International Journal of Pure and Applied Mathematics
Volume 116 No. 10 2017, 331-339
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

algorithm

Dr.Ashit kumar dutta
Associate professor
Department of Computer Science
Shaqra University
Kingdom of Saudi Arabia

Abstract

Text mining(TM) is the extraction of useful pattern from the text corpus. The process of knowledge discovery from the text corpus depends on the natural language processing (NLP) techniques. Accent and contextual understanding of the languages made automation process more complex and needs the help of fuzzy methods. Fuzzy approaches have the reputation to deal with the vague and uncertainty problem. Arabic language is complex in nature; categorisation of text in the language is very tedious. The aim of the research is to classify the large dataset in Arabic language. The proposed work employs enhanced fuzzy c – means (EFCM) to cluster the text corpus into different categories. J48, fuzzy c – means and k-means were compared to show the effectiveness of the research work.

Keywords: Text mining, fuzzy, text corpus, pre-process, Arabic text classifier

1. Introduction

TM is the text analysis refers to the extraction of meaningful information from large dataset. The process of text mining involves steps like parsing of text, feature extraction and the removal of meaningless text from the corpus. Text categorisation, Text clustering, and entity extraction are the tasks of text mining. Text is unstructured and difficult to retrieve from the storage. The text mining is used to indicate a system that investigate massive set of natural language text and find linguistic pattern to derive useful information. TM is largely used in the field of statistics, information retrieval and machine learning.[1][2][3]

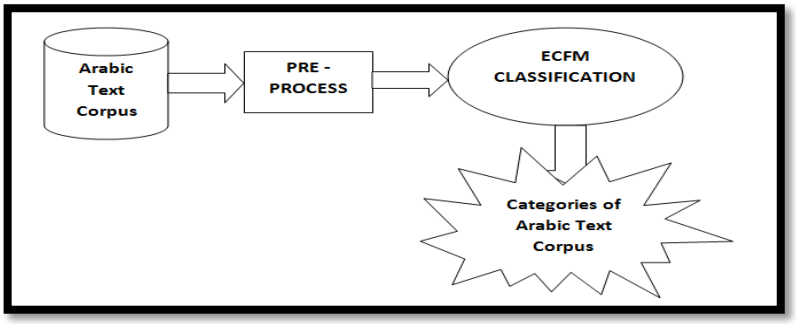

Figure 1.1 – Classification of Text corpus

TM uses bag of words or word stemming for the process of information retrieval and modern TM make use of NLP. The following part will discuss the applications of TM.

1.1. Application of Text mining

1.1.1. Social network monitor

Numbers of software based on TM are available in the market to monitor social networks and analyze text used in media and blogs. In Egypt, social networks were banned based on the report generated by national security applications during civil war in the country.

1.1.2. Bio – medical applications

Pubgene and Gopubmed are the popular TM tools used in the field of bio-medical research to visualize the image according to the text. Co –occurance networks used to visualize the relationships between terms and facilitate information retrieval for set of articles issued by the network.

1.1.3. Sentiment analysis

Affective computing is the area effectively uses TM. Student evaluations, movie reviews, opinion polls, children stories and science fiction stories are the applications of affective computing. WordNet and ConcepNet are the familiar applications of sentiment analysis.

1.2 Fuzzy clustering technique

Fuzzy clustering is also called soft clustering form clusters having common data points. Cluster analysis is the process of allocating data points to clusters. Distance, connections and intensity are the similarity measures to evaluate clusters. Fuzzy C – means (FCM) is the familiar fuzzy clustering algorithm similar to the k – means algorithm. Fuzzy clustering has wide varieties of applications and field of bio – informatics, image analysis and customer relationship are extensively using it.

The aim of the research work is to classify the large dataset in Arabic language into categories like Agriculture, Education, Entertainment, Politics, Geography and others. This paper is prepared as follows. The section 2 discusses related work in Tm using fuzzy approach. The section 3 discusses the experimental results of classification of dataset and comparison of results with other methods. Finally, section 4 concludes the paper and discusses future directions of this research.

