Mining Fuzzy Domain Ontology from Textual Databases

Raymond Y.K. Lau
Department of Information Systems
City University of Hong Kong
Tat Chee Avenue, Kowloon
Hong Kong
E-mail: raylau@cityu.edu.hk

Yuefeng Li and Yue Xu
School of Software Engineering and Data Communication
Queensland University of Technology
GPO Box 2434, Brisbane, Qld 4001
Australia
E-mail: {y2.li, yue.xu}@qut.edu.au

Abstract

Ontology plays an essential role in the formalization of common information (e.g., products, services, relationships of businesses) for effective human-computer interactions. However, engineering of these ontologies turns out to be very labor intensive and time consuming. Although some text mining methods have been proposed for automatic or semi-automatic discovery of crisp ontologies, the robustness, accuracy, and computational efficiency of these methods need to be improved to support large scale ontology construction for real-world applications. This paper illustrates a novel fuzzy domain ontology mining algorithm for supporting real-world ontology engineering. In particular, contextual information of the knowledge sources is exploited for the extraction of high quality domain ontologies and the uncertainty embedded in the knowledge sources is modeled based on the notion of fuzzy sets. Empirical studies have confirmed that the proposed method can discover high quality fuzzy domain ontology which leads to significant improvement in information retrieval performance.

Keywords: *Fuzzy Domain Ontology, Fuzzy Sets, Text Mining, Semantic Web.*

1 Introduction

The success of Semantic Web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding [11]. Although there is not a universal consensus on the definition of ontology, it is generally accepted that ontology is a specification of conceptualization [4]. Ontology can take the simple form of a taxonomy (i.e., knowledge encoded in a minimal hierarchical structure) or a vocabulary with standardized machine interpretable terminology supplemented with natural language definitions. On the other hand, the notion of ontology can also be used to describe a logical domain theory with very expressive, complex, and meaningful information. Ontology is often specified in a declarative form by using semantic markup languages such as RDF and OWL [3]. Ontology provides a number of potential benefits in representing and processing knowledge, including the separation of domain knowledge from application knowledge, sharing of common knowledge of subjects among human and computers, and the reuse of domain knowledge for a variety of applications.

As domain ontology captures domain (context) dependent information, an effective discovery method should exploit contextual information in order to build relevant ontologies. On the other hand, since the taxonomy relations discovered from a text mining method often involve uncertainty, an uncertainty management mechanism is required to address such an issue. The notions of Fuzzy set and Fuzzy Relation are effective to represent knowledge with uncertainty [23]. Therefore, a fuzzy ontology rather than a crisp ontology is discovered by the proposed text mining method.

Definition 1 (Fuzzy Set) *A fuzzy set $\mathcal{F}$ consists of a set of objects drawn from a domain X and the membership of each object x_i in $\mathcal{F}$ is defined by a membership function $\mu_{\mathcal{F}} : X \mapsto [0, 1]$. If Y is a crisp set, $\varphi(Y)$ denotes a fuzzy set generated from the traditional set of items Y.*

Definition 2 (Fuzzy Relation) *A fuzzy relation is defined as the fuzzy set $\mathcal{G}$ on a domain $X \times Y$ where X and Y are two crisp sets.*

From the text mining perspective, a keyword is an object and it belongs to different concepts (a linguistic class) with various memberships. The subsumption relations among linguistic concepts are often uncertain and are characterized by the appropriate fuzzy relations.

0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.20

Definition 3 (Fuzzy Ontology) *A fuzzy ontology is a quadruple* $Ont =< X, C, R_{XC}, R_{CC} >$*, where X is a set of objects and C is a set of concepts. The fuzzy relation* $R_{XC} : X \times C \mapsto [0, 1]$ *maps the set of objects to the set of concepts by assigning the respective membership values, and the fuzzy relation* $R_{CC} : C \times C \mapsto [0, 1]$ *denotes the fuzzy taxonomy relations among the set of concepts* C*.*

The main contribution of our research work presented in this paper is the development of a novel fuzzy domain ontology discovery method which exploits contextual information embedded in textual databases. By combining lexico-syntactic and statistical learning approaches, the accuracy and the computational efficiency of the ontology discovery process is improved [12]. The remainder of the paper is organized as follows. Section 2 highlights previous research in the related area and compare these research work with ours. The computational details of the proposed ontology mining method are then illustrated in Section 3. Section 4 reports the empirical testing of our fuzzy domain ontology mining method. Finally, we offer concluding remarks and describe future direction of our research work.

