A Fuzzy Ontological Knowledge Document Clustering Methodology

Amy J. C. Trappey, Charles V. Trappey,
Fu-Chiang Hsu, and David W. Hsiao

***Abstract*—This correspondence presents a novel hierarchical clustering approach for knowledge document self-organization, particularly for patent analysis. Current keyword-based methodologies for document content management tend to be inconsistent and ineffective when partial meanings of the technical content are used for cluster analysis. Thus, a new methodology to automatically interpret and cluster knowledge documents using an ontology schema is presented. Moreover, a fuzzy logic control approach is used to match suitable document cluster(s) for given patents based on their derived ontological semantic webs. Finally, three case studies are used to test the approach. The first test case analyzed and clustered 100 patents for chemical and mechanical polishing retrieved from the World Intellectual Property Organization (WIPO). The second test case analyzed and clustered 100 patent news articles retrieved from online Web sites. The third case analyzed and clustered 100 patents for radio-frequency identification retrieved from WIPO. The results show that the fuzzy ontology-based document clustering approach outperforms the K-means approach in precision, recall, F-measure, and Shannon's entropy.**

***Index Terms*—Fuzzy inference control, hierarchical clustering, ontology schema, patent analysis, text mining.**

I. Introduction

Companies nowadays face great uncertainty and time constraints during product development because designs are more complex and the product life cycles have become shorter. Furthermore, companies also face challenges from competitors that hold robust patent portfolios which are used to rapidly create new product designs. Facing such challenges, companies must continuously analyze patent knowledge to avoid infringement and to define intellectual property boundaries. In addition, companies are directing research and development efforts with an explicit goal of filing new patents that enrich and enlarge their strategic intellectual property portfolios. In the field of patent knowledge management, patent clustering plays a critical role to help define future research and development directions. However, current research on patent clustering depends on statistical methodologies which use keywords and phrases that do not adequately represent the knowledge contained in the patent documents. To provide a better solution to patent knowledge clustering, this correspondence adopts the technique of ontological knowledge representation and fuzzy logic control. Ontological knowledge representation enables domain experts to define knowledge in a consistent way and to improve the efficiency of knowledge interchange using a standard format (such as XML, resource description framework (RDF), or OWL). Fuzzy logic is then used on the linguistic expressions to derive the similarity measures among patent documents for clustering. With the support of these two techniques, a deeper knowledge of a patent's meaning can be derived and the similarity among patents can be reliably defined. In the following sections, related literatures from the field of text mining, ontology, fuzzy logic control, and clustering methodologies are surveyed. Afterward, the fuzzy ontological knowledge document clustering method is proposed. Three cases are provided to cluster chemical mechanical polishing (CMP) patents, patent news content, and radio-frequency identification (RFID) patents to demonstrate the efficiency and effectiveness of the proposed methodology.

II. Literature Review

Several areas of research are frequently cited when deriving methods for document clustering. These areas of research include text mining, ontology creation, fuzzy logic, and mathematical clustering. Text mining infers the structure of a document from sequences in the natural language text and is defined as the process of analyzing text to extract metadata or higher level information [1], [2]. There are a number of well-known research domains in text mining, such as information retrieval and natural language processing.

Information retrieval addresses the problem of finding relevant information from large sources of stored data such as the World Wide Web, intranets, and digital libraries. The informational retrieval approaches frequently use key phrases to index and retrieve documents. For example, Hou and Chan [3] present a methodology to extract document key phrases and then calculate frequencies and derive relationships between the phrases. Nevill-Manning *et al.* [4] present an interactive means to infer a document hierarchical structure where users select words from the larger phrases in which they appear. Witten [5] presents an algorithm, SEQUITUR, to extract a hierarchical phrase structure from text. The algorithm uses Naïve Bayes statistics, text term frequency, inverse document frequency, and placement distance to identify key-phrase sequences which are, in turn, used to infer the document structure. Sanchez *et al.* [6] use a feature data-mining algorithm, called One Clause At a Time, for the classification of text documents into disjoint classes. Feng and Croft [7] use a Markov model and the Viterbi algorithm for phrase extraction and demonstrate that this approach is more effective than a technique which uses tagged parts of speech.

Natural language processing uses the computer to study how humans process and understand language. The common approach is to analyze natural language using grammar and semantics. Computer programs parse the natural language of a sentence using the rules of grammar. However, determining the meaning of a sentence is a difficult and complicated problem that tends to be domain and language specific. Thus, researchers are beginning to combine different approaches and create ontologies (structured frameworks) which represent the underlying structure of knowledge to improve text mining and natural language processing.

A body of knowledge in an area of interest is represented by the objects, concepts, entities, and the relationships among them. The World Wide Web can be thought of having an ever expanding body of knowledge that requires a structured framework, i.e., ontology, to describe it and make it available for use. Thus, the RDF was created by the World Wide Web Consortium to model metadata about web resources and to form the ontology. RDF consists of the RDF model,

Manuscript received April 17, 2008; revised November 25, 2007, April 9, 2008, and July 29, 2008. This work was supported by academic grants from the Taiwan National Science Council and the Ministry of Economic Affairs. This paper was recommended by Associate Editor F. Karray.

A. J. C. Trappey is with the Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei 10608, Taiwan, and also with the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: trappey@ie.nthu.edu.tw).

C. V. Trappey is with the Department of Management Science, National Chiao Tung University, Hsinchu 30050, Taiwan (e-mail: trappey@cc.nctu.edu.tw).

F.-C. Hsu is with Avectec, Inc., Hsinchu 300, Taiwan (e-mail: rayon@avectec.com.tw).

D. W. Hsiao is with the Department of Industrial Engineering and Engineering Management, National Tsing Hua University, Hsinchu 30013, Taiwan (e-mail: d937801@oz.nthu.edu.tw).

Digital Object Identifier 10.1109/TSMCB.2008.2009463

1083-4419/$25.00 © 2009 IEEE

its fundamental syntax, the semantic aspects, the concepts, and the corresponding vocabulary. The base element of the RDF is a triple: A resource (the subject) is linked to another resource (the object) through an arc labeled with a third resource (the predicate). This means that the ⟨subject⟩ has a property ⟨predicate⟩ valued by an ⟨object⟩. Computers can readily share knowledge via the RDF, and several researchers have used this ontology as a means to improve data-mining techniques. Ontologies provide an interesting approach to improve text analysis, and many researchers have made significant contributions. Wu and Palmer [8] present a distance-based algorithm to compute the similarity values of pairwise keywords in the ontology. Kung [9] presents an algorithm that automatically generates the ontology and classifies information using fuzzy neural networks. Kao [10] presents a document classification methodology using an automatically constructed ontology but also uses document key term frequencies for classification. Fuzzy logic provides a means for researchers to mimic the classification rules of experts. Gruninger and Fox [11] proposed a methodology to facilitate ontology design and evaluation and implement it via the TOVE (TOronto Virtual Enterprise) modeling project.

