퍼지 및 지능시스템학회 논문지 2006, Vol. 16, No. 6, pp. 766-771

Building Domain Ontology Based on Linguistic Patterns

Kweonyang Kim* and Sooyeon Lim**

* School of Computer Engineering, Kyungil University

** Department of Computer Engineering, Kyungpook National University

Abstract

In this paper, we focus on the building domain ontology from corpus by extracting concepts and properties relationships based on linguistic patterns. The pharmacy field is selected as an experiment domain and we present an algorithm to extract hierarchical structure for terminology based on the noun/suffix patterns of terminology in domain texts. In order to show usefulness of our domain ontology, we compare a typical keyword based retrieval method with an ontology based retrieval method which uses related information in an ontology for a related feedback. As a result, our method shows the improvement of precision by 4.97% without losing recall.

Key Words : ontology, semantic relation, linguistic patterns, concept, terminology

1. Introduction

The success of the semantic Web depends on the proliferation of ontology which requires that the construction and update of ontology be completed quickly and easily. Ontology learning technique greatly helps ontology engineers to build and update ontology. Another critical thing is text mining technology that extracts concepts of the concerned domain and semantic relationship among the concepts[10]. Ontology learning constructs ontology semi-automatically from unstructured, semi-structured or fully structured data. Even if an automatized method is used, however, the core part of ontology, which is conceptual system, should be made manually in order to construct high quality semantic knowledge base. As semiautomatic methods of constructing ontology, there are largely those using existing resources such as thesauruses and dictionaries[7] and those constructing base ontology and expanding it using the distribution of words obtained from analyzing texts without using existing resources[13]. While the former can build knowledge base that can be utilized immediately without additional dictionary process as they use dictionaries containing a large volume of concepts, while the latter can expand concepts easily. Both methods are quite important for extracting high quality patterns of semantic relationships. In order to extract a larger number of patterns of semantic relationships, it is necessary to process terminology obtained from analyzing the form of terms appearing in texts within a domain corpus.

We focus on the building a domain ontology by extracting a semantic group and hierarchical structure after classifying and analyzing the linguistic patterns of terminology that appeared in a Korean document as a type of compound noun[12]. Our method classifies terminology words related to a specific subject by using a hierarchical structure to improve the performance of retrieving a document. A retrieval engine can be used as a base of inference to use the retrieval function using the concept and rule defined in the ontology. In order to experiment this retrieval engine, the texts that exists in a set of documents related to the pharmacy field are used.

The existing methods for extracting terminology can be largely classified by a rule based method and statistics based method. A rule-based method[3,4,6,8,10] builds a configuration pattern of terminology by hand or learning corpus and recognizes terminology by building a recognition pattern automatically. A rule-based method shows a relatively exact result because people directly describe the heuristic rules using a noun or suffix dictionary. A statistics based method[5,13] uses some kind of knowledge, such as the hidden Markov model, maximum entropy model, word type, and vocabulary information in order to learn the knowledge base for the recognition from a learning corpus. Our method extracts terminology by using a rule-based method and configures a rule to extract terminology by analyzing the linguistic patterns of the terminology.

2. Building a domain ontology

A development of an ontology indicates a configuration area of an ontology by considering the characteristics of domain and usage of ontology, and then defines the detailed items of the ontology based on a list of competency questions that a knowledge base based on the ontology should be able to answer . In order to perform this process, it is a definition of the concepts and structure through a conference with the specialists of domain is required and builds by using these results. In an actual application system, it is necessary to build an ontology that includes a specific knowledge for each

접수일자 : 2006년 10월 1일
완료일자 : 2006년 11월 30일

domain.

We selected the pharmacy field as an experiment domain and limited document. That is, a more detailed building of an ontology will be performed in the domain of a pharmacy. In addition, it is aimed to build a pharmacy ontology to retrieve a related name of drugs or documents.