2 Related Research

Cimiano et al. have presented an automatic taxonomy learning algorithm to extract concept hierarchies from a text corpus [2]. In particular, their taxonomy learning method is based on formal concept analysis [21]. Formal concept analysis is a systematic method for deriving implicit relationships among objects described by a set of attributes. Formal concept analysis can be seen as a conceptual clustering techniques at it provides intensional descriptions for the abstract concepts. The fuzzy ontology discovery method illustrated in this paper employs a novel subsumption based mechanism rather than the formal concept analysis approach to generate concept lattice. Semantically richer context vectors are used to represent concepts in our approach as opposed to the simple verb-based features employed by formal concept analysis. In addition, our concept hierarchy represents a fuzzy taxonomy of relations rather than a crisp taxonomy as proposed in [2].

The FOGA framework for fuzzy ontology generation has been proposed [20]. The FOGA framework consists of fuzzy formal concept analysis, fuzzy conceptual clustering, fuzzy ontology generation, and semantic representation conversion. Essentially, the FOGA method extends the formal concept analysis approach, which has also been applied to ontology extraction, with the notions of fuzzy sets. The notions of formal context and formal concept have been fuzzified by introducing the respective membership functions. In addition, an approximate reasoning method is developed so that the automatically generated fuzzy ontology can be incrementally furnished with the arrival of new instances. The FOGA framework is evaluated in a small citation database. Our method discussed in this paper differs from the FOGA framework in that a more compact representation of fuzzy ontology is developed. The proposed method is based on previous work in computational linguistic and with the computational mechanism built on the concept of fuzzy relations. We believe that the proposed method is computationally more efficient and be able to scale up for huge textual databases which typically consists of millions of records and thousands of terms. Finally, our proposed method is validated in a standard benchmark textual database which is considerably larger than the citation database used in [20].

A fuzzy ontology which is an extension of the domain ontology with crisp concepts is utilized for news summarization purpose [8]. In this semi-automatic ontology discovery approach, the domain ontology with various events of news is pre-defined by domain experts. The standard triangular membership function is used for computing membership values. The method discussed in this paper is a fully automatic fuzzy domain ontology discovery approach. There is no pre-defined fuzzy concepts and taxonomy of concepts, instead our text mining method will automatically discover such concepts and generate the taxonomy relations. In addition, there is no need to set the artificial threshold values for the triangular membership function, instead our membership function can automatically derive the membership values based on the lexico-syntactic and statistical features of the terms observed in a textual database.

An ontology mining technique is proposed to extract patterns representing users' information needs [9]. The ontology mining method consists of two parts: the top backbone and the base backbone. The former represents the relations between compound classes of the ontology. The latter indicates the linkage between primitive classes and compound classes. The Dempster-Shafer theory of evidence model is adopted to model the relations among classes. The presented method can effectively synthesizing taxonomic relation and non-taxonomic relation in a single ontology model. In addition, a novel method is proposed to capture the evolving patterns in order to refine the discovered ontology. Finally, a formal model is developed to assess the relevance of the discovered ontology with respect to the user's information needs. The ontology mining method is validated based on the Reuters RCV-1 benchmark collection. The research work presented in this paper focuses on fuzzy domain ontology discovery rather than the discovery of crisp ontology representing users' information needs.

An ontology based text mining system that extracts fuzzy relations from biological texts is present [1]. This approach preserves the basic structured knowledge format for storing domain knowledge, but allows for update of information at

the same time. The document processor parses the text documents and removes the tags pertaining to the biological domain. The strength of association between a tag pair E_i and E_j representing two biological entities is computed according to a fuzzy conjunction operator. Basically, the membership values of the relations are functions of frequency of co-occurrence of concepts. The fuzzy relations between the biological terms are used to guide information retrieval from a medical document collection called GENIA. The ontology discovery method presented in this paper deals with general textual databases rather than specifically tagged biological documents. Concept extraction in our approach is based on the lexico-syntactic characteristic of tokens appearing in a corpus rather than the pre-defined semantic of specific biological tags.