The world of knowledge is represented and stored using language which is governed by rules and conventions. Experts can find and process knowledge given that they understand the language and know the rules and conventions. Since experts are not always consistent in the interpretation of knowledge, then they process knowledge with different levels of accuracy. However, if the rules and conventions of experts are transformed into mathematics, then the computer can be programmed to mimic experts and process knowledge with consistency. For example, Lee *et al.* [12] use a predefined ontology to extract news content and apply a fuzzy inference model to derive the similarity of the news and generate news summaries.

Clustering is a general method to create sets that are fairly homogeneous within groups but significantly heterogeneous between groups. One mathematical principle of clustering maximizes the variance between the groups and minimizes the variance within the groups. Clustering approaches have been successfully applied to text processing. Runkler and Bezdek [13] clustered the text of web pages and the sequences of web pages visited by users (web logs). The Levenshtein distance algorithm and the fuzzy c-mean algorithm were jointly applied to generate the clusters. Another example of clustering for text analysis and synthesis was demonstrated by Hsu *et al.* [14] who used the K-means approach for clustering patent documents.

As shown by the previous research, phrases extracted from documents are frequently used to establish similarity relationships between document texts, and these similarity relationships are used as the basis to group documents. However, the statistical analysis of key phrases cannot fully represent the underlying knowledge. Consequently, this correspondence presents a method to analyze and cluster patents and related knowledge documents using a domain ontology schema rather than a key-phrase text-mining approach. The methodology requires experts to construct an ontological schema, i.e., a knowledge framework for the domain, and then train the system using a sample set of patents. Natural language processing is adapted to infer the ontology of patent documents, and fuzzy logic is used to derive the ontological similarity between the documents for clustering.

III. System Methodology

The methodology for fuzzy ontological document clustering (FODC) is described as follows. Initially, domain experts define the domain ontology using a knowledge ontology building and RDF editing tool called Protégé [15], and the words and phrases (e.g., speech, chunks, and lemmas) of the patent documents are mapped to the corresponding domain ontology concepts. The experts also create a training set of patents using a free and easy-to-use natural language processing and tagging tool called MontyLingua [16]. Afterward, the probabilities of the concepts in given document chunks are computed. The concept probabilities calculated in any given patent document are then used for clustering the patents with fuzzy logic inferences. Hence, the hierarchical clustering algorithm is refined by adapting fuzzy logic to the process of ontological concept derivation. The detailed FODC method is described step-by-step in the following sections.

TABLE I
Parts of Speech Using the Penn Treebank Tagset

POS tag	Description
NN	Noun, singular or mass
NNS	Noun, plural
NP	Proper noun, singular
NPS	Proper noun, plural

A. Building a Patent Ontology

The first step of the FODC methodology requires the use of a knowledge-based RDF editing tool called Protégé. The tool assists the domain experts in defining an ontology schema using a graphical interface. Noy and McGuinness [17] were among the first to propose the use of a knowledge-engineering methodology for ontology building. Protégé is a free open-source ontology editor and a knowledge acquisition system. Similar to Eclipse, Protégé is a framework on which various other software plug-ins can easily be added and linked. This application is written in Java and uses Swing to create the user interface [18]. Owing to these characteristics, Protégé is considered a suitable computer-aided tool for developing the ontology. The ontological web can be automatically transformed into standard data formats (XML, RDF, or OWL) for further manipulation and interpretation for knowledge analysis and synthesis.

B. Natural Language Processing and Terminology Training

In order to measure the knowledge contained in patent documents with respect to the defined ontology schema, the system is trained using a set of patent documents. The sentences from the training documents are tagged to extract the parts of speech, chunks, and lemmas using the MontyLingua natural language processing tool. The definitions for the parts of speech [19] are listed in Table I. Afterward, knowledge engineers map the extracted words to the concepts of the ontology. By using the example sentence "A chemical mechanical polishing apparatus and method for polishing semiconductor wafers...," the phrase *chemical mechanical polishing apparatus* and *method* maps to the concept *CMP_method* $(n.)$, *polishing* represents the concept *polish* $(v.)$, and *semiconductor wafers* represents the concept *substrate* $(n.)$ in the ontology schema. The system records the probabilities of the concepts that a word implies in the patent. The conditional probability, P(The patent concept | The word W in chunk C of the corpora), is derived during the training session. For example, we have ten training patents that contain the word polishing, and the chunk of polishing is NX in these data. To map polishing to the ontological concept, consider that the CMP_method concept is referred to in five patents, and the polish_pad concept is referred to in another five patents. Thus, P (The concept is CMP_method | The word polishing is in the NX corpora chunk) $= 0.5$ and P (The concept is polish_pad | The word polishing is in the NX corpora chunk) $= 0.5$.

To maintain the completeness of the FODC system, the research also includes an iterative relearning mechanism to include new words that are not part of the current terminology database. When a new term is

TABLE II
PARTIAL TERMINOLOGY OF CMP PATENT

Lemma	Chunk	Concept	Prob.
chemical	NX	CMP_method	1
cleaning	NX	clean	1
compound	NX	comprise	0.33
compound	NX	compound	0.33
control	NX	position	0.5
method	NX	CMP_method	1
polishing	NX	CMP_method	0.5
polishing	NX	polish_pad	0.5

TABLE III
ANALYSIS OF ONTOLOGY CONCEPT PROBABILITIES

NLP tag	Ontology concepts (probability)
chemical/NN/NX/chemical	CMP_method (1)
mechanical/JJ/NX/mechanical	CMP_method (1)
polishing/NN/NX/polishing	CMP_method (0.5), polish_pad(0.5)
apparatus/NN/NX/apparatus	CMP_method (1)
method/NN/NX/method	CMP_method (1)

detected, it is first stored in the terminology database. Afterward, the system manager assigns a corresponding ontological concept to this term to enable the system to automatically recalculate and update the terminology-ontological concept knowledgebase.

C. Terminology Analyzer

After natural language processing and terminology training, all of the sentence concepts are inferred. Hence, the probabilities of the concepts for each chunk (Table II) are computed. For example, Table III shows that the probabilities of deriving concepts "CMP_method" and "polish_pad" from parsing and analyzing the sentence "chemical mechanical polishing apparatus and method" are $0.9 = (1 + 1 + 0.5 + 1 + 1)/5$ and $0.1 = 0.1/5$, respectively.

