

Fuzzy k-mode clustering a Document Summarization System for Document Retrieval

Dr. Poonam Yadav¹

¹D.A.V College of Engineering & Technology, India

poonam.y2002@gmail.com

ABSTRACT

Document retrieval has received much attention by the data mining community since the growth of World Wide Web and tremendous information storage happened in IT field. Various algorithms are developed for document retrieval using different document features and similarity measure. But, computational time incurred for the algorithm is increasing whenever a new set of feature and similarity measures are included in the retrieval process. The computational effort should be minimized without much affecting the effectiveness of the results. This can be new direction of research happening in the data mining community recently. Accordingly, summarization system is used for shorten the text document and the retrieval process is done based on the contents presented in the summary. In our work, clustering process is also utilized along with summarization system for further reducing the complexity in retrieval process. Accordingly, fuzzy k-mode clustering and keyword-based features are utilized here for the document retrieval algorithm. The advantage of this method is that the matching can be done only with the representative keywords identified from the clustering algorithm. Finally, the experimentation is done with 100 documents and the results are compared with existing algorithm using precision, recall and F-measure.

Key Words: Document retrieval, World Wide Web, fuzzy k-mode clustering, Summarization, Similarity measure.

1. INTRODUCTION:

Due to the availability of immense data for the past fast years, a lot of time gets wasted for the user for browsing the Internet as well as retrieving for the information as per the user's desire. This leads to the task of searching, accessing, displaying, integrating and preserving the data tougher. The ultimate aim of the document retrieval systems is to discover the suitable documents to a user based on their query, where the query is a set of keywords. Document Retrieval [1, 4] is commonly called as Information Retrieval, where the user's request are matched with an automatically created index of the textual content of documents in the system for obtaining a record that consists of the list of documents in relevant to the user's query. Currently, search engines are being widely utilized for retrieving information from various resources around the world. Search engines are inefficacious in document clustering and indicating the significant level of the documents extracted from the databases.

In this paper, an algorithm for document retrieval is developed using fuzzy k-mode clustering and keyword-based features. At first, input documents are pre-processed using stop word removal and stemming. The important keywords identified from the previous step are

then given for the summarization system where, term weight, cue phrases and thematic word-based features are utilized to select the importance sentences. Then, summaries of the documents are given for fuzzy k-mode clustering to do the grouping process and cluster centroids are selected as representatives for indexing the documents. For a user query, representatives are matched with the user query and the relevant clusters are identified. The document belonging to the relevant clusters are given as the retrieved documents for the user. The paper is organized as follows; Section 2 explains the existing work and section 3 presents the proposed document retrieval process. Section 4 provides the experimental results and conclusion is given in section 5.

2. EXISTING WORK:

Recently, summarization system [2, 3] are applied to information retrieval for reducing the computation overhead of the process. Accordingly, Zhou and Li [1] have presented the use of multi-document summarization as post-processing step in information retrieval (IR). They have analyzed the differences between requirements for general multi-document summarization and necessities when it is applied for IR, and highlighted the requirements for information retrieval, which is much supportive to the users for

browsing and searching comparative results. They have utilized text summarization for generating the summary from the input document and the query words are matched with summary for information extraction. By taking this as motivational one, clustering process [5-10] is also included in this paper for reducing the computational time in the matching process because the existing algorithm matched user query with the summary but here, matching of user query is done only with the representatives of cluster. Further, every summary related to document is used for matching process in the existing algorithm. But, representatives belonging to clusters are only used for matching in the proposed algorithm.

3. PROPOSED METHOD: FUZZY K-MODE CLUSTERING AND DOCUMENT SUMMARIZATION SYSTEM FOR DOCUMENT RETRIEVAL:

This paper presents an effective document retrieval method using fuzzy k-mode clustering and document summarization. The document summarization generates the summary of the document which can be then utilized for clustering process. This avoids the computation overhead of the algorithmic process. Also, for a query, the documents are not matched directly even with the summary but with the representatives obtained from the fuzzy k-mode clustering process. The overall block diagram of the proposed method of document retrieval is explained in figure 1.