3 Text Mining for Fuzzy Ontology Discovery

It is believed that the main challenge in mining taxonomy relations from textual databases is to filter out the noisy relations[10, 12]. Accordingly, our text mining method is specifically designed to deal with such an issue. Standard document pre-processing such as stop word removal, POS tagging, and word stemming are applied [17]. Then, a *windowing process* is conducted over the collection of documents. The windowing process can help reduce the number of noisy term relationships. For each document (e.g., Net news, Web page, email, etc.), a *virtual window* of δ words is moved from left to right one word at a time until the end of a textual unit (e.g., a sentence) is reached. Within each window, the statistical information among tokens is collected to develop collocational expressions. Such a windowing process has successfully been applied to text mining before [7]. The windowing process is repeated for each document until the entire collection has been processed. According to previous studies, a text window of 5 to 10 terms is effective [5, 15], and so we adopt this range as the basis to perform our windowing process. To improve computational efficiency and filter noisy relations, only the specific linguistic pattern (e.g., Noun Noun, and Adjective Noun) defined by an ontology engineer will be analyzed. If a word has an association weight lower than a pre-defined threshold value, it will be discarded from the context vector of the concept. This is equivalent to the α-cut operation for fuzzy sets.

For statistical token analysis, several information theoretic methods are employed. Mutual Information has been applied to collocational analysis [15, 19] in previous research. Mutual Information is an information theoretic method to compute the dependency between two entities and is defined by [18]:

$$MI(t_i, t_j) = \log_2 \frac{Pr(t_i, t_j)}{Pr(t_i)Pr(t_j)} \tag{1}$$

where $MI(t_i, t_j)$ is the mutual information between term t_i and term t_j. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(t_i)$ is the probability that a term t_i appears in a text window. The probability $Pr(t_i)$ is estimated based on $\frac{|w_t|}{|w|}$ where $|w_t|$ is the number of windows containing the term t and $|w|$ is the total number of windows constructed from a textual database (i.e., a collection). Similarly, $Pr(t_i, t_j)$ is the fraction of the number of windows containing both terms out of the total number of windows.

We develop *Balanced Mutual Information* (BMI) to compute the degree of association among tokens. This method considers both term presence and term absence as the evidence of the implicit term relationships.

$$\begin{aligned}\mu_{c_i}(t_j) &\approx BMI(t_i, t_j) \\ &= \beta(Pr(t_i, t_j)\log_2(\tfrac{Pr(t_i,t_j)}{Pr(t_i)Pr(t_j)}) + \\ &\quad Pr(\neg t_i, \neg t_j)\log_2(\tfrac{Pr(\neg t_i,\neg t_j)}{Pr(\neg t_i)Pr(\neg t_j)})) - \\ &\quad (1-\beta)(Pr(t_i, \neg t_j)\log_2(\tfrac{Pr(t_i,\neg t_j)}{Pr(t_i)Pr(\neg t_j)}) + \\ &\quad Pr(\neg t_i, t_j)\log_2(\tfrac{Pr(\neg t_i,t_j)}{Pr(\neg t_i)Pr(t_j)}))\end{aligned} \tag{2}$$

where $\mu_{c_i}(t_j)$ is the membership function to estimate the degree of a term $t_j \in X$ belonging to a concept $c_i \in C$. $\mu_{c_i}(t_j)$ is the computational mechanism for the relation R_{XC} defined in the fuzzy ontology $Ont =< X, C, R_{XC}, R_{CC} >$. The membership function $\mu_{c_i}(t_j)$ is indeed approximated by the BMI score. $Pr(t_i, t_j)$ is the joint probability that both terms appear in a text window, and $Pr(\neg t_i, \neg t_j)$ is the joint probability that both terms are absent in a text window. The weight factor $\beta > 0.5$ is used to control the relative importance of two kinds of evidence (positive and negative). In Eq.(2), each MI value is then normalized by the corresponding joint probabilities. For the special case where $Pr(t_i, t_j) = 1$ is true, the joint probability value is replaced by a large positive integer because terms t_i, t_j have the strongest association. An α-cut is applied to discard terms from the potential concept if their membership values are below the threshold α. After computing all the BMI values in a collection, these values are subject to linear scaling such that each membership value is within the unit interval $\forall_{c_i \in C, t_j \in X} \mu_{c_i}(t_j) \in [0, 1]$. It should be noted that the constituent terms of a concept are always belonging to the concept with the maximal membership 1. Other measures that can be used to estimate the membership values of $t_j \in c_i$ include Jaccard (JA), conditional probability (CP), Kullback-Leibler divergence (KL), and Expected Cross Entropy (ECH) [6]:

$$\begin{aligned}\mu_{c_i}(t_j) &\approx Jacc(c_i, t_j) \\ &= \tfrac{Pr(c_i \wedge t_j)}{Pr(c_i \vee t_j)}\end{aligned} \tag{3}$$