D. Knowledge Extraction

After analyzing the terminology, we compute the concept probabilities for each chunk. The chunks implying concepts as predicates are the first to enter into the ontology. Fig. 1 shows that chunk 5 implies two concepts (candidates) as predicates in the ontology. The next step is to select chunks that imply the concepts as the subject in the ontology from the previous sentence to the next sentence. Therefore, the concepts that chunk 1, chunk 4, and chunk 8 imply are the candidates for the subject. The same process is used to determine the object candidates. If there are ten candidates for subject, two candidates for predicate, and ten candidates for object, then there are 200 (10^*2^*10) candidates for the statement. Statements that do not exist in ontology are eliminated. Finally, the output is generated using the probability derived from the following equation:

$$\underset{\text{for all statements based on chunk 5}}{\mathrm{Max}} \times \frac{\text{prob(subject)} + \text{prob(predicate)} + \text{prob(object)}}{3}. \quad (1)$$

The process described earlier is used for chunks that imply the concepts of the predicate in the document ontology. Thus, a document is transformed into a set of statements in the ontology. These statements are viewed as indices of the document and are the basis of similarity comparisons with other documents.

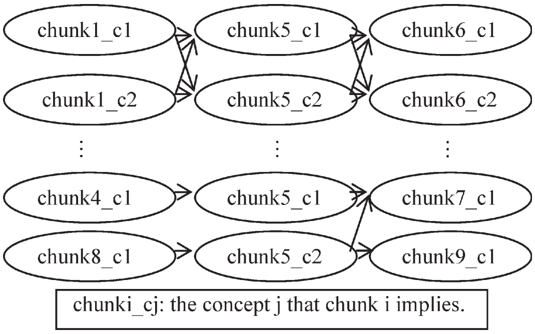


Fig. 1. Filtering ontology statements.

TABLE IV
FUZZY RULES FOR PATENT DOCUMENT SIMILARITY DERIVATION

No.	If two patent documents consisting of sentences (derived as ontology statements) with___.	then, the overall similarity of these two patent documents is ___.
1	Many matches (*mm*) of main concepts and Many matches of detailed descriptions	High
2	Many matches of main concepts and Some matches (sm) of detailed descriptions	High
3	Many matches of main concepts and Few matches (fm) of detailed descriptions	Medium
4	Some matches of main concepts and Many matches of detailed descriptions	High
5	Some matches of main concepts and Some matches of detailed descriptions	Medium
6	Some matches of main concepts and Few matches of detailed descriptions	Medium
7	Few matches of main concepts and Many matches of detailed descriptions	Medium
8	Few matches of main concepts and Some matches of detailed descriptions	Low
9	Few matches of main concepts and Few matches of detailed descriptions	Low

E. Patent Similarity Match

In order to compute the similarity between patent documents, fuzzy logic is used to derive the similarity measure. First, the contents of patent documents are partitioned into the set of main concepts and the set of details. The domain experts then reach an agreement on nine fuzzy rules (Table IV).

Before input to the inference model, the patent documents are translated into an ontological format including main concepts and details. The main concepts consist of higher triples, and the details consist of the lower triples

$$X = \frac{ST}{TT} \quad (2)$$

where

X similarity measure of document 1 and document 2;
ST the same triples in document 1 and document 2;
TT sum of triples in document 1 and document 2.

Fig. 2 shows that between document 1 and document 2, there are two of the same triples in the total of four triples from the main concepts and two of the same triples in the total of five triples from the detailed descriptions. Thus, $TT_m = 4$, $ST_m = 2$, and $X_m = 1/2$ form the similarity measure for the main concepts, and $TT_d = 5$, $ST_d = 2$, and $X_d = 2/5$ form the similarity measure for the detailed descriptions.

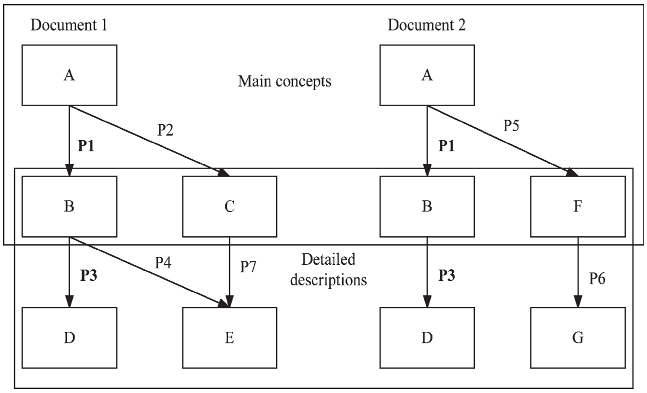


Fig. 2. Ontological comparison of two documents.

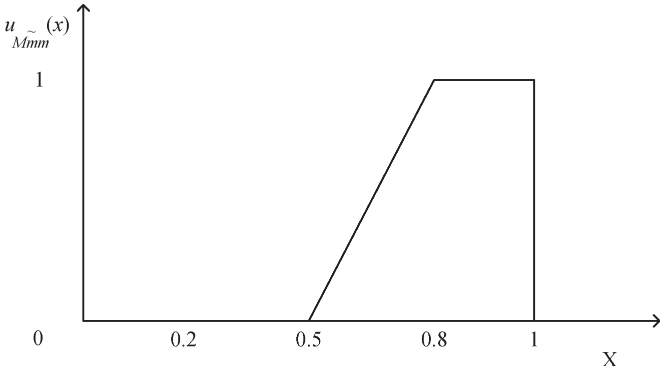


Fig. 3. Membership function for the concept "many matches."

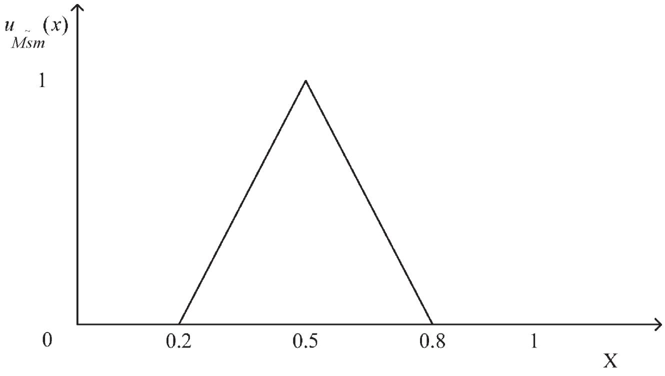


Fig. 4. Membership function for the concept "some matches."

The fuzzy logic representations of "many matches," "some matches," and "few matches" are defined by the membership functions shown in Figs. 3–5, respectively.

