A fuzzy clustering approach for finding similar documents using a novel similarity measure

Article *in* Expert Systems with Applications · October 2007

DOI: 10.1016/j.eswa.2006.06.002 · Source: DBLP

3 authors:

Rıdvan Saraçoğlu
Yuzuncu Yil University
8 PUBLICATIONS **113** CITATIONS
SEE PROFILE

Kemal Tütüncü
Selcuk University
8 PUBLICATIONS **100** CITATIONS
SEE PROFILE

Novruz Allahverdi
Karatay Univeristy
91 PUBLICATIONS **719** CITATIONS
SEE PROFILE

Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 33 (2007) 600–605
Expert Systems with Applications
www.elsevier.com/locate/eswa

A fuzzy clustering approach for finding similar documents using a novel similarity measure

Rıdvan Saraçoğlu *, Kemal Tütüncü, Novruz Allahverdi

Department of Electronic and Computer Education, Selçuk University, 42031 Konya, Turkey

Abstract

Searching for similar documents has a crucial role in document management. This paper aims for developing a fast and high quality method of searching similar documents based on fuzzy clustering in large document collections. In order to perform these requirements, a two layers structure is proposed. Formerly, finding the similarity in documents is based on the strategy that uses word-by-word comparison. The proposed method in this study uses two layers structure and lets the documents pass through it to find the similarities. In this system, predefined fuzzy clusters are used to extract feature vectors of related documents for finding similar documents of them. Similarity measure is estimated based on these vectors. To do this, a distance based similarity measure is proposed. It has been seen in empirical results that the proposed system uses new similarity measure and has better performance compared with conventional similarity measurement systems.

Keywords: Text mining; Document similarity; Fuzzy clustering; Fuzzy similarity measure; Distance based similarity

1. Introduction

With advanced technology, huge a amount of data is processed, transferred and kept in variety of sectors. Financial, medicine and information fields can be shown as examples of these sectors. The subject of the data mining is to process great amount of data and to extract useful information from them. As days passes importance of data mining is increasing rapidly. Big significant amount of data is text data. Books, papers, and journals are transferred to electronics environment and millions of web pages can be shown as examples.

Today, varieties of solution methods for managing and organizing huge amount of text documents and obtaining useful information from these data (text mining) are in searching progress. Apart from these, different techniques are used in the fields like natural language processing, extracting main theme from text and building a summary of a text (Guzman, 1998; Martinez-Trinidad, Beltran-Martinez, & Ruiz-Shulcloper, 2000).

These text mining applications generally include stages like pre-processing, classification and clustering. In order to implement these stages, a wide variety of techniques such as statistical, machine learning and other decision making techniques have been used. The most important of them are decision trees, artificial neural networks, genetic algorithms, rude clustering and content learning (Thuraisingham, 1999).

Searching of similar documents has an important role in text mining and document management. Varieties of intelligence searching techniques that also include artificial intelligence have been used for this aim.

Basically two different approaches exist for searching similar documents (Han & Kamber, 2001). The first is to extract keywords from document (keyword based approach) (Klose, Nürnberger, Kruse, Hartmann, & Richards, 2000; Weng & Lin, 2003; Weng & Liu, 2004). Later on,

* Corresponding author. Tel.: +90 332 223 33 35; fax: +90 332 241 21 79.
E-mail addresses: ridvan@selcuk.edu.tr (R. Saraçoğlu), ktutuncu@selcuk.edu.tr (K. Tütüncü), noval@selcuk.edu.tr (N. Allahverdi).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.06.002

via this, extraction similarity is searched. Other approach is based on using all the words in the document. Property vector of the document is determined with the help of all the terms that exist in the document and similarity is searched (Dhillon, Fan, & Guan, 2001; Kou & Gardarin, 2002; Widyantoro & Yen, 2000). This property between all terms and all documents is presented by using vector space model.

Although keyword extracting makes decision making process faster, constituting property vector from all terms makes the search more certain.