Algorithm FuzzyOntoMine(*D, Para, Ont*)
Input: corpus D and vector of threshold values $Para$
Output: a fuzzy domain ontology Ont
Main Procedure:

1. $Ont = \{\}$
2. Foreach document $d \in D$ Do
   (a) Construct text windows $w \in d$
   (b) Remove stop words sw from w
   (c) Perform POS tagging for each term $t_i \in w$
   (d) Apply Porter stemming to each term t_i
   (e) Accumulate the frequency for $t_i \in w$ and the joint frequency for any pair $t_i, t_j \in w$
   (f) IF $lower \leq Feq(t_i) \leq upper$, $X = X \cup t_i$
3. End for
4. Foreach term $t_i \in X$ Do
   (a) compute its context vector c_i using BMI, MI, JA, CP, KL, or ECH
   (b) $C = C \cup c_i$
5. End for
6. Foreach $c_i \in C$ Do /* Concept Pruning - α-cut */
   (a) IF $\forall t_i \in c_i : \mu_{c_i}(t_i) < \alpha$
   (b) THEN $C = C - c_i$
7. End for
8. Foreach pair of concepts $c_i, c_j \in C$ Do
   (a) Compute the taxonomy relation $R(c_i, c_j)$ using $Spec(c_i, c_j)$
   (b) IF $\mu_{C\times C}(c_i, c_j) > \lambda$, $R = R \cup R(c_i, c_j)$
9. End For
10. Foreach $R(c_i, c_j) \in R$ Do /* Taxonomy Pruning */
    (a) IF $\mu_{C\times C}(c_i, c_j) < \mu_{C\times C}(c_j, c_i)$
    (b) THEN $R = R - R(c_i, c_j)$
    (c) IF $\exists P(c_i \rightarrow c_x, \ldots, c_y \rightarrow c_j)$
    (d) AND $\mu_{C\times C}(c_i, c_j) \leq \min(\{\mu_{C\times C}(c_i, c_x), \mu_{C\times C}(c_x, c_y), \ldots, \mu_{C\times C}(c_y, c_j)\})$
    (e) THEN $R = R - R(c_i, c_j)$
11. End For
12. Output Ont

Figure 1. The Fuzzy Domain Ontology Discovery Algorithm

$$\begin{aligned} \mu_{c_i}(t_j) &\approx Pr(c_i|t_j) \\ &= \frac{Pr(c_i,t_j)}{Pr(t_j)} \end{aligned} \quad (4)$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx KL(c_i||t_j) \\ &= \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \quad (5)$$

$$\begin{aligned} \mu_{c_i}(t_j) &\approx ECH(t_j, c_i) \\ &= Pr(t_j) \sum_{c_i \in C} Pr(c_i|t_j) \log_2 \frac{Pr(c_i|t_j)}{Pr(c_i)} \end{aligned} \quad (6)$$

To further filter the noisy concept relations, only the relatively prominent concepts for a domain will be further explored. We adopt the TFIDF [17] like heuristic to filter non-relevant domain concepts. Similar approach has also been used in ontology learning [14]. For example, if a concept is significant for a particular domain, it will appear more frequently in that domain when compared with its appearance in other domains. The following measure is used to compute the relevance score of a concept:

$$Rel(c_i, D_j) = \frac{Dom(c_i, D_j)}{\sum_{k=1}^{n} Dom(c, D_k)} \quad (7)$$

where $Rel(c_i, D_j)$ is the relevance score of a concept c_i in the domain D_j. The term $Dom(c_i, D_j)$ is the domain frequency of the concept c_i (i.e., number of documents containing the concept divided by the total number of documents in the corpus). The higher the value of $Rel(c_i, D_j)$, the more relevant the concept is for domain D_j. Based on empirical testing, we can estimate a threshold rel for a particular domain. Only the concepts with relevance score greater than the threshold will be selected. For each selected concept, its context vector will be expanded based on the synonymy relation defined in WordNet [13]. This is in fact a *smoothing* procedure [2]. The intuition is that some words that belong to a particular concept may not co-occur with the concept in a corpus. To make our ontology discovery method more robust, we need to consider these missing associations. For instance, our example context vector for "chief executive" will be expanded with the feature "presidency" based on the synonymy relation of WordNet, and a default membership value will be applied to such a term.