The Mamdani fuzzy inference model applies legacy if–else rules to fuzzify the input and output. The ease of formulating the model, the simple calculation, and the clarity in presenting human linguistics support the selection of this approach. Thus, the Mamdani fuzzy inference model using a min–min–max [21] operation considering two rules is adopted and modified. The original Mamdani min–min–max operation only considers a two-rule approach, but this correspondence considers nine rules simultaneously. The detailed procedures for the modified Mamdani fuzzy inference model are shown in Figs. 6–8. The steps for the procedure are as follows.

1) Calculate the similarity of the documents matched in main concepts (X_{mc}) and the similarity of the documents matched in detailed descriptions (X_{dd}).

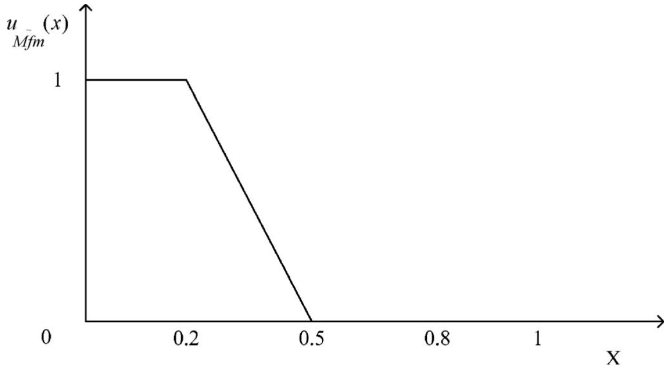


Fig. 5. Membership function for the concept "few matches."

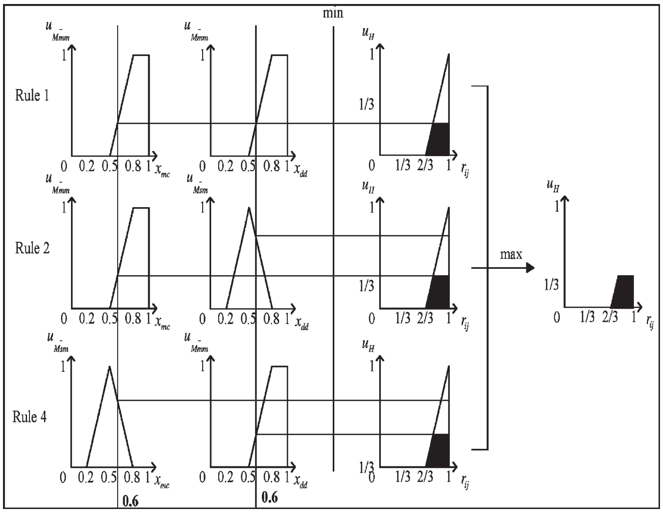


Fig. 6. Procedure for calculating the membership of "High Similarity" given the input $(X_{\mathrm{mc}}, X_{\mathrm{dd}}) = (0.6, 0.6)$.

2) Evaluate X_{mc} and X_{dd} using the rules (Table IV) to derive the corresponding memberships.
3) Compare the memberships and select the minimum membership from these two sets to represent the membership of the corresponding concept (high similarity, medium similarity, and low similarity) for each rule.
4) Collect memberships which represent the same concept in one set.
5) Derive the maximum membership for each set, and compute the final inference result.

F. Defuzzification and Patent Clustering

From the aforementioned inference procedures, the representing membership values of different similarity levels are generated. However, these values are still fuzzy values, and they require dedicated defuzzification processes to help generate the values representing the similarity of patent documents. The defuzzification processes consist of two steps. The first step is to decide which similarity ("High Similarity," "Medium Similarity," and "Low Similarity") best represents the relationship between these two documents. The second step focuses on transforming the value from the similarity membership. Detailed transformation of the similarity value from its similarity membership is depicted in the following three cases. Fig. 9 shows the defuzzification procedures based on the previous example where $(X_{\mathrm{mc}}, X_{\mathrm{dd}}) = (0.6, 0.6)$.

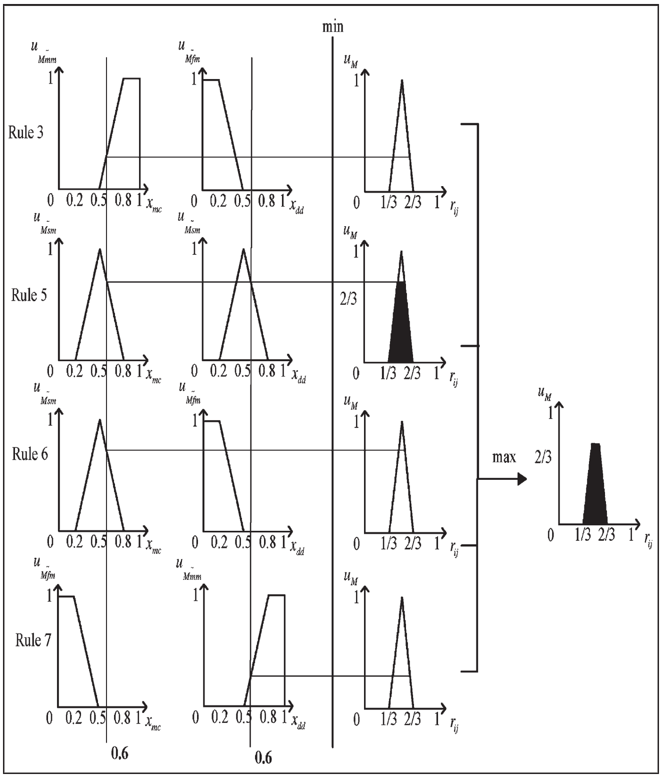


Fig. 7. Procedure for calculating the membership of “Medium Similarity” given the input $(X_{\mathrm{mc}}, X_{\mathrm{dd}}) = (0.6, 0.6)$.