2.1 Ontology building process

The building process of the proposed ontology consists of four steps in Figure 1. First, the web documents that exist in the related web are collected and structurized through a document transformation process. Second, the verbs that express the relation between concepts and the concepts are extracted after passing a simple natural language processing step. Third, the terminology will be extracted from the extracted concepts and produce a hierarchical structure from the results of the structure analysis. Finally, the extracted relations will be added to the existing ontology. The relation that will append the concepts of the pharmacy ontology, which will be built using the results of the documents that exist in the pharmacy database (http://www.druginfo.co.kr) and the results will be configured. The collected documents will form concepts according to the tag that is attached after the transformation process to fit the configured structure using a semi-structurized(tagging) document. The concepts that existed in the built ontology and its relation is represented as OWL(Web Ontology Language).

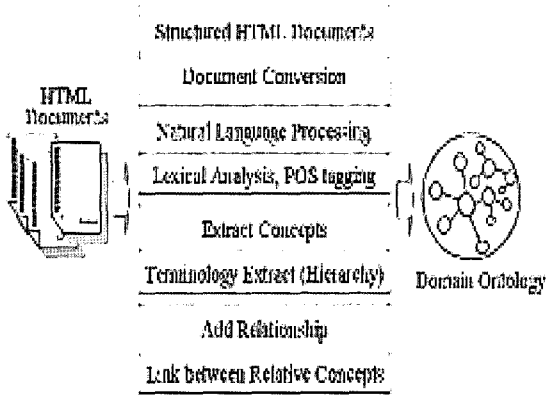

Fig. 1 Ontology building process

2.1.1 Base ontology

For building an ontology, an ontology engineer decides the most general concepts in the domain, that are located in the upper level, and the subsequent specialization of the concepts. These concepts are called by the base ontology. Our method configures a base ontology using 48 words as shown in Figure 2.

In order to configure this scheme, the 4 concepts such as the name of diseases, symptoms, drugs are defined as the hypernym node and its 45 subsequent hyponym nodes. The hyponym nodes consist of each node group; 20 nodes, which are defined by following the classification of a specific noun or suffices that form the name of a disease or symptom that exist in the pharmacy domain, and 15 nodes for the configured structures, and 10 nodes, which express general nouns that have a high frequency of appearance.

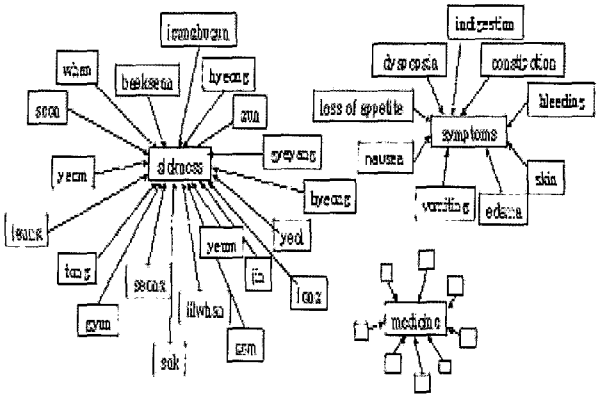

Fig. 2 Configuration of a base ontology

2.1.2 Concepts extraction

The stop words included in a document will be removed after the processes of morpheme analysis and part of speech tagging. Then, all nouns and verbs are extracted from the sentences of a document. In order to perform this process, the stop words list was made by using 181 words by considering the morphological characteristics of Korean language in which a part of suffices was excluded in the process of stemming because these suffices will be nicely used to extract terminology. The extracted nouns from an ontology that is a kind of network with a lot of words mean the concept of ontology. The tags and verbs present the relation between the concepts and act as a network, which plays a role in the connection of the concepts to each other.

There are some proper nouns and compound nouns in the documents what we have an interest in, such as the name of a disease, symptoms, ingredients, and other things. The proper nouns that present a major concept are processed the same way as a general noun. In addition, the terminology that appeared as a compound noun in the domain is extracted and hierarchicalized, and then is added into the ontology.

2.1.3 Adding relations

This paper introduces two methods for extracting relations between concepts. The one is using the value of tag attached at the front of text, and the other is using the verb by extracting it from the text. The left part of Table 1 shows 15 types of semantic relations according to the value of the attached tag.

