World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:8, 2013

A Study on Finding Similar Document with Multiple Categories

R. Saraçoğlu and N. Allahverdi

Open Science Index, Computer and Information Engineering Vol:7, No:8, 2013 publications.waset.org/16229/pdf

Abstract**—**Searching similar documents and document management subjects have important place in text mining. One of the most important parts of similar document research studies is the process of classifying or clustering the documents. In this study, a similar document search approach that includes discussion of out the case of belonging to multiple categories (multiple categories problem) has been carried. The proposed method that based on Fuzzy Similarity Classification (FSC) has been compared with Rocchio algorithm and naive Bayes method which are widely used in text mining. Empirical results show that the proposed method is quite successful and can be applied effectively. For the second stage, multiple categories vector method based on information of categories regarding to frequency of being seen together has been used. Empirical results show that achievement is increased almost two times, when proposed method is compared with classical approach.

Keywords**—**Document similarity, Fuzzy classification, Multiple categories, Text mining.

I. Introduction

AS the result of developing technology and information sharing that becomes widespread, huge amount of data in wide varieties of fields is appeared in each day. To process such a large amount of data and extraction of useful information from this data, are the main aims of data mining.

Regarding to the structure of these data, it can be seen that majority of data consist of textual data. The journals, articles, billions of web pages and books which are transferred to electronic media, can be shown as textual data examples. So, the organizing and management of semi-structured or unstructured data and the knowledge or rule discovery from these (text mining), has became an important study field in recent years. Using text classification techniques for replaying the e-mails [1], the studies about question answering [2] can be given as examples.

The searching operation on the textual documents is very important nowadays. Both the searching in a very large field (internet search engines) and the searching in a particular field (like searching in a library) can be met frequently in daily life. So, increasing the speed or success of the searching operations, has became a major problem in the text mining process.

R. Saraçoğlu is with the Department of Electrical Electronic Engineering, Yuzuncu Yil University, 65000, Van, Turkey (corresponding author to provide phone 90-432-225-1725 ; e-mail: ridvansaracoglu@yyu.edu.tr).

N. Allahverdi is with the Department of Computer Engineering, Faculty of Technology, Selçuk University, 42003 Konya, Turkey (e-mail: noval@selcuk.edu.tr).

The core of finding similar documents consists of document clustering and classification. Majority of the previous studies have focused on this subject. There exist many of methods related with classification and clustering. Inductive decision tree [3], [4], Bayesian [5], [6], neural net [7], k-nearest neighbor [8]-[10], neighbor-weighted k-nearest neighbor [11], spherical k-means [12], support vector machine [13], [14], self-organizing map [15]-[17] and fuzzy logic [18], [19] can be given as examples.

The main aim of this study is to create a search framework by using all the words that textual documents with multiple categories contain. This process was discussed in three stages. These stages are preprocessing, clustering and extracting feature vector [20] and similarity measure [21], [22].

Fuzzy logic based methods started gaining importance apart from many other methods related with classification and clustering. In this study, multiple categories state is discussed for similar document research subject. So, the problem need to be solved first is to be able to classify the documents with multiple categories. FSC is emphasized for this. This method is tried to be developed so as to make it has property of solving multiple categories problem.

The paper is organized as follows: Section II has brief summary of former studies about this subject. In Section III, the proposed methods that refer to multiple categories problem are explained in details. The empirical results and their analyses are given in Section IV. The related discussions and future studies exist in Section V and that is the last section as well.

II. Related Work

The field that text mining works on it is semi-structured or unstructured textual documents. The study where Widyantoro and Yen [18] proposed a fuzzy similarity approach for textual classification can be given as an example of studies related with unstructured textual documents. The study where Garofalakis et al. [23] formed a Document Type Descriptor (DTD) based extraction model that processes on the semi-structured XML documents, can be given as an example of studies related with semi-structured textual documents.

The important part of document management and Information Retrieval (IR) studies are done for clustering and classification methods. In their study Zhao and Karypis [24], made a comparison of partitioned and agglomerative clustering approach. And then, they put forward a new clustering approach where both these methods were used.