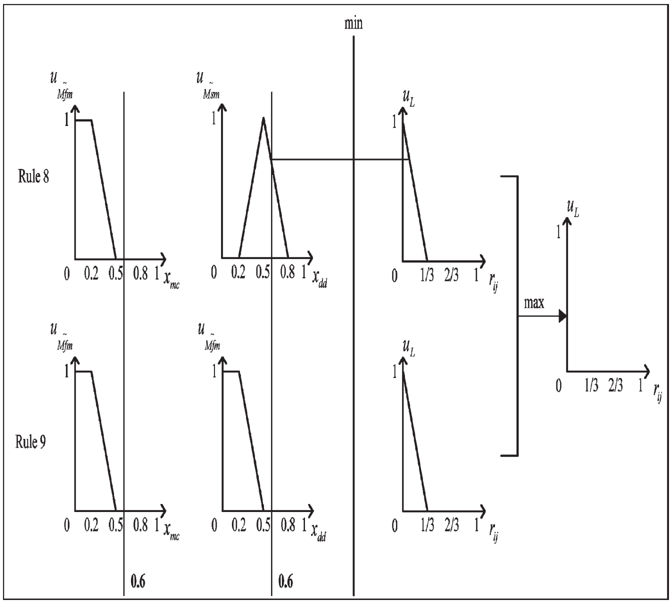
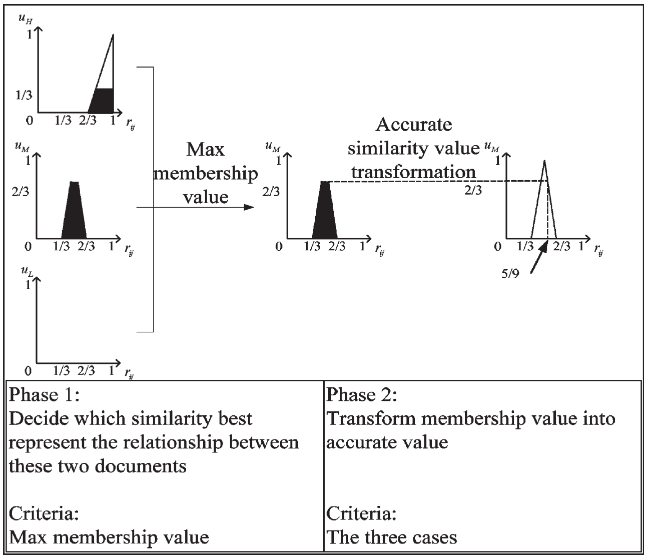


Fig. 8. Procedure for calculating the membership of “Low Similarity” given the input $(X_{\mathrm{mc}}, X_{\mathrm{dd}}) = (0.6, 0.6)$.

Case 1—The Similarity Is High $(U_H > U_L \text{ and } U_H > U_M)$: If the value calculated from the aforementioned procedure (Mamdani fuzzy inference) comes from the “High Similarity” concept, the following equation is used to determine the similarity value of documents i and j (defuzzification):

$$r_{ij}(U_H) = \left\{ \frac{2 + U_H}{3} \right\} \tag{3}$$



Phase 1: Decide which similarity best represent the relationship between these two documents	Phase 2: Transform membership value into accurate value
Criteria: Max membership value	Criteria: The three cases

Fig. 9. Defuzzification processes.

where

U_H membership function for high similarity;
U_M membership function for medium similarity;
U_L membership function for low similarity.

with

$$0 \le U_H, U_M, U_L \le 1.$$

Case 2—The Similarity Is Medium $(U_M > U_H \text{ and } U_M > U_L)$: If the value calculated from the aforementioned procedure (Mamdani fuzzy inference) comes from “Medium Similarity,” the following equation is used to determine the similarity value. When determining the similarity value for “Medium Similarity,” the relationship between “High Similarity” and “Low Similarity” affects the shift of the defuzzification value. As a result, three equations are used to fit different relationships between “High Similarity” and “Low Similarity.”

$$r_{ij}(U_M) = \begin{cases} \frac{2+U_M}{6}, & \text{if } U_L > U_H \\ \frac{4-U_M}{6}, & \text{if } U_H > U_L \\ \frac{1}{2}, & \text{if } U_H = U_L. \end{cases} \tag{4}$$

Case 3—The Similarity Is Low $(U_L > U_H \text{ and } U_L > U_M)$: If the value calculated from the aforementioned procedure (Mamdani fuzzy inference) comes from the “Low Similarity” concept, the following equation is used:

$$r_{ij}(U_L) = \left\{ \frac{1 - U_L}{3} \right\}. \tag{5}$$

After all measures of similarities between patents are calculated, the similarity matrix is generated. The hierarchical clustering algorithm is then used to sequentially search for different clusters according to the different degrees of relations between objects as expressed in matrix

$$\begin{bmatrix} 1 & r_{12} & \cdots & r_{1i} \\ r_{21} & \cdots & \cdots & r_{2i} \\ \cdot & \cdots & \cdots & \cdot \\ r_{i1} & \cdots & \cdots & 1 \end{bmatrix} \tag{6}$$

where r_{ij} is the similarity of document i and document j; hence, the value of r_{ij} is equal to r_{ji}.

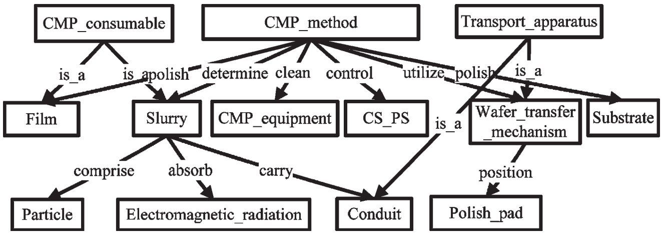


Fig. 10. Ontology schema for the CMP technical domain.

The hierarchical clustering algorithm is applied as follows.

1) Find the $\max(r_{ij})$ in the matrix, and group the documents i and j into a new cluster.
2) Calculate the relationship between the new cluster and other documents by using the average-linkage method.
3) Go to Step 1), until there is only one cluster left.

IV. Case Examples and Experiment

The methodology is demonstrated using three cases: the clustering of a collection of patents related to CMP machines, the clustering of a collection of patent news articles, and the clustering of RFID patents. First, 50 CMP patent documents were collected and downloaded from the World Intellectual Property Organization (WIPO) patent pool as the training documents for ontology building and terminology training. An additional 50 CMP patent documents were collected from the WIPO as the test set. These 50 test documents came from three subcategories (types) of CMP patent documents. Type 1 focuses on the mechanical aspects of CMP machines, Type 2 considers new chemical compositions for the polishing slurry, and Type 3 covers innovative cleaning methods for CMP machinery. Among the 50 patents, 20, 15, and 15 patents belong to Type 1, Type 2, and Type 3, respectively.

The CMP domain knowledge was derived by CMP experts and defined in the ontology schema as shown in Fig. 10. For example, "CMP_ method" is a subject, "polish" is a predicate, and "substrate" is an object in the CMP ontological schema. The domain experts expected that the system would automatically derive three CMP patent document technical clusters (Type 1, Type 2, and Type 3) from the WIPO test set.

After the CMP ontology schema was defined, the system automatically transformed the knowledge documents into ontological expressions. Every knowledge document uses an ontology instance to express what the document means in the RDF format. Finally, the fuzzy logic controller infers the similarity between documents and the hierarchical clusters. The ontology-based fuzzy clustering results show that the FODC methodology and its software automatically grouped 18, 18, and 14 patents, respectively, into Cluster 1 (mechanical design), Cluster 2 (composition of slurry), and Cluster 3 (cleaning method).

