World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:2, No:1, 2008

Target Concept Selection by Property Overlap in Ontology Population

Seong-Bae Park, Sang-Soo Kim, Sewook Oh, Zooyl Zeong, Hojin Lee, and Seong Rae Park

***Abstract*—**An ontology is widely used in many kinds of applications as a knowledge representation tool for domain knowledge. However, even though an ontology schema is well prepared by domain experts, it is tedious and cost-intensive to add instances into the ontology. The most confident and trust-worthy way to add instances into the ontology is to gather instances from tables in the related Web pages. In automatic populating of instances, the primary task is to find the most proper concept among all possible concepts within the ontology for a given table. This paper proposes a novel method for this problem by defining the similarity between the table and the concept using the overlap of their properties. According to a series of experiments, the proposed method achieves 76.98% of accuracy. This implies that the proposed method is a plausible way for automatic ontology population from Web tables.

***Keywords*—**ontology population, domain knowledge consolidation, target concept selection, property overlap

I. Introduction

An ontology is a formal specification of concepts for a domain of interest, and is considered as one of most important elements in semantic web. In addition, many research communities pay great attention to it as a tool for representing the domain knowledge of their tasks. In spite of its wide applications, the construction of an ontology is a difficult task since it requires the inspection of the target domain from the human experts. As a result, in many real-world applications, only the schema structure of an ontology is designed. The lack of instances in the ontology leads to loss of its reality. That is, many instances are needed for such an ontology to be used in the real-world applications.

In spite of the importance of instances in the ontology construction, it is a tedious, time-consuming, laborious work to add instances into the ontology [5]. One must find the information source containing the information for the concepts in the ontology, classify the concept according to its contents, and merge several kinds of information as the consolidated knowledge. As a result, recently the automation of this process is paid attention to by many researchers [4]. A term 'automatic ontology population' is generally used for this task [3].

The automatic insertion of instances into the ontology from unstructured normal texts is practically impossible due to poor performance of many natural language processing techniques. The identification of proper instances from a natural language sentence requires syntax analysis, word sense disambiguation, named-entity recognition, relation extraction, coreference solution, and so on. However, these techniques all suffer from severe ambiguities and the resolution of the ambiguities is not yet satisfactory.

When the analysis of natural language is insufficient, the candidate information source is Web pages, semi-structured documents. Especially, the Web pages contains a number of tables, and a table is an arranged form of human knowledge. Since it is relatively easy to extract knowledge from tables, it is practical to add instances from Web tables.

The primary problem using HTML tables for ontology population is to find the most proper concept within the ontology for a given table. In this paper we propose a novel method to select the target concept among the concepts within the ontology. The target concept is defined as the concept which is most similar to the table instance. The similarity between a table and a concept is measured by the overlap of properties between them.

The performance of the proposed method is evaluated with a set of Web pages about IT products. An ontology for cataloging IT products is used as a base ontology. This ontology is a part of the *National IT Ontology* supported by Korean government. The proposed method achieves 76.98 % of accuracy, which implies that it is a plausible method for automatic ontology population.

The rest of this paper is organized as follows. Section 2 proposes the overall structure for the automatic ontology population and Section 3 explains how the target concept is selected in ontology population. Section 4 presents the experimental results. Finally, Section 5 draws conclusions.

II. Ontology Population

Even though the proposed method can be applied to any ontology, for the explanation of the method let us consider an ontology for IT products which is a part of the *National IT Ontology* supported by Korean government[1]. Like many other ontologies available on World Wide Web, this ontology has just a few instances for each concept. In order to use it in the real-world applications, it is required to add more instances into the ontology.

The most common and ubiquitous source for gathering knowledge is the unstructured normal texts written in a natural language. However, since the natural language processing techniques are unsatisfactory, the normal texts are not suitable for ontology population. The most probable candidate is then the Web pages. If there are the Web pages with the specifications on IT products, they can be used as the knowledge

Seong-Bae Park and Sang-Soo Kim are with Department of Computer Engineering, Kyungpook National University, Daegu 702-01, Korea. Sewook Oh, Zooyl Zeong, Hojin Lee, and Seong Rae Park are with KTF Network Laboratory, Seoul 138-240, Korea.

[1] http://coreonto.kaist.ac.kr/core_eng/index.asp

(a) A Web page with digital television specifications

Warranty Term - Parts	1 year
Warranty Term - Labor	1 year
Product Height	29-3/5"
Product Width	40-3/5"
Product Weight	60.6 lbs.
Product Depth	11-9/10"
TV Type	LCD Flat-Panel
Screen Size	42"
Aspect Ratio	16:9
Display Type	Flat-panel LCD
Built-In DVR	No
...	...

