Fuzzy clustering and categorization of text documents

Conference Paper · December 2013

DOI: 10.1109/HIS.2013.6920493

CITATIONS
5

READS
430

3 authors:

Heba Ayeldeen
Cairo University
17 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Aboul Ella Hassanien
Cairo University
1,093 PUBLICATIONS **15,305** CITATIONS
SEE PROFILE

Aly Fahmy
Cairo University
73 PUBLICATIONS **1,285** CITATIONS
SEE PROFILE

Fuzzy clustering and categorization of text documents

Heba Ayeldeen
Scientific Research Group in Egypt (SRGE)
Faculty of Computers and Information
Cairo, Egypt
heba.ayeldeen@gmail.com

Aboul Ella Hassanien
Scientific Research Group in Egypt (SRGE)
Faculty of Computers and Information
Cairo, Egypt
aboitcairo@gmail.com

Aly Aly Fahmy
Faculty of Computers and Information, Cairo University
Cairo, Egypt
aly.fahmy@gmail.com

Abstract— The fuzzy Euclidean distance clustering algorithm has been well studied and used in information retrieval society for clustering documents. However, the fuzzy logic algorithm poses problems in dealing with large amount of data. In this paper we proposed results for clustering theses documents based on Euclidean distances and cluster-dependent keyword weighting. The proposed approach is based on the Fuzzy Euclidean distance clustering algorithm. The cluster dependent keyword weighting help in partitioning and categorizing the theses documents into more meaningful categories.

Keywords— ***Lexical similarity, Medical Ontology, MeSH, Fuzzy Euclidean distance Algorithm***

I. INTRODUCTION

To support operations, management and decision-making, organizations are in need of information systems and strategic tools. Organizations and individuals use different types of systems for different purposes and for taking the right decision at the right time. Depending upon the different levels within the organization as well as the business needs of the systems, Information systems differs [1][2].

In the late 1960s, decision support systems (DSS) evolved through time and became a practical model-oriented after the information system. DSS is another type of information system that assists higher management to make long term decisions. It helps the decision makers as well as strategic management staff making the right decision and getting the optimum solutions on the right time. Making the right decision and getting the optimum solution can be provided by information, models or analysis tools [3][4][5].

Data mining tools are considered as important part of any DSS. Of the main characteristics of DSS that make it special than any other information system that DSS are flexible, adaptable to the changes within the organization as well as changes in the environment. The user controls inputs and outputs. They support the decision process and often are sophisticated modeling tools so managers can make simulations and predictions [6].

One of the important tasks performed as part of many text mining and information retrieval systems is clustering. Clustering can be used for efficiently finding the nearest neighbors of a document, for improving the precision or recall in information retrieval systems, for aid in browsing a collection of documents, and for the organization of search engine results, and lately for the personalization of search engine results [7][8].

Most current document clustering approaches work with the vector-space model, where each document is represented by a vector in the term-space [9]. Documents generally consist of the keywords which are considered as the main features of the document. Generally speaking the word frequencies (WF) in a given document can be used to form a vector model for this document [10].

In the past, the clustering techniques have been widely applied in a variety of scientific areas such as pattern recognition, information retrieval and so forth. In general, cluster analysis aims in collecting the alike individuals and classifying groups based on certain feature for measuring the degree of similarity of each individual [9].

There are many fuzzy clustering methods being introduced. Fuzzy clustering algorithm is one of most important and popular fuzzy algorithms which is extensively used in feature analysis, pattern recognition, image processing, classifier design, and others [11][12][13].

Fuzzy Search functionality enables finding strings that match a pattern approximately (rather than exactly). The word fuzzy search means that a database query returns records even if the search term (the user input) contains additional or missing characters or other types of spelling error, which in return saves time and money [13].

The rest of this paper is organized as follows. Section 2 discusses the Fuzzy logic algorithm. Section 3 states the semantic similarity by using the medical Ontology and its importance to improve the relatedness of medical data. While section 4 states steps involved in applying the fuzzy logic algorithm to text mining. Section 5 shows a case study while applying the fuzzy Euclidean Distance Algorithm (FED). Last but not least is section 6 includes the interpreting results of the case study. The last section, presents conclusion and the future works.