2. Review of Literature

Lan H.Witten[4] introduced the concept of TM and its application in the field of computer science. TM attempts to glean information from natural language content. Text is un-structured, amorphous and tedious to deal with algorithmically. TM denotes any system that finds large quantities of text and detects lexical or linguistic patterns to extract information. Text is a raw data and process of computation needs more time and storage Text summarizer used to produce a condensed representation of input for computation in human readable form. Information retrieval is same as document retrieval where the documents are processed to extract the particular information for the user.

In [5] a survey on TM techniques and applications. TM automatically extracts information from different resources. The aim of TM is to discover unknown information. TM and NLP train computers to analyze and understand natural languages. A topic tracking system keep user profiles to predict the user interest content. Some TM tools allow user to select particular categories of interest automatically based on the user visited information and click through information. Lexical chaining is the process of grouping lexically related terms into lexical chains. Summarization tools search for titles and sub topics in order to identify the key points for the document. Categorization is the way to identify the topic of a document by placing it into pre-defined labels.

In [6], proposed a text document classification based on fuzzy rules. The extraction of information from text corpus becomes vital for companies to take decisions. The fuzzy approach used to develop powerful mechanisms to represent knowledge from large dataset. The fuzzy set theory has the ability to represent and retrieve texts for the user. The importance of text documents in groups through linguistic terms allows the discovery of fuzzy rules used in the recovery of textual information. The research implemented FCM to cluster text documents into different categories. Dataset of ACM digital library were used to show the effectiveness of the research and result shown that the proposed work outperformed all other methods.

In [7] proposed a TM based on FCM. Document clustering is the prime application of TM. Hard and soft clustering are the two types of clustering in TM. The research work forms clusters recursively for a set of fuzzy clusters and the supplement cluster centers that indicate the system of the data as best as possible. FCM clustering uses distance measures to control the degree of membership of each data to a particular cluster. An FCM cluster shows better results than k-means for experimented dataset.

3. Research Methodology

The research work employs fuzzy approach for the process of classification from the Arabic dataset.

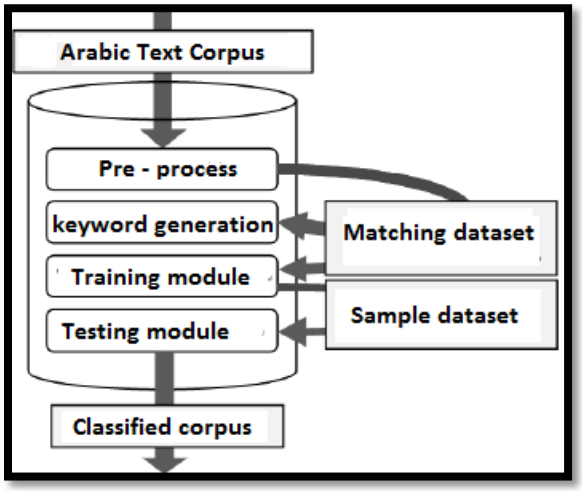

Figure 3.1 – Process of Classification of Arabic corpus

Figure 3.1 shows the process involved in the classification of Arabic corpus. The text corpus pre – processed and necessary keywords / bag of words generated from it[8][9]. The training module handles the sample dataset and train the classifier to perform well with the testing module. The other sample dataset generated from the Arabic corpus supplied to the testing module and the values will be recorded to evaluate the performance of the proposed work.

The following procedure shows the process of TM.

Step 1: Pre – process of Arabic dataset – Removal of stop words from the dataset.

Step 2: Pre – process stage – Word stemming

Step 3: Find the frequency of terms in dataset.

Step 4: Selection of terms from dataset.

Step 5: Generation of vectors to generate clusters of classified dataset.

Step 6: Apply similarity evaluation.

Step 7: Input pre – processed dataset into proposed and other methods.

Step 8: Classified dataset.

The following procedure shows the EFCM algorithm and the proposed work altered the memory and tuned the clustering process in FCM algorithm.