Text categorization can be described as grouping documents into previously determined number of classes or

Open Science Index, Computer and Information Engineering Vol:7, No:8, 2013 publications.waset.org/16229/pdf

categories. The need of automatic text classification and the studies related to this field are increasing in each day. A text categorization process usually consists of training and test stages. In the training stage, a model is prepared by using the text whose categories are known. In the test stage, the class of a new text is determined by using the prepared model. One of the well-known methods that are used for text categorization and classification is Rocchio algorithm.

The most critical parts in text mining studies can be usually seen as clustering and classification. Today as many fields, fuzzy logic is also used in text mining and text classification. Fuzzy approaches are seen as fuzzy clustering [19] and FSC [18].

In fuzzy clustering, the values of term-document matrix are used to form fuzzy multi clusters. Then uniqueness measure between these fuzzy multi clusters is defined. Clustering operation is completed with respect to fuzzy c-means algorithm by using this measurement [19].

Fuzzy clustering which is also used in this study is based on fuzzy similarity. While implementing the clustering, a supervised learning process is considered, firstly. Therefore, the categories of training data are known beforehand and clustering is made with the help of this information. A relationship is defined between terms and categories to constitute fuzzy term category relations.

Considering finding similar documents, usually, an approach that has three stages namely preprocessing, clustering and similarity measure, is used [25], [9], [1], [26]. In the previous studies, improving one or a few of this stages and obtaining a more successful clustering approach or a more efficient similarity measure were tried to be put forward.

Some of the previous studies about searching similarity are the approaches which intend to enrich the research method. The studies about concepts of documents can be given as examples. In their study Weng and Lin [9] proposed an approach that uses multiple concepts and the distribution of concepts in a document for determination of similarity. The multiple concepts information has been extracted by using the similarities between document vectors and concepts vector. Then the distribution of concepts in the document has been formed. By combining the information of concepts similarity and concepts distribution, the similarity of documents has been determined.

It is also possible to meet the document similarity in the question answer systems (QASs). In the traditional QASs, where the document similarity is used, [25], [27], [28] question is taken and analyzed as a sentence or passage and separated into parts. Also in some systems, input document studied on is matched with a question in the QAS and investigated by separating it into terms [9].

III. Algorithm

In this study, multiple categories problem for finding similar documents is discussed. The multiple categories problem can be summarized as determination of the categories when a test document belongs to more than one category. Multiple categories problem can be discussed in two parts.

The first one is determination of which documents belong to more than one category. The second one is determination of the categories if a document belongs to more than one category. These two parts can be thought as independent from each other. The proposed methods for these two parts are α-threshold Fuzzy Similarity Classification Method (α-FSCM) and Multiple Categories Vector Method (MCVM) respectively [26].

Proposed α-FSCM for solution of the problem of which documents belong to more than one category, takes FSC as basis. Explanation of the method is as follows:

Traditionally, FSC contains calculation of cluster center vectors as the same number as categories. Training data is used for this calculation. These cluster center vectors express clustering. Every category is matched to a cluster. Fuzzy similarity is applied to test document with all these cluster center vectors. Thus, the values of belonging to clusters for the test document are found. Since cluster vectors and categories are matching one to one, the membership (belonging) values to categories for the test document are obtained.

A similar approach exists in the basis of α-FSCM that is proposed for the problem of determination of documents which belong to more than one category. A cluster center calculation (clustering) that includes two clusters (single and multiple category clusters) will be carried out. As a difference from the FSC approach, cluster centers do not match with two categories (single category class - multiple categories class) directly. Single category cluster value and multiple category cluster value that are calculated via cluster vectors will determine category after being passed through additional operations.

It is impossible to solve this single-multiple category classification by using classical FSC. Because, document collection is distributed into two classes but class differentiation cannot be appeared directly and clearly via this distribution. Besides, the documents which belong to multiple category class form a very small part of whole document. So, this makes difficult for classical FSC to distinguish classes clearly. Also the results obtained from the applications done for this study verify this.