II. FUZZY LOGIC: REVIEW

The main aim of the Fuzzy Euclidean distance Algorithm is to minimize the objective function by categorizing documents based on the fuzziness of the words within the document given by [12][13]:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{c} \mu_{ij}^{m} \left\| X_i - C_j \right\|^2 \quad (1)$$

$1 \leq m < \infty$

Where in equation 1:

m denotes any real number greater than 1

μ_{ij} denoted membership degree of X_i in the Jth cluster

X_i is the ith dimension of the measured data

C_i Is the ith dimension of the cluster center

Similarly, a mathematical optimization procedure was used to minimize J with respect to the center to obtain

$$C_{ik} = \begin{cases} 0 \\ \frac{\sum_{j=1}^{N} (\mu_{ij})^{m} x_{jk}}{\sum_{j=1}^{N} (\mu_{ij})^{m}} \end{cases} \quad (2)$$

The partition Matrix values are updated using the formula:

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{d_{ij}^2}{d_{kj}^2} \right)^{\frac{1}{m-1}}} \quad (3)$$

where in equation 2:

$d_{ij}^2 = \sum_{k=1}^{n} v_{ik} (x_{jk} - C_{ik})^2$, V is the vector of word (4)

is the weighted aggregate Euclidean distance of words to each cluster center.

The Fuzzy Euclidean distance (FED) algorithm's iteration stops when the maximum change in the values of the Fuzzy Partition Matrix is less than a certain predefined value which is a termination criterion with the value between 0 and 1.

The steps followed in the Fuzzy Euclidean distance Algorithm are given below:

1. Initialize the fuzzy partition matrix, U=$[\boldsymbol{\mu_{ij}}]$ matrix, $\mathbf{U^{(0)}}$
2. At K-step: calculate the center's vectors $\mathbf{C^{(k)}}$= $[\mathbf{c_j}]$ with $\mathbf{U^{(k)}}$ using equation (2)
3. Update $\mathbf{U^{(k)}}$, $\mathbf{U^{(k+1)}}$ using equation (3)
4. If $\left\| \mathbf{U^{(k+1)}} - \mathbf{U^{(k)}} \right\| < \boldsymbol{\partial}$ then stop, otherwise repeat step 2

III. Ontology-based Semantic similarity: MEsh

Knowledge and text representation as well as estimation of the semantic similarities of biomedical data has become a great challenge for relatedness and text clustering. Understanding and transforming the data is the key strategy for any organization to achieve competitive advantage. The need of text mining and strategic tools for analysis has become very essential for organizations to cope with the environmental and technological changes [14]. Ontologies have been applied to a number of DSS where ontologies present a common understanding of the knowledge domain using major concepts and terms applied in that domain and identify the relationships between these concepts [15][16]. Here in the paper, we will mainly focus on the medical ontologies. A more general approach was proposed to exploit biomedical sources. In 2009, Al- Mubaid and Nguyen proposed a methodology calculate the similarity measure of biomedical sources by using medical Ontologies such as SNOMED CT and MeSH [17][18]. The Medical Subject Headings (MeSH) contains a hierarchy of medical and biological terms defined by the US National Library of Medicine. This classification was initially created to catalogue books and other library materials, and to index articles for inclusion in health related databases including MEDLINE. In MeSH tree, there are 102 basic categories, with more than 22,000 concepts. We used the latest 2012 MeSH XML files available for download.

http://www.ncbi.nlm.nih.gov/mesh. The MESH Ontology consists of classes, properties or definitions, relationships between classes and individuals as well as categories and subcategories [19].

IV. Steps involved in applying fuzzy logic algorithm to text mining

Test data was collected from the digital library of Faculty of Medicine, Cairo University. About 4,878 theses data was collected and about 16,000 keyword in the theses data.

