Fuzzy Domain Ontology Discovery for Business Knowledge Management

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

Abstract

Ontology plays an essential role in the formalization of business information (e.g., products, services, relationships of businesses) for effective human-computer interactions. However, engineering of domain ontologies turns out to be very labor intensive and time consuming. Recently, some machine learning methods have been proposed for automatic discovery of domain ontologies. Nevertheless, the accuracy and computational ef ciency of the existing methods need to be improved to support large scale ontology construction for real-world business applications. This paper illustrates a novel fuzzy domain ontology discovery algorithm for supporting real-world business ontology engineering. By combining lexico-syntactic and statistical learning methods, the accuracy and the computational ef ciency of the ontology discovery process is improved. Empirical studies have con rme d that the proposed method can discover high quality fuzzy domain ontology which leads to signi cant improvement in information retrieval performance.

Keywords: *Domain Ontology, Fuzzy Sets, Text Mining, Information Retrieval, Knowledge Management.*

1 Introduction

Knowledge has been recognized as the most important corporate asset and it is the key for organizations to achieve sustainable competitive advantage. Knowledge management is a collection of processes that govern the creation, dissemination, and utilization of knowledge [25, 26]. To be able to effectively manage the intellectual capital, businesses need an effective approach to identify and capture information and knowledge about business processes, products, services, markets, customers, suppliers, and competitors, and to share this knowledge to improve the organizations' goal achievement. Ontologies allow domain knowledge such as products, services, markets, etc. to be captured in an explicit and formal way such that it can be shared among human and computer systems.

The notion of ontology is becoming very useful in various elds such as intelligent information extraction and retrieval, cooperative information systems, electronic commerce, and knowledge management [38]. Since Tim Berners-Lee, the inventor of the World Wide Web (Web), coined the vision of a Semantic Web [3], the proliferation of ontologies has been under tremendous growth. The success of Semantic Web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding [19]. Although there is not a universal consensus on the de niti on of ontology, it is generally accepted that ontology is a speci cat ion of conceptualization [9]. Ontology can take the simple form of a taxonomy (i.e., knowledge encoded in a minimal hierarchical structure) or as a vocabulary with standardized machine interpretable terminology supplemented with natural language de nitions. Ontology provides a number of potential bene ts 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. Ontology is often speci ed in a declarative form by using semantic markup languages such as RDF and OWL [6]. igure 1 shows an example of the domain ontology extracted from the Reuters RCV1 corpus [16] and Figure 2 depicts the corresponding OWL statements.

Domain ontologies specify the knowledge for a particular type of domain [7]. This kind of ontologies generalize over application tasks in such domains such as medical, tourism, banking, nance, etc. A well-known example is the Uni ed Medical Language System (UMLS) and its component parts such as the Medical Subject Heading (MeSH). Although domain ontologies are useful in many

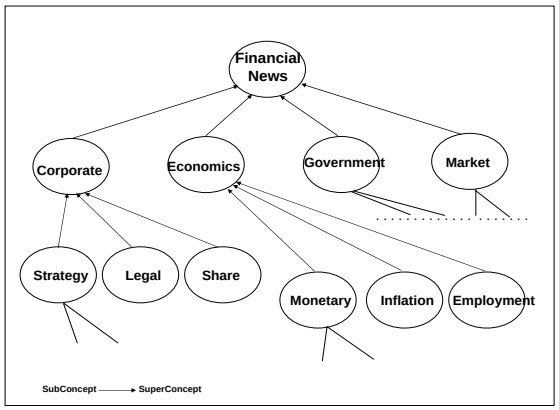


Figure 1. A Crisp Domain Ontology from the RCV-1 Corpus

```
<?xml version="1.0"?>
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns\#"
xmlns:xsd="http://www.w3.org/2001/XMLSchema\#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema\#"
xmlns:owl="http://www.w3.org/2002/07/owl\#"
xmlns="http://www.owl-ontologies.com/unnamed.owl\#"
xml:base="http://www.owl-ontologies.com/unnamed.owl">
  <owl:Ontology rdf:about=""/>
  <owl:Class rdf:ID="FinancialNews"/>
  <owl:Class rdf:ID="Corporate">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Economics">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Government">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Market">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#FinancialNews"/>
    </rdfs:subClassOf>
  </owl:Class>
  <owl:Class rdf:ID="Strategy">
    <rdfs:subClassOf>
      <owl:Class rdf:ID="\#Corporate"/>
    </rdfs:subClassOf>
  </owl:Class>
  .....
</rdf:RDF>
```

Figure 2. OWL for the Financial News Ontology

areas, engineering of these ontologies turns out to be very labor intensive and time consuming. Therefore, many automatic or semi-automatic ontology engineering techniques have been proposed. Although fully automatic construction of perfect domain ontology is beyond the current state-of-the-art, we believe that the automatic ontology mining method illustrated in this paper can assist ontology engineers to build domain ontology quicker and more accurately.