Moreover, the verbs, which connect the nouns, that appears in the surrounding context are also extracted. As a result, the frequency of 35 verbs that have a frequency of over 200 was 11,250. It covers 47.97% of the frequency of the entire verbs. This paper classifies these

results as a semantic pattern and defines it 18 semantic relations as presented in the second column of Table 1. The relationship between the extracted verbs and nouns can be verified by the co-occurrence information. If there is a relationship that exists between the nouns and verbs, it will compare a relationship between other nouns and verbs.

Tab. 1 Types of the semantic relations

Semantic Relations based on Tag Value		Semantic Relations based on Extracting Verb	
productedBy	hasContra	appear	use
hasInsCode	hasSideEffect	beWorse	cause
hasComCode	byMean	inject	accompany
hasClsCode	byAmount	noTake	prevent
hasColor	byUnit	reduce	control
hasKind	byAge	return	infect
hasForm		rise	cure
hasEffect		take	improve
hasMethod		relax	maintain

2.1.4 Terminology extraction

Terminology is a set of lexical unit that has a specific meaning in a given domain and characterizes a subject by expressing the concept used in a domain. A language resource for terminology is important to perform effectively and precisely, such as a machine translation or information retrieving for a specific domain because this terminology is a necessary element to understanding a domain.

This paper analyzes an appearance type of terminology in order to extract the information automatically. The shape combining of terminology shows very varied ways. Almost all of the terminology appeared in an appropriate domain presented as a type of compound noun and can be classified as two types as follows. The one is a singleton term, that is, it has a simple shape of combining with one word that has no spacing words. The other is a multi-word term that has spacing words and is a kind of compound noun with two more words in which it has a semantic relation with the front element of a word.

The nouns and suffices that are configured by a singleton terminology are classified by 20 kinds, such as yeom, jeung, tong, gyun, seong, jilwhan, sok, yeomjeung, jin, gam, jong, byeong, yeol, gweyang, seon, baekseon, jeunghugun, hyeong, hwan, gun in which it is appended by "hyponymOf" because it is almost a lower level word of a specific noun term.

A multi-word term terminology has almost a relation of modifier and keyword like "acute bronchitis" in which there are many cases that a keyword consists of terminology, which has a singleton term. This paper configures 5 semantic patterns and defines the semantic relation of an ontology according to these patterns[12].

2.2 Ontology extension

A built ontology can be extended by other resources, such as other ontologies, thesaurus, and dictionaries. It is possible to reduce the time and cost to extend an ontology by applying the predefined concepts and rules by using the existing resources.

The process consists of 3 steps, such as the import, extract, and append. The step of import means the bringing and using of external resources. This paper uses two external resources(http://www.nurscape.net /nurscape/dic/frames.html, http://www. encyber.com/) In this step, the extraction was applied to terminology and its hyponym concepts for the results of the appropriate concept retrieving because the range of text was so wide. Then, the extracted concepts will be appended at a proper location by considering the hypernym and hyponym relations in the ontology.

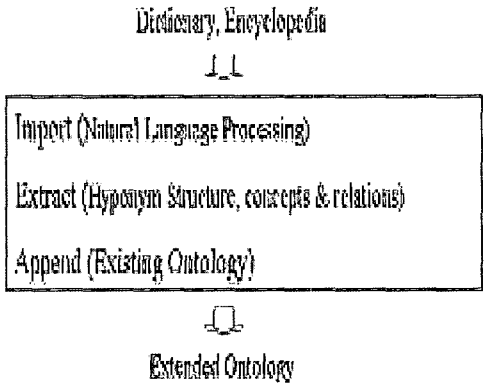

Fig. 3 Ontology extending process

3. Ontology application for document retrieval

This study applies ontology as a means to improve the accuracy of document retrieval in order to process users' queries effectively. Through the major set of documents in a specific field, concepts and relations are extracted by analyzing these documents. The objective of concept extraction is to extract the nouns that represent the documents as well as possible. Especially, in the case of a retrieval or question-answering system using an ontology given by weights, it is helpful to the judgment of a user by presenting selected minor information according to the weights.

