A Demo for Constructing Domain Ontology from Academic Papers

Feiliang REN
Northeastern University, Shenyang, 110819, P.R.China

renfeiliang@ise.neu.edu.cn

ABSTRACT

Traditional construction methods of domain ontology usually have following two limits. First, these methods usually depend on some high cost resources. Second, these methods are easily to result in error propagation because of the errors introduced in the concept identification step. In this paper we present a demo that constructs domain ontology with an easy method. And three main features distinguish our method from traditional methods. First, the proposed method uses academic papers to construct domain ontology. Second, the proposed method carefully selects some keywords in these academic papers as domain concepts. Thus error propagation is reduced accordingly. Third, the proposed method mines hierarchical relations among concepts with a graph generation and conversion method. The effects of our proposed method are evaluated from two perspectives in an IT domain ontology which is constructed with the proposed method: the quality of domain concepts and the quality of concept's relations. And evaluation results show that both of them achieve high qualities.

KEYWORDS : Domain Ontology; Graph Generation; Graph Conversion; Ontology Concept; Hierarchical Relation

Proceedings of COLING 2012: Demonstration Papers, pages 369–376,
COLING 2012, Mumbai, December 2012.

1 Introduction

Domain ontology is a kind of domain related ontology knowledge which usually contains three basic items: domain concepts, concept relations and concept interpretations. Because it is well known that domain ontology can reduce or eliminate conceptual and terminological confusion for many hot research areas such as semantic web, informational retrieval, question and answering, machine translation, and so on, a lot of researchers have been devoting to constructing various domain ontologies for decades. Compared with general purpose ontology like WordNet, domain ontology has following two features.

First, all of the ontology items must be related to the same domain. It is easily to understand that the domain concept is crucial to domain ontology because the other two ontology items will center around it. To achieve high quality domain ontology, the domain concepts must be accurately identified.

Second, domain concepts are dynamic: new concepts are constantly emerging.

Because of these features, two barriers are put up for the construction of domain ontology. One is how to identify domain concepts accurately. And the other is how to update domain ontology timely when new concepts emerge? To construct a practical and useful domain ontology, these barriers must be overcome effectively.

Traditional construction methods of domain ontology cannot overcome these barriers effectively. First, because of the technology limits, many errors are introduced during the process of identifying domain concepts. Second, traditional method usually cannot respond to concept's change timely. Even more serious, these methods usually depend on some high cost resources like other general purpose ontology, right concept tagged corpus, right relation tagged corpus and so on. However, these resources are not always acquired easily, especially for those resource-lack languages.

In this paper, we present a demo that construct domain ontology with an easy way, and it can overcome above barriers effectively. The proposed method takes academic papers as data source and selects some keywords in these academic papers as domain concepts. And the hierarchical relations among concepts are mined with a graph generation and conversion method. When new concepts emerge, domain ontology can be completely reconstructed easily.

2 Our Basic Idea and System Architecture

Usually academic papers are easily acquired even for those resource-lack languages. Among these papers there are three implicit but widely acknowledged facts which are useful for domain ontology construction. First, it is certain that authors will submit their paper to those journals that are related to their research fields. Thus we can say that academic papers have been classified into appropriate domains before submitted as these research fields are nature classification of domains. So it is easily to collect some papers in a specific domain according to journals' research scopes. Second, keywords are usually used to discover paper theme in a concise way and they usually contain rich information that is related to a specific domain. Thus keywords are born concepts in the domain where they belong to. Third, there are usually two kinds of keywords in an academic paper. One is more related to paper's domain, while the other is more related to paper theme. So if we use a directed graph to describe a domain ontology, keyword

frequency can be used to reveal a kind of hierarchical relation among keywords: high frequency keywords are usually more related to domain and should be placed in the higher levels of the ontology graph; while low frequency keywords are usually more related to paper themes and should be placed in the lower levels of the ontology graph.

These facts indicate that domain ontology can be constructed in such an easy way: using academic papers as data source, selecting some keywords as domain concepts, and mining hierarchical relations among concepts based on their frequencies.