The final stage towards our ontology discovery method is fuzzy taxonomy generation based on subsumption relations among extracted concepts. Let $Spec(c_x, c_y)$ denotes that concept c_x is a specialization (sub-class) of another concept c_y. The degree of such a specialization is derived by:

$$\begin{aligned} \mu_{C\times C}(c_x, c_y) &\approx Spec(c_x, c_y) \\ &= \frac{\sum_{t_x \in c_x, t_y \in c_y, t_x = t_y} \mu_{c_x}(t_x) \otimes \mu_{c_y}(t_y)}{\sum_{t_x \in c_x} \mu_{c_x}(t_x)} \end{aligned} \quad (8)$$

where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (specificity) of c_x to c_y is based

on the ratio of the sum of the minimal membership values of the common terms belonging to the two concepts to the sum of the membership values of terms in the concept c_x. For instance, if every object of c_x is also an object of c_y, a high specificity value will be derived. The $Spec(c_x, c_y)$ function takes its values from the unit interval $[0, 1]$ and the subsumption relation is asymmetric. When the taxonomy is built, we only select the subsumption relations such that $Spec(c_x, c_y) > Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ where λ is a threshold to distinguish significant subsumption relations. The parameter λ is estimated based on empirical tests. If $Spec(c_x, c_y) = Spec(c_y, c_x)$ and $Spec(c_x, c_y) > \lambda$ is established, the *equivalent* relation between c_x and c_y will be extracted. In addition, a pruning step is introduced such that the redundant taxonomy relations are removed. If the membership of a relation $\mu_{C\times C}(c_1, c_2) \leq \min(\{\mu_{C\times C}(c_1, c_i), \ldots, \mu_{C\times C}(c_i, c_2)\})$, where $c_1, c_i, \ldots, c_2$ form a path P from c_1 to c_2, the relation $R_(c_1, c_2)$ is removed because it can be derived from other stronger taxonomy relations in the ontology. The fuzzy domain ontology mining algorithm is summarized and shown in Figure 1.

4 Evaluation

Since one of the most important applications of domain ontology is for intelligent information retrieval, our context-sensitive fuzzy ontology mining method is evaluated within the context of information retrieval. Our first experiment is similar to the routing tasks used in the Text REtrieval Conference (TREC) (`http://trec.nist.gov/`) which is a well-known international benchmark forum for information retrieval systems. The Reuters-21578 standard corpus with the Lewis-Split subset which contains 19,813 documents is used in our experiments. The training set consists of 13,625 documents and the test set consists of 6,188 documents. Our fuzzy domain ontology is automatically constructed based on the training set only. It takes 19 minutes only to complete the ontology mining process on a Pentium-4 2.2GHz PC. In this experiment, a window size of 5, a term size of 1, a single Noun pattern, and the (BMI) computational method with $\beta = 0.7$ are used.

For our ontology extraction method, a concept's relevance score defined in Eq. 7 is computed with respect to a variety of domains. Therefore, several other corpora are constructed based on the Web documents retrieved under different Yahoo categories such as "computer", "entertainment", "education" etc. For the Reuters-21578 corpus, a set of queries are composed based on the pre-defined Reuters topics and the top five (weighted by TFIDF) terms from one relevant document of the training set. For each Reuters subject code such as "acq", the corresponding subject description such as "acquisitions or mergers" is retrieved from the Reuters-21578 category description file. Each query is then applied to the testing set and the documents are ranked with respect to their relevance to the query. The vector-space model [16] is employed in this routing task. The routing tasks are performed with (the experimental group) and without (the control group) the help of our automatically constructed fuzzy domain ontology. Basically, the domain ontology is used for query expansion [22] for the routing task. For instance, each term in the original query is expanded with respect to the domain ontology to obtain a equivalent, a broader, or a more specific term. Standard performance measures [17] such as precision, recall, and F-measure are then computed based on the top 100 documents retrieved in both groups.