Although some learning techniques have been applied to the extraction of domain ontology [4, 7, 31], these methods are still subject to further enhancement in terms of computational ef cienc y and accuracy. One of the ways to improve automated domain ontology discovery is to exploit contextual information from the knowledge sources. 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 [42]. Therefore, a fuzzy ontology rather than a crisp ontology is discovered by the proposed text mining method.

De nition 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 de ned 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.*

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

Figure 3 highlights the fuzzy domain ontology corresponding to the one depicted in Figure 2. The current OWL syntax can easily be extended to represent fuzzy domain ontology using approach similar to [8]. However, we will only focus on the mining of fuzzy concepts and fuzzy taxonomy relations in this paper. The term $\mu_{C\times C}(c_2, c_1)$ in Figure 3 denotes the membership value of the taxonomy relation from subclass c_2 to superclass c_1. 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.

De nition 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 (e.g., product description databases). By combining lexico-syntactic and statis-

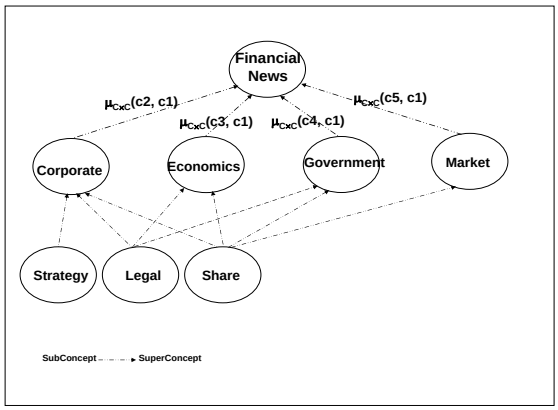


Figure 3. A Fuzzy Domain Ontology from the RCV-1 Corpus

tical learning approaches, the accuracy and the computational ef cienc y of the ontology discovery process is improved [20]. 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. Section 3 gives an overview of our text mining methodology. The cognitive and linguistic foundations of the proposed context-sensitive ontology discovery method is described in Section 4. The computational details of the proposed ontology mining method are then illustrated in Section 5. Section 6 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

With the increasing importance of product information management in eCommerce environment, it is vital that precise de nition of product and services readily available in sharable, manageable, e xible, and scalable form, that is in the form of an ontology. Although the idea of utilizing ontology for e-Catalogs has been proposed long ago, an operational product ontology system for a speci c domain is not yet available. Lee et. al. [15] developed an operational product ontology system called KOCIS for the government procurement service. It consists of the ontology construction and management sub-system to build the ontology database from the product databases and to manage the the real-time processing for update operations while maintaining a consistency of the ontology data. In addition, the ontology search sub-system can retrieves and navigates the product ontology information. The search subsystem addresses the problem of ranking keyword search results by modeling the product ontology as a Bayesian belief network. Although, the KOCIS system addresses the operational aspects of an ontology management and search system, it does not support automated or semi-automated discovery of product ontology from information sources. This paper focuses on the development of a fuzzy ontology discovery algorithm for automatic business knowledge management; the proposed method can be readily applied to discover product ontology for eCommerce.

Cimiano et al. have presented an automatic taxonomy learning algorithm to extract concept hierarchies from a text corpus [5]. In particular, their taxonomy learning method is based on formal concept analysis [40]. 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. Central to formal concept analysis is the notion of a context which is essentially the prominent attributes or features common to a set of objects of the same class. A formal context is a triple $K = (G, M, I)$ where G and M represent a set of objects and attributes respectively and I is a binary relation between G and M. Thereby, a formal concept (A, B) is de ned by $A = \{g \in G | \forall_{m \in M}(g, m) \in I\}$ and $B = \{m \in M | \forall_{g \in G}(g, m) \in I\}$. In order to derive attributes from a certain corpus, part-of-speech tagging and linguistic analysis are performed to extract verb/prepositional phrase complement, verb/object and verb/subject dependencies. For each noun appearing as head of the extracted syntactic structures, the corresponding verbs are taken as the attributes for building the formal context. Their approach is evaluated by comparing the automatically generated concept hierarchies with hand-crafted taxonomies in a tourism and a nance domain. 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 [5].