Document clustering and classification is not only being used in searching of similar documents but also in other text mining applications. Therefore, previous studies mainly focused on clustering and classification fields. Varieties of methods have been used for clustering and classification. The most important of them are inductive decision tree (Apte, Damerau, & Weiss, 1998; Quinlan, 1986), Bayesian (Sahami, Heckerman, & Horvitz, 1998; Tzeras & Hartmann, 1993), neural net (Wiener, Pederson, & Weigend, 1995), k-nearest neighbor (Masand, Linoff, & Waltz, 1992; Tan, 2005, in press; Weng & Lin, 2003), neighbor-weighted k-nearest neighbor (Tan, 2005a, 2005b), spherical k-means (Dhillon et al., 2001), support vector machine (Dumais, Platt, Heckerman, & Sahami, 1998; Joachims, 1997), self-organizing map (Klose et al., 2000; Yang & Lee, 2004; Yang & Lee, 2005) and fuzzy logic (Miyamoto, 2001; Widyantoro & Yen, 2000) approaches.

In this study, a clustering method that is implemented by using fuzzy logic is used. Later on, document similarity value is determined via property vector that is obtained from this clustering.

This paper is organized as follows:

In Section 2, framework of this study and related background theory are presented. In Section 3, methodology of the study and system architecture are explained. In Section 4, results of the experiments and analysis of these results are presented. In Section 5, summary of the subject and discussion related with future studies are given.

2. Study environment and background theories

In this chapter, definition of the problem and existing methods are presented.

2.1. Problem description

Generally, an index structure based on keywords is used in search mechanisms. Great amount of collection can be searched by this way. As it can be seen in this study, if all terms in the documents are included in the search operation then search space size gains importance. In order to do an effective search, a structure for speeding up the research operation is required. The aim of this study is to extract property vectors for documents with the help of fuzzy clustering. Later on similarity measure will be done over this vector. A new proposed similarity measure is suggested to increase efficiency of the current system instead of traditional similarity measures.

2.2. Fuzzy clustering

One of the important parts of the similarity document search is to form document classification or clustering. In this field, wide varieties of methods have been used for so many years but recently, fuzzy clustering (Miyamoto, 2001) and classification based on fuzzy similarity (Masand et al., 1992) attract attention.

Term-document values in fuzzy clustering matrix are used to form fuzzy multi clusters. Later on, uniqueness measure between these fuzzy multi clusters is defined. Clustering operation with respect to fuzzy c-averages algorithms is completed by using these measures (Miyamoto, 2001).

Fuzzy clustering that is also used in this study depends on fuzzy similarity. Firstly, a supervised learning process is considered at the stage of implementation of this clustering. Categories that subject training data belong to are known or determined forehand. Clustering is made by using this information. In order to constitute fuzzy term category relations, a relationship between terms and categories is defined.

2.2.1. Fuzzy term-category relations

Determining the relations between $T = \{t_1, t_2, \ldots, t_n\}$ terms and $C = \{c_1, c_2, \ldots, c_m\}$ categories is the basic problem here. Term-category relation is defined as follows (Miyamoto, 1990).

Term-category matrix denoted as R ($R = T \times C \to [0, 1]$) will be constituted with the help of a given D training cluster. Every member of this matrix will show the membership degree of appropriate term that belongs to appropriate category. $\mu_R(t_i, c_j)$ shows the membership degree where t_i is the ith term and c_j is the jth category. Training cluster that will determine this membership is denoted as $D = \{(d_1, c(d_1)), (d_2, c(d_2)), \ldots, (d_n, c(d_n))\}$ and includes n documents and d_i shows ith document and $c(d_i)$ shows the categories that the ith document belongs to. Every document in the training cluster is a kind of term-frequency pair's cluster. t_i shows the term number that being exists in the document and w_i shows value in the formula $d = \{(t_1, w_1), (t_2, w_2), \ldots, (t_m, w_m)\}$.

$\mu_R(t_i, c_j)$ membership values are calculated as follows (Widyantoro & Yen, 2000):

$$\mu_R(t_i, c_j) = \frac{\mathrm{dist}(t_i, c_j)}{\max_{q,r}(t_q, c_r)}$$
$$\mathrm{dist}(t_i, c_j) = \frac{\sum_{w_i \in d_k \wedge d_k \in D \wedge c(d_k) = c_j} w_i}{\sum_{w_i \in d_k \wedge d_k \in D} w_i} \quad (1)$$

All training documents are grouped according to their categories. If documents belong to more than one category then they will take place in the group of each of their category. For the terms, numbers of that exists in the group will be added. The value of dist(t_i, c_j) will be found by dividing total number of t_i that exists in c_j category to number of

t_i that exists in all training clusters. If this term is only seen in one category, it is obvious that value of dist(t_i, c_j) will be 1.0. The more number belonging category causes less value for dist(t_i, c_j). Finding final value of the term $\mu_R(t_i, c_j)$ will be as by dividing every dist(t_i, c_j) that is found for each term to the maximum value between them.

