ELSEVIER

Expert Systems with Applications 31 (2006) 1–12

Expert Systems with Applications

www.elsevier.com/locate/eswa

Ontology construction for information classification

Sung-Shun Weng [a,*], Hsine-Jen Tsai [a], Shang-Chia Liu [b], Cheng-Hsin Hsu [a]

[a] *Department of Information Management, Fu Jen Catholic University, 510 Chung-Cheng Road, Hsin-Chuang City, Taipei 242, Taiwan, ROC*
[b] *Graduate Institute of Business Administration, Fu Jen Catholic University, 510 Chung-Cheng Road, Hsin-Chuang City, Taipei 242, Taiwan, ROC*

Abstract

Following the advent of the Internet technology and the rapid growth of its applications, users have spent long periods of time browsing through the ocean of information found in the Internet. This time-consuming hunt, however, makes searching, retrieving, displaying, integrating and maintaining data such arduous tasks. One way to solve this problem is to study the concept behind the Semantic Web in accordance with the principles of ontology. Apart from facilitating the process of information search in the Semantic Web, ontology also provides a method that will enable computers to exchange, search and identify text information. But establishing the ontology necessitates a great deal of expert assistance; manually setting it up would entail a lot of time, not to mention that there are only a handful of experts available. For this reason, using automatic technology to construct the ontology is a subject worth pursuing. This research uses the theory of formal concept analysis to serve as the groundwork in assembling the different levels of ontological concepts in an automated fashion. An ontology diagram will be presented to show the correlation of concepts and their corresponding significance. Moreover, the experiments of this research select a collection of different concepts in an attempt to classify the relationships between documents and concepts. The objective is to develop an automated technology of ontology construction that will support the present information classification system, as well as to upgrade the ontological aspect of the Semantic Web.

Keywords: Ontology; Semantic Web; Information classification; Formal concept analysis

1. Introduction

Following the rapid widespread of computer networks, data handling had slowly grown dependent on computer networks and servers particularly in carrying out automated exchange tasks. In other words, data storage, data management, data transmission and even analysis rely on computer and network technology. At the same time, as a result of the vigorous developments and accessibility of the World Wide Web (WWW), a great quantity of information was suddenly made available to people. However, due to the enormousness of the data, users waste a lot of time browsing the Internet and searching for the information they need; it makes the tasks of searching, accessing, displaying, integrating and maintaining data more laborious. With the aim of solving this difficulty, Berners-Lee and Fischetti (1999) conceived the concept of the Semantic Web. Based on this concept, ontology and intelligent agents constitute the foundations of the Semantic Web. Ontology is made capable to 'describe metadata' in order to build one complete glossary that will clearly define the data found in the World Wide Web (WWW). The sharing and reusing of ontology enable users to communicate with computers by using accurate syntax and semantics. The powerful capabilities of the Semantic Web and its full development depend on the proper knowledge and handling of the agents. Aside from being able to fathom the user's demands, intelligent agents must also be able to find similar or related resources from the Semantic Web as constructed by ontology, and furthermore, to build trust in the whole architecture of the Semantic Web. This can be achieved through the mutual collaboration of different agents to test, verify, and assure the reliability of information.

Semantic Web is anything but a new kind of network; it is built within the existing network environment and provides a highly readable data without modifying or altering any of the contents. Simply put, a Semantic Web is an integration of numerous metadata. These data serve to describe documents, web pages and general concepts. It may be impossible for computers to understand the context of any document but with the help of the Semantic Web, exchanging, searching and recognizing the meaning of characters become possible.

Therefore, the development and success of the Semantic Web will greatly depend on how fast and efficiently ontology is established. New websites may possibly consider creating its blueprint for ontology during its early developmental phase but

[*] Corresponding author. Tel.: +886 2 29052717; fax: +886 2 29052812.
E-mail address: im1032@mails.fju.edu.tw (S.-S. Weng).

0957-4174/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2005.09.007

because data or information can quickly become obsolete, the integration between the ontology and the website can be expected to gradually diminish. When this occurs, there is a need to build a whole new ontology. Furthermore, data found in current websites have completely changed or evolved other than those in the time when the Internet was just beginning. For this reason, the automatic construction of ontology as a stimulus to promote great advancement of the Semantic Web is a topic worthy of an in-depth analysis.

These days, the study of ontology in the Semantic Web suggests the setting up of ontology that will represent Web data during the preliminary stages of constructing the website. The current information classification systems are likewise dependent on the classification framework defined by the experts. Since, these measures may encounter difficulties with the current information classification systems, this research aims to provide new theories that will help solve the above-mentioned problems. We believe that there are several difficulties in manually defining a classification framework. Overcoming these challenges would be most beneficial to the existing information classification systems and ontology in the Semantic Web. The following are brief explanations of these problems:

1.1. Lack of flexibility in the classification framework

Whether it is the current information classification systems or ontology of the Semantic Web, modifying the existing classification framework is very difficult, if not impossible. Whether the website appears as a database or in a format of table of contents, it is imperative to adhere to the pre-defined classification framework in categorizing information. But with the rapid addition to or modifications in the data, it becomes clear that there is an apparent lack of flexibility in manually defining the classification framework. If we can base the classification framework on the data content and automatically generate the corresponding classification, then the tremendous inflow and/or modifications in the data will no longer be a problem. Although the current ontological construction technology can achieve a partially automated classification framework, still there pose several limitations. It is therefore, one of the serious aims of this research to make a significant breakthrough and achieve a fully automated classification framework.

1.2. Conceptual relationships and significance of resources

The classification relationships manifested in the existing classification framework is generally absolute. Its incapability to show the corresponding significance of identical classification levels will result in the omission of vital information when searching for a general concept. Therefore, identifying the significance of various classifications and resources in a proper manner can increase the accuracy rate of the search function.

On account of the above-mentioned grounds, this research intends to provide a automatic construction technology of ontology that will help solve the difficulties enumerated above. Solving these problems will not only upgrade the existing systems but will also support the construction of ontology in the Semantic Web.

2. Review of related literature

Up until the present times, extensive and general construction methods are not found within the domain of ontology learning. Although in other fields such as linguistics, information retrieval, machine learning, data mining and software engineering, etc. there are plenty of studies and related technologies that can apply the benefits of ontology learning.