Our method extracts the concepts using the extraction method proposed in this study and configures it as a node in the ontology. At this moment, the objective of this process is retrieving not only the concepts that are an inputted query in the ontology but also its hyponym concepts. A relevance feedback is well known as an effective way to process a reformation of a query that is an important part to access information. In the case of

applying the relevance feedback to improve the traditional method of tf*idf , it is well recognized that it can improve a precision rate if it is applied to a set of sample document of small scale.

This paper uses a hierarchical relation for the process of user relevance feedback in the ontology. The query will be extended by using the terminologies, which appeared as a hyponym information in the ontology that related to the inputted queries, and calculates the weights for the rewritten query.

In this process, the hyponym retrieval level to retrieve the nodes in the ontology was set by 2. For instance, let us assume that the hyponym nodes of 'exudative otitis media' and 'acute exudative otitis media' for the node of 'otitis media' exist in the ontology. If a query [otitis media] is inputted as an input, the set of queries will be extended as [otitis media, secretory otitis media, acute secretory otitis media] after retrieving the ontology. Then, it will recalculate the similarity based on its weights. The most widely known method of (term-weights allocation strategy) will be used to grant the weights. The calculated weights will increase the retrieving speed and precision by storing it in order of the arrangement with a document number to the appeared document. Figure 4 presents the configuration of a document retrieval system mentioned above in which it largely consists of a preprocessing module and retrieval module.

First of all, a morphological analysis process will be done in order to configure a set of index terms for the object documents in the preprocessing. From the results, the nouns are only extracted as a set of index terms in which the nouns will be mainly applied to get statistical information that represents a document for retrieving information and its classification.

In this system, the ontology will be used as a set of index terms. In order to compare the retrieval performance, a set of the upper 30 correct answered documents for the 10 queries by introducing the 5 specialists' advices about 430 documents was configured. Then, the rates of recall and precision for each query were produced based on this configuration.

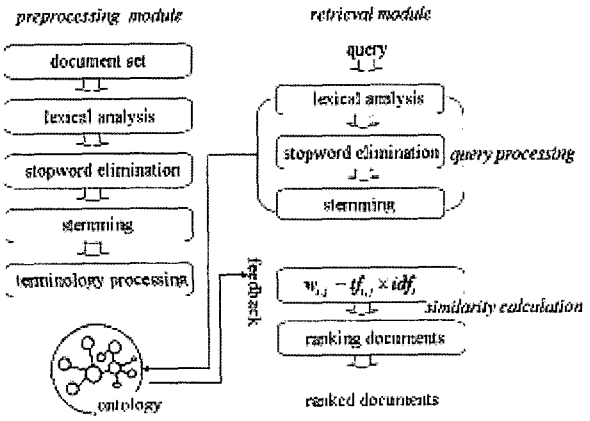

Fig. 4 Configuration of the document retrieval system

4. Experiments and evaluation

From the results of the morphological analysis of text in the formed corpus, the terminologies that were applied by specific expressions or patterns will be selected from the extracted nouns. In the case of the failure of analyzing a morpheme due to the errors of word spacing or typing, the errors of the analyzing was modified.

In order to verify the effectiveness of the built ontology, a keyword based document retrieval system that gives weights by using the traditional method of tf*idf will be compared and analyzed with an ontology based document retrieval system that uses a hyponym information that exists in the ontology to a relevant feedback and recalculates the weights.

In order to present the effectiveness for retrieving a document using the proposed method, this study compares the two methods. The one is a keyword based retrieval method by using the traditional method of, and the other is an ontology based retrieval method by using the hierarchical information that exists in the ontology to a relevance feedback.

An experiment reference collection and evaluation scale is used to evaluate an information retrieval system. An experiment reference collection consists of a set of literatures, information query examples, and set of relevant literatures for each information query. This paper collects 430 health/disease information documents from the home page of Korean Medical Association (http://www.kma.org) in order to configure a reference collection and configures an information query using 10 queries as follows. The experiment was carried for the 430 documents extraction. The objective of the experiment was to produce the recall and precision for the 10 queries in which a set of correct answers for each inputted query was defined in order of the documents set by the specialists.

