, we have performed several test cases on the Semantic Service Platform of Chinese Academic Journals and Articles (SSPCAJA). For the purpose of comparison, two leading linked data service platforms, NPG and DBLP, have been tested at the same time.

Table 1 shows the summary of the functional tests. Symbol “o” means the function or information is provided, while symbol “x” means the function or information is not provided. Maximum kinds of supported queries, as well as the personal name disambiguation mechanism, are the main advantages of the LinkedCJ knowledge base.

Table 1. The results of the functional test on query types.

Query type		SSPCAJA	NPG	DBLP
	SPARQL	○	○	×
By title	Show start page	×	○	×
	Show end page	×	○	×
	Show secondary subject term	○	×	×
	Show project	○	×	×
	Show creator's working unit	○	×	×
By author	Search author	○	○	○
	Personal name disambiguation	○	×	×
By journal	Search journal	○	×	○
	Sorted by volume and number	×	×	○
	By keywords	○	×	×

Other test cases have revealed that LinkedCJ is superior to NPG and DBLP in representing information of Chinese academic journals.

5 Conclusions

In this paper, we present the design and implementation of LinkedCJ, the knowledge base of Chinese academic journals. LinkedCJ knowledge base is evaluated using the SSPCAJA. Results of the test cases show that information of Chinese academic journals and articles is correctly represented by the knowledge base.

Acknowledgements. This work is supported by CCF Opening Project of Chinese Information Processing (Grant No. CCF2013-02-02) and the National Natural Science Foundation of China (Grant No. 61100049).

References

1. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3), 1-22 (2009)
2. IFLA Study Group on the Functional Requirements for Bibliographic Records.: *Functional Requirements for Bibliographic Records: Final Report*. UBCIM publications, München (1998)
3. Peroni, S., Shotton D.: FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*. 17, 33-43 (2012)
4. Shotton, D.: CiTO, the Citation Typing Ontology. *Biomedical Semantics*. 1(S-1) S6 (2010)
5. Kurtz, M.: Dublin Core, DSpace, and a brief analysis of three university repositories. *Information Technology and Libraries*. 29(1), 40-46 (2013)

A Review and Design of Framework for Storing and Querying RDF Data using NoSQL Database

Chanuwas Aswamenakul¹, Marut Buranarach²,
and Kanda Runapongsa Saikaew^{1*}

¹ Department of Computer Engineering, Faculty of Engineering,
Khon Kaen University, Khon Kaen, Thailand
chanuwas.a@kkumail.com, krunapon@kku.ac.th

² Language and Semantic Technology Laboratory
National Electronics and Computer Technology Center (NECTEC), Pathumthani, Thailand
marut.bur@nectec.or.th

Abstract. This paper reviews existing systems and describes a design of RDF database system that uses NoSQL database to store the data which aims to enhance performance of the Semantic Web applications. RDF data is a standard of data in the form of Subject-Predicate-Object called Triples and stored in database called Triple Store. Typically RDF database system uses SPARQL query language to query the RDF data from Triple Store database, e.g. Jena TDB. Our design of RDF database system uses NoSQL database, i.e., MongoDB, to store the data in JSON-LD format and query by using query API of NoSQL database. We will use the Berlin SPARQL Benchmark to compare the performance of Triple Store and NoSQL systems.

Keywords: Semantic Web application framework, RDF database, NoSQL

1 Introduction

Currently the amount of data has increased excessively with a variety of formats. The Semantic Web technology aims to provide standards and facilitate analyzing such big data. The Semantic Web uses RDF data to describe the data on the web in form of Subject-Predicate-Object called “triples” [1] that makes the data to have the standard data model.

In the present, there are many approaches to store and query RDF data. One approach to store RDF data is Triple Store designed for storing the triples format of RDF data [2] and queried by using SPARQL query language. However, from the Berlin Benchmark results [3], Triple Stores show poor performance when compared to the relational database systems. NoSQL database removes some features of relational databases and uses other data models to improve the performance of database. This has motivated many works to store RDF data by using NoSQL database.