Henceforth, term-category matrix and indirect membership degree of terms with respect to categories are determined.

2.2.2. Fuzzy similarity measure

The aim is to find out category of the document. So that, test document is processed with term-category matrix. In this operation, the main aim is just the measurement of similarities of test document and category cluster centers. If C_j is accepted as center of cluster category j then all terms membership degrees in j category exists in R term-category vector form c_j.

If we accept test document as $d = \{(t_1, \mu_d(t_1)), (t_2, \mu_d(t_2)), \ldots, (t_m, \mu_d(t_m))\}$ then $\mu_d(t_i)$ shows membership degrees of t_i that belong to d. An effective approach for determining value of $\mu_d(t_i)$ is to accept it as one, if term exists in document otherwise zero, another approach can be the ratio of every term to term that is most frequently existing in the document.

For a given $R(T \times C)$ term-category vector, similarity between d document and c_j category cluster center can be calculated as follows (Widyantoro & Yen, 2000):

$$\mathrm{sim}(d, c_j) = \frac{\sum \mu_R(t, c_j) \otimes \mu_d(t)}{\sum \mu_R(t, c_j) \oplus \mu_d(t)} \tag{2}$$

$\otimes$ and $\oplus$ are fuzzy conjunction and fuzzy disjunction operators, respectively. Different formulas exist for calculating these fuzzy operators. The most important of them can be seen in Table 1. The best performance between operators exists in Table 1 which belongs to Algebraic operator pairs in classification (Widyantoro & Yen, 2000).

In the upper formulas, $\mu_d(t_i)$ shows the membership degree of t_i terms belonging to d. To find out this value some approaches can be used. It can be found as by using term frequency inverse document frequency (TFIDF) values or accepted as 1 if the term exists in document, otherwise 0.

The cluster that test document belongs to is determined by choosing biggest value of similarities of d document to every cluster center.

2.3. Similarity measure

The most commonly used similarity measures in similar document search are cosines and dice similarity measure.

Table 1
Some fuzzy operators

	t-norm(x, y)	t-conorm(x, y)
Einstein	$\frac{xy}{2-(x+y-xy)}$	$\frac{xy}{1+xy}$
Algebraic	x, y	$x + y - xy$
Min–max	Min(x, y)	Max(x, y)

For example, cosine similarity regards the angle between two points as basis. In addition, for higher dimensional data, a popular measure is the Minkowski metric (Ichino & Yaguchi, 1994). These measurements are presented as follows:

$$\mathrm{CSim}(x_i, x_j) = \frac{\sum_{k=1}^{m} x_{ik} x_{jk}}{\sqrt{\sum_{i=1}^{m} (x_{ik})^2 \sum_{i=1}^{m} (x_{jk})^2}} \tag{3}$$

$$\mathrm{DSim}(x_i, x_j) = \frac{2\sum_{k=1}^{m} x_{ik} x_{jk}}{\sum_{i=1}^{m} (x_{ik})^2 + \sum_{i=1}^{m} (x_{jk})^2} \tag{4}$$

$$\mathrm{Minkowski}_p(x_i, x_j) = \left(\sum_{k=1}^{m} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \tag{5}$$

where x is property vector, m is the dimension of the data.

In the Minkowski metric, the *Euclidean* distance is a special case where $p = 2$, while *Manhattan* metric has $p = 1$. However, there are no general theoretical guidelines for selecting a measure for any given application.

3. System architecture

The aim of similar document search system is to find documents that are most similar to input document. Input document means a question in traditional question–answer system (Clarke, Cormack, Kisman, & Lynam, 2000; Elworthy, 2000). In this study, input document means a document that has same shape with searched documents.