Maedche and Staab (2001) mentioned that ontology learning can be divided into four parts: extract, prune, refine, import or reuse. We will, for now, direct our attention to the extraction methods upon which we will base our research findings. There are four categories in the construction methods of ontology learning, these are: dictionary-based, text clustering, association rules, and knowledge base. These categories are further explained below:

2.1. Dictionary-based construction method

Using the compilation concept of a traditional dictionary, the hierarchy of concepts is automatically formed. Traditional dictionaries present entries together with their synonyms, root words, etymology, etc. The definitions and relationships presented in the dictionary are used to determine the hierarchy relationships of concepts (Khan & Luo, 2002; Kietz, Maedche, & Volz, 2000; Tan, Han, & Elmasri, 2000).

The dictionary-based construction method normally is the groundwork of other construction methods. The other three methods are somehow related to the dictionary-based construction method either in the preliminary construction phase or in the final pruning and verification stage. This is so because the dictionary-based method has its own limitations and will only be effective when paired with another kind of method. It is never used independently. The limitations are listed as follows:

(1) The ontology formed using the dictionary-based method has a general description and is not at all domain specific. Only when it is combined with another method does it provide a more significant and valuable ontological framework.

(2) The dictionary-based method is generally restricted to the volume size of the dictionary and can thus form domains having different scopes. Using this method alone will not only pose possible setbacks due to the quality of the dictionary, it will also prove incapable of adapting to the incessantly changing environment

2.2. Text-clustering-based construction method

Using the text-clustering-based method to computerize the establishment of conceptual hierarchy is based on related terms grouped together according to their synonyms. Every cluster is represented by a particular word or term that is believed to be more frequently used. Thus, repeating the exercise can derive the hierarchy of the terminologies. Right now, there are still several problems found in using this method which restrict its usability as explained below (Hotho, Maedche, & Staab, 2001):

(1) Text clustering is generally regarded as an objective type of method that generates well-defined results considered optimum in certain aspects. However, this is the exact opposite of what goes on in reality. Different users have different requirements for clustering because they have varying opinions and perspectives in viewing a particular document. Thus, we ought to adopt individual standards so as to accommodate differing perspectives in accomplishing the task of text clustering.
(2) Text clustering is normally required in high dimensional spaces to perform clustering computations since, every word or term is seen as an attribute of the entire document. However, experiments and mathematical analysis confirmed that clustering calculations in high dimensional spaces are inefficient because every data point possesses tendency of similar distances with other existing data points.
(3) Text clustering by itself is ineffective unless it is combined with a specific domain. The common solution is to formulate clustering regulations which unfortunately produces another set of problems such that clustering regulations generate too many features.

2.3. Construction method based on association rules

Using association rules to achieve an automatic construction of concept hierarchies is derived from the idea that association rules with stronger support, confidence and more extensive conceptual relationships can be placed on the upper level of ontology (Maedche & Staab, 2000). For example, we compute the support and confidence, respectively, on (region, accommodation facilities) and (regions, hotels). If the result of support and confidence in (region, accommodation facilities) is clearly higher than (region, hotels), ‘accommodation facilities’ is placed higher than ‘hotels’. Using association rules as a basis for construction method still causes several restrictions which are further explained below (Maedche & Staab, 2000; Wei, Bressan, & Ooi, 2000):

(1) Using association rules can generate combinations of different conceptual relationships, for example, the combination of (Person, Person, HIT) and (Person, Person, LOVE). However, if we treat them as two separate concepts, it will be difficult to arrive at the needed support.
(2) When the document is composed of transactions required by the computation of association rules, different tactics of combination can result in different outcomes. For instance, there are 100 documents about ‘Hotel’ with contents giving detailed descriptions about room types and facilities. After natural language processing, it might result in 10,000 concepts. If a document becomes one transaction, Hotel → Address and Room → Bed with support of 100% can be derived. However, the computation complexity can most likely be high. If every sentence in the document makes up one transaction, then the vital relationship between Hotel and Address may be inadvertently omitted.

2.4. Construction method based on knowledge base

Using the knowledge base as a basis for the construction method requires the prior construction of knowledge bases in related domains. The knowledge base must include basic rules and simple examples. When a user enters keywords to search for information, the rules in the knowledge base is used in order to filter data, while similar examples are displayed to make a possible comparison. When the required result is picked out, rules in the knowledge base again are used to establish related ontology as well as giving the summary and results. This type of method is different from the above-mentioned three methods since, the rules in the knowledge base can be regarded as a kind of ontological manifestation. The rules in the knowledge base are used to assemble related ontology (Alani et al., 2003).

2.5. Formal Concept Analysis

Formal concept analysis (FCA) is a method for data analysis, knowledge representation and information management. It was proposed by Rudolf Wille in 1982 (Wille, 1982). In recent years, formal concept analysis has grown into an international research community with applications in many disciplines, such as linguistics, software engineering, psychology, medicine, AI, database, library science, ecology, information retrieval, etc. One of the distinctive features of formal concept analysis is its ability to generate graphic visualization from the structure of any given data set, more particularly in social science where it is almost impossible to perform a quantitative analysis. Formal concept analysis serves to augment formal analysis methods and to complement statistics and conceptual analysis. FCA also presents a lot of benefits to the field of information science. The exercise of FCA in mathematics can be used to explain classification systems. Formal classification system is capable of analyzing based on the consistence among relationships (Ganter & Wille, 1999). Stumme (2002) explains that FCA shifted emphasis to applications in computer science partly due to a merger with the conceptual graphs community (Sowa, 1984). An overview of the relationship between conceptual graphs and FCA is provided by Mineau, Stumme, and Wille (1999).

3. Research methodology

This research uses ontology learning technology to construct conceptual maps of documents in order to provide an effective reference to users as they perform information searches. The main three problems are given as follows:

(1) How can document data produce the conceptual maps of ontology construction technology?
(2) How can the degree of relationships among various concepts be expressed and studied?
(3) How can users break through conceptual maps in order to perform quick and accurate searches?