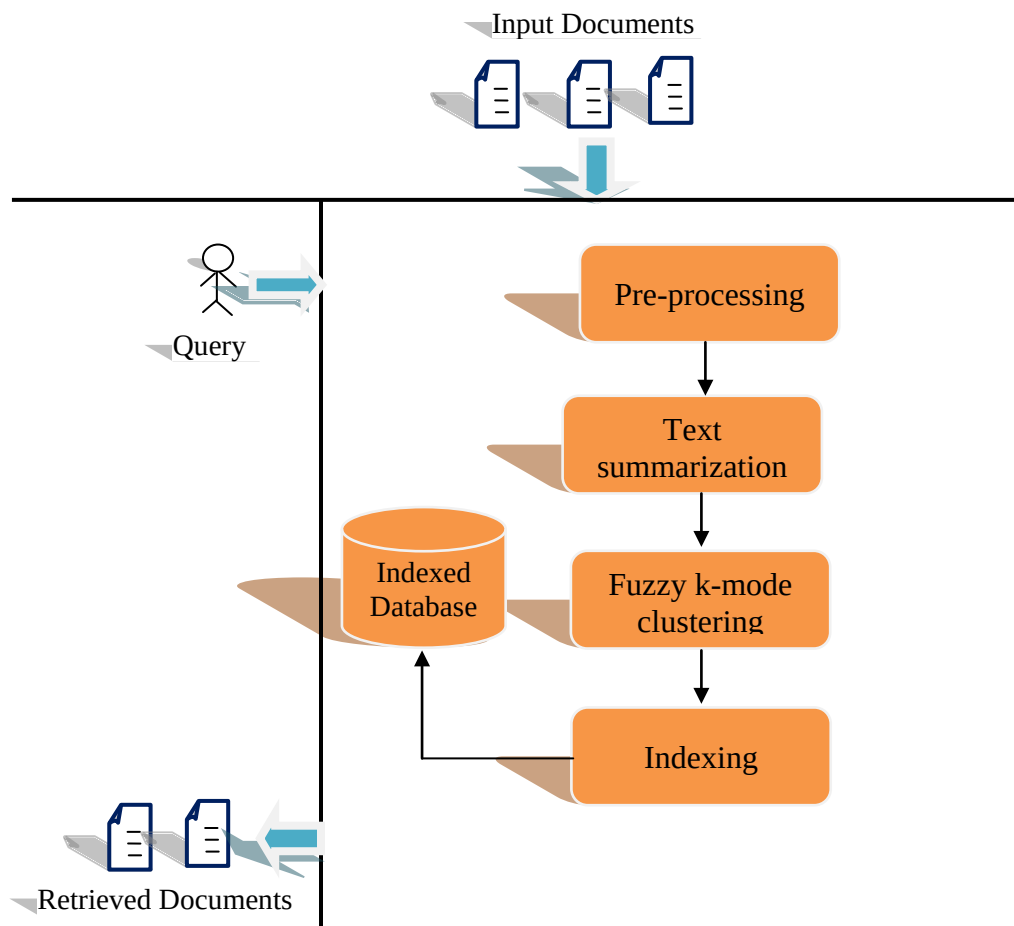


Figure 1: Block diagram of Fuzzy k-mode clustering and document summarization system for document retrieval

A. Pre-processing:

At first, input documents are read out through programming and the indexing should be done for matching the user query. Once the documents are read out from the document, keywords are extracted by removing the delimiters. Then, the words which are not

giving any meaning if it is considered separately are removed from the list. The common stop words are like, 'a', 'an', 'the', 'may', 'might' and so on. Again, derived words are rooted to its original form if those are in the different tense format. After this process, important

keywords are identified and stored as the contents of the respective document.

B. Document summarization using keyword-based features:

The keywords represented for every document are then given for document summarization process which selects only the most significant sentences from the document through different keyword-based feature. The reason of bringing this summarization system is lessen the computation overhead happened when the clustering process is done. If the clustering is performed with summary instead of document, the processing time is less because the length of document is less. The summarization of the input documents is performed using three different keyword-based features. Every sentence are obtained a score value using these three features and then, final score value of every sentence is obtained by taking average of score values. The three features considered in this work are term weight, thematic feature and cue-phrases.

Term weight-based score: Term weight of a sentence is computed for every sentence by taking the ratio of the sentence factor to the maximum sentence factor in the given text document. The weight factor is computed by multiplying the term frequency and inverse document.

$$TW(S) = \frac{S_F}{\max_{j \in D} (S_F(j))} \quad (1)$$

$$S_F = \sum_{i=1}^n WF_i \quad (2)$$

$$WF = TF \times ISF \quad (3)$$

$$ISF = \log (N_S / N_S(w)) \quad (4)$$

Where, S_F is sentence factor, WF is weight factor of the word in a sentence, n is number of words in a sentence, TF is number of occurrences of the term or word in a text document, ISF is Inverse Sentence Frequency, N_S is Total number of sentences in a document and $N_S(w)$ is total number of sentences that contain the word (w).