Firstly, the system pre-process input document. Document property vector is obtained from all the terms that the document includes. Later on, for the document that is subjected to clustering, clusters that it belongs to can be found with respect to pre-determined threshold value. Documents in this cluster are called as *candidate documents*. Similarity is determined by comparing category property vector that includes degrees of documents belonging to clusters and input document category property vector. Distance based similarity measurement is defined for this comparison.

The system has architecture as shown in Fig. 1. The aim in pre-processing stage is to determine terms and the frequency of existence of these terms with respect to the documents. Currently four methods exist for extraction of terms according to Murata et al., 2000 (Weng & Lin, 2003). These are; only usage of shortest terms, usage of all terms pattern, usage of a lattic and down-weighting methods. First method is chosen so as to simplify the process. The document is broken into as words and these words are taken as terms in this method. Later on, during the information extraction stage, unimportant words (stop-words) are selected and taken out from the document. Remaining words are stemmed. Latest obtained words are accepted as terms. Frequency of terms in each document is calculated. Fuzzy clustering system that is trained by former document collection exists. Input document whose terms frequency is calculated is directed to candidate documents via this clustering system. So, first stage is completed.

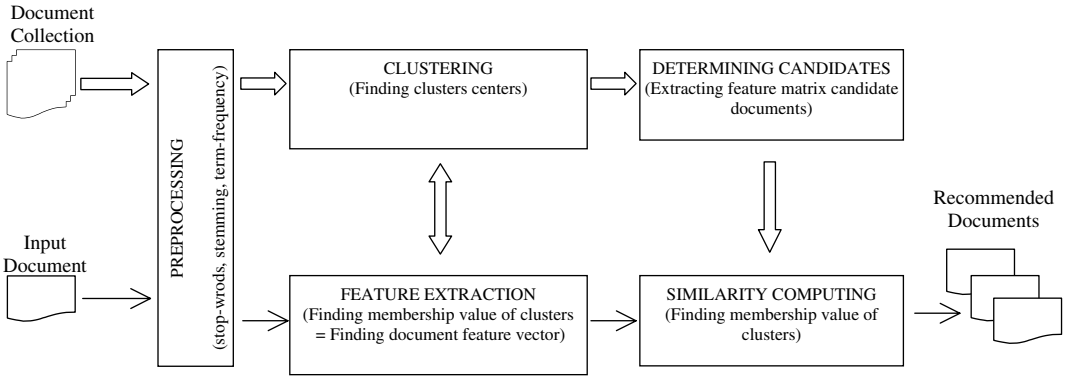

Fig. 1. System architecture.

The next stage's aim is to order candidate documents according to their similarities. Determined candidate documents and input document are subjected to similarity measurement that depends on distance. Thus, similarity values are found. Candidate documents that pass previously determined threshold value are suggested as similar documents of input document.

Here, suggested distance based similarity measure (Dimension Root similarity) is as follows:

$$\mathrm{DRSim}(x_i, x_j) = \left(\frac{\sum_{k=1}^{m} |x_{ik} - x_{jk}|^2}{m} \right)^{\frac{1}{m}} \quad (6)$$

where x_i and x_j show feature vector of compared documents. Dimension of the property vector is denoted by m.

As a result, the document whose similarity is searched for is not compared with all current collection, but is directed to pre-determined candidate group under the framework of previously prepared clustering. Later on, it is compared with previously determined limited number of clusters for similarity at the level of the belonging degree instead of comparison at the level of term that may be hundreds or more. Thus, similarity determination will be done in a short time via using property vector that is obtained from all terms.

4. Evaluation

Two different document collections are used in this study for application aims. The first one is Routers-21,578 distribution 1.0 that is commonly used in text mining. This collection includes 21,578 documents that have more than 135 subjects. Some of the current 135 topics or subjects exist in very few numbers of documents. So, 10 topics whose frequency takes place is at most has been chosen among 135 topics. There are 8595 documents belonging to the chosen topics and 6456 of them have been used as training data and the rest of them for test.

Second collection constitutes 4020 abstracts that belong to three different categories. These are computer science related "computer collection" that includes 1587 abstracts; medical related "Medlars collection" that includes 1033 abstracts and aerodynamics related "Cranfiled collection" that includes 1400 abstracts. There are 4020 documents belonging to these collections and 3015 of them have been used as training data and the rest of them for test.