(b) An extracted table from the Web page

Fig. 1. An example Web page containing digital television specifications and the table form transformed from the Web page by a Wrapper.

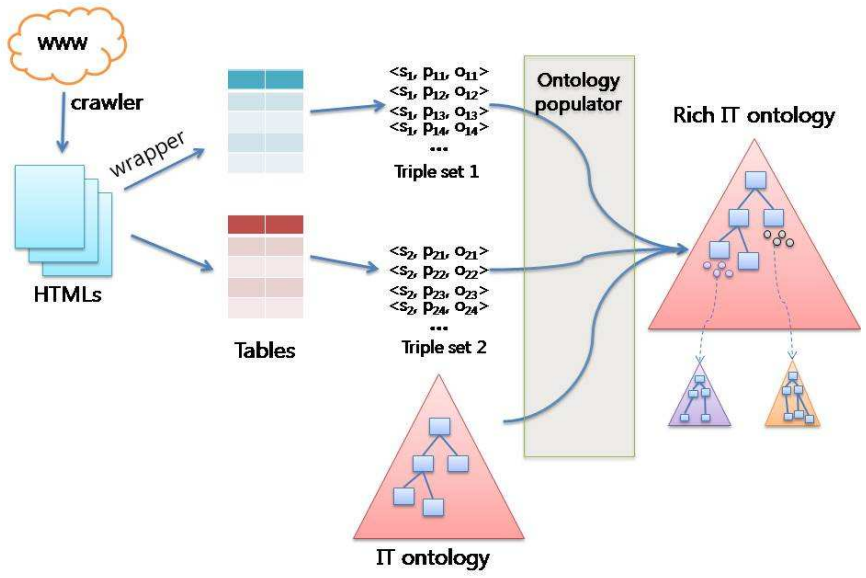

Fig. 2. The process of automatic population from Web pages with tables.

source for ontology population since the target ontology is for IT products.

Assume that a crawler collects a great number of Web pages containing information on IT products. If such Web pages contain the specific information for IT products as tables, the information can be extracted by a Web wrapper [8]. For instance, let us assume that we have a Web page about a digital television as shown in Figure 1-(a). This kind of tables can be automatically collected from World Wide Web by the focused Web crawler. Then, a Web wrapper extracts the tabular information within the page as in Figure 1-(b).

The IT ontology, an ontology for cataloging IT products is then populated from a set of such tables. Figure 2 shows the process how the ontology is populated from the Web pages. The tables for IT product specifications are prepared by a focused crawler and a wrapper. Each entry in a table is a pair of a property name and the corresponding value. The property is considered as a predicate and the value is regarded as an object of the predicate. Thus, each table can be considered as a set of triples, where the common subject is missing. This anonymous and unknown entity is the subject for all predicates, since a set of triples represents an entity.

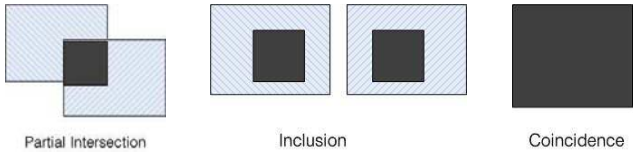

Fig. 3. The level of overlap in domain knowledge between the table and the ontology.

The entity represented as a set of triples from a table has to be matched with a concept within the IT ontology in order to populate it into the ontology. For instance, an entity in Figure 1-(b) should be matched with a concept DTV within the IT ontology, since the figure is a table of DTV specifications. That is, DTV is the most proper concept for the given table. Then, all information in the table is inserted into the IT ontology as instances of the concept DTV.

Since the domain knowledge of the ontology designer could be different from that of the table designer, the predicates used in the table are not perfectly matched with those in the ontology. There are three possibilities in adding instances from tables into the ontology according to the level of overlap in the domain knowledge [1] as shown in Figure 3. The first case is partial intersection between them, the second is inclusion where one of them is totally included into the other, and the third is coincidence where they two matches exactly. Since the ontology is our target, the information which is given in the table but can not be expressed in the ontology is all discarded.

III. Target Concept Selection

In order to find the most proper concept in the ontology for a product given as a table, a metric for properness should be first defined. The properness of a product against a concept in the ontology is computed by the information similarity between the product and the concept. Therefore, the most proper concept c^* for a given product T is

$$c^* = \arg\max_{c \in C} Sim(T, c), \tag{1}$$

where C is a set of concepts within the ontology and $Sim(T, c)$ is the similarity between T and c.