Thematic features: Thematic words are the words which are occurred in a document more frequently. The number of thematic words signifies the words or terms with maximum possible relativity. The score of this feature for every sentence is calculated by taking the ratio of

thematic words presented in a sentence to total number

$$\text{of thematic words. } T(S) = \frac{N_T(S)}{N_T(D)} \quad (5)$$

Where, $N_T(S)$ is the number of thematic words presented in sentence and $N_T(D)$ is the number of thematic words presented in a document.

Cue-phrases: Commonly, the phrases such as “in summary”, “in conclusion”, and superlatives such as “the best”, “the most important”, “according to the study”, “hardly” are used to mention the importance of the document. So, this measure considers the score value based on the presence of the cue-phrases in the document.

$$C(S) = \frac{N_{CP}(S)}{N_{CP}(D)} \quad (6)$$

Where, $N_{CP}(S)$ is number of cue-phrases in the sentence and $N_{CP}(D)$ is number of cue-phrases in the document.

C. Indexing word selection using Fuzzy k-mode clustering:

The document summaries are then given for the clustering process not only for grouping the documents but also, to select the indexed words for the grouped document. Fuzzy k-mode clustering is categorical clustering algorithm used to group the summaries based on the representatives (centroid) which are then used for indexing the document. Let we take a database $D = \{S_1, \dots, S_n\}$ of having ‘n’ set of summaries belong to ‘n’ document. The summaries are to be clustered, where each object $S_i = (s_{i,1}, \dots, s_{i,m})$, $1 \leq i \leq m$ is defined by m keywords. The process is explained as follows:

Step 1: Take initial representatives as random way by selecting ‘m’ keywords for every cluster.

Step 2: Then, membership matrix, u_{kl} is computed as,

$$u_{kl} = \frac{1}{\sum_{t=1}^C \left(\frac{\|x_k - c_l\|}{\|x_k - c_t\|} \right)^{\frac{2}{m-1}}} \quad (7)$$

The dissimilarity of two data points is computed based on the following equations.

$$\|x_j - c_l\| = \sum_{j=1}^m \delta(x_j, c_j) \quad (8)$$

$$\delta(x_j, y_j) = \begin{cases} 0; & \text{if } x_j = c_j \\ 1; & \text{otherwise} \end{cases} \quad (9)$$

Step 3: The keywords of every attributes within the cluster C_i are arranged in accordance with their relative frequency. The keywords with high relative frequency of every attributes are chosen for the new representative.

Step 4: The processes given in step 2 and step is repeated until it satisfies the condition,

$$\| (u^{k+1}), (u^k) \| < T$$

The final representative obtained for every cluster is taken as indexed word for the corresponding cluster.

D. Retrieval of documents using similarity measure:

This step find the score value of every indexed word with user query Q and the retrieval process is done as per the score value of the cluster.

$$C_s = \frac{1}{Q_w} \sum_{i=1}^{S_w} \left(\frac{M_{Qw}}{S_w} \right) \quad (10)$$

Where, Q_w is the number of query words and M_{Qw} equal to one when the query word is matched with the indexed word. S_w is the number of indexed words. This score value is found out for all the indexed words stored in the database with respect to the query words. Then, clusters are sorted based on the score value and documents belonging to the cluster which are having the highest score value is given for the user in a ranked way as an output.

4. EXPERIMENTAL RESULTS:

This section presents experimental results and discussion of the proposed Fuzzy k-mode clustering and document summarization-based system for document retrieval. The

proposed algorithm is implemented with 100 documents having two groups, one is related with sports articles and other one is related with politics' related articles. Every group contains 50 documents and it is given as input to the algorithm for generating the summary. Then, two query words, such as 'politics' and 'sports' are given to the algorithm for retrieval process. The evaluation of information retrieval system is done using Precision, Recall and F-measure.

$$\text{Precision, } P = \frac{RT \cap RL}{RT} \quad (11)$$

$$\text{Recall, } R = \frac{RT \cap RL}{RL} \quad (12)$$

$$F - \text{measure, } F = \frac{2 * P * R}{(P + R)} \quad (13)$$

Where, RT is retrieved document and RL is the relevant document

A. Comparative analysis:

The comparison of the proposed method with the existing algorithm is done using three metrics namely, precision, recall and f-measure. For the experimentation, two queries such as, 'politics' and 'sports' are given to the proposed retrieval system and the retrieved documents are compared with the original relevant documents using the above given formulae. The figure 2 shows the precision plot of the proposed retrieval system. Here, proposed method obtained the value of 78% for the 'sports' query and 74% for 'politics' query. Recall value reached by the proposed method is 76% and for the existing method, it is 70% as per figure 3. From the figure 4, the proposed method obtained the f-measure values of 77% for the 'sports query' and 72% for the 'politics' query while compared with the existing algorithm.