Because of the reasons above, classical FSC approach is improved and adapted to multiple categories problem. Single and multiple cluster values are obtained from classical FSC. Then, multiple cluster values are divided to single cluster values and compared with a pre-determined threshold value of α. The ones which exceed this threshold are assigned to multiple category class, and the ones which do not exceed are assigned to single category class.

In proposed α-FSCM, determining threshold value gains importance. The generalized method that we propose for this threshold is explained as follows:

Let D is a training dataset, M and S are discrete subsets of D (M is the set of documents that belong to more than one category; S is the set of documents that belong to only one category).

$$M \cup S = D$$
$$M \cap S = \varnothing$$

$|M|$ and $|S|$ respectively denotes the number of documents that M and S document sets contain. Firstly, by applying classical FSC, $sim(d,c_M)$ and $sim(d,c_S)$ values (the membership value of multiple category cluster and the membership value of single category cluster respectively) are obtained for each d document in the document set D. $sim(d,c_M)$ values are divided to $sim(d,c_S)$ values and documents are listed in decreasing order according to these ratios. As the result of this ordering, first $|M|$ items of documents are the documents that will be assigned to multiple category class according to proposed α-FSCM. The ratio value of $|M|$th document in ordering is chosen as α threshold value that we search for. Because, when this value is chosen as threshold, documents will be separated into two groups accordingly. $|M|$ items of document which are equal and greater than this threshold will belong to multiple categories. Rest $|S|$ items of documents which are less than this threshold will belong to single category. The threshold value that is obtained via training dataset will also be used when classification is being done for test documents.

When a document belongs to more than one category, determination of these categories is the second part of the multiple categories problem.

In classical FSC method, for a document, the category that has the maximum value in document-category vector was the category which the document belongs to. If document belongs to more than one category, the categories which the document belongs to are the categories which have the maximum values respectively in the document-category vector. For example, if document belongs to two categories, the first category which the document belongs to is the one that has the maximum value and the second category is the one that has the maximum value after the first category. This is a classical approach. But it is seen that the classical approach was given unsuccessful results in the application made for this study.

MCVM is proposed instead of this classical approach. Determining the relationships between each other categories is the main problem here. Category-category relationship is used for solution. The detail of this method can be seen in [26].

IV. Empirical Study

In this study, a search approach that handles the state of being belonged to more than one category has been improved. In this section, firstly, the document collection that was used in empirical study and empirical methodology has been explained. Then the analyses of obtained results have been given place.

A. Document Collection

Most of the text collections which are used for text mining research have a hierarchical structure. The 20 Newsgroups data set (it was originally collected by Ken Lang), Industry Sector [29], Cora dataset [30] and etc. can be given as examples. Because of this hierarchical structure, documents don't belong to more than one category. There are high level and low level categories. This causes the membership of documents to categories to have crisp values. So, these kinds of text collections are not appropriate for being used in multiple categories problem.

In this study, multiple categories problem is discussed with a fuzzy approach. Therefore, hierarchical document collections are not appropriate for the structure of this study. The text document collection used here is Reuters-21578 distribution 1.0 that is often used for text mining researches. The property of this collection is to have the categories that aren't hierarchical. Since some documents in the collection belong to more than one category at the same time, the collection is appropriate for the nature of research. This collection contains 21578 documents having over 135 topics. Some of the current 135 topics exists in a very few numbers of documents. So, 10 of 135 topics whose frequency of existing is at most have been chosen for being used in the present study. There are 8595 documents belonging to these chosen topics and 6456 of them have been used as training data, whereas 2139 have been used as test data. 622 items of training data documents and 221 items of test data documents belong to more than one topic.

Chosen training data have 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 preferred [31]. As a result of this, documents will be clustered according to the stem words, in other words, with respect of the terms that they include.

It is necessary to evaluate performance of the new classification method that is improved in the present study. Precision-recall method has been used for this evaluation.