The FOGA framework for fuzzy ontology generation has been proposed [37]. 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 fuzzi ed 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 in-

stances. 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 ef cient 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 [37].

A fuzzy ontology which is an extension of the domain ontology with crisp concepts is utilized for news summarization purpose [14]. In this semi-automatic ontology discovery approach, the domain ontology with various events of news is pre-de ned by domain experts. A document preprocessing mechanism will generate the meaningful terms based on the news corpus and a Chinese news dictionary pre-de ned by the domain experts. The meaningful terms are classi ed according to the events of the news by a term classi er . Basically, every fuzzy concept has a set of membership degrees associated with the various events of the domain ontology. The main function of the fuzzy inference mechanism is to generate the membership degrees (classi- cation) for each event with respect to the fuzzy concepts de ned in the fuzzy ontology. The standard triangular membership function is used for the classi cation purpose. The method discussed in this paper is a fully automatic fuzzy domain ontology discovery approach. There is no pre-de ned 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 arti c ial 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 [17]. 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 re ne 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.

Personalized Abstract Search Services (PASS) is a domain speci c search engine providing abstracts of papers from IEEE Transactions sponsored by the IEEE Neural Network Council [39]. The system uses a fuzzy ontology of term associations to support semantic based information retrieval. The fuzzy ontology is automatically built using information obtained from the system's document collection. The system extracts a set of two or three consecutive words exhibiting some linguistic patterns such as "noun noun", "adj noun", etc. from a corpus. The system then eliminates the phrases that contain at least one stop word from a pre-de ned control le. The notions of narrower and broader term relations are introduced and a fuzzy conjunction operator is applied to compute the membership values of the term relations. By evaluating the users' searching activities, it was found that the fuzzy ontology of term relations significantly contributes to the information retrieval process. Our work presented in this paper differs from the PASS system in the fuzzy concepts (instead of terms) are rst identi ed and the taxonomy relations of concepts are then developed. In addition, our fuzzy ontology mining approach has been evaluated based on a bench-mark collection in the eld of information retrieval.

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 speci cally 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-de ned semantic of speci c biological tags.

A semiautomatic ontology engineering environment called OntoEdit has been developed [19, 20]. The workbench supports ontology import, extraction, pruning, re nement, and evaluation. Merging existing semantic structures or de ning mapping rules between these structures allows importing and reusing available ontologies. Ontology ex-

traction is one of the main tasks of ontology engineering, which deals with learning the appropriate ontologies from the domain sources. The initial ontology which results from import, reuse, and extraction, is then pruned to better t the purpose of the particular application. Traditional text processing techniques such as n-gram [30] is used to extend the set of lexical entries L based on source documents. Hierarchical clustering is applied to learn the taxonomy relations H_C. In addition, morphological analysis and generalized association rule mining are applied to learn the relations R among some concepts C. Our work presented in this paper focuses on the ontology extraction stage of the ontology engineering cycle. Moreover, a subsumption-based computational method rather than the traditional clustering method is used for the extraction of concept lattice.

3 An Overview of the Text Mining Methodology

Figure 4 depicts the proposed text mining methodology for the automatic discovery of fuzzy domain ontology from a textual database (corpus). A text corpus is parsed to analyze the lexico-syntactic elements. For instance, stop words such as "a, an, the" are removed from the source documents since these words appear in any contexts and they cannot provide useful information to describe a domain concept. For our implementation, a stop word le is constructed based on the standard stop word le used in the SMART retrieval system [29]. Lexical pattern is identi ed by applying Part-of-Speech (POS) tagging to the source documents and then followed by token stemming based on the Porter stemming algorithm [28]. We refer to the WordNet lexicon [21] to tag each word during this process. During the linguistic pattern ltering stage, certain linguistic patterns are extracted based on the speci c requirements speci ed by the ontology engineers. For example, the ontology engineers may only focus on the "Noun Noun" and "Adjective Noun" patterns instead of all the linguistic patterns. This is in fact a good way to gain computational ef cienc y by reducing the number of patterns for further statistical analysis. In addition, to extract relevant domain speci c concepts, the appearances of concepts across different domains should be taken into account. The basic intuition is that a concept frequently appears in a speci c domain (corpus) rather than many different domains is more likely to be a relevant domain concept. The statistical Token Analysis step employs the information theoretic measure to compute the co-occurrence statistics of the targeting linguistic patterns. Finally, taxonomy of domain concepts is developed according to the fuzzy conjunction operator. The details of the proposed ontology mining method will be discussed in Section 5.