For benchmarking the FODC result, the key-phrase-based K-means approach was applied to the same 50 CMP test patents for comparison. The key phrases were extracted using the term frequency and inverse document frequency (TF*IDF) methodology [22]. The K-means clustering algorithm places 15, 21, and 14 patents into Clusters 1, 2, and 3, respectively. The reason for the higher degree of error in clustering using key-phrase-based K-means is due to the inclusion of insignificant key phrases (e.g., structure, method, substrate, and water) which are applied as the basis of the clustering criteria. These key phrases often appear in CMP patent documents. For example, some CMP patents belonging to mechanical control (Type 1) may contain less significant key phrases which cause the K-means approach to place the patents into the wrong clusters.

To further demonstrate the feasibility of the methodology, a second case related to patent news content is provided. One-hundred patent-news-related documents were collected and downloaded as the training set for ontology building and terminology training. Additional 100 documents were collected as the test set. The test documents come from three subcategories (types) of patent news documents. Type 1 news focuses on patent infringement and discusses cases of "one company accusing another company of infringing on their patent rights." Type 2 news articles cover patent trade and refer to "companies licensing their patents to other companies or selling their patents in the technology marketplace." Type 3 news considers the application of new patents to make products after "companies acquire new patents." Among the 100 news articles, 51, 18, and 31 documents belong to Type 1, Type 2, and Type 3, respectively.

The patent news domain knowledge is studied by experts that define the ontology schema as shown in Fig. 11. For example, "patent right owner" is a subject, "get" is a predicate, and "patent" is an object in the patent news ontological schema. The experts expected that the system would automatically derive three patent news document technical clusters (Type 1, Type 2, and Type 3) based on the proposed methodology.

After the patent news ontology schema is defined, the system automatically transforms the knowledge documents into an ontological expression. Every knowledge document is transferred into an ontology instance to express what the document means in the RDF format. Finally, the fuzzy logic controller is applied to infer the similarity between documents and the hierarchical clusters. The ontology-based fuzzy clustering result using the FODC methodology and its software implementation automatically grouped 49, 16, and 35 patents,

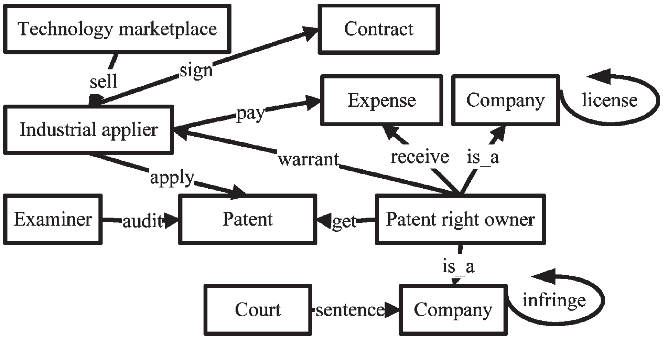


Fig. 11. Ontology schema for the patent news domain.

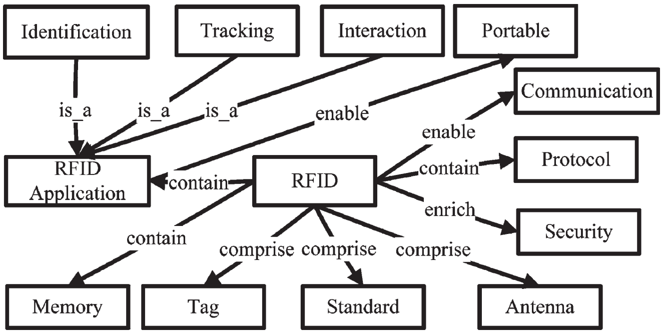


Fig. 12. Ontology schema for the RFID domain.

TABLE V
CONFUSION MATRIX

Confusion matrix		System inferred	
		The same cluster	Different cluster
Actual	The same cluster	a	b
	Different cluster	c	d

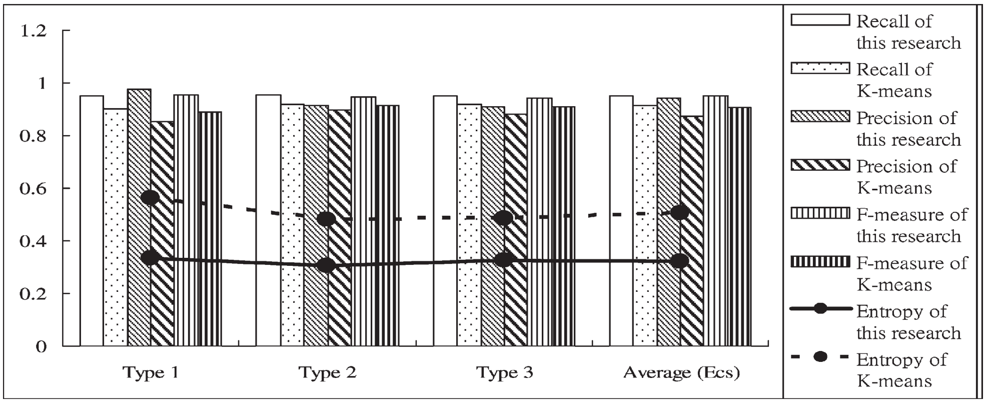


Fig. 13. CMP case results comparing this correspondence and the K-means approach.

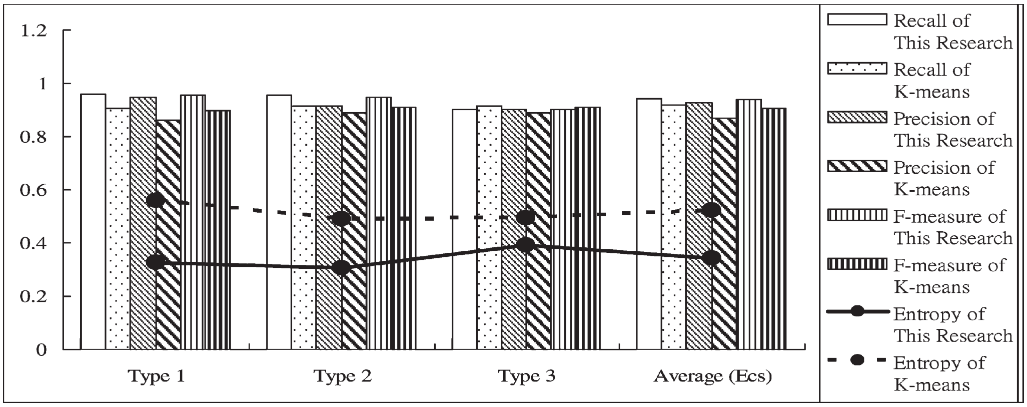


Fig. 14. Patent news case results comparing this correspondence and the K-means approach.

respectively, into Cluster 1 (patent infringement), Cluster 2 (patent licensing), and Cluster 3 (new patent).