Query 1: {otitis media, symptom, kinds}
Query 2: {otitis media, cure}
Query 3: {otitis media, cure, drug}
Query 4: {otitis media, property}
Query 5: {otitis media}
Query 6: {chronic otitis media diagnosis, cure}
Query 7: {acute otitis media}
Query 8: {otitis media, infection route}
Query 9: {fever, disease}
Query 10: {ear, disease}

An index for the texts was produced to increase the retrieval speed. It consists of two elements, such as the vocabulary and frequency. The vocabulary is a set of all words that exist in the text in which it has an appeared document vector for each word. An appeared document vector stores the location of the appeared document including the weights considered by the frequency.

Figures 5 and 6 present the comparison of the dis-

tributions of the precision and recall for the inputted queries using the two methods mentioned above. The average recall and precision for the two retrieval methods were respectively produced. The results are shown in Table 2. From the results, the ontology based document retrieval system that uses a hyponym information existed in the ontology to extend a query and grants the weights presented a high recall and precision by 0.78% and 4.79% respectively compared with the traditional method of tf*idf . This means that a hierarchical relation in the ontology that is used as a relevant feedback in a retrieval system will not largely affect to the recall but will affect the increase of the precision.

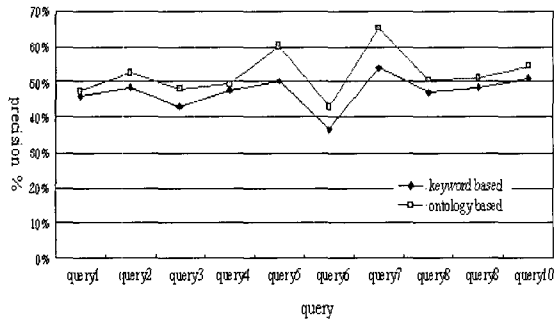

Fig. 5 Comparison of the precision

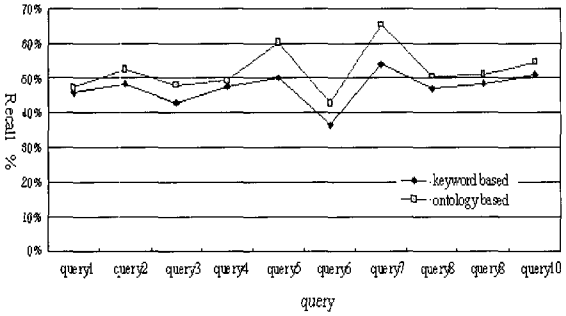

Fig. 6 Comparison of the recall

Tab. 2 Average recall and precision

	Keyword-based retrieval	Ontology-based retrieval
recall	46.34%	47.12%
precision	47.28%	52.25%

5. Conclusions

We propose a semiautomatic method to build a domain ontology using the results of text analysis and applies it to a document retrieval system. The experiment domain defined by the field of pharmacy. In addition, a corpus was formed by collecting a relevant document for drugs from the web. The structure of ontology can be configured by analyzing the structure of the text in the corpus and sets the types of relation to extract the concepts and relations. This paper especially proposes a method to process the ontologies that were combined with a specific nouns or suffices in order to extract the concepts and relations for building the pharmacy ontology after analyzing the types of ontologies appeared in the relevant documents. The proposed method is a kind of semi-automatic method using a text mining and can reduce the time and cost for building an ontology.

In addition, a keyword based document retrieval system that gives weights by using the frequency information compared with an ontology based document retrieval system that uses relevant information existed in the ontology to a relevant feedback in order to verify the effectiveness of the hierarchical relation in the ontology. From the evaluation of the retrieval performance, it can be seen that the precision increased by 4.97% while the recall was maintained as similar to other one. This means that if the hierarchical relation in the ontology is used as relevant feedback information, the precision will be improved.

It is necessary to modify and extend the ontology in order to process a more precise processing of an query. On the other hand, 33 semantic relations defined in the present pharmacy ontology may be short. In this case, it is necessary to redefine other semantic relations required in a specific domain and extend the pharmacy ontology. Moreover, it is required to study for using the pharmacy ontology in general purposes because it designed for a specific domain. That is, it is necessary for building an ontology for various domains and to study for a method that applies the proposed method of building ontology to the average domain.