B. Implementation of α-FSCM

For preparation of application programs in the present study, Matlab 7.0 software package has been used. As mentioned in the proposed method, first the membership degrees of test documents to single-multiple category clusters were calculated. The ratio of multiple category value to single category value was calculated by using these values. Here, the last step for classification of documents is to determine a threshold value. First, effect of the threshold value on classification was investigated experimentally. Then, as mentioned in the proposed method α-FSCM, threshold value must be determined. As it is indicated before, training data consists of 6456 documents. 622 items of these documents belong to more than one category.

The membership (belonging) values of these training documents to two classes were found by using the proposed method. From these values, the ratio of multiple cluster value to single cluster value was founded and documents were ordered with respect to these ratios. According to the proposed method, ratio value of 622nd document has been chosen as threshold. This threshold has been determined as 0.317. This value will be used for classification of test data.

The proposed α-FSCM was compared with Rocchio algorithm and naive Bayes method that are commonly used as classification methods in text mining. But it can be seen in

Open Science Index, Computer and Information Engineering Vol:7, No:8, 2013 publications.waset.org/16229/pdf

Open Science Index, Computer and Information Engineering Vol:7, No:8, 2013 publications.waset.org/16229/pdf

Table I that these two methods are so insufficient for single-multiple category classification. Also, it can be seen from this table that proposed method is so successful with respect to both Rocchio algorithm and naive Bayes method.

C. Implementation of MCVM

The subject that we investigated in next step is determination of categories to which the documents with multiple categories belong. But, in the test stage, multiple categories were accepted as two and three categories. In other words, if a document has multiple categories, it is accepted that as number of categories to which the document belongs is 2 and 3 items. Determination of these categories is aimed. The results obtained by application of MCVM using Formulas 1 and 2 are shown in Table II.

TABLE I
COMPARISON OF CLASSIFICATION PERFORMANCE FOR A-FSCM AND ROCCHIO ALGORITHM

Method	Single Category			Multiple Categories			Average
	Recall	Precision	F-Measure	Recall	Precision	F-Measure	F-Measure
Rocchio	0.708	0.897	0.791	0.299	0.105	0.155	0.473
Bayes	0.511	0.945	0.663	0.742	0.149	0.248	0.455
α-FSCM	0.962	0.964	0.963	0.692	0.680	0.686	0.824

TABLE II
THE DETERMINATION VALUES OF CATEGORIES USING MCVM

Number of Categories	Total	CR	CM	MCVM	Increase (%)
2	221	153	29	81	179.31
3	26	21	2	7	250

The explanations of the columns in Table II are as follows:

Total is the total number of documents that belong to more than one category in test document collection. CR (Classification Results) is the total number of documents which are determined correctly by using proposed α-FSCM in test collection documents. CM (Classical Method) is the total number of documents whose categories are determined correctly by using classical method. MCVM (Multiple Categories Vector Method) is the total number documents whose categories are determined correctly by using proposed MCVM. Increase is the percentage value of the increment between the results of classical method and proposed method.

As can be seen in Table II, when number of category is 2 or 3, there is an important increase of success between proposed method and classical method in determination of the categories which documents belong to. This table does not contain the documents having 4 categories because only one item exists for this situation. Considering the situation of belonging more than one category of documents is important for finding similar documents systems. This importance can be seen clearly in the example given in Table III. Here, A and B categories denotes "money-fx" and "interest" categories respectively in the Reuter 21578 document collection. 12 documents that belong to these categories were chosen as examples from training data. 4 items of these chosen documents only belong to A category, 4 items only belong to B category and the rest 4 items belong to both A and B categories. These documents which were chosen from training documents were compared with the documents which were chosen as samples from the test data. Here, the results obtained for the sample document (test document numbered as 6452) chosen from the test data can be seen in Table III. This test document belong to both A and B categories.