The $F_{\eta=1}$ measure and the recall results of 15 randomly selected Reuters topics are depicted in Table 1. The first column in Table 1 shows the topic names of the Reuters-21578 collection; the second column shows the number of true relevant documents for each topic. The remaining two columns are the $F_{\eta=1}$ and the recall results achieved when domain ontology is applied to expand initial query. The last two columns show the $F_{\eta=1}$ and the recall figures when domain ontology is not used for query expansion. Except for the topic of "coffee", the IR performance is improved with the help of the fuzzy domain ontology for query expansion. The reason why there is no improvement for the "coffee" topic is that the automatically generated domain ontology does not provide additional knowledge to expand the initial query. The difference of IR performance (both F-measure and Recall) between these two groups is statistically significant ($p < 0.01$) according to a paired one tail t-test. The average improvement of the $F_{\eta=1}$ measure is 58.3%. Therefore, we can conclude that the automatically discovered fuzzy domain ontology is with good quality and it is useful for enhancing information retrieval performance.

In our second experiment, various information theoretic measures are tested for the purpose of extracting domain concepts from a corpus. The same routing task is conducted except the use of different computational methods such as BMI, MI, JA, CP, and KL to estimate the membership of a term for a concept. The topic "carcass" is used to illustrate the typical performance of these methods. The precision-recall graph of these runs is plotted in Figure 2. The x axis indicates the various recall levels and the y axis shows the precision values obtained at the corresponding recall level. For example, the recall level 0.1 indicates the Nth position where 7 relevant documents (there are 68 relevant records for this topic) are found from the ranked list, and the corresponding precision values indicate the retrieval effectiveness of various methods (e.g., the best precision 0.36 is achieved by BMI). In general, the higher the precision curve, the better performance the information retrieval system is. As can be seen, the BMI method leads to the best

		With Ontology		No Ontology	
Topic	$\|Rel\|$	$F_{\eta=1}$	Recall	$F_{\eta=1}$	Recall
acq	2366	0.026	0.014	0.018	0.009
trade	426	0.076	0.047	0.057	0.035
livestock	61	0.273	0.361	0.180	0.295
bop	72	0.209	0.250	0.105	0.125
carcass	68	0.286	0.353	0.155	0.191
cocoa	73	0.254	0.301	0.139	0.164
coconut	6	0.075	0.667	0.019	0.167
coffee	139	0.268	0.230	0.268	0.230
copper	65	0.364	0.462	0.182	0.231
corn	237	0.196	0.139	0.101	0.072
gas	39	0.187	0.333	0.129	0.231
cotton	39	0.230	0.410	0.173	0.308
cpi	93	0.218	0.226	0.114	0.118
lei	4	0.035	0.133	0.017	0.067
crude	478	0.125	0.075	0.083	0.050
Average		0.188	0.267	0.119	0.153

Table 1. Comparative IR Performance with/without Fuzzy Domain Ontology

performance because it can take into account both positive and negative term co-occurrences. It implies that the BMI method leads to the generation of a higher quality fuzzy ontology. The ECH method is the closest to the BMI method at the expense of extra computational cost. The precision curve at the bottom of Figure 2 shows the worst retrieval performance when no fuzzy domain ontology is applied to refine the original query.

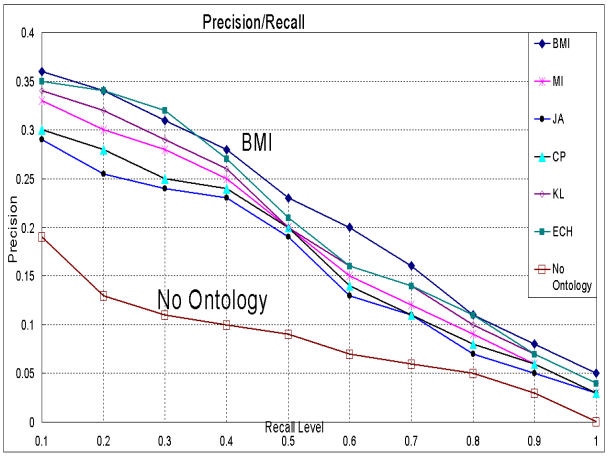

Figure 2. Comparative Performance of Various Computational Methods

5 Conclusions

Domain ontology plays an important role in many fields such as intelligent information retrieval, knowledge management, and the semantic Web. However, it is a very labor intensive and time consuming process for a purely manual construction of domain ontologies. In addition, as uncertainty often presents in real-world applications, it is less likely that domain ontologies with crisp concepts and relations can satisfy these applications. This paper proposes a novel fuzzy domain ontology discovery algorithm to facilitate the ontology engineering process. In particular, contextual information of a domain is exploited so that higher quality fuzzy domain ontologies can be automatically constructed. Our preliminary experiments show that the automatically generated fuzzy domain ontology can significantly improve the performance of information retrieval. Future work involves comparing the accuracy and the computational efficiency of our fuzzy ontology mining method with that of the other approaches. In addition, larger scale of quantitative evaluation of our fuzzy ontology mining method will be conducted.