References

[1] Baeza-Yates, R. and Robeiro-Neto, B.: Modern Information Retrieval. ACM Press, New York, NY, USA, 1999.

[2] Bettina, B., Andreas, H., Gerd, S.: Towards Semantic Web Mining. International Semantic Web Conference, 2002.

[3] Gyeong-Hee Lee, Ju-HO Lee, Myeong- Choi, Gil-Chang Lim, "Study on Named Entity Recognition in Korean Text," Proceedings of the 12th Conference on Hangul and Korean Information Processing, pp. 292-299, 2000.

[4] Hyo-Shik Shin, Young-Soo Kang, Key-Sun Choi, Man-Suk Song, Computational Approach to Zero Pronoun Resolution in Korean Encyclopedia, Proceedings of the 13th Conference on Hangul and Korean Information Processing, pp. 239-243, 2001.

[5] JongHoon Oh, KyungSoon Lee, KeySun Choi, "Automatic Term Recognition using Domain

Similarity and Statistical Methods," Journal of the Korea Information Science Society, Vol. 29, No. 4, pp. 258-269, 2002.
[6] Jung-Oh Park, Do-Sam Hwang, "A Terminology extraction system,"Proceedings of Korea Information Science Society Spring Conference(2001),Vol 27, No 1, pp. 381-383, 2000.
[7] Kang, S. J. and Lee, J.H.: Semi-Automatic Practical Ontology Construction by Using a Thesaurus, Computational Dictionaries, and Large Corpora. ACL 2001 Workshop on Human Language Technology and Knowledge Management, Toulouse, France, 2001.
[8] Klavans, J. and Muresan, S., "DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from On-line Text," Proceedings of AMIA Symposium, pp. 201-202, 2000.
[9] Michele M., Paola V. and Paolo F., "Text Mining Techniques to Automatically Enrich a Domain Ontology", Applied Intelligence 18, 322-340, 2003.
[10] Missikoff, M., Velardi, P. and Fabriani, P., "Text Mining Techniques to Automatically Enrich a Domain Ontology,"Applied Intelligence, Vol. 18, pp. 322-340, 2003.
[11] S. Y. Lim, Koo, S. O., Song, M. H., Lee, S, J., "Hub_word based on Ontology Construction for Document Retrieval", IC-AI'03, Las Vegas, USA, 2003.
[12] Soo-Yeon Lim, Mu-Hee Song, Sang-Jo Lee, "Domain-specific Ontology Construction by Terminology Processing," Journal of the Korea Information Science Society(B), Journal of the Korea Information Science Society Vol. 31, No. 3, pp. 353-360, 2004.
[13] Yi-Gyu Hwang, Bo-Hyun Yun, "HMM-based Korean Named Entity Recognition," Journal of the Korea Information Procissing Society(B), vol.10, No. 2, pp. 229-236, 2003.

저 자 소 개

김권양(Kweon Yang Kim)
1983년 : 경북대학교 전자공학과(학사)
1990년 : 경북대학교 전자공학과(석사)
1998년 : 경북대학교 컴퓨터공학과(박사)
1983~1988년 : ETRI 연구원
1999년~2000년 : University of Central Florida 방문교수
1991년~현재 : 경일대학교 컴퓨터공학부 교수

관심분야 : 시멘틱웹, 한글공학
Phone : 053-850-7287
Fax : 053-850-7609
E-mail : kykim@kiu.ac.kr

임수연(SooYeon Lim)
1988년 2월 : 경북대 전자공학과(학사)
1991년 2월 : 경북대 컴퓨터공학과(석사)
2004년 8월 : 경북대 컴퓨터공학과(박사)
2004년9월~2005년8월 : 경북대 박사후연구원
2006년3월~현재 : 경북대 전산교육부 초빙교수

관심분야 : 기계학습, 정보검색, 시멘틱웹, 자연어처리