TABLE III
THE SIMILARITY BETWEEN THE TEST DOCUMENTS AND TRAINING DOCUMENTS

Test document no.	First Category of test doc.	Second Category of test doc.	Training document no.	Similarity	Category or categories of training doc.
6452	A	B	19201	0,42279	A-B
			21343	0,42051	A
			19237	0,41775	A-B
			19512	0,41582	B
			21285	0,40890	B
			21508	0,39984	A-B
			20500	0,39711	A-B
			19529	0,38071	B
			15364	0,37444	A-B
			21539	0,36893	A
			21556	0,35933	A
			19557	0,34262	B
			21561	0,34166	A

The utility of determination of the multiple categories is clearly seen here. For example, as it can be seen in Table III, if the test document numbered as 6452 was accepted to belong to only category A (like in the previous studies), either the documents belong to only category A or both A and B categories would be chosen as candidate documents. However, the documents (for example 19512 and 21285) that only belong to category B and are similar to the current test document in a great ratio were ignored. So this shows that it is necessary to choose the documents as candidates that belong to all these categories, if this test document belongs to more than one category.

As the result, number of documents that are candidates to comparison similarity has been increased by using the proposed method. So, the documents that have high similarity ratio but ignored in single category approach can be considered by this method. Thus, more number of similar documents will be found.

V. CONCLUSION

Multiple categories problem for finding similar documents that is discussed in the present study is the subject that hasn't been mainly handled in the former studies. Solution of the problem has been investigated in two stages as determination of the documents with multiple categories and determination of the categories that these documents belong to. In the first stage, a fuzzy logic based approach has been embraced to determine the documents with multiple categories, since the documents belonging to more than one category is a situation to be considered. FSC method that has important role in text classification has been improved as to be adapted to the

problem. When compared with commonly used Rocchio algorithm and naive Bayes method for IR and Text Mining studies, the new proposed method (α-FSCM) has been quite successful. In the second stage, information of categories regarding to frequency of being seen together has been used for determination of the categories which documents with multiple categories belong to. MCVM that is proposed for this stage has been quite successful with respect to the classical approach related to this subject.

This study has shown the importance of multiple categories problem for finding similar documents. In the former approaches in which test documents are assumed to belong to only one category, great amount of candidate documents are ignored. Proposed similarity search method aimed to solve this problem and succeed it in a great ratio. And the method has a multi stages structure. So it can be possible to be improved for each stage. As an example, refinement of determination of parameter α in the first stage can be given. Also in the second stage of problem, performance increase can be investigated by using different parameters in forming multiple categories matrix.

ACKNOWLEDGMENT

This study is supported by Scientific Research Projects Unit of Yuzuncu Yil University.