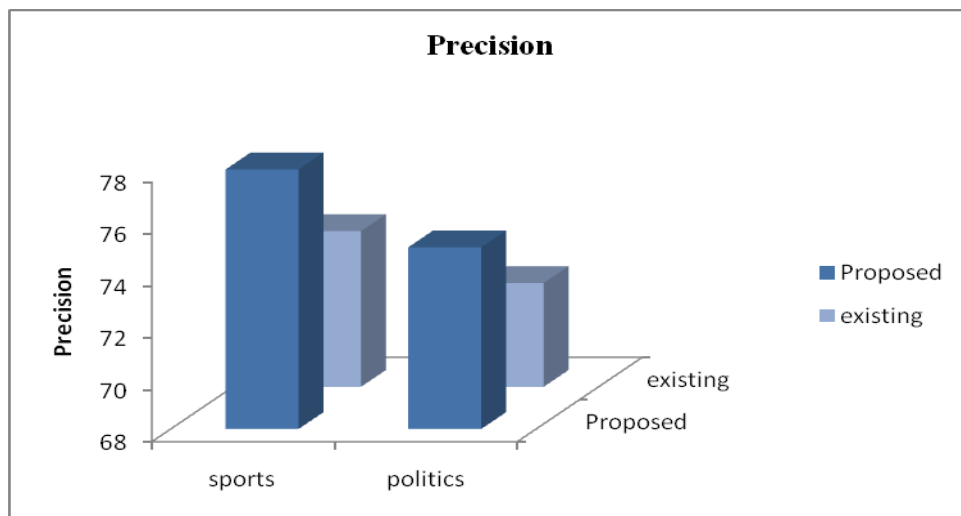


Figure 2: Precision graph

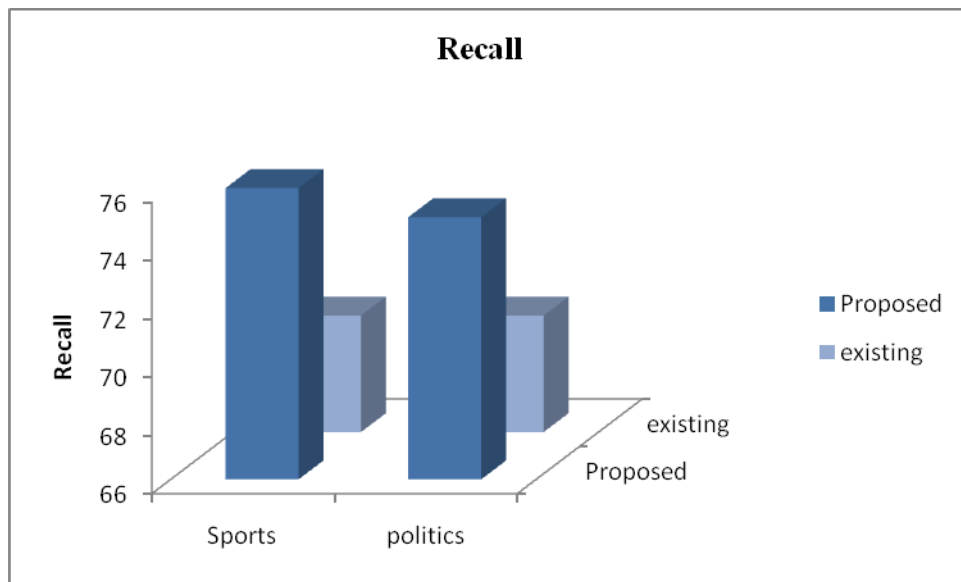


Figure 3: Recall graph

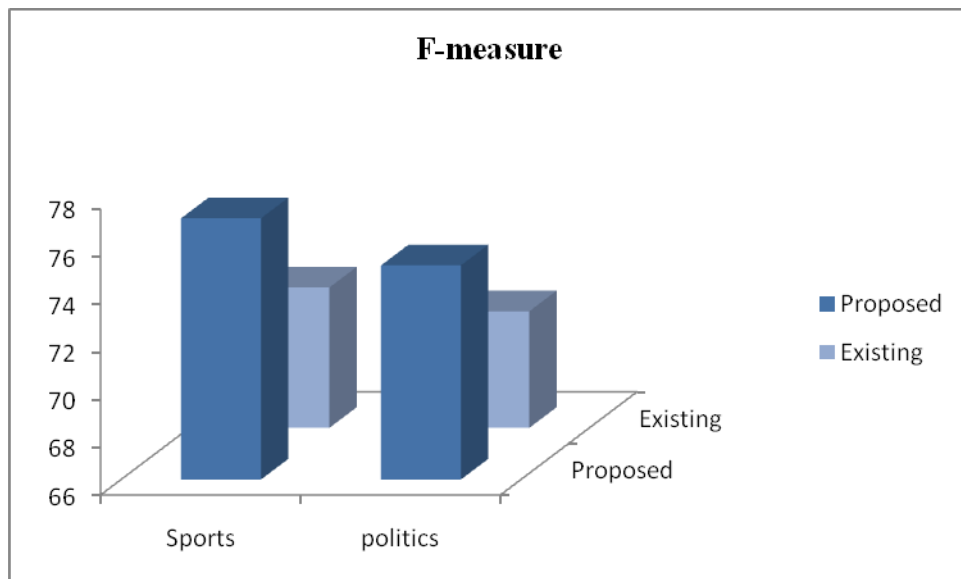


Figure 4: F-measure graphs

5. CONCLUSION:

This paper presents a fuzzy k-mode clustering and document summarization system for document retrieval. The input documents were indexed as representative keywords which are identified from the fuzzy k-mode clustering. Before that, clustering process obtained summary as an input which was generated from the summarization system using keyword-based features. The representatives of a set of documents are then utilized to match the user query and the relevant cluster is identified based on the similarity score value. The documents related to the particular cluster are given for the users as the final output. Finally, experimentation is

done with 100 document and the results are compared with existing algorithm using precision, recall and F-measure. From the results, the proposed method obtained the f-measure values of 77% for the 'sports query' and 72% for the 'politics' query while compared with the existing algorithm.

6. REFERENCES:

1. Zhou, D. ; Lei Li, "Multi-Document Summarization as Applied in Information Retrieval", in proceedings of International Conference on Natural Language Processing and Knowledge Engineering, pp. 203 - 208, 2007.

2. Chien Chin Chen and Meng Chang Chen, "TSCAN: A Content Anatomy Approach to Temporal Topic Summarization", IEEE transactions on knowledge and data engineering, Vol. 24, No. 1, pp. 170 – 183, 2012.
3. Hien Nguyen, Eugene Santos and Jacob Russell, "Evaluation of the Impact of User-Cognitive Styles on the Assessment of Text Summarization", IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 41, No. 6, pp.1038 -1051, 2011.