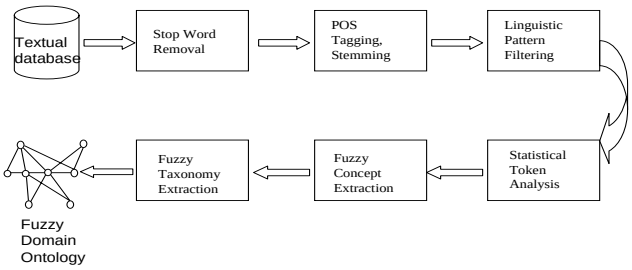


Figure 4. Context-Sensitive Fuzzy Ontology Discovery Process

4 The Linguistic Foundations

The proposed context-sensitive fuzzy ontology discovery method is based on the *distributional hypothesis* which assumes that terms (concepts) are similar according to the extent that they share similar linguistic contexts [10]. In particular, we borrow the notion of *collocational expressions* from computational linguistic to identify the semantics of some lexical elements such as concepts from text corpora. For computational linguistic, a term refers to one or more tokens (words) and a term is also a concept if it carries recognizable meaning speci c to a domain [23]. Collocational expressions are groups of words related in meaning, and the constituent words of an expression are frequently found in a near loci of a few adjacent words in a textual unit [33, 35]. The collocational expressions are indeed providing the underlying context of a given concept embedded in natural language text such as Web documents.

Contextual information has long been recognized as one of the major contributors to concept learning in the eld of computer science [43]. Nevertheless, to automatically detect the semantics (meanings) of a concept is not a trivial task since the meanings of a concept is context (domain) dependent. For example, the concept "bank" can refer to a nancial institute such as a "commercial bank, or refer to the raised shelf of ground such as the "river bank". Therefore, to accurately extract domain ontologies from text, contextual information must be exploited to disambiguate different senses. In this regard, *static* lexicons (i.e., generic linguistic ontologies) such as WordNet [21] with meanings (senses) computed *a priori* may not be able to capture the speci c semantics of concepts pertaining to a particular application domain. However, WordNet can be used to boot-

strap the performance of information extraction when domain ontologies are built [22, 24]. Our general approach is that the collocational expressions are rst extracted from the source documents; these collocational expressions which carry context-sensitive semantics are then used to de ne the meanings of the concepts.

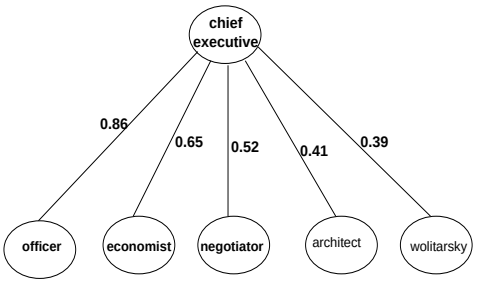


Figure 5. Domain Specific Semantics of the Concept "Chief Executive"

In the eld of information retrieval (IR), the notion of *context vectors* [11, 32] has been proposed to give computer-based representations of concepts. In this approach, a concept is represented by a vector of words and their numerical weights. The weight of a word indicates the extent to which the particular word is *associated* with the underlying concept. For example, the concept "chief executive" is represented by the words such as of cer , negotiator, economist, etc. as depicted in Figure 5, which is an interesting example by parsing the Reuters-21578 corpus (http://www.daviddlewis.com/resources/testcollections/). The context vector of "chief executive" is shown as follows:

Concept: chief executive
Context Vector:
$\{(officer, 0.72), (economist, 0.65), (negotiator, 0.63), (architect, 0.61), (wolitarsky, 0.44)\}$

The context vector can be seen as a point in a multi-dimensional geometric information space with each dimension representing a property term. It should be noted that the meanings (senses) of "chief executive" is "head of state" or "presidency" as de ned in WordNet [21], which is quite different from that discovered by our context-sensitive text mining method. The last term in the example context vector is "wolitarsky" which is the name of the chief executive of a nancial institution often mentioned in the Reuters nancial news in that period. So, our method can really discover domain speci c relation such as "wolitarsky" is a chief executive. Static lexicons such as WordNet can only capture the lexical knowledge of a concept, but fails to represent domain speci c non-lexical knowledge. A linguistic concept such as "chief executive" can be taken as a class (set) with respect to the fuzzy set framework. A term such as "wolitarsky" will then be treated as an object which belongs to the set with certain degree.