Chosen training data has been pre-processed. Firstly, 350 stop-words have been taken out from these documents. For stemming of words, commonly used Porter Stemmer algorithm has been chosen (Jones & Willett, 1997). As a result of this, documents have been clustered according to the stem words that they include.

Selected training data are subjected to clustering operation with fuzzy similarity method like Formulas (1) and (2). MatLab 7.0 is software package that is used for preparing for application programs.

The following method is applied for comparison of similarity measure (Formula (3)–(6)). A collection is built by choosing hundred text pairs that belong to the same cluster. The scores of this collection are calculated according to several similarity measures.

In the same way, another collection is built by choosing in random hundred text pairs that belong to different clusters. Same operation is done for this collection as well.

In order to compare similarity measure, every method's value obtained from "same cluster collection" is divided by same method's value obtained from "different cluster collection".

The values used in time comparison are total comparison duration of documents belonging to the "same" or "different" category. So each value is the duration of results of the total 200 comparisons.

These operations are repeated 10 times and average values are obtained.

Having done the experiment or test has shown that proposed similarity measurement method has higher similarity ratio than current methods. Average similarity values of chosen document can be seen in Tables 2 and 3.

Compared similarity methods are dimension root similarity (DRSim), dice similarity (DSim), cosine similarity (CSim), Manhattan similarity (MSim), Euclidean similarity (ESim), Minkowski similarity p is chosen as dimension (MDSim), Minkowski similarity p is chosen as 20 (M20Sim), Minkowski similarity p is chosen as 50 (M50Sim).

Table 2
Results of the experiment that belong to Router collection

Average of similarities	Same cluster	Different cluster	Ratio between different and same cluster	Total time (ms)
DRSim	0.405599	0.260450	1.5573	7.673
DSim	0.952949	0.734766	1.2969	10.407
CSim	0.955275	0.739114	1.2925	10.736
MSim	0.929763	0.819675	1.1343	6.986
ESim	0.911628	0.767758	1.1876	7.455
MDSim	0.856086	0.617108	1.3873	12.891
M20Sim	0.840893	0.577592	1.4559	13.202
M50Sim	0.830160	0.549523	1.5107	13.735

Table 3
Results of experiment belonging to second collection

Average of similarities	Same cluster	Different cluster	Ratio between different and same cluster	Total time (ms)
DRSim	0.784886	0.472734	1.6603	7.080
DSim	0.979118	0.766672	1.2771	7.469
CSim	0.980513	0.768626	1.2757	7.023
MSim	0.907535	0.655913	1.3836	7.078
ESim	0.895316	0.613923	1.4583	7.220
MDSim	0.887064	0.590788	1.5014	7.938
M20Sim	0.860155	0.524840	1.6389	8.734
M50Sim	0.851423	0.512209	1.6623	9.718

DRSim similarity measurement is more effective than other similarity measurements and traditional Cosine similarity measurement. This efficiency is seen obviously in "ratio" column of Table 2. In this column, the ratio of similarity values of documents belonging to same cluster to similarity values of documents belonging to different clusters is presented.

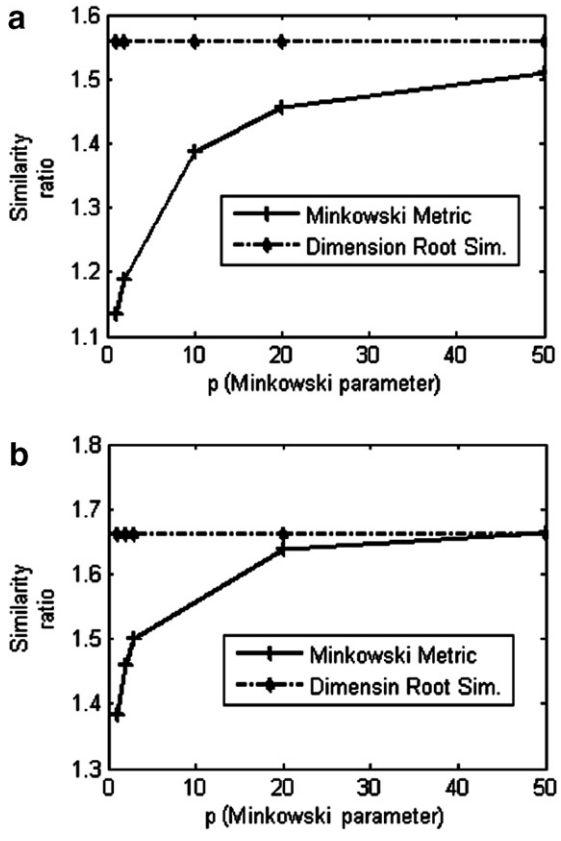

Fig. 2. (a) Reuter collection (10 categories); (b) second collection (3 categories). Similarity ratios between Minkowski metric and DRSim.