Based on these analyses, we design our domain ontology construction method whose system architecture is shown in Figure 1.

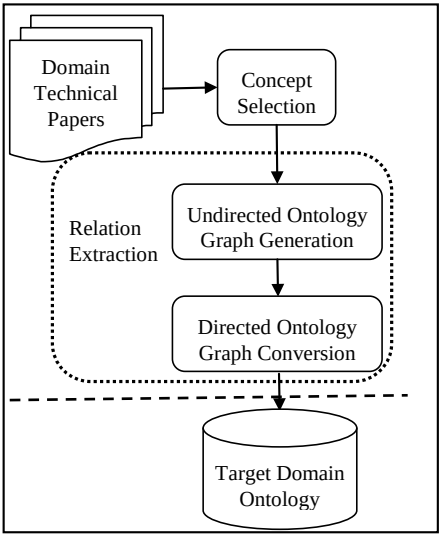

FIGURE 1. SYSTEM ARCHITECTURE

There are three main components in our domain ontology construction method: concept selection, undirected ontology graph generation and directed ontology graph conversion. And the latter two components can been viewed as a relation extraction model.

3 Our Method

Concept Selection

We have pointed out that keywords are born concepts in the domain where they belong to. But we should also notice that some keywords are so common that will appear in several completely different domains. Obviously these keywords are not appropriate for taking as concepts in a specific domain. Here we use two methods to select some keywords as appropriate concepts.

The first one is the *tf*idf* method. We think a keyword will be appropriate for taking as domain concept if it has a high frequency in a target domain but a low frequency in other domains. Based on this idea, using *tf*idf* value to select concept from keywords is a natural choice.

The second one is to remove all of the abbreviation keywords that are made up of capital letters because it is hard to understand the real meaning of an abbreviation keyword. For example, there is such keyword like "*TMT*", none can understand its real meaning is "*Trustable Machine Translation*" if there are not any contexts provided. Thus these abbreviations are more likely to introduce confusions than to eliminate confusions in some applications that need domain ontology. So we remove all of those abbreviation keywords from the concept list that is generated by the *tf*idf* method.

3.1 Relation Extraction

In an ontology graph, we take those selected concepts as vertexes and use directed edges to represent the hierarchical relations among concepts. After concept selection, two steps are taken to mine this kind of hierarchical relation. The first step is to construct an undirected graph based on co-occurrence information among concepts. The second step is to convert this undirected graph into a directed graph.

Step 1: Undirected Ontology Graph Construction

In this step, our basic idea is that if two concepts appear in the same paper's keyword list, there will be a potential hierarchical relation among them. And we use undirected edges to describe these potential relations among concepts. Thus the aim of this undirected ontology graph construction step is to find all of these potential relations. Specifically, if two concepts appear in the same paper's keyword list, an undirected weighted edge will be added between them. In the final undirected ontology graph, the weight of an edge is the co-occurrence frequency of this edge's two adjacent concepts. The detail of this construction algorithm is shown in FIGURE 2.

Input: Concept set *CSet* and keyword list set $C=\{S_1,S_2,\ldots,S_n\}$

Output: Undirected Ontology Graph G=(V(G),E(G))

Algorithm:

1. **Initialize** G, set V=*CSet*, E(G)= Φ
2. **For** *i=1* to *n* **Do**

   For each two keywords $k_{im},k_{in} \in S_i \cap V$

   If edge $(k_{im},k_{in}) \in$ E(G) **then** Set $W(k_{im},k_{in})=W(k_{im},k_{in})+1$;

   Else Add edge (k_{im},k_{in}) in *E(G)* and Set $W(k_{im},k_{in})= 1$;
3. **Remove** all edges whose weights less than a given threshold θ
4. **Return** *G*

FIGURE 2. Undirected Ontology Graph Construction

In FIGURE 2, $W(k_{im},k_{in})$ is the weight of undirected edge (k_{im},k_{in}), V(G) and E(G) are vertex set and edge set of the ontology graph respectively.