5 Text Mining for Fuzzy Ontology Discovery

It is believed that the main challenge in mining taxonomy relations from textual databases is to lter out the noisy relations[18, 20]. Accordingly, our text mining method is speci cally designed to deal with such an issue. After standard document pre-processing such as stop word removal, POS tagging, and word stemming [30], 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 [13]. 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 [11, 27], and so we adopt this range as the basis to perform our windowing process. To improve computational ef cienc y and lte r noisy relations, only the speci c linguistic pattern (e.g., Noun Noun, and Adjective Noun) de ned by an ontology engineer will be analyzed. The following is an example segment of a news article in the Reuters-21578 collection:

```
<REUTERS OLDID="5545" NEWID="2"><TEXT>
<TITLE>STANDARD OIL TO FORM FINANCIAL
UNIT</TITLE>
<BODY>Standard Oil Co and BP North
America Inc said they plan to form
a venture to manage the money market
borrowing and investment activities
of both companies.
</BODY></TEXT> </REUTERS>
```

After parsing the main body of the news article, our ontology extraction program will remove the stop words, apply POS tagging and stem the words. So, the result will look like:

```
standard (Adj) oil (N) co (N)
```

```
bp (N) north (Adj) america (N)
inc (N) said (V) plan (V) form (V)
venture (N) manage (V) money (N)
market (N) borrow (V) investment (N)
activit (N) compan (N) .
```

Assuming that the window size of 5 is used and the ontology engineer speci es the "Noun Noun" linguistic pattern as the only focus, the potential concepts "Oil Co" and "Co BP" will be extracted from the rst virtual text window. The concept "Oil Co" might be represented by the features such as "standard", "bp", and "north". After parsing the whole corpus, the statistical data (by statistical token analysis) about the potential concepts can be collected. If a word has an association weight lower than a pre-de ned 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 [27, 36] in previous research. Mutual Information is an information theoretic method to compute the dependency between two entities and is de ned by [34]:

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

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.

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

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} de ned in the fuzzy ontology $Ont =< X, C, R_{XC}, R_{CC} >$. The membership function $\mu_{c_i}(t_j)$ is

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 6. The Fuzzy Domain Ontology Discovery Algorithm

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) [12]:

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

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

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

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

To further lter the noisy concept relations, only the relatively prominent concepts for a domain will be further explored. We adopt the TFIDF [30] like heuristic to lter non-relevant domain concepts. Similar approach has also been used in ontology learning [24]. For example, if a concept is signi cant 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:

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

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 de ned in WordNet [21]. This is in fact a *smoothing* procedure [5]. 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 na l 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:

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

where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (speci city) 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 speci city 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 signi can t 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 6.

6 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 rst 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 de ned 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-de ned Reuters topics and the top ve (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 le. 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 [29] is employed in this routing task. For instance, the standard TFIDF term weighting scheme is used to compute the term weights of a document and a query respectively, and the cosine similarity measure is used to rank each document:

$$sim(\overrightarrow{q}, \overrightarrow{d}) = \frac{\sum_{i=1}^{n} w_q(k_i) \times w_d(k_i)}{\sqrt{\sum_{i=1}^{n}(w_q(k_i))^2} \times \sqrt{\sum_{i=1}^{n}(w_d(k_i))^2}} \quad (9)$$

where $\overrightarrow{q}$ and $\overrightarrow{d}$ are the query vector and the document vector respectively. The term $w_q(k_i)$ represents the weight of the ith keyword k_i in the query vector $\overrightarrow{q}$, and the term $w_d(k_i)$ represents the weight of the ith keyword k_i in the document vector $\overrightarrow{d}$.