A. Preprocessing of text

Certain limitations of this work must be mentioned to facilitate the interpretation of the results. The main limitation of this research is the relatively unorganized keywords in the theses we compiled to create the test set. To address this limitation, we focused on the main goal which was to compare several established measures of similarity and relatedness that are based on medical ontological knowledge sources (MeSH) to lexical similarity method. In the context of these comparisons, we felt it was justified to take keywords from theses data that were collected from the digital library of the Faculty of Medicine, Cairo University. This step involves cleaning up the documents/text like removing the hyphens, affiliations etc. Based on the collected data we came up with 68 categories from the MESH ontology.

B. Feature generation

The documents are represented in the “Bag of Words” method. Stop words like “the”, “is”,”about” etc. are removed.

C. Feature selection

In this step, the features that we will use are selected in order to cluster separately, the documents relating to certain category or group. However, presence of words like “Eye” , “Lenses” ,”Eye lashes” etc. will help relate the documents to the MESH category of "Anatomy". Similarly, words like “Blood vessels”, “Artery”, “Aorta” will relate the documents to "Cardiovascular System" category.

In some cases, it will be hard to manually find out the features which distinguish documents that should belong to different categories. In such situations, we need to find out what features we should use for our clustering. To do this, we can take sample documents which are already known to belong to our required categories. Then after applying several analysis steps on these documents, the information gained from the analysis can be then applied to point out the differences among documents belonging to different clusters. The weight of the

word (word frequency), normalized by total document length, can be calculated for these documents. We choose the words with a significant word frequency variation between different document types.

D. Clustering

Using Fuzzy Logic and text-mining we can cluster similar documents together. Document Clustering is used by a computer application to group documents into meaningful groups.

For the Fuzzy Euclidean Algorithm, every observation here has a membership value associated with each of the clusters which is related inversely to the distance of that observation from the centre of the cluster.

E. Evaluation and Interpretation of results

According to the membership value of a document for a given cluster, we can conclude to what degree a given document belongs to each cluster/category. For instance if the membership value is high compared to other value, the document can be said to strongly belong to that cluster. However, if any of the membership values corresponding to a specific document are almost same among other clusters, it would imply that the document does not quite strongly belong to any of those clusters.

V. CASE STUDY

The main target of the study is to increase the scientific research field in the different faculties of Cairo University. To do so, we focused in the theses mining concept for instance for the Faculty of Medicine, Cairo University. Data was collected from the digital library of the Faculty of Medicine and then cleansed for further process. The data collected was theses documents including the title of the theses and the abstract with keywords. The theses documents are classified into Master and doctorate theses. Documents are tracked within the last 10 years separated and categorized based on the departments within the Medicine school. Data collected was 4,878 theses with 15,808 keyword across all departments.

Faculty of Medicine in Cairo University is classified into 35 departments. The aim of the work is to find out the departments that can work together easily to increase the research within the faulty. As well as making it easy for departments to find a way to increase the research within each department.

Different steps were performed with different results and a significant accuracy that varies in each step.

First step includes measuring the keywords similarity within each theses document based on the lexical similarity measures and the keywords occurrences within the documents. With the huge amount of data collected, knowledge is then extracted stating that certain departments have potential impact in working together for the purpose of increasing the level of scientific research and helping students in the information retrieval though the theses mining.

Although the documents are classified based on the departments we focused on applying a cluster analysis and then applying the Euclidean distance to get more accurate combination of departments that can potentially work together. In the second step, we started using medical ontologies to help researchers classify and categorize the theses documents on specific bases. As Faculty of Medicine in Cairo Universities does not have medical ontologies to work on we suggested to use the MESH ontology as the standard to work with. Results were much effective than just measuring the keyword similarity in each document. As in this step we measured the semantic similarity of the keywords within the documents based on the MESH ontology. The MESH ontology includes several categories to classify medical words (dictionary).

We came up based on the keywords we have extracted from the theses documents that we have 68 categories. Each category includes keywords; and based on these categories knowledge was inducted showing correlation between departments with a percentage of accuracy higher than that was in the first two steps.