Efficiency comparisons of proposed similarity measurement are seen in Figs. 2 and 3. As can be seen in Fig. 2, the more p parameter values, the more similarity measurement efficiency (increasing p parameter value causes to increase efficiency of similarity).

DRSim and Minkowski metric approximately shows very close results when value of p becomes 50. But when p parameter value of Minkowski is taken as value of the category in collection (10 for Reuter, 3 for second collection), proposed DRSim shows better results. This result is better as 12% for Reuter and 10% for second collection.

In Fig. 3, time comparison is presented. As can be seen clearly for all the values where p parameters of Minkowski metric are greater then 2, DRSim gave better results. Again, DRSim is faster as 68% for Reuter and 12% for second collection, when p parameter of Minkowski is based on numbers of categories of both collections.

Value of p is chosen till 50 in Minkowski measure. Proposed new similarity measure presents better results even if value of p ranges from 1 to 50.

5. Conclusion and future works

Similar document search takes important role in the fields of text mining and document management. Former studies have focused on comparison of clustering and classification methods. In some studies, effectiveness of these algorithms have been tried to be increased. In this study, a system depends on fuzzy clustering when similar

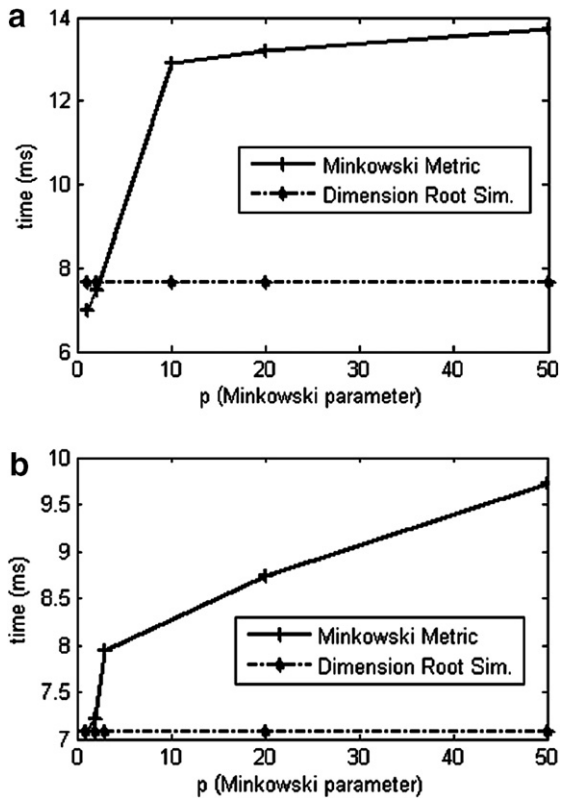

Fig. 3. (a) Reuter collection (10 categories); (b) second collection (3 categories). Time comparison between Minkowski metric and DRSim.

document search is proposed. More, distance based similarity measurement for this system is defined to be used in the similarity measurement. According to experimental results, performance of the proposed DRSim similarity measurement is better than traditional similarity measurement performance.

Current method, a document which is at the process of training in cluster operation can belong to more than one category. But for the document under test stage it is assumed to belong to only one cluster. Final comparison operation is done by documents in this determined cluster. For the future works, it can be stressed on the possibility of belonging to more than one cluster for this document. From this point of view, a new method can be developed.

References

Apte, C., Damerau, P., & Weiss, S. (1998). Text mining with decision rules and decision trees. In *Proceedings of the conference automated learning and discovery*, CMU.

Clarke, C. L. A., Cormack, G. V., Kisman, D. I. E., & Lynam, T. R. (2000). Question answering by passage selection. In *The ninth text retrieval conference, Gaithersburg*.