This paper reviews existing systems and designs a framework to store RDF data in NoSQL database. One of the main goals is to design a Semantic Web application framework that uses RDF data with NoSQL database, i.e., MongoDB. The ultimate

* Corresponding author

objective is to provide a better support for researchers in developing the Semantic Web applications.

2 Review of NoSQL-based RDF Database

This section reviews some of RDF database systems that use NoSQL to store the RDF data including Neo4j [4] , AllegroGraph [5] , H2RDF [6] , Oracle NoSQL [7] , MonetDB [8] and CumulusRDF [9]. The comparison is based on some criteria of database software such as Implementation language, Database Model, SPARQL1.0, SPARQL1.1, Trigger, Transaction Concept, Secondary Index, Consistency Concept, Partitioning Method, Replication Method, Concurrency, Map Reduce, Durability and Security. Table 1 provides a review summary of RDF database systems that use NoSQL database.

Table 1. Review summary of RDF database systems that use NoSQL database

Name	Neo4j	AllegroGraph	H2RDF	Oracle NoSQL	MonetDB	CumulusRDF
Implementation language	Java	Common Lisp	Java	Java	C	Java
Database Model	Graph Database	Graph Database, Document store Database	Column Store Database	Key-Value Database	Column Store Database	Column Store Database
SPARQL 1.0	Yes	Yes	Yes	Yes	Yes	Yes
SPARQL 1.1	Yes	Yes	Yes	Yes	No	Yes
Trigger	Yes	No	Yes	No	Yes	Yes
Transaction Concept	ACID	ACID	Configure ACID + Visibility	ACID	ACID	Configure ACID(Lightweight Transaction)
Secondary Index	Yes	Yes	Yes	No	Yes	Yes
Consistency Concept	Eventual consistency	Strong consistency	Strong consistency	Several consistency policies	Strong consistency	Tunable consistency
Partitioning method	Cache Sharding	Sharding	Sharding	Sharding	None	Sharding
Replication method	Master-slave	Master-slave	Master-slave	Master-slave	None	Selectable replication factor
Concurrency	Yes	Yes	Yes	Yes	Yes	Yes
MapReduce	No	No	Yes	Yes	Yes	Yes
Durability	Yes	Yes	Yes	Yes	Yes	Yes
Security	Security Rule	Filter per User and/or Role	Access Control List (ACL)	User and Role Permission	fixed user and password by admin	Object Permission

3 Framework Design

This section describes our design for an application framework representing system architecture that compares the Triple Store-based implementation with the NoSQL-based implementation. We also provide query translation that represents some example translation of basic SPARQL queries adapted from the Berlin Benchmark [3] to MongoDB queries.

In a system architecture based on the OAM framework [10], we compare between Triple Store based implementation and NoSQL based implementation. The Triple store based implementation uses Jena TDB to store the RDF data and OAM API that uses SPARQL to query the data from Jena TDB. In NoSQL based implementation, we use RDF to JSON-LD Converter to convert RDF data format to JSON-LD format, which is JSON-based format designed for Linked data [11], and use JSON-LD Parser to parse and import JSON-LD data to MongoDB. The OAM API then uses MongoDB query API to query the data from MongoDB.

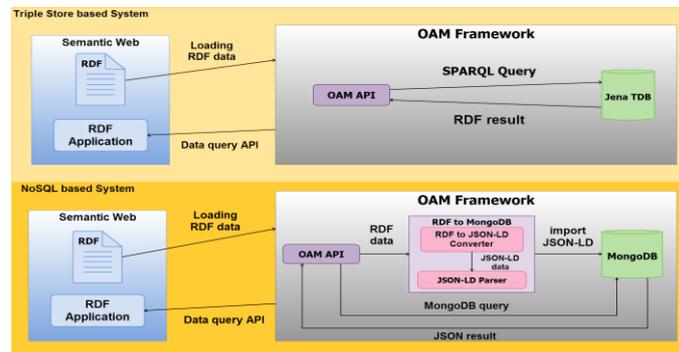


Fig. 1. Architecture of the OAM framework using Triple Store vs. NoSQL RDF database

Table 2 illustrates some query translation based on the Berlin SPARQL benchmark. In Table 2, query 1 shows an example of query using FILTER, ORDER and LIMIT. Query 2 shows an example of query using OPTIONAL. Query 3 shows an example of query using regular expression.