The document conceptual maps developed by ontology learning technology in this research are intended to provide a good reference for users when they perform information searches (Fig. 1). The entire system architecture in this study can be divided into three major subsystems and other relative components. These subsystems are enumerated and explained below:

3.1. Term parsing subsystem

When documents of various data sources are entered, they must pass through different preprocessing methods in order for them to qualify in subsequent requirements. The five steps encountered in this phase are:

(1) Elimination of document layout: as a result of having various data sources, document layout can also appear in various forms. Thus, the first step is to disregard all irrelevant information, such as: typesetting format, annotation and other additional information. The output of this phase is a data stream of characters.
(2) Lexical analysis: lexical analysis is the process of transforming the data stream of characters into a data stream of terms (Baeza-Yates,& Ribeiro-Neto, 1999). English lexical analysis makes use of a space symbol or punctuation marks to convert data streams into a set of terms.
(3) Elimination of stop words: in the second stage of the lexical analysis, we noticed that the most frequently used terms normally do not have distinguishing or recognizing property. In fact, in one document, more than 80% of the terms are meaningless and are often filtered out during the analysis. The terms referred here usually are articles, prepositions, conjunctions, and other terms that do not constitute the main idea or concept of the document. Examples are a, as, and, etc. Eliminating these stop words not only saves memory space but also decreases complicated calculations.
(4) Elimination of stemming words: different writers exercise different writing styles. For this reason, the chances of having slight variations in the context due to how a particular term was used are inevitable. Plurality, verbal nouns, and tenses can alter the basic form of the word. The standard form of the word or root word is used to replace its different forms. Say for instance the word ‘connect’, variations of this word are connecting, connection, connections, etc. Using the root word to replace all other variations of the same word can free up memory space and reduce complicated calculations.
(5) Thesaurus: it is probable that different words can actually mean the same thing, thus, the thesaurus is used to disregard redundant terms (Baeza-Yates,& Ribeiro-Neto, 1999).

3.2. Ontology construction subsystem

Upon changing the document content into a set of terms, the ontology construction subsystem adopts the ontology construction technology to produce document conceptual maps. This subsystem consists of two major components:

3.2.1. Establish a set of conceptual relationships and hierarchy among terms

Here, we use the idea of formal concept analysis (Buchli, 2003; Ganter & Wille, 1999) to establish a set of conceptual relationships and hierarchy of terms. We believe that while the more general abstract concepts appear more often in

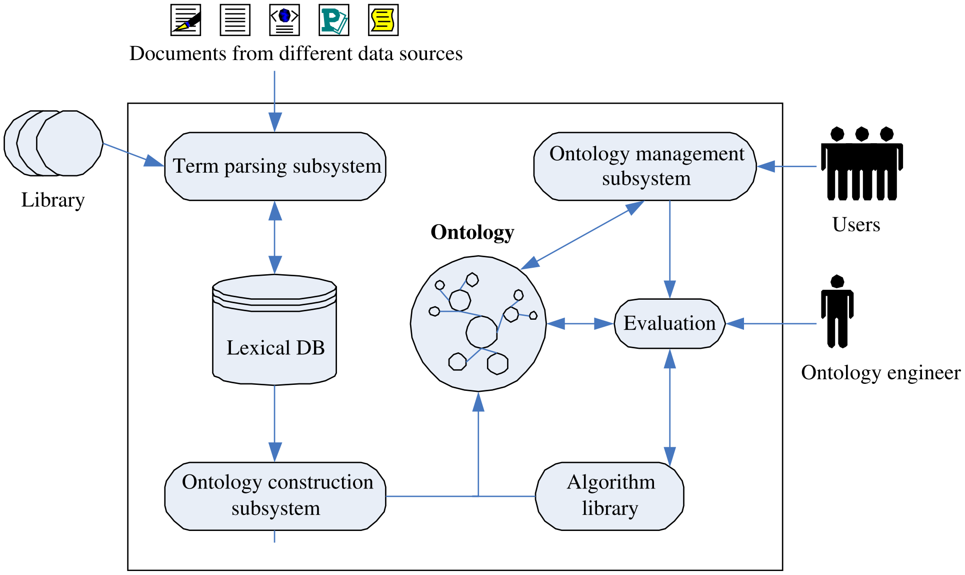

Documents from different data sources
Library
Term parsing subsystem
Ontology management subsystem
Users
Ontology
Lexical DB
Evaluation
Ontology engineer
Ontology construction subsystem
Algorithm library

Fig. 1. System architecture.

documents of some related subject, the technical concepts in more details appear less often. Generally speaking, there are three kinds of relationships that exist among concepts. These are independence, intersection and inheritance. In order to establish the concept relationships and hierarchy of the different terms, there are five steps to follow:

Step 1: Produce the binary relation matrix between documents and terms

In every document, a term that will best represent the main concept of the document must be obtained in the term retrieval subsystem. We do this by referring to the document set and the term set. If a term appears in a document, the corresponding entry of the matrix is labeled as '*X*'. From here we can generate the binary relation matrix between documents and terms. The initial starting point in using FCA is setting up a context (Buchli, 2003). A context is a triple: $L=(D,T,L)$. In this research, the context of the ontology is identified as L, the related document set of the ontology is represented by D, the related term set of the ontology is marked as T, and lastly, I is a binary relation between D and T: $I \subseteq D \times T$.

Step 2: Generate the concept set C

If we let X be the partial set of D, and Y as the partial set of T, therefore, $X \subseteq D$, $Y \subseteq T$. The mappings:

$$\sigma(X) = \{t \in T | \forall\, d \in X : (t,d) \in I\},$$

the common terms of X, and

$$\tau(Y) = \{d \in D | \forall\, t \in Y : (t,d) \in I\},$$

the common documents of Y. Based on the above definitions, a concept is defined. A concept is a pair of sets: a set of documents and a set of terms (X,Y) such that: $Y=\sigma(X)$ and $X=\tau(Y)$. Therefore, a concept is a maximal collection of documents sharing common terms. Thus, taking concept c as an example, it means that the biggest document set that contains the common terms is in the maximal rectangle constituted by all the relationships I in the binary relation matrix. The set of all the concepts of c is represented by C.

Step 3: Calculate hierarchy relationship of concepts

The set of all the concepts of a given context forms a complete partial order. Thus, we define that a concept (X_0,Y_0) is a sub-concept of concept (X_1,Y_1), denoted by $(X_0,Y_0) \subseteq (X_1,Y_1)$. In the event that the document set X_1 of a term set Y_1 is contained in the document set X_2 of another term set Y_2, denoted by $X_1 \subseteq X_2$, (X_1,Y_1) becomes the sub-concept of (X_2, Y_2), denoted by $(X_1,Y_1) \subseteq (X_2,Y_2)$. For concept set C, it means $c_1(X_1,Y_1)$ becomes the sub-concept of $c_2(X_2,Y_2)$.