3.1 EFCM algorithm

Step 1 - Initialize N=[nij] matrix, N(0), L=1

Step 2 - At J/L-step: calculate the centers vectors C(k)=[cj] with N(k)

$$c_j = \frac{\sum_{i=1}^{N} N_{ij}^{m} \cdot x_i}{\sum_{i=1}^{N} N_{ij}^{m}}$$

Step 3 - Update N(k) , N(k+1), L++;

$$N_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|} \right)^{\frac{2}{m-1}}}$$

Step 4 - If || N(k+1) - N(k)||< ε then STOP; otherwise return to step 2.

4. Results and Discussion

The experiment conducted in i7 processor, 2.4 GHz with 4 GB ram in Windows 10 environment. Dataset collected from Aracorpora website (www.aracorpus.e3rab.com) having 1.3 million words with number of occurrences. The pre-processed dataset given as input to the system and the output is a classified Arabic dataset with categories Education, Agriculture, Politics, Entertainment, Geography and others. The literals C – 1, C – 2 , C – 3, C – 4 ,C – 5,and C – 6 denote the clusters for each categories. Table 4.1 shows the Arabic words used to train the proposed and other methods.

Categories	Arabic words used for Training
Education	العلم,أعلم,علم,العلماء,العلمية,علم, كِتَاب,علماء,علمت,علماوعلم,والعلماالعلمي,بالعلم,ويعلم,علموا,فعلمعلمهم,المعلم,وعلم
Agriculture	لزراعة المائية, التنويع الزراعي,الاقتصاد الزراعي,الهندسة الزراعية,التسويق الزراعي,علم الحيوان,تربية,الحيوان, تغذية الحيوان, علم النبات, تسميد النبات,الهندسة الحيوية,الهندسة الوراثية,تصنيف الحيوان,البستنة,علم الغذاء
Politics	سياسية التعليم,الممارسة السياسية,السياسة العصبية,السلوك السياسي,الأحزاب السياسية,سياسات الشتات,سياسات الهوية,الانتخابات السياسية,النشاط السياسي,الجماعات السياسية,الأفكار السياسية,النظام السياسي,الدستور,العلاقات الدولية,الحريات العامة,حقوق الإنسان
Entertainment	الرسوم المتحركة,السينما,المسرح,الكوميديا,الرسوم الهزلية,القراءة,الألعاب,المتنزهات,الموسيقى,السباحة,رسوم الكارتون,الكاريكاتير,ألعاب الأرقام,ألعاب الأحرف,الحدائق,الشواطئ
Geography	جغرافيا ثقافية,جغرافيا سياسية,جغرافيا اقتصادية,جغرافيا نقدية,جغرافيا طبيعية,حتمية جغرافية,جغرافيا حيوية,جغرافيا سياحية,جغرافيا بشرية,جغرافيا متكاملة,خرائط جغرافية,فقه جغرافي,الغابات,التضاريس,المناخ,الطقس

Table 4.1 Training words for the classification process

Table 4.2, 4.3, 4.4 and 4.5 shows the accuracy for clusters having metrics Cosine similarity, Euclidean distance and Dice co-efficient for each experimented methods with the dataset. The accuracy of EFCM is better than the other methods.

Clusters / Algorithms	C – 1	C – 2	C – 3	C – 4	C – 5	C – 6
Cosine Similarity	28.6	18.4	20.4	18.7	28.6	19.6
Euclidean Distance	25.3	23.4	21.3	19.3	31.6	18.3

Dice Coefficient	26.8	11.3	27.3	26.4	27.9	17.6

Table 4.2 Accuracy for clusters using J48

Clusters / Algorithms	C – 1	C – 2	C – 3	C – 4	C – 5	C – 6
Cosine Similarity	34.6	23.5	23.4	21.8	31.4	21.4
Euclidean Distance	31.2	27.3	22.6	23.4	33.8	22.3
Dice Coefficient	29.3	19.3	28.3	27.6	29.6	24.9

Table 4.3 Accuracy for clusters using FCM

Clusters / Algorithms	C – 1	C – 2	C – 3	C – 4	C – 5	C – 6
Cosine Similarity	20.3	14.6	21.3	14.6	25.3	17.3
Euclidean Distance	18.3	13.8	19.3	19.3	24.6	16.5
Dice Coefficient	23.7	17.3	17.6	20.5	28.3	18.3