The third step is also focusing in increasing the level of the scientific research in the faculty of medicine Cairo University. But that time by using the previous steps, we are building a medical dictionary for each department to facilitate the process of theses mining and keyword clustering based on the fuzzy logic of the word within the documents.

This step is based on the categories from the MESH ontology by which we applied the Fuzzy logic algorithm to measure the fuzziness of each document towards the clusters we have.

TABLE 1: BAGS OF WORDS REPRESENTATION OF THE THESES DOCUMENTS (TITLES)

Word	*Occurrences*
Ultrasound	87
Biopsy	16
Pacemaker	3
⋮	⋮
Catheters	1
Stents	3

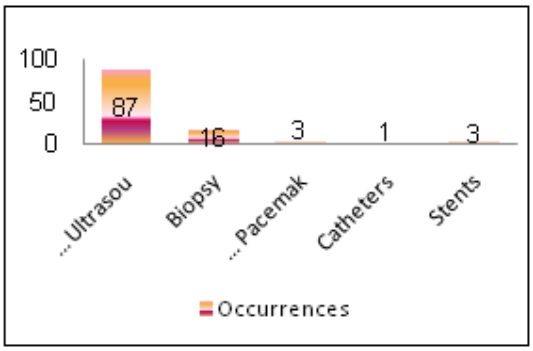

Fig. 1. WORDS FREQUENCIES IN TITLES

Other than a single word, we applied also the use of a phrase (terms) for better results. For example like "Primary care", "Abdominal wall management" and "Diabetes mellitus". So the bag of word can be treated either as a single or a term as a whole. In the case study, we assigned the weights as the number of occurrences of the word in the title of the document normalized by the document title length and multiplied by 10,000 to get the word frequency in each document.

$$\mathrm{WF} = \left(\frac{\text{Word Count}}{\text{Total Words in the Document)}}\right) \times 10{,}00 \qquad (5)$$

TABLE 2: WORD FREQUENCIES FOR 2 CATEGORIES (E04 AND E07)-RELATED DOCUMENTS' TITLES

Word	*E04 Surgical procedures, operative*	*E07 Equipment and supplies*
Ultrasound	25.026	21.245
Biopsy	3.952	3.084
Pacemaker	1.054	1.371
Catheters	0.263	0.343

By looking at table 2, for simplicity we selected two of the MESH categories in the example with random selection of words within these categories. The table shows that values of word frequencies varies for both categories with respect to the words selected. The table shows that the word " Ultrasound" WF (25.026 v/s 21.245), word " Biopsy" WF (3.952 v/s 3.084), word " Pacemaker" WF (1.054 v/s 1.371) and the word " Catheters" WF (0.263 v/s 0.343). On basis of these 4 words, we clustered the given documents into category E04 and E07 categories.

Now for category E04 and category E07 we get all documents which are clustered based on the feature we selected which is the keywords in that category. We treat those documents as the total documents for that certain category and we apply the equation (5) to normalize those documents and get the total count of words within the document as well as the word frequency for those documents. We start our clustering process.

Let $D = \{d_1, d_2, d_3 \ldots d_n\}$ represent the 'n' documents in each category to be clustered

Now each of these documents have specific selected features d_i, in our case the keyword in a category:

$d_i = \{d_{i1}, d_{i2}, d_{i3} \ldots d_{im}\}$

Every d_i in the D is normalized in the same way we did in the feature selecting by using eq(5). Then we start building our fuzzy matrix putting in mind a constraint that all entries should be within the interval [0,1] and the column total =1.