For benchmarking the FODC result, the key-phrase-based K-means approach was applied to the same test set for comparison. The key phrases were extracted using the TF-IDF methodology. The K-means clustering algorithm placed 45, 23, and 32 patents into Clusters 1, 2, and 3, respectively.

A third case related to RFID content is provided. One hundred RFID-related patents were collected and analyzed as the training set for ontology building and terminology training. Additional 100 documents were collected as the test set. The test documents come from three subcategories (types) of RFID patents. Type 1 focuses on data detecting and data presenting. Type 2 covers the field of digital data processing. Type 3 considers the application of signal devices and interaction devices. Among the 100 RFID patents, 40, 30, and 30 documents belong to Type 1, Type 2, and Type 3, respectively.

The RFID domain knowledge is studied by experts that define the ontology schema as shown in Fig. 12. For example, “RFID” is a subject, “enable” is a predicate, and “communication” is an object in the RFID ontological schema. The experts expected that the system would automatically derive three RFID patent technical clusters (Type 1, Type 2, and Type 3) based on the proposed methodology.

After the RFID ontology schema is defined, the system automatically transforms the knowledge documents into an ontological expression. Every knowledge document is transferred into an ontology instance to express what the document means as the RDF format. Finally, the fuzzy logic controller is applied to infer the similarity between documents and the hierarchical clusters. The ontology-based fuzzy clustering result using the FODC methodology and its software implementation automatically grouped 37, 35, and 28 patents, respectively, into Cluster 1 (data detecting and data presenting), Cluster 2 (digital data processing), and Cluster 3 (signal devices and interaction devices).

For benchmarking the FODC result, the key-phrase-based K-means approach was applied to the same test set for comparison. The key phrases were extracted using the TF-IDF methodology. The K-means clustering algorithm placed 30, 43, and 27 patents into Clusters 1, 2, and 3, respectively.

In order to statistically compare the clustering results between FODC and K-means, this correspondence uses Recall, Precision [23], and the F-measure [24] as the evaluation rules. Table V presents the confusion matrix used to generate the recall value and precision value. After computing the recall and precision values, the F-measure is derived.

$$\text{Recall} = \frac{a}{a+b} \tag{7}$$

$$\text{Precision} = \frac{a}{a+c} \tag{8}$$

$$F = \frac{1+k^2}{\frac{k^2}{\text{Recall}} + \frac{1}{\text{Precision}}} \tag{9}$$

where k is an instant, and usually, $k = 2$.

Shannon’s entropy [25], [26] is a well-known indicator to measure the clustering capability. First, the entropy values of different clusters are calculated

$$E_j = -\sum p_{ij} \log_2(p_{ij}) \tag{10}$$

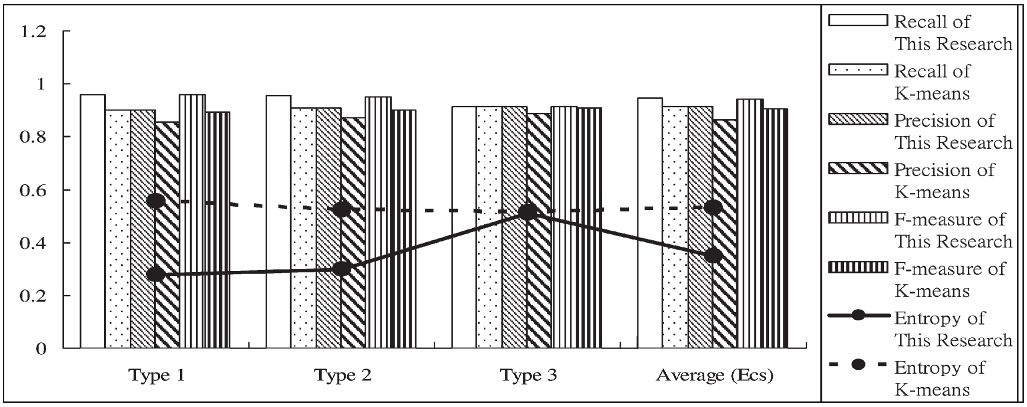


Fig. 15. RFID case results comparing this correspondence and the K-means approach.

TABLE VI
DIFFERENCES BETWEEN FODC AND KEY-PHRASE K-MEANS CLUSTERING

No.	This research (FODC)	Key phrase based K-means approach
1	Extracts representative information	Extracts general phrases
2	Ontological structure is applied to present knowledge	Key phrase sentence fragments are used to present knowledge
3	Documents provide various views including the main concepts and details	Documents provide a key phrase view point (all key phrases are equal)
4	Ontology carries meanings and relations	Key phrases are less meaningful
5	The ontologies derived by experts can be reused and store different types of domain knowledge	Experts need to carefully monitor the key phrase extraction process and delete meaningless phrases
6	The iterative re-learning mechanism provides a better way to maintain the terminology base and ontological structure	Key phrases are extracted for every operation
7	FODC's F-measure is higher than the key phrase based K-means approach	K-mean's F-measure is lower than the FODC approach.
8	The FODC outperforms K-means as demonstrated using Shannon's Entropy and the Ecs evaluation	The K-means method is not as effective as FODC

where P_{ij} is the probability that a number of cluster j belongs to class i. After the Shannon's entropy values of different clusters are calculated, the entropy value of the entire clustering solution (E_{cs}) is generated as

$$E_{\text{cs}} = \sum_{j=1}^{m} \frac{n_j * E_j}{n} \tag{11}$$

where n_j is the size of cluster j, m is the number of clusters, and n is the total number of patent documents. A methodology with a smaller E_{cs} indicates superior clustering capabilities than the one with a larger E_{cs}.

Fig. 13 shows the comparison between the ontology-based fuzzy clustering method and the key-phrase-based K-means approach of CMP case. The results show that the FODC approach outperforms the K-means approach in precision, recall, F-measure, Shannon's entropy, and E_{cs}. Furthermore, K-means clustering causes the Type 2 cluster (composition of slurry) to include some patents that belong to the Type 1 cluster (mechanical design). Fig. 14 shows the result of patent news content case. The errors in patent news clustering using the K-means approach are due to the inclusion of insignificant key phrases (e.g., infringement, license, trade, and new patent) applied as the basis for clustering. For example, patent news belonging to Type 2 clusters (patent licensing) may contain fewer key phrases, which results in the placement of news articles into incorrect clusters. Finally, the result of the RFID case is shown in Fig. 15. Based on the experiments, the differences between the FODC and the K-means approach are summarized in Table VI, with all cases demonstrating the superior clustering results of using the FODC approach.