Step 4: Generate the entire hierarchy of concepts

It is possible for concept c to have various father concepts as well as sub-concepts. For this reason, computing various hierarchy relationships for different concepts is required in order to obtain the entire hierarchy of concepts. Each node in the hierarchy represents a concept. Given two elements (D_1,T_1) and (D_2,T_2) in the concept hierarchy, their supremum or join is defined as (Buchli, 2003):

$$(D_1,T_1) \cup (D_2,T_2) = (\tau(T_1 \cap T_2), T_1 \cap T_2).$$

Let $c_1(X_1,Y_1)$ and $c_2(X_2,Y_2)$ be two concepts, the supremum of the two concepts is computed in order to determine their respective positions in the concept hierarchy.

Step 5: Generate the inter-relationships of concepts

After constructing the hierarchy relationships among concepts, we now identify the inter-relationships of concepts. Let $c_1(X_1,Y_1)$ and $c_2(X_2,Y_2)$ are two concepts, if $Y_1 \subset Y_2$ and $Y_2 \subset Y_1$, since the two concepts are partially contained by one another, it allows us to identify the inter-relationship between c_1 and c_2.

3.2.2. Calculate the degree of relevancy among concepts

After establishing the relationships between concepts, we can begin to calculate the degree of relevancy among concepts which are not directly inherited. In this regard, let us examine the method of calculation formulated by Kang, Huh, Lee, & Kim (2000) in computing for the correlation between concepts. The formula and related variables are as follows:

$$f_{jk} = \text{relevancy}(T_j, T_k) = \frac{\sum_{i=1}^{n} \mathrm{d}_{ijk}}{\sum_{i=1}^{n} \mathrm{d}_{ij}} \times \text{WeightingFactor}(T_k) \qquad (1)$$

$$d_{ijk} = \mathrm{tf}_{ijk} \times \log_{10}\left(\frac{N}{\mathrm{d}f_{jk}} \times w_j\right) \qquad (1.1)$$

$$d_{ij} = \mathrm{tf}_{ij} \times \log_{10}\left(\frac{N}{\mathrm{d}f_j} \times w_j\right) \qquad (1.2)$$

$$\text{WeightingFactor}(T_k) = \frac{\log_{10}\frac{N}{\mathrm{df}_k}}{\log_{10}N} \qquad (1.3)$$

Formula (1) describes the degree of relevancy between two terms. All degrees of relevancy have a corresponding direction. The significances computed by different terms as central points are different. For example, in formula (1), the central point of the calculation is T_j as the correlation between T_k and T_j is being established. Incidentally, formula (1) can be broken down into three other equations as seen in formulas (1.1), (1.2), and (1.3). Note that formulas (1.1) and (1.2) make use of the TF-IDF concept (Salton & McGill, 1983). In formula (1.1), d_{ijk} is decided by the frequency that both T_k and T_j appear simultaneously and the inverse document frequency. tf_{ijk} represents the frequency that T_j and T_k both appear in document i, df_{jk} represents the total document number that T_j and T_k appear together. Consequently, when both terms have higher relevancy, the frequency of T_k and T_j appearing in the same document should also be high and they should centralize in some specific documents. The same concept applies to formula (1.2). On the other hand, as seen in formula (1.3), WeightingFactor(T_k) corresponds to the specificity of T_k against the documents. As the term T_k becomes more general, the value of the WeightingFactor(T_k) decreases. The description of variables in formula (1) is shown in Table 1.

To further explain, let us use Fig. 2 as an example. Fig. 2 shows the frequency of the terms in every document. The result

Table 1
Description of variables in formula (1)

Variables	Description	Variables	Description
N	Total number of keywords	tf_{ij}	Frequency of term *j* in document *i*
tf_{ijk}	Co-occurrence of term *j*,*k* in document *I*	df_j	Document frequency of term *j*
w_j	Weight for inverse document frequency	df_{jk}	Document frequency of term *j*,*k*

calculated by formula (1) and the conceptual hierarchy generated by the formal concept analysis (FCA) constitute the ontology conceptual map in Fig. 3. In the same figure, the arrows with full lines serve to show the inheritance among concepts, while the dotted lines show the mutual relationships between two concepts. The number seen on the dotted lines is the significance of concepts, which follows a single direction. This is so because different concepts have their own related concepts and thus, have different significances.

3.3. Ontology management subsystem

Ontology management subsystem has two major parts. First, for construction builders, the most important consideration is the absence of error in the hierarchy relationship among concepts rather than the correlation among concepts. As a result, it is necessary for builders to be involved and determine the validity of the ontological construction result. In the event that an error occurs in the hierarchy relationship or there was a failure in forming the necessary hierarchy relationship, the builder should examine the adequacy of data representation or other possible causes. From the standpoint of users, an error in the hierarchy will result in users' misunderstanding about the concepts. When this happens, users will rather find another sources or methods for learning. Secondly, as far as the Semantic Web is concerned, erroneous ontology will lead to numerous errors when the intelligent agents perform an information search. In the blueprint of the Semantic Webs, intelligent agents search and examine different ontology, and then generate the final result. Thus, errors in the ontology will correspond to errors in the results provided by the intelligent agents. This happens because errors can be found if ontology is used by human while intelligent agents judge by the results so that the ontology with errors cannot be used by intelligent agents. With regard to the second part of this subsystem, it is the aim of this research to supply users with an effective searching interface that will facilitate their search. In addition, users can select a concept based on the conceptual map in Fig. 3. It will lead them to uncover a related concept, or they can simultaneously select several concepts to get the relationships among them as well as documents with relative significance.

	Data mining	Ontology	Classification	Ontology learning	Association rule	Decision tree	Text knowledge engineering	Apriori	OIL	SMES	C5.0	Clustering	DAML
Document 1	5		3				7					2	
Document 2	3		7		6					7			
Document 3	4			8	7							4	
Document 4	5			3	6			2					
Document 5	5		6		4					4			
Document 6	9		9			14					7		
Document 7		6		5	3					1			
Document 8		9		7	5					2			
Document 9		10		8	4								2
Document 10		8		3			4		3				

Fig. 2. Matrix denoting frequencies of terms appearing in documents.

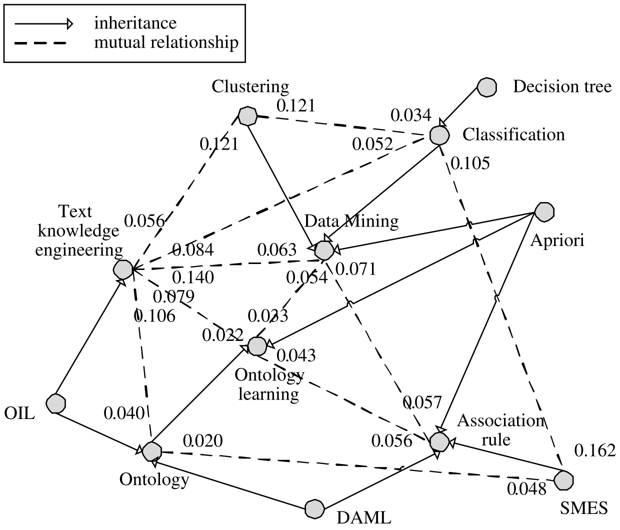


Fig. 3. Ontology conceptual map from Fig. 2.