4. Comfort T. Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe and Ifio J. Udo, "A Fuzzy-Ontology Based Information Retrieval System for Relevant Feedback," International Journal of Computer Science, Vol. 8, No.1, pp. 382 – 389, 2011.
5. S.R. Kannana, S. Ramathilagam and P.C. Chung, "Effective fuzzy c-means clustering algorithms for data clustering problems", Expert Systems with Applications, Vol. 39, No.7, pp.6292–6300, June 2012.
6. Jinchao Ji, Wei Pang, Chunguang Zhou, Xiao Han and Zhe Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", Journal of Knowledge-Based Systems, Vol. 30, pp.129-135, June 2012.
7. Swagatam Das, Ajith Abraham and Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, pp. 218-237, Jan. 2008.
8. José A. Castellanos-Garzón and Fernando Diaz, "An evolutionary computational model applied to cluster analysis of DNA microarray data", Expert Systems with Applications, Vol. 40, No.7, pp.2575–2591, Jun.2013.
9. Jianbin Huang, Heli Sun ; Qinbao Song ; Hongbo Deng ; Jiawei Han, "Revealing Density-Based Clustering Structure from the Core-Connected Tree of a Network", IEEE Transactions on Knowledge and Data Engineering, Vol.25, No.8, pp. 1876-1889, Aug. 2013.
10. Jie Cao, Zhiang Wu, Junjie Wu and Hui Xiong, "SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering ", IEEE Transactions on Cybernetics, Vol.43, No.2, pp. 570-584, Apr.2013.



Dr. Poonam Yadav obtained B.Tech in Computer Science & Engg. From Kurukshetra University Kurukshetra and M.Tech in Information Technology from Guru Govind Singh Indraprastha University in 2002 and 2007 respectively. She had Awarded PhD. In Computer Science & Engg. From NIMS University, Jaipur. She is currently working as Principal in D.A.V College of Engg. & Technology, Kanina (Mohindergarh). Her research interests include Information Retrieval; Web based retrieval and Semantic Web etc. Dr. Poonam Yadav is a life time member of Indian Society for Technical Education and her email id is poonam.ir@gmail.com.