V. CONCLUSION

Traditionally, methodologies process knowledge documents using key phrases. However, a phrase can represent many meanings, and many different phrases can represent the same meanings. In this correspondence, we analyze the grammar of the sentences and derive the ontology of documents. Then, the relationships between documents are inferred, and the document similarities and differences are compared. A fuzzy ontology-based methodology for clustering knowledge documents (the FODC methodology) is presented and compared to the frequently used key-phrase K-means approach. The benchmarking results demonstrate that the FODC approach outperforms the K-means clustering approach and provides R&D managers with a new and beneficial approach for IP and innovation management.

ACKNOWLEDGMENT

The authors would like to thank L. T. Chang for the coding contribution and the reviewers and the associate editor for providing the comments and suggestions.

REFERENCES

[1] M. Fattori, G. Pedrazzi, and R. Turra, "Text mining applied to patent mapping: A practical business case," *World Pat. Inf.*, vol. 25, no. 4, pp. 335–342, Dec. 2003.

[2] R. N. Kostoff, D. R. Toothman, H. J. Eberhart, and J. A. Humenik, "Text mining using database tomography and bibliometrics: A review," *Technol. Forecast. Soc. Change*, vol. 68, no. 3, pp. 223–253, Nov. 2001.

[3] J. L. Hou and C. A. Chan, "A document content extraction model using keyword correlation analysis," *Int. J. Electron. Bus. Manag.*, vol. 1, no. 1, pp. 54–62, 2003.

[4] C. G. Nevill-Manning, I. H. Witten, and G. W. Paynter, "Lexically-generated subject hierarchies for browsing large collections," *Int. J. Digit. Libr.*, vol. 2, no. 2/3, pp. 111–123, Sep. 1999.

[5] I. H. Witten, "Adaptive text mining: Inferring structure from sequences," *J. Discret. Algorithms*, vol. 2, no. 2, pp. 137–159, Jun. 2004.

[6] S. N. Sanchez, E. Triantaphyllou, and D. Kraft, "A feature mining based approach for the classification of text documents into disjoint classes," *Inf. Process. Manag.*, vol. 38, no. 4, pp. 283–604, Jul. 2002.

[7] F. Feng and B. W. Croft, "Probabilistic techniques for phrase extraction," *Inf. Process. Manag.*, vol. 37, no. 2, pp. 199–220, Mar. 2001.

[8] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proc. 32nd Annu. Meeting Assoc. Comput. Linguist.*, Las Cruces, NM, Jun. 27–30, 1994, pp. 133–138.

[9] C. C. Kung, "Personalized XML information service system with automatic object-oriented ontology construction," M.S. thesis, Dept. Comput. Sci. Inform. Eng., Nat. Cheng Kung Univ., Tainan, Taiwan, 2000.

[10] C. C. Kao, "Personalized information classification system with automatic ontology construction capability," M.S. thesis, Dept. Comput. Sci. Inform. Eng., Nat. Cheng Kung Univ., Tainan, Taiwan, 2000.

[11] M. Grüninger and M. S. Fox, "Methodology for the design and evaluation of ontologies," in *Proc. Workshop Basic Ontological Issues Knowl. Sharing—In International Joint Conference on Artificial Intelligence*, Montreal, QC, Canada, 1995.

[12] C. S. Lee, Y. J. Chen, and Z. W. Jian, "Ontology-based fuzzy event extraction agent for Chinese E-news summarization," *Expert Syst. Appl.*, vol. 25, no. 3, pp. 431–447, Oct. 2003.

[13] T. A. Runkler and J. C. Bezdek, "Web mining with relational clustering," *Int. J. Approx. Reason.*, vol. 32, no. 2/3, pp. 217–236, Feb. 2003.

[14] F. C. Hsu, A. J. C. Trappey, C. V. Trappey, J. L. Hou, and S. J. Liu, "Technology and knowledge document cluster analysis for enterprise R&D strategic planning," *Int. J. Technol. Manag.*, vol. 36, no. 4, pp. 336–353, Jul. 2006.

[15] Accessed 11/04/2007. [Online]. Available: http://protege.stanford.edu/plugins/ontoviz/ontoviz.html

[16] MontyLingua: A Free, Commonsense-Enriched Natural Language Understander for English. Accessed 11/04/2007. [Online]. Available: http://web.media.mit.edu/~hugo/montylingua/

[17] N. F. Noy and D. L. McGuinness, "Ontology development 101: A guide to creating your first ontology," Knowl. Syst. Lab., Stanford Univ., Stanford, CA, Tech. Rep. KSL-01-05, 2001.

[18] Protégé, Wiki. Accessed 13/11/2007. [Online]. Available: http://en.wikipedia.org/wiki/Main_Page

[19] M. Marcus, B. Santorini, and M. Marcinkiewicz, "Building a large annotated corpus of english: The penn treeBank," *J. Comput. Linguist.*, vol. 19, no. 1, pp. 313–330, Jun. 1993.

[20] E. H. Mamdani, "Application of fuzzy logic to approximate reasoning using linguistic synthesis," in *Proc. 6th Int. Symp. Multiple-Valued Logic*, Logan, UT, 1976, pp. 196–202.

[21] E. H. Mamdani, "Application of fuzzy algorithm for control of simple dynamic plant," *Proc. Inst. Elect. Eng.*, vol. 121, no. 12, pp. 1585–1588, Dec. 1974.

[22] A. J. C. Trappey, C. V. Trappey, and B. H. S. Kao, "Automated patent document summarization for R&D intellectual property management," in *Proc. CSCWD*, Nanjing, China, May 3–5, 2006, pp. 1–6.

[23] C. J. van Rijsbergen, *Information Retrieval*. London, U.K.: Butterworth, 1979.

[24] G. Salton, *Automatic Text Processing: The Transformation Analysis, and Retrieval of Information by Computer*. New York: Addison-Wesley, 1989.

[25] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques, text mining workshop," in *Proc. 6th ACM SIGKDD Int. Conf. Knowl. Discov.* Data *Mining KDD*, Boston, MA, Aug. 20–23, 2000, pp. 1–20.

[26] H. Li, K. Zhang, and T. Jiang, "Minimum entropy clustering and applications to gene expression analysis," in *Proc. 3rd IEEE Comput. Syst. Bioinform. Conf.*, Stanford, CA, 2004, pp. 142–151.