Table 2. Sample query translation based on the Berlin SPARQL Benchmark

Query Description	SPARQL	MongoDB query
1. Find products for given product type and value of property numeric1 must be greater than 318 then results ordered by value of label and limit number of results by 10.	SELECT ?product ?label WHERE { ?product label ?label ?product a ProductType56 ?product PropertyNumeric1 ?value FILTER (?value > 318) } ORDER BY ?label LIMIT 10	db.collection.find({label : {\$exists : true}, types : 'ProductType56', PropertyNumeric : {\$gt : 318}} ,{label : 1}).sort({label : 1}).limit(10)
2. Retrieve the basic information of products and products may not have property numeric2 (OPTIONAL in SPARQL).	SELECT ?label ?comment ?propertyTextual1 ?propertyNumeric2 WHERE { Product127 label ?label Product17 comment ?comment Product1277 PropertyTextual1 ?propertyTextual1 OPTIONAL { Product1277 PropertyNumeric2 ?propertyNumeric2 } }	db.collection.find({_id : 'Product1277', label : {\$exists : true}, comment : {\$exists : true}, PropertyTextual : {\$exists : true}} , {_id : 0, label : 1, comment : 1 , PropertyTextual1 : 1 , PropertyNumeric2 : 1})
3. Find products having a label that contain given string by using regular expression.	Select ?product ?label where { ?product label ?label ?product type Product FILTER regex(?label, "dung") }	db.collection.find({label : {\$regex : 'dung*'} , '@type' : 'Product'} , {label : 1})

4 Conclusions and Future Work

This paper has proposed the design of RDF database system by using MongoDB to store the data in JSON-LD format and its query API. In the future, we will conduct the performance comparison of Triple Store, MongoDB RDF Database, and relational database using the Berlin SPARQL Benchmark. Several techniques will be investigated to improve the performance of the MongoDB RDF Database.

Acknowledgement

The financial support from Young Scientist and Technologist Programme, NSTDA (YSTP: SP-56-NT03) is gratefully acknowledged.

References

1. RDF [Online]. Available: <http://www.w3.org/RDF/>
2. Triple Store [Online]. Available: <http://www.w3.org/wiki/RdfStoreBenchmarking>
3. Bizer, C., Schultz, A.: The berlin sparql benchmark. *International Journal on Semantic Web and Information Systems (IJSWIS)* 5(2), 1–24 (2009).
4. Neo4j [Online]. Available: <http://docs.neo4j.org/chunked/2.0.4/>
5. AllegroGraph [Online]. Available: <http://franz.com/agraph/allegrograph/>
6. Papailiou, N., Konstantinou, I., Tsoumakos, D., Koziris, N.: H2RDF: Adaptive Query Processing on RDF Data in the Cloud. In *WWW*, 2012.
7. Oracle NoSQL database [Online]. Available: http://docs.oracle.com/cd/E26161_02/html/RDFGraph/
8. MonetDB [Online]. Available: <https://www.monetdb.org/Home>
9. Cudré-Mauroux, P., Enchev, I., Fundatureanu, S., Groth, P. T., Haque, A., Harth, A., Keppmann, F. L., Miranker, D. P., Sequeda, J. & Wylot, M. (2013), *NoSQL Databases for RDF: An Empirical Evaluation*. *International Semantic Web Conference (2)*, Springer, pp. 310-325 .
10. Buranarach, M., Thein, Y., Supnithi, T.: A Community-Driven Approach to Development of an Ontology-Based Application Management Framework. In: Takeda, H., Qu, Y., Mizoguchi, R., and Kitamura, Y. (eds.) *Semantic Technology*. pp. 306–312. Springer Berlin Heidelberg (2013).
11. JSON-LD [Online]. Available: <http://json-ld.org/>

RDB2Graph: A Generic Framework for Modeling Relational Databases as Graphs

Kang Min Yoo, Sungchan Park, and Sang-goo Lee

Intelligent Data Systems Laboratory
Seoul National University
{kangminyoo, baksalchan, sglee}@europa.snu.ac.kr