Table 4.4 Accuracy for clusters using K – means

Clusters / Algorithms	C – 1	C – 2	C – 3	C – 4	C – 5	C – 6
Cosine Similarity	38.3	26.3	33.6	26.4	36.7	40.2
Euclidean Distance	36.2	24.5	38.5	24.8	32.4	39.6
Dice Coefficient	37.6	23.8	39.7	29.6	31.9	38.4

Table 4.5 Accuracy for clusters using EFCM

Precision, Recall and F1 measure are the metrics used to indicate the performance of an algorithm and Table 4.6, 4.7 and 4.8 shows the performance of the proposed and other methods. The proposed method has better score than other methods. EFCM has the ability to perform well in complex situation. Even the Arabic language is hard and complex; the EFCM has learnt well about the language and produced good results. Figure 4.1 shows the relevant graph of F1 measure. FCM and EFCM have performed well comparing to J48 and K – means. FCM and EFCM were derived from K – means but they can be scaled for the large dataset.

Clusters /Methods	**C – 1**	**C – 2**	**C – 3**	**C – 4**	**C – 5**	**C – 6**	**Overall Average**
J48	78.5	82.3	84.6	79.8	81.3	83.4	81.65
FCM	82.3	86.4	86.3	83.4	82.3	84.3	84.17
K – means	76.5	79.6	79.4	76.2	73.4	79.5	77.43
EFCM	89.3	84.3	87.3	86.3	84.9	85.7	86.3

Table 4.6 Precision for classification

Clusters /Methods	***C – 1***	***C – 2***	***C – 3***	***C – 4***	***C – 5***	***C – 6***	***Overall Average***
J48	**79.6**	**78.6**	**81.2**	**75.6**	**76.4**	**77.4**	78.13

FCM	**86.3**	**85.3**	**85.3**	**86.3**	**85.3**	**86.3**	85.8
K – means	**79.5**	**79.4**	**76.3**	**78.3**	**79.6**	**84.3**	79.57
EFCM	**85.9**	**86.3**	**88.2**	**87.9**	**88.6**	**87.5**	87.4

Table 4.7 Recall for classification

Clusters /Methods	***Precision***	***Recall***	***F1 Measure***
J48	81.65	78.13	79.85
FCM	84.17	85.8	84.98
K – means	77.43	79.57	78.49
EFCM	86.3	87.4	86.85

Table 4.8 Precision, Recall and F1 Measure

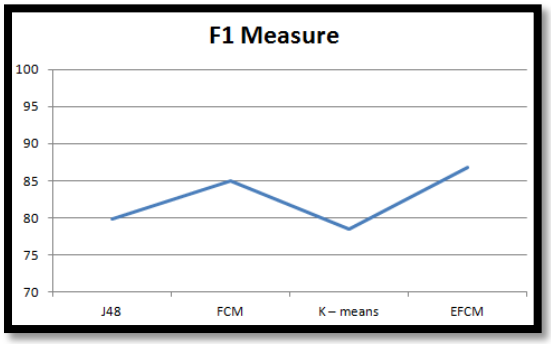

Figure 4.1 F1 measure of classification

Accuracy of the results also matters than the performance. To evaluate the accuracy the dataset were classified manually and Table 4.9 shows the accuracy of the methods. EFCM have more than 85% accuracy than other methods. Figure 4.2 shows the accuracy graph relevant to the Table 4.9.

Clusters /Methods	***C – 1***	***C – 2***	***C – 3***	***C – 4***	***C – 5***	***C – 6***
J48	**81.4**	**79.5**	**74.6**	**79.4**	**84.2**	**82.6**
FCM	**86.2**	**83.3**	**84.3**	**81.3**	**85.4**	**87.4**
K – means	**79.4**	**76.3**	**79.4**	**78.6**	**81.3**	**80.4**
EFCM	88.3	89.3	90.2	91.3	90.8	92.3