REFERENCES

[1] S.S. Weng and C.K. Liu, Using text classification and multiple concepts to answer e-mails, *Expert Systems with Applications* 26(4) ,529-543, 2004.
[2] D. Elworthy, Question answering using a large NLP system, *The Ninth Text Retrieval Conference*, Gaithersburg, 2000.
[3] C. Apte, P. Damerau and S. Weiss, Text Mining with Decision Rules and Decision Trees, *In Proceedings of the Conference Automated Learning and Discovery*, CMU, 1998.
[4] J.R. Quinlan, Induction of Decision Trees, *Machine Learning Journal* 1 81-108, 1986.
[5] M. Sahami, S. Dumais, D. Heckerman and E. Horvitz, A Bayesian Approach to Filtering Junk e-mail, AAAI 98, *Workshops on Text Categorization*, 1998.
[6] K. Tzeras and S. Hartmann, Automatic Indexing Based on Bayesian Inference Networks, *In Proceedings of the 16th Annual ACM/SIGIR Conference on Research and Development in Information Retrieval*, 22-34, 1993.
[7] E. Wiener, J. Pederson and A. Weigend, A Neural Network Approach to Topic Spotting, *Fourth Annual Symposium on Document Analysis and Information Retrieval*, 1995.
[8] G. Guo, H. Wang, D. Bell, Y. Bi and K. Greer, Using *kNN* model for automatic text categorization, *Soft Computing* 10,423-430, 2006.
[9] S.S. Weng and Y.J. Lin, A Study On Searching For Similar Documents Based On Multiple Concepts And Distribution Of Concepts, *Expert Systems with Applications* 25(3) 355-368, 2003.
[10] B. Masand, G. Linoff, and D. Waltz, Classifying News Stories Using Memory Based Reasoning, *In Proceedings of the 15th Annual*, 1992.
[11] S. Tan, Neighbor-weighted K-nearest neighbor for unbalanced text corpus, *Expert Systems with Applications*, 28, 667-671, 2005.
[12] I.S. Dhillon, J. Fan and Y. Guan, Efficient Clustering of Very Large Document Collections, In Data Mining for Scientific and Engineering Applications, Kluwer Academic Publishers 357-381, 2001.
[13] S. Dumais, J. Platt, D. Heckerman and M. Sahami, Inductive Learning Algorithm and Representations for Text Categorization, *In Proceedings of the 1998 ACM 7th International Conference on Information and Knowledge Management* 148-155, 1998.
[14] T. Joachims, Text Categorization with Support Vector Machines: Learning with Many Relevant Features, *In Proceedings of the 10th European Conference on Machine Learning* 1, 137-142, 1998.
[15] A. Klose, A. Nürnberger, R. Kruse, G. Hartmann, and M. Richards, Interactive Text Retrieval Based on Document Similarities, *Phys. Chem. Earth* (A), 25(8), 649-654, 2000.
[16] [.C. Yang and C.H. Lee, A text mining approach on automatic generation of web directories and hierarchies, Expert Systems with Applications, 27, 645-663, 2004.
[17] H.C. Yang and C.H. Lee, A text mining approach on automatic construction of hypertexts, *Expert Systems with Applications* 29(4), 723-734, 2005.
[18] D.H. Widyantoro, and J. Yen, A Fuzzy Similarity Approach in Text Classification Task, IEEE, 2000.
[19] S. Miyamoto, Fuzzy Multisets and Fuzzy Clustering of Documents, *In Proc. of the IEEE International Conference on Fuzzy Systems*, FUZZ-IEEE, 2001.
[20] G. Salton, and C. Buckley, Term Weighting Approaches in Automatic Text Retrieval, *Information Processing and Management*, 24(5), 513-523, 1998.
[21] R. Saraçoğlu, K. Tütüncü and N. Allahverdi, A Fuzzy Clustering Approach for Finding Similar Documents Using a Novel Similarity Measure, *Expert Systems with Applications*, 33(3), 600-605, 2007.
[22] X Wan, A novel document similarity measure based on earth mover's distance, *Information Sciences*, 177, 3718-3730, 2007.
[23] M. Garofalakis, A. Gionis, R. Rastogi, S. Seshadri and K. Shim, XTRACT: Learning Document Type Descriptors from XML Document Collections, *Data Mining and Knowledge Discovery*, 7, 23–56, 2003.
[24] Y. Zhao and G. Karypis, Hierarchical Clustering Algorithms for Document Datasets, Data Mining and Knowledge Discovery, 10, 141-168, 2005.
[25] C.L.A. Clarke, G.V. Cormack, D.I.E. Kisman and T.R. Lynam, Question answering by passage selection, *The Ninth Text Retrieval Conference*, Gaithersburg, 2000.
[26] R. Saraçoğlu, Searching for Similar Documents Using Fuzzy Clustering, PhD Thesis, Institute of the Natural and Applied Sciences, Selçuk University, 2007.
[27] S. Kim, D. Baek, S. Kim, H. Rim, Question Answering Considering Semantic Categories and Co-occurrence Density, *The Ninth Text Retrieval Conference*, 2000.
[28] T.S. Morton, Using Coreference in Question Answering, *The Eighth Text Retrieval Conference*, 1999.
[29] C. Elkan, Deriving TF-IDF as a Fisher Kernel, *Proceedings of the International Symposium on String Processing and Information Retrieval* (SPIRE'05), Buenos Aires, Argentina, 296-301, 2005.
[30] A. McCallum, K. Nigam, J. Rennie and K. Seymore, Automating the Construction of Internet Portals with Machine Learning, *Information Retrieval Journal*, 3, 127-163, 2000.
[31] S. Jones and P. Willett, Readings in information retrieval, Morgan Kaufmann Publisher, 1997.

Open Science Index, Computer and Information Engineering Vol:7, No:8, 2013 publications.waset.org/16229/pdf