4. Experiment design and results

This research is conducted with the principal aim to upgrade the existing Internet applications. Data in the experiment come from Internet sources. The system proposed in this research is implemented in the Internet. Furthermore, it is seen from the system architecture in Fig. 1 that the system requires the use of some function library. For this reason, this research has chosen the Java language as the implementation language.

4.1. Experiment evaluation design

The ultimate objective of the ontological construction technology in this research is to build a map for related ontological concepts that will help users in their search for relevant information. With the present ontological construction technology, errors could not be completely avoided in establishing the hierarchy relationship no matter with the techniques of dictionary-based, text clustering, association rules or the knowledge base. In this regard, we use the hierarchy relationship by comparing the concept nodes to

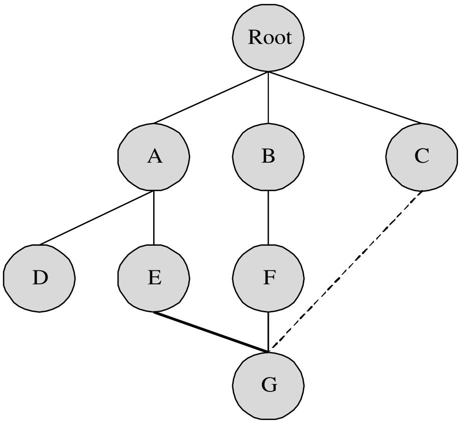


Fig. 4. Diagram of a conceptual hierarchy.

obtain the accuracy rate of the entire ontology. Thus, in measuring the efficiency of the construction method, this research adopted the most commonly used measures in data mining, namely, precision and recall, for the general assessment (Han & Kamber, 2001). This is further illustrated in the following equations:

$$\text{Precision} = \frac{|\{\text{Relevant}\}\&\{\text{Retrieved}\}|}{|\{\text{Retrieved}\}|} \qquad (2)$$

$$\text{Recall} = \frac{|\{\text{Relevant}\}\&\{\text{Retrieved}\}|}{|\{\text{Relevant}\}|} \qquad (3)$$

In measuring the conceptual hierarchy in ontology, precision refers to the accurate ratio of conceptual hierarchies that are automatically constructed, while recall refers to the accurate ratio of conceptual hierarchies that should be generated. Fig. 4 shows a diagram of a conceptual hierarchy. As seen in the diagram, the hierarchy has a total of eight concept nodes and eight conceptual relationships. Among them is an erroneously constructed conceptual relationship represented by the full line. On the other hand, the dotted line shows conceptual relationship that should exist but was not automatically constructed. Thus, applying formulas (2) and (3) in measuring the conceptual hierarchy will yield a precision of $(8-1)/8=87.5\%$ and a recall of $7/(8-1+1)=87.5\%$.

4.2. Experiment 1

The usage scope of the experiment materials in experiment 1 is smaller while the contents are more identical in order to test the efficiency and accuracy of the construction method in this research. The experiment materials used were the recorded dissertations found in the 'Dissertation and thesis abstract system' (http://datas.ncl.edu.tw/theabs/1/), provided the titles included the terms 'data mining'. A total of 187 documents were collected.

Wu, Day, and Hsu (2001) pointed out that subject words and keywords are usually composed of noun–verb and noun–noun terms. Thus, by the syntactical functions and morphological features of speech and the sifted speech base can filter out majority of the irrelevant terms. And due to the possibility that two or more terms can mean exactly the same thing, particularly proper nouns of foreign languages having multiple meanings and translations, it becomes necessary to establish a dictionary of synonyms that will facilitate the accurate translation of terms. Accomplishing this will certainly lead to a higher rate of efficiency. Finally, we come to the stop word list. Although by means of the properties of speech and synonyms can gather up most of the terms based on nouns, it is not the case that all nouns have the distinguishing meanings. For this reason, there is a need filter out the stop words in order to increase the efficiency rate.

Table 2 shows the original set of terms collected from experiment 1, the set of terms after usage of properties of speech, synonyms, and stop words, as well as the final collection of terms and the filtering rate of the term set. If we take a closer look at the table, we will notice that the set after usage of properties of speech has the highest filtering rate. This is because, we only sieve out certain nouns and verbs to represent concepts. It also proves that describing terms and sentences have the highest number in any given document. The filtering rate of the synonyms and stop words may appear lower but its effect on the overall efficiency must not be overlooked. The final collection of terms is only 15% of the initial collection.

Different numbers of term collection gave rise to different ontological content and efficiency levels. The amount of terms filtered out decides their capability to represent the ontology of data content. Too many conceptual nodes will generate noises, while too few conceptual nodes may not be adequate enough. This research stands by the fact that the hierarchy rate of ontology conceptual hierarchy can be used to denote the level of disorganization in data contents. Supposing a single conceptual node falls under the root node and there exists no other nodes below it, when such is the case, we believe that this particular node has a low correlation with other existing nodes and it is no longer considered as part of the conceptual hierarchy. It is commonly known as an independent node, see node C (filled with oblique lines) in Fig. 5. Thus, we have come to a definition of the hierarchy ratio as shown

Table 2
Number of term set in different situations

	Original set of terms	Set after usage of properties of speech	Set after usage of synonyms	Set after usage of stop words	Final collection of terms
Number of term set	4468	865	764	676	676
Filtering rate (%)	100	19	17	15	15

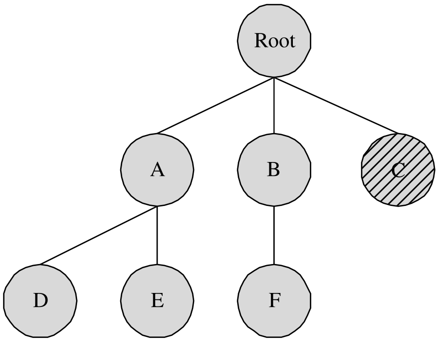


Fig. 5. Diagram that shows hierarchy ratio and the independent node.