Abstract. Graph data mining is highly versatile, as it applies not only to graph data but to relational data, as long as it can be represented as pairs of relationships. However, modeling RDBs as graphs using existing methods is limited in describing semantics of the relational data. In this paper, we propose a two-phased graph-modeling framework that converts any RDB to a directed graph with richer semantics than previously allowed. We implemented the framework and used it for analyzing medical records of diabetes patients.¹

Keywords: relational database, graph, graph modeling

1 Introduction

Graph data mining is a well-studied area of research because of numerous applications in fields such as social data mining and biochemical analysis [1]. Recently, there has been a growing interest in applying graph mining techniques to relational databases, as viewing them as graph data exposes inherent semantics [3] [4]. Since relational databases are a de facto standard in data warehousing, applying graph mining techniques to mine latent information from the massive relational data seems even more attractive.

However, modeling relational databases as graphs is not straight-forward, because it is challenging to devise appropriate models that expose underlying semantics. W3C formalized the graph-modeling process by defining a new language, R2RML [2], but the language is limited in describing some semantic aspects of graph conversion. For example, it cannot be used to describe a vertex that combines several attributes (fig. 1). It also fails at describing semantics not apparent in relational schemata, such as events that could be connected in chronological order (fig. 2).

We propose a new framework to solve the current problem of modeling RDBs as graphs. The framework converts any RDB into a directed graph given some conversion rules (sec. 2). We have also implemented a program based on RDB2Graph (R2G) to analyze medical records of diabetes patients.

¹ This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea Government(MSIP) (No. 20110030812).

2 RDB2Graph

Given a **relational database** D and a set of modeling rules, the framework generates a directed graph. The output of the framework is **edge set** E and **vertex set** V . Each vertex is a set of key-values, i.e. $\{(k_1, l_1), (k_2, l_2), \dots, (k_n, l_n)\}$, where each key in k_1, \dots, k_n corresponds to some attribute and each l in l_1, \dots, l_n is some value in D . An edge e is directed and denoted by (v_s, v_t) , where v_s is the source vertex and v_t is the target vertex. Both v_s and v_t are in V . Given some rule set Γ and D , framework must return E and V such that elements satisfy the specifications of Γ .

2.1 The Two-Phase Conversion

The conversion takes place in two phases — the first phase discovers relationships within each tuple, while the second phase establishes additional relationships among the vertices constructed from the first phase. Thus, Γ is an union of Γ_1 and Γ_2 , the rules for both phases respectively. Splitting in two phases is necessary because it allows incorporation of implicit relationships not apparent from the relational database itself.

Phase I: Tuples to Edges In order to create new vertices and edges from the relational database, a set of rules Γ_1 must specify the followings:

1. a target relation (r)
2. two sets of attributes (C_s, C_t) from relation r to indicate which values of each tuple in r are stored as l in key-value pairs of source or target vertices
3. two sets of key aliases (K_s, K_t) to indicate k in key-value pairs of source or target vertices.
4. a bit (b) to indicate whether the edge is bidirectional or not
5. a selection predicate (p) to filter tuples.

For each c in C_s or C_t , the corresponding value u in each tuple is paired with the corresponding k in K_s or K_t to form a key-value pair (k, u) , which is added to the generated source or target vertex. An example of edge generation is presented in figure 2.1.

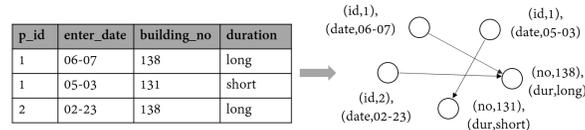


Fig. 1. Phase I edge construction based on the rule $C_s = \{p_id, enter_date\}$, $C_t = \{building_no, duration\}$, $K_s = \{id, date\}$, $K_t = \{no, dur\}$, and $b = 0$.

The purpose of specifying the alias sets K_s and K_t is to combine semantically identical attributes together. For example, consider a case where relation r_1 has a foreign key constraint that references a primary key of another relation r_2 . Their attribute names might be different, but they are identical semantically. By having the ability to give a common key for the attributes, we are able to generate common vertices. It is apparent from figure 2.1 that some vertices have common key configurations (e.g. $\{id, date\}$). We call such key configurations **vertex schemata**.