The similarity between a table and an ontology is computed under the assumption that both are represented as a set of triples. Since an ontology is generally understood in RDF and the RDF metadata model is based upon the idea of making statements about resources in the form of triples, the assumption that an ontology is a set of triples is true. In addition, since we assume that the Web tables can be considered as a set of triples, the assumption that a Web table is a set of triples is also true.

Since we just focus on ontology population, we do not consider, in this paper, the cases where the same properties or predicates have different types or values. That is, the values of the properties do not matter from now on. As a result, both tables and ontologies can be thought of as a set of properties or predicates. Then, the similarity between a table and an ontology is defined as an overlap of their properties.

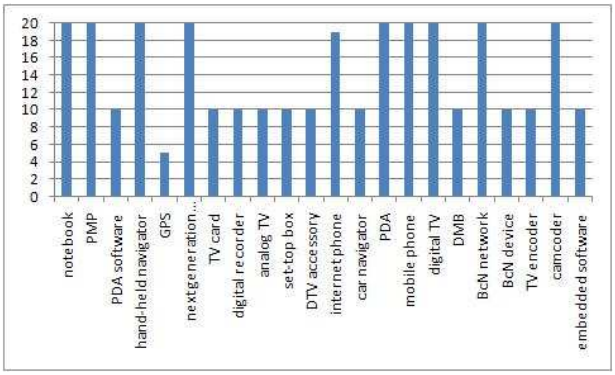

Fig. 4. Distribution of Web pages according to the product types.

An ontology has in general a relatively long depth in the concept hierarchy. That is, it is unfair to compare a table with general concepts which appear as the upper concept nodes in the ontology. Thus, for each concept $c \in C$, its depth is limited by k, an user-defined parameter. In the all experiments below, we set $k = 2$ because this value results in the best performance.

Let a table T is represented as a set of properties $< p_{T_1}, p_{T_2}, \ldots, p_{T_n} >$, and a concept in the ontology c is also given as a set of properties $< p_{c_1}, p_{c_2}, \ldots, p_{c_m} >$. The size of T or c is defined as the number of properties within. That is, $|T| = n$ and $|c| = m$. Then, the similarity between T and c in Equation (1) is

$$Sim(T, c) = \frac{2 \cdot |T \cap c|}{n + m}. \tag{2}$$

This definition of similarity implies intuitively that the similarity between T and c increases as the number of shared properties does. In addition, it has a good property that its value is bounded. That is, $0 \leq Sim(T, c) \leq 1$ for all possible T and c. Actually, the difference in property names is compensated by

$$Sim(T, c) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} lsim(p_{T_i}, p_{c_j})}{n \cdot m}, \tag{3}$$

where $lsim(p_{T_i}, p_{c_j})$ is the Levenshtein distance [7]. The Levenshtein distance is defined to be the number of deletions, insertions, or substitutions required to transform one string into the other string. Thus, the similarity measure is considered as an expected similarity of property names. This distance is very simple, but shows as high performance in many word matching tasks as other complex measures based on WordNet [2] like *Jiang-Coranth's distance* [6].

IV. Experiments

A. Data Sets

For the evaluation of the method proposed in this paper, we collected the Web pages which contain the specifications of IT products. The number of Web pages satisfying the condition is 278, and they contain the specifications of 22 kinds of products: *notebook, PMP, PDA software, hand-held navigator, GPS, next generation communicator*, *TV card, digital recorder,*

TABLE I
A SIMPLE STATISTICS ON THE DATA SET USED IN THE EXPERIMENTS.

Data set property	Value
No. of Web pages	278
No. of Product Kinds	22
No. of Total Properties	306
Average Web pages per Product	12.6
Average Properties per Product	13.9

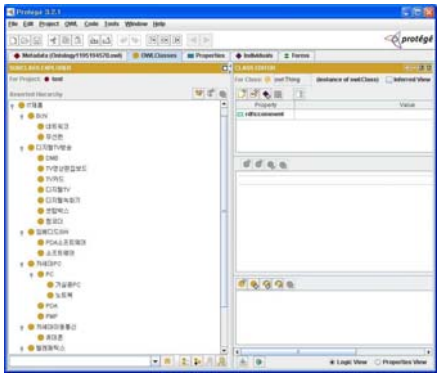

Fig. 5. A snapshot of Portégé for the IT ontology.

analog TV, set-top box, DTV accessory, *internet phone, car navigator, PDA, mobile phone, digital TV, DMB, BcN network, BcN device*, *TV encoder, camcoder*, and *embedded software*. Figure 4 shows how the categories of the IT products are distributed.