Step 2: Directed Ontology Graph Conversion

This step aims to make those potential hierarchical relations explicit. That is to say we need to convert those undirected edges into directed edges so as to reveal the farther-child relations among concepts. As we have pointed our previously, in an undirected edge, the high frequency concept is usually more related to the target domain and thus should be taken as a farther vertex; while the low frequency keyword is usually more related to paper theme and thus should be taken as a child vertex. Based on this idea, we design a conversion algorithm to make those potential hierarchical relations explicit. And the detail of this algorithm is shown in FIGURE 3.

Input: Undirected Ontology Graph G

Output: Directed Ontology Graph G'

Algorithm:

1. **For each** undirected edge $(c_i,c_j) \in E(G)$

   If $deg(c_i)$-$deg(c_j) > \theta$ **Change** (c_i,c_j) to $<c_i,c_j>$;

   Else If $deg(c_j)$-$deg(c_i) > \theta$ **Change** (c_i,c_j) to $<c_j,c_i>$;

   Else Change (c_i,c_j) to $<c_j,c_i>$ and $<c_i,c_j>$;

2. **Return** *G'*

FIGURE 3. Directed Graph Conversion

In FIGURE 3, $<c_i,c_j>$ denotes a directed edge from concept c_i to concept c_j. And $deg(c_i)$ denotes the degree of concept c_i.

After this step, we construct a directed ontology graph. In this graph, every vertex denotes a concept, and every directed edge denotes a hierarchical relation in which its starting concept denotes an upper father vertex and its ending concept denotes a lower child vertex.

After this step, the remaining directed edges and their adjacent concepts together constitute the final domain ontology graph.

4 Approach Evaluation

4.1 Data Preparation

For Chinese, almost all published academic papers can be downloaded from a Website (http://www.cnki.net/) and all of these papers have been classified into proper domains on the Web. From this Website, we downloaded more than four hundred thousand of academic papers in *Information Technology* (IT for short) domain that span almost the past thirty years.

With these data, we constructed an IT domain ontology with the proposed method and used it to evaluate the proposed approach. Specifically, we set the thresholds of *tf* and *idf* to 2 during the process of concept selection. That is to say, only those keywords whose *tf* values are greater than 2 and *idf* values are less than 2 will be selected as domain concepts. Finally, we constructed the target IT domain ontology that contains 383176 concepts and 239720 hierarchical relations.

4.2 Evaluation Strategy

In this paper, we use accuracy, recall and F1 value to evaluation method. All of these evaluations are performed by five experienced human experts who come from our research group. And we randomly select 500 concepts from our domain ontology. And we also randomly select 500 relations from our domain ontology as test relation set. And the evaluation results are shown in Table 1 and Table 2 respectively.

	Our Method
Accuracy	*93.6%*
Recall	84.2%
F1	*88.7%*

TABLE 1 – Concept Evaluation Results

	Our Method
Accuracy	88.4%
Recall	80.2%
F1	84.1%

TABLE 2 –Relation Evaluation Results

From our experimental results we can see that the domain ontology constructed with our methods achieves far higher quality. We think following reasons play major roles for this result. First, our method takes keywords as domain concepts. It is well known that most of the keywords are domain terms, so they are nature domain concepts. Thus our method effectively avoids the error propagation which will often trouble traditional domain ontology construction methods. Second, our concept relation discovery method is mainly based on the co-occurrence of two concepts and the frequency of each adjacent concept. From the experimental results it can be seen that our method well captures the writing habits of most researchers when they writing technical papers. From the experimental results we can also see that our method is very effective. It can construct a domain ontology with rich concepts and hierarchical relations.

5 Conclusions

In this paper, we propose a simple but effective domain ontology construction method. Our method uses academic papers as data source and selects some keywords in these academic papers as domain concepts. The hierarchical relations among concepts are mined based on a graph generation and conversion method.