Selection predicate p is similar to the counterpart of relational algebra. It can be directly used in SQL queries to filter out tuples. For each rule, the framework retrieves a set of tuples from T that satisfy the predicate and produces exactly one edge of (v_s, v_t) , which is added to E . The framework will also attempt to add v_s and v_t into V , if they do not exist already. At the end of phase I, E and V are generated and passed to phase II.

Phase II: Vertices to Edges Given the graph constructed from the previous phase, E and V , and phase II rules Γ_2 , the framework constructs additional edges that satisfy Γ_2 . In this phase, each rule of Γ_2 specifies the followings.

1. two sets of vertex schema (K_s, K_t)
2. a bit (b) to indicate whether the generated edge is bidirectional or not
3. a vertex selection predicate (q) to filter vertices

For each rule in Γ_2 , vertices with schema K_s or K_t are considered as source or target vertices of an additional edge. Given the vertices, the framework further filters them using vertex selection predicate q . An example of such edge generation is shown in figure 2.1.

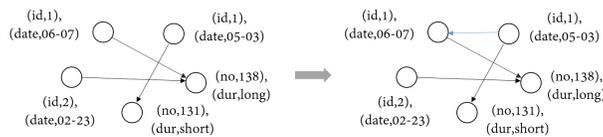


Fig. 2. Phase II edge construction based on the rule $K_s = \{id, date\}$, $K_t = \{id, date\}$, $b = 0$, and q selects source and target vertices such that values of id are equal but the target date is the nearest later date to the source date.

q is different from p of Γ_1 — in p , left-hand side variables are references to some attributes of a relation; in q , left-hand side variables are references to some keys of either K_s or K_t . Phase II produces E and V as well, but with additional edges that satisfy Γ_2 .

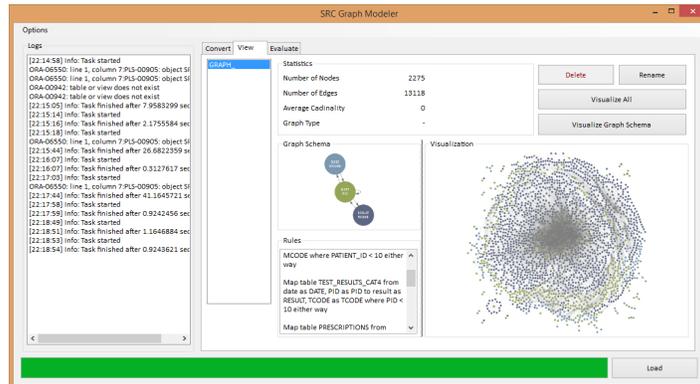


Fig. 3. SRCGraphModeler is an implementation of RDB2Graph. It was used to analyze medical records of diabetes patients.

2.2 Implementation

Our implementation of the framework, SRCGraphModeler (SGM), is written in C#, and it connects to Oracle 11g based RDB. SGM converts each rule into PL/SQL procedures in order to run them on the database server, improving runtime efficiency. Using SGM, we converted medical records of diabetes patients into various graph models, then we applied graph analysis on the graphs to extract correlation among medications and symptoms.

3 Conclusion

We have presented a graph-modeling framework that enables semantically richer conversions than that attempted by previous works. As future works, we plan to study how transformation of RDB data to graph data affects the information contained in it.

References

1. Charu C. Aggarwal and Haixun Wang. *Managing and Mining Graph Data*, volume 40 of *Advances in Database Systems*. Springer, 2010.
2. Souripriya Das, Seema Sundara, and Richard Cyganiak. R2RML: RDB to RDF Mapping Language. <http://www.w3.org/TR/r2rml>, September 2012.
3. Jaehui Park. *A Graph-based Framework for Processing Keyword Queries over Relational Databases*. PhD thesis, Seoul National University, 2012.
4. Subhesh Pradhan, Sharma Chakravarthy, and Aditya Telang. Modeling Relational Data as Graphs for Mining. In *International Conference on Management of Data*, 2009.