TABLE 3: FUZZY PARTITION MATRIX

	T1	*T2*	*T3*	*T4*	*T5*	*T6*	*T7*	*T8*
Cluster 1	1	0	1	0	0	0	1	1
Cluster 2	0	1	0	1	1	1	0	0

By randomly selecting documents' titles, our initial assumption based on table 2 is that Title 1 in Document thesis 1, T3, T7 and T8 belong to cluster 1 (E04) and T2, T4, T5 and T6 belong to cluster 2 (E07). Now we have to calculate the initial centers for both clusters C1 and C2 are the centers of cluster 1 and cluster 2 respectively. Based on the values in table 3, the centers can be calculated using eq(2).

$$C_{1j} = \left(\frac{(1)X_{1j} + (1)X_{2j}(1)X_{3j}(1)X_{4j}}{(1^2 + 1^2 + 1^2 + 1^2)}\right) = \frac{(X_{1j} + X_{2j} + X_{3j} + X_{4j})}{4}$$

Now we say our frequencies for each documents are:

TABLE 4: WORD FREQUENCIES IN EACH DOCUMENT (TO BE CLUSTERED)

	Ultrasound	*Biopsy*	*Pacemaker*	*Catheters*
T1	1250	0	0	0
T2	0	0	0	1250
T3	1000	0	11.11	0
T4	0	1428.57	769.23	1666.66
T5	833.3	1428.57	0	0
T6	0	0	1428.57	0
T7	833.3	1250	0	1.006
T8	476.19	0	11.37	0

After applying eq(2), we get the two clusters below:

$C_1 = \{889.88,\ 312.5,\ 2.84,\ 0.2515\}$ is the center of cluster1

$C_2 = \{208.33,\ 714.285,\ 549.45,\ 729.165\}$ is the center of cluster2

Now, calculating the Euclidean distances of each document from both center clusters, eq(4):

$$D_{11} = ((1250 - 889.88)^2 + (0 - 312.5)^2(0 - 2.84)^2 + (0 - 0.2515)^2)^{\frac{1}{2}} = 476.813$$

Similarly the distances below,

$D_{12} = 1565.702,\ D_{13} = 331.438,\ D_{14} = 2324.171,\ D_{15} = 1117.505, D_{16} = 1709.459, D_{17} = 939.21, D_{18} = 518.525$

After applying the same equation to the second cluster we get:

$D_{21} = 1558.479,\ D_{22} = 1061.494,\ D_{23} = 1399.441,$
$D_{24} = 1216.883,\ D_{25} = 1316.968,\ D_{26} = 1363.135,$
$D_{27} = 1228.709, D_{28} = 1184.552$

Now the step of updating the Matrix Partition Matrix based on the distances we got earlier, by using the eq(3). We assume that the value of m=3 so after applying the formula we get:

$U_{11} = 0.914, \quad U_{12} = 0.315, \quad U_{13} = 0.947, \quad U_{14} = 0.215, U_{15} = 0.581, U_{16} = 0.389, U_{17} = 0.613, U_{18} = 0.839$

$U_{21} = 0.086, \quad U_{22} = 0.685, \quad U_{23} = 0.053, \quad U_{24} = 0.785, U_{25} = 0.419, U_{26} = 0.611, U = 0.369, U_{28} = 0.161$

TABLE 5: UPDATED FUZZY PARTITION MATRIX AFTER THE FIRST ITERATION

	Cluster 1	*Cluster 2*
T1	0.914	0.086
T2	0.315	0.685
T3	0.947	0.053
T4	0.215	0.785
T5	0.581	0.419
T6	0.389	0.611
T7	0.631	0.369
T8	0.839	0.161

In our case let's set the threshold change/stopping condition to 0.001 where there is no big change in the values of each documents in the cluster. Comparing with the initial Fuzzy Partition Matrix un table 3, our max change here is 0.581 in T5 which is greater than 0.001. Hence we will continue the steps by calculating the new cluster centers by the updated fuzzy partition matrix values.

$$C_{1j} = \frac{(T1 + T2)}{(T3)}$$

Where after applying eq(2):

$$T_1 = \left(0.914\, x_{1j} + 0.315\, x_{2j} + 0.947\, x_{3j} + 0.215\, x_{4j}\right)$$
$$T_2 = \left(0.581\, x_{5j} + 0.389\, x_{6j} + 0.631\, x_{7j} + 0.839\, x_{8j}\right)$$
$$T_3 = (0.914^2 + 0.315^2 + 0.947^2 + 0.215^2 + 0.581^2 + 0.389^2 + 0.631^2 + 0.839^2)$$

After several iterations, we see that T1, T3, T5 and T7 are belonging to cluster 1 which is E02 Surgical procedures, operative and T2, T4, T6 and T8 are classified to cluster 2 which is E07 Equipment and supplies on basis of high membership values in both clusters.