below:

$$\text{Hierarchy ratio} = 1 - \frac{\text{number of independent nodes}}{\text{total number of nodes}} \tag{4}$$

After obtaining the hierarchical ratio, the optimum number of term sets we obtained was 107. As a result, we combined the term sets with the FCA algorithm to produce the complete ontological framework. The ontological results from experiment 1 are shown in Table 3. Since, the term sets were filtered out, the number of documents decreased, from 187 to 184 in this experiment. On the other hand, the depth and breadth of the hierarchy reveal the range of the content included in the ontology. The wider the hierarchy expands, the more diversified and general the concepts are. While the deeper the hierarchy goes, the more detailed the contents become. The number of hierarchy relationships manifests the degree of complication of the nodes. This experiment had a total of 107 nodes but only 132 hierarchy relationships were produced. This means that the relationships among nodes are not at all complicated. The precision and recall of ontology determined by the experts are 84.1 and 81.1%, respectively. In Table 3, the numbers in the parentheses following the precision and recall represent the number of errors in the hierarchy relationships and un-constructed relationships.

Based on the results of experiment 1, Fig. 6 is the result as the node with ‘Classification algorithm’ was chosen to become the inquiry output. From Fig. 6, we could see that the father node of ‘Classification algorithm’ is ‘Data mining’, while its sub-nodes are ‘Bayesian classification’, ‘Neural networks’, ‘Decision tree’, ‘Patent analysis’ and ‘Virus’. The five nodes connected to the node with ‘Classification algorithm’ as well as the papers related to ‘Classification algorithm’ and their respective significances are shown in Fig. 6. Therefore, we can represent the hierarchy relationships by using the relationships among nodes. Some other correlated nodes and documents are also used to convey the significance of resources.

4.3. Experiment 2

The primary objective of this experiment is to test the efficiency and accuracy of the data such that the experimental data contents are more general and diversified. The source of data is the timely news found in http://news.pchome.com.tw/. The primary subject is the news content divided into nine sections, namely: politics, society, finance and economics, science and technology, entertainment, sports, lifestyle, international news, and cross-strait. The news materials to be tested were divided into two groups. The first is under the ‘date’ unit. This group was tested for a total of seven days, from January 14 to January 20. Data were gathered every night after 8 pm. The second one belongs to the ‘category’ group and is composed of nine categories. Same length of time was spent in testing the materials of the same category. Each category consists of seven day’s worth of materials. The number and distribution of news materials are shown in Table 4 below.

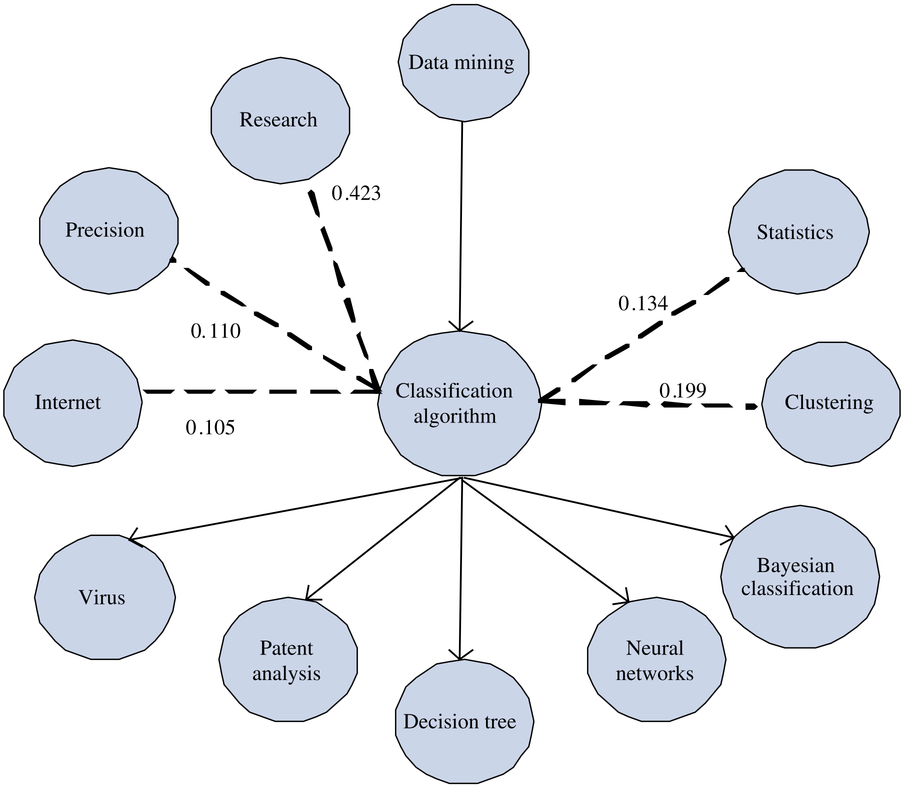
In each experiment, we were able to obtain the optimum collection of terms and performed the FCA algorithm to produce a complete ontological framework. Due to the filtering process done on the term sets, the number of documents decreased. Furthermore, we noticed that the number of ‘category’ nodes is fewer and there is a more reduction in the documents. The cause of this, as we have inferred, is the centralization of the range of ‘category’ data.

The depth and breadth of the hierarchy show the scope range of the ontology. The wider the hierarchy extends, the more diversified and general the concept is. On the other hand, the deeper the hierarchy goes, the more detailed the content becomes. We also noticed that the ‘category’ group has a higher average depth and a smaller width than the ‘date’ group. This is just as we expected since, the scope of daily news encompasses everything, while the data in the ‘category’ group merely include their respective categories.

The results of experiment 2 can be seen in Table 5, Figs. 7 and 8. We notice from the results that the ontology generated by the ‘category’ group showed a higher precision, particularly the ‘finance and economics’ category with 88.68%. The precision was at its lowest on January 17 with only 72.20%. Another observation worth noting is the big changes under recall. The lowest recall rate was 73.58%. This verifies that the method used in this research is indeed feasible and completely satisfies the requirements of dynamically generating classification frameworks.

Table 3
Ontology results of experiment 1

Experiment 1	Number of nodes	Number of documents	Depth of hierarchy	Breadth of hierarchy
	107	184	5	47
	Hierarchy ratio	Number of hierarchy relationships	Precision	Recall
	83.18%	132	84.1% (21)	81.1% (25)



Paper id	Paper title	Significance
D 176	Applications of data mining: mining purchasing features of customers	0.294
D 174	Applying data mining to construct adaptive web sites	0.265
D 135	Applying data mining to integrate purchase orders	0.168
D 98	Applying data mining on the analysis of traffic accidents	0.161
D 147	Using data mining techniques to process document recommendation	0.147
D 136	Integrate fuzzy neural networks and principal analysis on classification	0.131

Fig. 6. Result that the node with 'Classification algorithm' is chosen to become the inquiry output.