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 [41] 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 speci c term. In this experiment, the type of relations is selected manually from the fuzzy domain ontology to optimize the retrieval effectiveness. Standard performance measures [30] such as precision, recall, and F-measure are then computed based on the top 100 documents retrieved in both groups:

$$Precision = \frac{a}{a+b} \quad (10)$$

$$Recall = \frac{a}{a+c} \quad (11)$$

$$F_\eta = \frac{(1+\eta^2) Precision \times Recall}{\eta^2 Precision + Recall} \quad (12)$$

where a, b, c represent the number of retrieved relevant documents, the number of retrieved non-relevant documents, and the number of not retrieved relevant documents respectively. The $F_{\eta=1}$ measure and the recall results of 15 randomly selected Reuters topics are depicted in Table 1. The rst 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 gures 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 signi can t ($p < 0.01$) 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 7. 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 performance because it can take into account both positive

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) [12]:

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

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

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

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

To further lter the noisy concept relations, only the relatively prominent concepts for a domain will be further explored. We adopt the TFIDF [30] like heuristic to lter non-relevant domain concepts. Similar approach has also been used in ontology learning [24]. For example, if a concept is signi cant 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:

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

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 de ned in WordNet [21]. This is in fact a *smoothing* procedure [5]. 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.

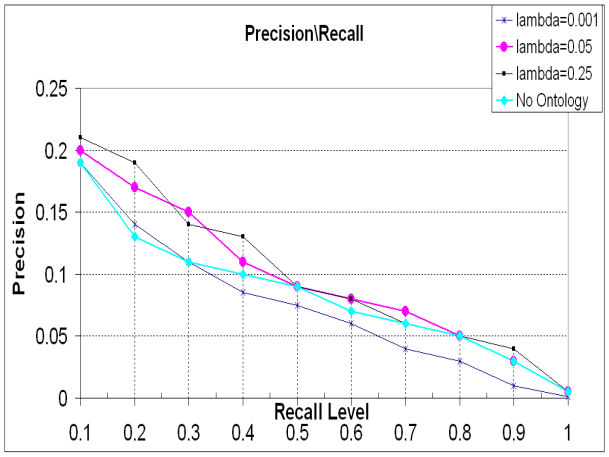
The na l stage towards our ontology discovery method is fuzzy taxonomy generation based on subsumption relations among extracted concepts. Let $Spec(c_x, c_y)$ denotes that concept c_x is a specialization (sub-class) of another concept c_y. The degree of such a specialization is derived by:

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

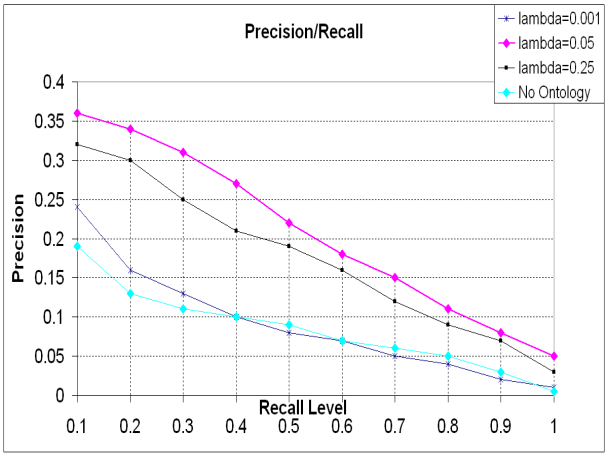
where $\otimes$ is a fuzzy conjunction operator which is equivalent to the min function. The above formula states that the degree of subsumption (speci city) of c_x to c_y is based on the ratio of the sum of the minimal membership values of the common terms belonging to the two concepts to the sum of 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 speci city 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 signi can t 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 6.

6 Evaluation

Since one of the most important applications of domain ontology is for intelligent information retrieval, our context-



a. Window Size 20 and windowing across sentences



b. Window Size 5 and windowing within sentences

Figure 8. The Impact of Windowing and Taxonomy Pruning

too many noisy taxonomy relations exist in the ontology which leads to poor query expansion. On the other hand, if the threshold λ is too high, many useful taxonomy relations are ltere d out such that the ontology is not useful for query re nement. When the threshold $\lambda = 0.001$ is used, a noisy ontology will be generated which leads to retrieval performance worse than the baseline where no ontology is used for query expansion. If an appropriate window size is employed and the windowing process is carried out within sentence boundary, a fuzzy domain ontology with higher quality is generated (as depicted in Figure 8.b).