Compared with other domain ontology construction methods, our method has following novel aspects. First, the proposed method can be used to construct domain ontology for many languages. In our method, the used data source is a kind of very common resource that can be acquired easily for many languages. Thus the proposed method has a large scope and can be easily transplanted to any languages even for those resource-lack languages. Second, the proposed method can construct some domain ontologies with high qualities in both concept quality and relation quality. Third, the proposed method is easily implemented. It doesn't use any complex technologies or high-cost resource. Any researchers can implement our work easily. Fourth, the proposed method is suitable to construct some large-scale domain ontologies.

Acknowledgements

This paper is supported by the National Natural Science Foundation of China (Grand No. is 61003159, 61100089, 61073140, and 61272376).

References

R. Navigli, P.Velardi, A.Gangemi, “Ontology Learning and Its Application to Automated Terminology Translation”, IEEE Intelligent Systems, 2003, pp.22-31.

Navigli, Roberto, and Paola Velardi. 2004. “Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites”. Journal of Computational Linguistics, volume 30: 151-179.

P.Cimiano and J.Volker. “Text2Onto-A Framework for Ontology Learning and Data-driven Change Discovery”, Proceedings of NLDB 2005.pp227-238.

Buitelaar, Paul, Daniel Olejnik, and Michael Sintek. 2004. A Protégé Plug-In for Ontology Extraction from Text Based on Linguistic Analysis. In the Proceedings of the 1st European Semantic Web Symposium:31-44

H.Poon and P.Domingos. “Unsupervised Ontology Induction from Text”, Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 2010, pp.296–305.

H.Tan and P.Lambrix. “Selecting an Ontology for Biomedical Text Mining”, Proceedings of the Workshop on BioNLP, June 2009, pp.55–62.

D.Estival, C.Nowak and A.Zschorn. “Towards Ontology-Based Natural Language Processing”. Proceeedings of the Workshop on NLP and XML, 2004, pp59-66.

J.Uwe Kietz and R.Volz. “Extracting a Domain-Specific Ontology from a Corporate Intranet”. Proceedings of CoNLL-2000 and LLL-2000, pp.167-175.

A.Maedche and S.Staab, “Ontology Learning for the Semantic Web”, IEEE Intelligent Systems, Vol.16, no.2, 2001, pp.72-79.

Shih-Hung Wu and Wen-Lian HSU, “SOAT: a semi-automatic domain ontology accuisition tool from Chinese corpus”, Proceedings of the 19th international conference on Computation linguistics (COLING 2002), pp1-5.

Sara Salem and Samir AbdeRahman, “A Multiple-Domain Ontology Builder”, Proceedings of the 23rd International Conference on Computation Linguistics (COLING 2010), pp967-975.

Pinar Wennerberg, “Aligning Medical Domain Ontologies for Clinical Query Extraction”, Proceedings of the EACL 2009 Student Research Workshop, pp79–87.

Jantine Trapman and Paola Monachesi, “Ontology engineering and knowledge extraction for crosslingual retrieval”, International Conference RANLP 2009, pp455–459.

Mihaela Vela and Thierry Declerck, “Concept and Relation Extraction in the Finance Domain”, Proceedings of the 8th International Conference on Computational Semantics, 2009, pp346–350.

Tingting He, Xiaopeng Zhang, Xinghuo Ye, “An Approach to Automatically Constructing Domain Ontology”, PACLIC 2006, pp150-157.

Chu-Ren Huang, Ya-Jun Yang and Sheng-Yi Chen, “An Ontology of Chinese Radicals: Concept Derivation and Knowledge Representation based on the Semantic Symbols of Four Hoofed-Mammals”, 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC 2008), pp189–196.

Chu-Ren Huang, “Text-based construction and comparison of domain ontology: A study based on classical poetry”, PACLIC 2004, pp17-20.

He Tan, Rajaram Kaliyaperumal, Nirupama Benis, “Building frame-based corpus on the basis of ontological domain kwowledge”, Proceedings of the 2011 Workshop on Biomedical Natural Language Processing, ACL-HLT 2011, pp74–82.