Table 4
The number of news materials and their distribution

	1/14	1/15	1/16	1/17	1/18	1/19	1/20	Total
Politics	58	48	33	39	40	53	48	319
Society	66	45	34	32	33	49	44	303
Finance and economics	67	48	35	42	53	46	36	327
Science and technology	52	30	32	47	37	33	38	269
Entertainment	64	40	49	61	78	68	51	411
Sports	61	41	40	46	46	47	42	323
Lifestyle	55	51	41	40	39	50	49	325
International	47	42	37	45	35	32	35	273
Cross-strait	61	33	31	35	34	30	33	257
Total	531	378	332	387	395	408	376	2807

Table 5
Results of experiment 2

	Number of nodes	Number of documents	Depth of hierarchy	Breadth of hierarchy	Hierarchy ratio (%)	Number of hierarchy relationships	Precision	Recall
1/14	139	424	2	58	79.14	178	75.8%(43)	80.78%(36)
1/15	167	323	2	61	84.43	220	78.6%(47)	81.36%(41)
1/16	118	362	2	55	73.73	120	80.8%(19)	77.50%(27)
1/17	277	405	4	81	88.45	435	72.2%(121)	88.74%(49)
1/18	204	424	2	54	88.73	304	75.7%(74)	87.50%(38)
1/19	154	439	2	64	76.62	170	77.6%(38)	80.00%(34)
1/20	107	319	2	73	89.16	234	81.62%(43)	83.76%(38)
Politics	47	218	3	45	100	53	86.79%(7)	75.47%(13)
Society	100	210	3	33	87.00	134	81.34%(25)	86.57%(18)
Finance and Econ.	48	244	4	46	100	53	88.68%(6)	73.58%(14)
Science and tech.	99	198	2	39	84.85	163	82.21%(29)	87.12%(21)
Entertainment	57	340	3	34	100	67	86.57%(9)	86.57%(9)
Sports	77	234	3	35	100	59	83.05%(10)	80.66%(12)
Lifestyle	157	226	3	64	81.53	268	82.09%(48)	86.19%(37)
International	91	206	3	41	80.22	82	80.27%(17)	82.93%(14)
Cross-Strait	81	193	4	32	81.48	79	82.28%(14)	81.01%(15)

5. Conclusion

As a result of the extensive developments in the Internet, sharing knowledge with each other has finally become a reality. Unfortunately, it is for the same reason that we are facing an overflow of data and information. Nevertheless, the Semantic Web concept proposed by Berners-Lee and Fischetti (1999) paved the way to the formulation of possible and effective solutions.

The most vital tools in searching for information and related resources in a Semantic Web are the ontology and intelligent agent. In the field of ontology, ontological framework is normally formed using manual or semi-automated methods requiring the expertise of developers and specialists. This is highly incompatible with the developments of World Wide Web as well as the new E-technology because it restricts the process of knowledge sharing.

Consequently, this research adopted the formal concept analysis algorithm to study the automation of developing ontological framework and with the hope of fully satisfying the requirements of such. Since, the Semantic Web technology is still in its research stage, it is still difficult to thoroughly assess the cost and efficiency of an automatic ontology development. In this research, analyzing the news material was an attempt to develop the ontology of the news website http://news.pchome.com.tw/. The objective is to apply it in the future Semantic Webs so that users can find the information they need at a much faster pace, as well as to recommend users with the most related content.

The experiment materials used in this research consist of dissertation papers and news information. The precision and recall of the ontological framework generated by the different data were computed. The ontological framework tallied with the data content concept and is capable of solving problems associated with the flexibility of the classification framework and conceptual relationships. From the results of experiments 1 and 2, we knew that the methods used in this research complied with the conceptual hierarchy and the relationships in the

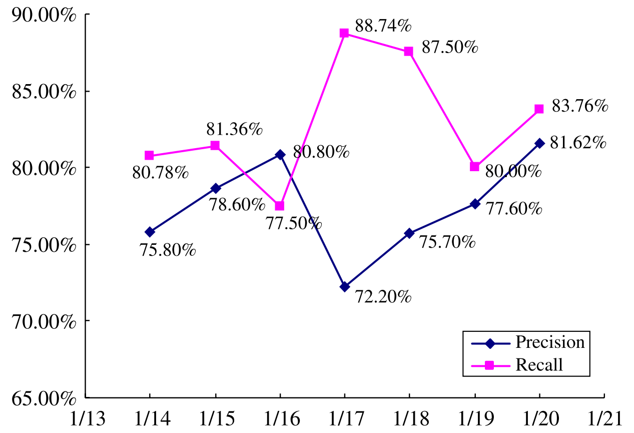


Fig. 7. Result of experiment 2 based on dates.

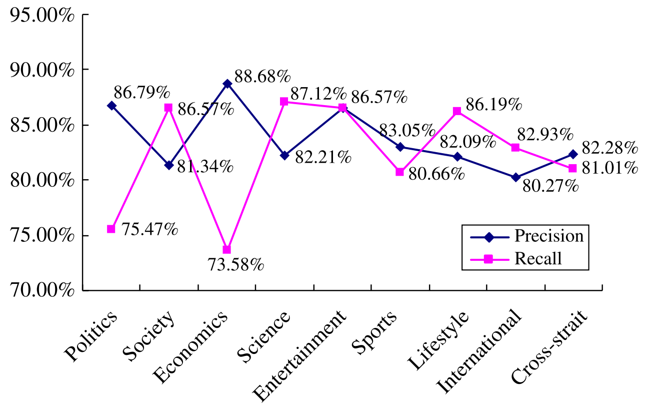


Fig. 8. Result of experiment 2 based on news categories.