VI. CONCLUSION

In this paper, we showed mathematically how texts can be clustered by the fuzzy logic Euclidean distance equation on documents. By taking an example where theses documents were clustered into two MESH ontology categories which are: Surgical procedures, operative; and Equipment and supplies.

Other algorithms can be considered as well for future work, like applying the genetic programming; neural networks and comparing the results simultaneously.

REFERENCES

[1] C. Wiseman, "Strategic Information Systems, "Knowledge Management, Information Today Medford, NJ.: Irwin, Home-Wood, IL. hidebound systems, in Srikantaiah, T.K. and Koenig, M.E.D. (Eds), 1988.

[2] Bentley, L. D. and J. L. Whitten, "Systems Analysis and Design for the Global Enterprise", Mcgraw-Hill Publ. Comp (7): 43 – 46, 2007.

[3] Singh, S. K., "Database Systems: Concepts, Designs and Applications", Pearson Education India, 2006.

[4] Y. Malhotra, "Information management to knowledge management: beyond 'hi-tech hidebound' systems," Knowledge Management for the Information Professional, Information Today, 2000.

[5] LeBlanc, L. A. and M. T. Jelassi, "DSS software selection: A multiple criteria decision methodology", Information & Management, 1989.

[6] Rainer, R. K., C. A. Snyder et al., "Decision Support systems". 8(4): 333- 341, 1992.

[7] Jayanthi Ranjan, " Managing student data: a data mining-based framework for business schools," Int. J. Information and Operations Management Education, Vol. 4, No. 1, 2011.

[8] Jiawei Han and Micheline Kamber, Data Mining: Concepts and Techniques, Second Edition ed.: Elsevier Inc., 2006.

[9] J. H. a. M. Kamber, Data Mining: Concepts and Techniques: Addison-Wesley, 2001.

[10] A.K. Jain, M.N. Murty, and P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[11] Ross, T. J., Fuzzy Logic with Engineering Applications, Third Edition, John Wiley & Sons, Ltd, Chichester, UK, 2010.

[12] F. K. Höppner, F, Krise R, Runkler T, "Fuzzy Cluster Analysis: Methods for Classification," Data Analysis and Image Recognition, pp. 5-114, 2000.

[13] R. E. a. W. F. James C. Bezdek, "FCM: THE FUZZY c-MEANS CLUSTERING ALGORITHM," Computers & Geosciences, vol. 10, pp. 191-203, 1984.

[14] F. B. Pari Delir Haghighi, Arkady Zaslavsky b and Paul Arbon "Development and evaluation of ontology for intelligent decision support in medical emergency management for mass gatherings," Decision Support Systems, vol. 54, pp. 1192–1204, 2013.

[15] H. a. N. Al-Mubaid, A., "Measuring semantic similarity between biomedical concepts within multiple ontologies," IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, vol. 39, pp. 389–398, 2009.

[16] Donald Metzler, Susan Dumais and Christopher Meek, "Similarity Measures for Short Segments of Text," Proceeding ECIR'07 Proceedings of the 29th European conference on IR research, pp. 16-27, 2007.

[17] H. A, "Semantic similarity measures in the MESH ontology and their application to information retrieval on medline," Diploma Thesis, Dept. of Electronic and Computer Engineering, Technical Univ. of Crete (TUC), Crete, Greece, 2005.

[18] D. S. a. M. Batet, "Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective," Journal of Biomedical Informatics archive, vol. 44, pp. 749-759, 2011.

[19] J. D. a. H. B. Nelson SJ, "Relationships in medical subject headings, relationships in the organization of knowledge," K.A. Publishers, 2001.