Table 4.9 Accuracy of proposed and other methods

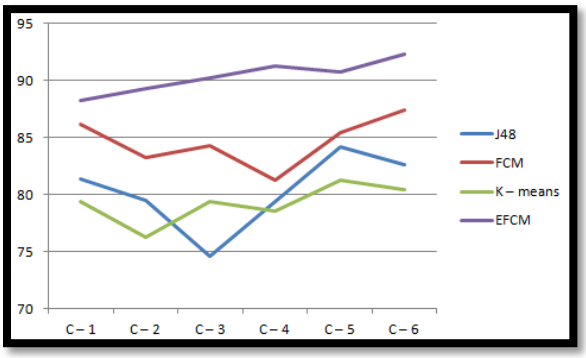

Figure 4.2 Accuracy of proposed and other methods

Conclusion

The fuzzy approach is the familiar method to deal complex and vague problems. TM is used to extract interest pattern from the large text content. The proposed work has classified a large text corpus into categories like Education, Agriculture, Politics, Entertainment, Geography and others. The EFCM algorithm performance is better than the other existing methods. Precision, Recall and F1 measure values shown that the EFCM have processed the dataset efficiently than other methods. The accuracy of the proposed method is far better than the other method. The future scope of the research is to further scale to large dataset in the same language.

References

[1.] L. A. Zadeh, "Fuzzy sets," Information and Control, vol. 8, no. 3, pp. 338–353, 1965.

[2.] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: theory and applications, 1st ed. Prentice-Hall, 1995.

[3.] E. M. Rodrigues and L. Sacks, "Learning topic hierarchies from text documents using a scalable hierarchical fuzzy clustering method," in International Conference on Recent Advances in Soft Computing, 2005, pp. 269–274.

[4.] Lan H. Witten, "Text Mining"

[5.] Vishal Gupta, and Gurpreet S.Lehal," A survey of text mining techniques and applications",Journal of emerging technologies in web intelligence,Vol 1.No. 1, Aug 09.

[6.] Tatiane M.Nogueria, Solange O. Rezende, Hebisa A.Canargo, " On the use of fuzzy rules to text document classification", 10th International conference on hybrid intelligent system, 2010, pp. 19 – 24

[7.] Deepa B.Patil, and Yashwant V.Dongre, " A fuzzy approach for text mining",International journal of mathematical sciences and computing, 2015, pp. 34 – 43.

[8.] Y.-J. Horng, S.-M. Chen, , Y.-C. Chang, and C.-H. Lee, "A new method for fuzzy information retrieval based on fuzzy hierarchical clustering and fuzzy inference techniques," IEEE Transactions on Fuzzy Systems, vol. 13, no. 2, pp. 216–228, 2005.

[9.] G. Bordogna, M. Pagani, and G. Pasi, Soft Computing for Information Retrieval on the Web. Springer Verlag, 2006.

[10.] E. Herrera-Viedma, "Modeling the retrieval process for an information retrieval system using an ordinal fuzzy linguistic approach," Journal of the American Society for Information Science and Technology, vol. 52, no. 6, pp. 460–475, 2001.

[11.] Lopez-Herrera, E. Herrera-Viedma, and F. Herrera, "Applying multi-objective evolutionary algorithms to the automatic learning of extended boolean queries in fuzzy ordinal linguistic information retrieval systems," Fuzzy Sets and Systems, vol. 160, pp. 2192–2205, 2009.

[12.] R. Saracoglu, K. TuTuncu, and N. Allahverdi, "A fuzzy clustering approach for finding similar documents using a novel similarity measure," Expert Systems with Applications, vol. 33, pp. 600–605, 2007.

[13.] L. Wang and J. Mendel, "Generating fuzzy rules by learning from examples," IEEE Transaction on Fuzzy Systems, Man and Cybernetics, vol. 22, pp. 414–1427, 1992.

[14.] Z. Chi, H. Yan, and T. Pham, Fuzzy Algorithms with Applications to Image Processing and Pattern Recognition. World Scientific, 1996.

[15.] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," SIGKDD Explorations, vol. 11, no. 1, 2009.

[16.] M. V. B. Soares, R. C. Prati, and M. C. Monard, "PRETEXT II: Description of restructuring tool preprocessing of texts,"ICMC-USP, Tech. Rep. 333, 2008, (in portuguese).