Some of them are distinguished one another easily, but some are difficult to discriminate. On average each product has 12.6 Web pages and 13.9 properties. Table I shows the simple statistics on the gathered data set.

The IT ontology used as a target ontology has 29 concepts and 159 properties (see Figure 5). Thus, a concept has 5.5 properties on average. Table II summarizes the simple statistics on the ontology.

B. Experimental Results

Table III summarizes the performance of the proposed method. 'Random Selection' in this table is to choose the target concept at random. Since there are 29 concepts in the IT ontology, the accuracy of this method is just $\frac{1}{29} \cdot 100 = 3.45\%$. 'Property Overlap with Edit Distance' is the method proposed in this paper.

The accuracy of 'Property Overlap with Edit Distance' is 76.98%, since it predicts the target concept correctly for 214 cases among 278. The accuracy is high enough to trust this method in target selection of ontology population. In Korean, there are no reliable thesaurus available online or for free. Thus, the edit distance is the only feasible candidate for computing the similarity between two lexicons in ontology population.

V. CONCLUSION

In this paper, we have proposed a method for selecting the most proper concept in ontology population. It finds the most proper concept by the overlap of information within two knowledge sources: an *ontology* and a *Web table*. According to the experimental results, the proposed method achieves 76.98% of accuracy. The experimental result implies that the proposed method is plausible for automatic ontology population.

The main advantage of the proposed method is that it does not depend on the external knowledge so much in populating instances from the Web tables. It helps the ontology written in minor languages without linguistic resources being automatically populated from world wide web. In addition, it is general enough to be applied to any ontology beyond the IT ontology only if the information for the instances is given in tabular form.

The drawback of the proposed method is that it looses the structural information residing in the table or in the ontology. Since both representations have a topological meaning, the consideration such topological information will increase selection accuracy. However, it is not simple to explore the relevant features to handle a complex structure. Thus, a method to compute the similarity between complex structures without explicit enumeration of relevant features is needed, and it is out future work.

TABLE II
A SIMPLE STATISTICS ON THE IT ONTOLOGY.

Ontology Characteristics	Value
No. of Concepts	29
No. of Properties	159
Average No. of Properties per Concept	5.5

TABLE III
THE EXPERIMENTAL RESULTS OF THE TARGET CONCEPT SELECTION.

Method	Accuracy
Random Selection	3.45%
Property Overlap with Edit Distance	76.98%

ACKNOWLEDGMENT

This work was supported by the IT R&D program of MIC/IITA [2007-S-048-01, Development of Elementary Technologies for Fixed Mobile IP Multimedia Convergence Services on Enhanced 3G Network].

REFERENCES

[1] J. Barrasa, Ó. Corcho, and A. Gómez-Pérez, "R_2O, an Extensible and Semantically Based Database-to-Ontology Mapping Language," In *Proceedings of the 2nd Workshop on Semantic Web and Databases*, 2004.
[2] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Semantic Distance," *Computational Linguistics*, Vol. 32, No. 1, pp. 13–47, 2006.
[3] S. Castano, S. Espinosa, A. Ferrara, V. Karkaletsis, A. Kaya, S. Melzer, R. Möller, S. Montanelli, and G. Petasis, "Ontology Dynamics with Multimedia Information: The BOEMIE Evolution Methodology," In *Proceedings of International Workshop on Ontology Dynamics*, 2007.
[4] H. Davulcu, S. Vadrevu, S. Nagarajan, and I.V. Ramakrishnan, "OntoMiner: Bootstrapping and Populating Ontologies from Domain-Specific Web Sites," *IEEE Intelligent Systems*, Vol. 18, No. 5, pp. 24–33, 2003.
[5] S. Handschuh, R. Volz, and S. Staab, "Annotating for the Deep Web," *IEEE Intelligent Systems*, Vol. 18, No. 5, pp. 42–48, 2003.
[6] J. Jiang and D. Conrath, "Semantic Similarity based on Corpus Statistics and Lexical Taxonomy," In *Proceedings of the 10th International Conference on Research in Computational Linguistics*, pp. 19–33, 1997.

[7] E. Sang, S. Canisius, A. Bosch, and T. Bogers, "Applying Spelling Error Techniques for Improving Semantic Role Labelling," In *Proceedings of the 9th Conference on Computational Natural Language Learning* pp. 229–232, 2005.

[8] T. Sugibuchi and Y. Tananka, "Interactive Web-Wrapper Construction for Extracting Relational Information from Web Documents," In *Proccedings of the 14th International Conference on World Wide Web*, pp. 968–969, 2005.