Table 6
Different ontology construction methods and comparison of limitations

Limitations of the method	Construction method				
	Dictionary-based	Text clustering	Association rules	Knowledge base	FCA
Construction method depends on the specific data	X			X	
Construction method cannot be used independently	X			X	
Preprocessing of data can influence the construction method			X		
Faces a difficult choice between efficient management and data loss[a]		X	X		
Output is incapable of being interpreted		X			

[a] An enormous collection of terms can result in lower efficiency. Although reducing the number of term collection may raise up the efficiency level, it runs the risk of losing portions of the information. Thus, a choice must be made between the two factors.

hierarchy produced by the data content. Unlike in the existing ontology development technology, it does not require, for example, a dictionary with a specific domain for dictionary-based construction method and does not have the difficult problems encountered in text clustering construction method. Thus, the methods applied in this research have overcome the limitations of the existing methods with regard to automatically generating conceptual hierarchy. If we take a look again at the results of experiments 1 and 2, we will see that the methods used in the experiments were able to obtain a higher precision and recall rate particularly in smaller data scope and in data with more identical content. Furthermore, the methods produced favorable results in its experiment on the general news information. In connection with the ontology conceptual hierarchy, it was observed that smaller data scope or data with more identical content will produce deeper conceptual hierarchies with narrower breadths. When general data have shallower conceptual hierarchy, the breadth of the hierarchy is wider. This is because the content of general data are more dispersed so they tend to produce a more flat conceptual hierarchy, while the specific data content generate a more complete ontological conceptual hierarchy. All these prove that the methods adopted in this research are more suitable to data with smaller scopes.

We believe that this research is able to make the following contributions:

(1) It solves the limitations of the existing ontological construction for classification framework.

At present, the ontological construction technology is divided into four kinds, namely: dictionary-based, text clustering, association rules, and knowledge bases. This research experiment attempted to work these limitations out and came out with comparative measures as shown in Table 6.

This research made use of the formal concept analysis and combined it with the conceptual relationships to construct the ontological concept diagram. Table 6 shows how the ontological construction method used in this research was able to overcome and solve the limitations of existing methods. In addition, the results of experiment 1 and 2 proved that the methods were able to fully satisfy the requirements for the automatic construction of ontological classification framework.

(2) It solves the problems of conceptual relationships and resources significance.

The categorical relationships manifested by the present classification framework are usually absolute but are in no way capable of showing the relative significance of identical classification level. This problem can result in a fixed path when searching for a concept and can easily omit vital information. The methods proposed in this research can construct the relationships between various concepts and sort the significance among concepts. Due to the less flexibility brought about by manually defining the classification framework, the correlation between concepts becomes more difficult to identify. Furthermore, the significance of different document resources against concepts can be expressed by different combinations of concepts. The document significances obtained by different conceptual combinations will not be exactly the same, thus, helping the users to increase the precision rate of their searches and reduce the time spent in searching for information.

Finally, in the present ontological construction methods, no extensive automated construction method exists that can absolutely fulfill the requirements of a Semantic Web. Above all is the enormous amount of information that had increased exponentially since, the beginning of the Internet technology. Therefore, the manual construction of ontology is definitely not the answer if we want to see the full development of Semantic Webs. What is needed is an automatic construction method that will solve the difficulties encountered in the present.

According to our experiment results, the method we proposed can achieve more favorable results compared to other methods under similar data content with smaller scopes. Therefore, if we can integrate it with other ontological construction methods, say for instance the text clustering method, the general data contents can be segmented and grouped into similar contents together, and then proceed to apply the methods suggested in this research. We believe that the system in this research can thus yield better results. Moreover, it will merit further investigation and analysis by future research studies.

References

Alani, H., Kim, S., Millard, D. E., Weal, M. J., Hall, W., Lewis, P. H., et al. (2003). Automatic ontology-based knowledge extraction from web documents. *IEEE Intelligent Systems*, *18*(1), 14–21.

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: Addison-Wesley.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor*. San Francisco, CA: HarperAudio.

Buchli, F. (2003). *Detecting software patterns using formal concept analysis*. Bern, Switzerland: University of Bern.

Ganter, B., & Wille, R. (1999). *Formal concept analysis: Mathematical foundations*. New York: Springer.

Han, J., & Kamber, M. (2001). *Data mining: Concepts and techniques*. New York: Morgan Kaufmann.

Hotho, A., Maedche, A., & Staab, S. (2001). *Ontology-based text clustering Proceedings of the IJCAI-2001 workshop text learning: Beyond supervision, Seattle*.

Kang, S. H., Huh, W., Lee, S., & Kim, Y. (2000). *Automatic classification of WWW documents using a neural network Proceedings of international conference on production research, Bangkok*.

Khan, L., & Luo, F. (2002). *Ontology construction for information selection Proceedings of the 14th IEEE international conference on tools with artificial intelligence, Washington, DC* pp. 122–127.

Kietz, J. U., Maedche, A., & Volz, R. (2000). *A method for semi-automatic ontology acquisition from a corporate intranet Proceedings of the EKAW'2000 workshop on ontologies and texts*. France: Juan-les-Pins.

Maedche, A., & Staab, S. (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems*, *16*(2), 72–79.

Maedche, A., & Staab, S. (2000). *Discovering conceptual relations from text Proceedings of the 14th European conference on artificial intelligence, Berlin* pp. 321–325.

Mineau, G., Stumme, G., & Wille, R. (1999). Conceptual structures represented by conceptual graphs and formal concept analysis. In W. Tepfenhart, & W. Cyre, *Conceptual structures: Standards and practices. Proceedings of the seventh international conference on conceptual structures* (pp. 423–441). Berlin: Springer.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. New York, NY: McGraw Hill.

Sowa, J. (1984). *Conceptual structures: Information processing in mind and machine*. Reading, MA: Addison-Wesley.

Stumme, G. (2002). Formal concept analysis on its way from mathematics to computer science. In U. Priss, D. Corbett, & G. Angelova (Eds.), *Conceptual structures: Integration and interfaces, 10th international conference on conceptual structures* (pp. 2–19). Berlin: Springer.

Tan, K. W., Han, H., & Elmasri, R. (2000). *Web data cleansing and preparation for ontology extraction using WordNet Proceedings of the first international conference on web information systems engineering, Hong Kong*, Vol. 2 pp. 11–18.

Wei, J., Bressan, S., & Ooi, B. C. (2000). *Mining term association rules for automatic global query expansion: Methodology and preliminary results Proceedings of the first international conference on web information systems engineering, Hong Kong*, Vol. 1 pp. 366–373.

Wille, R. (1982). Restructuring lattice theory: An approach based on hierarchies of concepts. In I. Rival (Ed.), *Ordered sets* (pp. 445–470). Dordrecht-Boston: Reidel.

Wu, S. H., Day, M. Y., & Hsu, W. L. (2001). *FAQ—centered organizational memory Proceedings of the IJCAI'2001 workshop on knowledge management and organizational memories, Seattle*.