References

[1] Muhammad Abulaish and Lipika Dey. Biological ontology enhancement with fuzzy relations: A text-mining framework. In Andrzej Skowron, Rakesh Agrawal, Michael Luck, Takahira Yamaguchi, Pierre Morizet-Mahoudeaux, Jiming Liu, and Ning Zhong, editors, *Proceedings of the 2005 IEEE / WIC / ACM International Conference on Web Intelligence (WI 2005)*, pages 379–385, Compiegne, France, September 19–22 2005. IEEE Computer Society.

[2] P. Cimiano, A. Hotho, and S. Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence Research*, 24:305–339, 2005.

[3] The World Wide Web Consortium. Web Ontology Language, 2004. Available from `http://www.w3.org/2004/OWL/`.

[4] T. R. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220, 1993.

[5] Hongyan Jing and Evelyne Tzoukermann. Information retrieval based on context distance and morphology. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Language Analysis, pages 90–96, 1999.

[6] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In Douglas H. Fisher, editor, *Proceedings of ICML-97, 14th International Conference on Machine Learning*, pages 170–178, Nashville, Tennessee, 1997. Morgan Kaufmann Publishers, San Francisco, California.

[7] R.Y.K. Lau. Context-Sensitive Text Mining and Belief Revision for Intelligent Information Retrieval on the Web. *Web Intelligence and Agent Systems An International Journal*, 1(3-4):1–22, 2003.

[8] Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang. A fuzzy ontology and its application to news summarization. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 35(5):859–880, 2005.

[9] Yuefeng Li and Ning Zhong. Mining ontology for automatically acquiring web user information needs. *IEEE Transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[10] A. Maedche, V. Pekar, and S. Staab. Ontology learning part onełon discovering taxonomic relations from the web. In N. Zhong, J. Liu, and Y. Yao, editors, *Web Intelligence*, pages 3–24. Springer, 2003.

[11] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, 2001.

[12] Alexander Maedche and Steffen Staab. Ontology learning. In *Handbook on Ontologies*, pages 173–190. 2004.

[13] G. A. Miller, Beckwith R., C. Fellbaum, D. Gross, and K. J. Miller. Introduction to wordnet: An on-line lexical database. *Journal of Lexicography*, 3(4):234–244, 1990.

[14] Roberto Navigli, Paola Velardi, and Aldo Gangemi. Ontology learning and its application to automated terminology translation. *IEEE Intelligent Systems*, 18(1):22–31, 2003.

[15] Patrick Perrin and Frederick Petry. Extraction and representation of contextual information for knowledge discovery in texts. *Information Sciences*, 151:125–152, 2003.

[16] G. Salton. Full text information processing using the smart system. *Database Engineering Bulletin*, 13(1):2–9, March 1990.

[17] G. Salton and M.J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, New York, 1983.

[18] C. Shannon. A mathematical theory of communication. *Bell System Technology Journal*, 27:379–423, 1948.

[19] Mark A. Stairmand. Textual context analysis for information retrieval. In *Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 140–147, 1997.

[20] Quan Thanh Tho, Siu Cheung Hui, Alvis Cheuk M. Fong, and Tru Hoang Cao. Automatic fuzzy ontology generation for semantic web. *IEEE Transactions on Knowledge and Data Engineering*, 18(6):842–856, 2006.

[21] Rudolf Wille. Formal concept analysis as mathematical theory of concepts and concept hierarchies. In Bernhard Ganter, Gerd Stumme, and Rudolf Wille, editors, *Formal Concept Analysis, Foundations and Applications*, volume 3626, pages 1–33. Springer, 2005.

[22] J. Xu and W.B. Croft. Query expansion using local and global document analysis. In Hans-Peter Frei, Donna Harman, Peter Schäuble, and Ross Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4–11, Zurich, Switzerland, 1996.

[23] L. A. Zadeh. Fuzzy sets. *Journal of Information and Control*, 8:338–353, 1965.