Dhillon, I. S., Fan, J., & Guan, Y. (2001). Efficient clustering of very large document collections. *Data Mining for Scientific and Engineering Applications*.

Dumais, S., Platt, J., Heckerman, D., & Sahami, M. (1998). Inductive learning algorithm and representations for text categorization. In *Proceedings of the 1998 ACM 7th international conference on information and knowledge management* (pp. 148–155).

Elworthy, D. (2000). Question answering using a large NLP system. In *The ninth text retrieval conference, Gaithersburg, 2000*.

Guzman, A. (1998). Finding the main themes in a Spanish document. *Expert Systems with Applications, 14*, 139–148.

Han, J., & Kamber, M. (2001). *Data mining: concepts and techniques*. Morgan Kaufman Publishers.

Ichino, M., & Yaguchi, H. (1994). Generalized Minkowski metria for mixed feature-ljpe data analysis. *IEEE Transactions On Systems, Man, and Cybernetics, 24*(4).

Jones, K., & Willett, P. (1997). *Readings in information retrieval*. Morgan Kaufmann Publishers.

Joachims, T. (1997). Probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. In *Proceedings of the international conference on machine learning (ICML'97)* (pp. 143–151).

Klose, A., Nürnberger, A., Kruse, R., Hartmann, G., & Richards, M. (2000). Interactive text retrieval based on document similarities. *Physics and Chemistry of the Earth (A), 25*(8), 649–654.

Kou, H., & Gardarin, G. (2002). Similarity model and term association for document categorization. In *Proceedings of the 13th international workshop on database and expert systems applications (DEXA'02)*.

Martinez-Trinidad, J. F., Beltran-Martinez, B., & Ruiz-Shulcloper, B. (2000). A tool to discover the main themes in a Spanish or English document. *Expert Systems with Applications, 19*, 319–327.

Masand, B., Linoff, G., & Waltz, D. (1992). Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual ACM/SIGIR conference on research and development in information retrieval* (pp. 59–65).

Miyamoto, S. (1990). *Fuzzy sets in information retrieval and cluster analysis*. Kluwer Academic Publisher.

Miyamoto, S. (2001). Fuzzy multisets and fuzzy clustering of documents. In *Proceedings of the IEEE International Conference on Fuzzy Systems, FUZZ-IEEE*.

Murata, M., Ma, Q., Uchimoto, K., Ozaku, H., Utiyama, M., & Isahara, H. (2000). Japanese probabilistic information retrieval using location and category information. In *Proceedings of the fifth international workshop on information retrieval with Asian language*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning Journal, 1*, 81–108.

Sahami, M., Dumais S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In *AAAI 98, workshops on text categorization*.

Tan, S. (2005a). Neighbor-weighted *K*-nearest neighbor for unbalanced text corpus. *Expert Systems with Applications, 28*, 667–671.

Tan, S. (2005b). An effective refinement strategy for KNN text classifier. *Expert Systems with Applications, 30*, 290–298.

Thuraisingham, B. (1999). *Data mining: Technologies, techniques, tools, and trends*. CRC Press.

Tzeras, K., & Hartmann, S. (1993). Automatic indexing based on Bayesian inference networks. In *Proceedings of the 16th annual ACM/SIGIR conference on research and development in information retrieval* (pp. 22–34).

Weng, S. S., & Lin, Y. J. (2003). A study on searching for similar documents based on multiple concepts and distribution of concepts. *Expert Systems with Applications, 25*(3), 355–368.

Weng, S. S., & Liu, C. K. (2004). Using text classification and multiple concepts to answer e-mails. *Expert Systems with Applications, 26*(4), 529–543.

Widyantoro, D. H., & Yen, J. (2000). *A fuzzy similarity approach in text classification task*. IEEE.

Wiener, E., Pederson, J., & Weigend, A. (1995). A neural network approach to topic spotting. In *Fourth annual symposium on document analysis and information retrieval*.

Yang, H. C., & Lee, C. H. (2004). A text mining approach on automatic generation of web directories and hierarchies. *Expert Systems with Applications, 27*, 645–663.

Yang, H. C., & Lee, C. H. (2005). A text mining approach on automatic construction of hypertexts. *Expert Systems with Applications, 29*, 723–734.