7 Conclusions

The manipulation and exchange of semantically enriched business intelligence (e.g., products, services, markets, etc.) can enhance the quality of an eCommerce system and offer a high level of inter-operability among different enterprise systems. Ontology certainly plays an important role in the formalization of business knowledge. However, the biggest challenge for the wide spread applications of ontologies is on the construction of these ontologies because it is a very labor intensive and time consuming process. 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 illustrates 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. The proposed discovery method combines lexico-syntactic and statistical learning approaches so as to reduce the chance of generating noisy concepts and relations. Empirical studies have been performed to evaluate the quality of the fuzzy domain ontology discovered by the proposed ontology mining algorithm. Our preliminary results show that the automatically generated fuzzy domain ontology can signi cantly improve the effectiveness in information retrieval. Future work involves comparing the accuracy and the computational ef cienc y of our fuzzy ontology mining method with that of the other approaches. In addition, larger scale of quantitative evaluation of our fuzzy ontology mining algorithm in the context of business information management 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] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo, editors, *VLDB'94, Proceedings of 20th International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile, September 12–15 1994. Morgan Kaufmann Publishers.

[3] T. Berners-Lee, J. Hendler, and O. Lassila. The semantic web. *Scienti c American*, 284(5):34–43, 2001.

[4] Shan Chen, Damminda Alahakoon, and Maria Indrawan. Background knowledge driven ontology discovery. In *Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service*, pages 202–207, 2005.

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

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

[7] Michael Dittenbach, Helmut Berger, and Dieter Merkl. Improving domain ontologies by mining semantics from text. In *Proceedings of the First Asia-Paci c Conference on Conceptual Modelling (APCCM2004)*, pages 91–100, 2004.

[8] Mingxia Gao and Chunnian Liu. Extending OWL by fuzzy description logic. In *Proceedings of the 17th IEEE International Conference on Tools with Arti cial Intelligence (ICTAI'05)*, pages 562–567. IEEE Computer Society, 2005.

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

[10] Z. Harris. *Mathematical Structures of Language*. Wiley, New York, 1968.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] Taehee Lee, Ig hoon Lee, Suekyung Lee, Sang goo Lee, Dongkyu Kim, Jonghoon Chun, Hyunja Lee, and Junho Shim. Building an operational product ontology system. *Electronic Commerce Research and Applications*, 5(1):16–28, 2006.

[16] D. Lewis, Y. Yang, T. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397, 2004.

[17] 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.

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

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

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

[21] 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.

[22] Michele Missikoff, Roberto Navigli, and Paola Velardi. Integrated approach to web ontology learning and engineering. *IEEE Computer*, 35(11):60–63, 2002.

[23] Christine A. Montgomery. Concept extraction. *American Journal of Computational Linguistics*, 8(2):70–73, 1982.

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

[25] I. Nonaka. A dynamic theory of organizational knowledge creation. *Organization Science*, 5(1):14–37, 1994.

[26] I. Nonaka and H. Takeuchi. *The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation*. Oxford University Press, New York, 1995.

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

[28] M. Porter. An algorithm for suf x stripping. *Program*, 14(3):130–137, 1980.

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

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

[31] M. Sanderson and B. Croft. Deriving concept hierarchies from text. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 206–213. ACM, 1999.

[32] Hinrich Schütze. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–124, 1998.

[33] Satoshi Sekine, Jeremy J. Carroll, So a Ananiadou, and Jun'ichi Tsujii. Automatic learning for semantic collocation. In *Proceedings of the third Conference on Applied Natural Language Processing*, pages 104–110, Trento, Italy, March 31–April 3 1992. Association for Computational Linguistics.

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

[35] G. Smith. *Computers and Human Language*. Oxford University Press, New York, New York, 1991.

[36] 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.

[37] 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.

[38] Christopher A. Welty. Ontology research. *AI Magazine*, 24(3):11–12, 2003.

[39] Dwi H. Widyantoro and John Yen. A fuzzy ontology-based abstract search engine and its user studies. In *The 10th IEEE International Conference on Fuzzy Systems*, pages 1291–1294. IEEE Press, 2001.

[40] 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.

[41] 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.

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

[43] Wlodek Zadrozny. Context and ontology in understanding of dialogs. In *Proceedings of the IJCAI'95 Workshop on Context in NLP*, May 15 1995.