International Journal of Computer Systems (ISSN: 2394-1065), Volume 02– Issue 04, April, 2015
Available at http://www.ijcsonline.com/

Document Features-Enabled Text Summarization System for Information Retrieval

Poonam Yadav

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

Abstract

Information retrieval has become a challenging problem due to the explosion of information availability. Retrieving of suitable documents for the user's query has become big problem in document retrieval tasks. This paper presents a method called, document features-enabled text summarization system for document retrieval. In this method, document database is applied to the summarization system which finds score value based on the important sentences in the document. Here, three different document features, such frequency-based, title-based and position-based are utilized to find the importance of the sentences. The sentences are then ranked in the order of their score values and the most informative sentences are selected for the summary. For a user query, summary is matched with query to find the documents which are most suitable to the user query. The proposed algorithm is implemented with 100 web documents and the performance is evaluated with precision, recall and f-measure. From the results, for a query 'sports', the proposed method achieved the highest value of 80% as compared with existing algorithm

Keywords: *Information retrieval; Summarization; Query; Precision; Recall; F-measure.*

I. INTRODUCTION

The increasing accessibility of online information has caused numerous problems for information access. Finding similar documents for the user's query is a major challenge in document retrieval tasks. Document Retrieval (usually referred to as Information Retrieval) [1-4] is the computerized process of generating a list of documents that are relevant to an inquirer's request by matching the user's request to an automatically produced index of the textual content of documents in the system. Usually, documents having one or more query words are retrieved to the user. Such methods will, nevertheless, lack to retrieve relevant documents that do not share words with users' queries. One motive for this is that the standard retrieval models (e.g., Boolean, vector space, probabilistic) take words as if they are pair-wise orthogonal or independent, even if it is moderately understandable that they are not. Commonly, keyword-based retrieval retrieves imprecise and imperfect outputs when various keywords are utilized to illustrate the same concept in the documents and in the queries.

In this paper, information retrieval for an input query is done using document features and document summarization. At first, input documents are summarized using a set of feature extraction techniques after performing the preprocessing techniques, such as sentence segmentation, stop word removal and stemming. Then, indexing of input documents are done based on the summary of the documents. For a user query, the relevant document can easily obtainable by matching the summary with query of user [11]. The paper is organized as follows: Section 2 presents existing algorithm and section 3 presents the proposed algorithm for information retrieval. Section 4 presents the experimental result and finally, conclusion is given in section 4.

II. EXISTING WORK

Recently, Natural Language Processing (NLP) techniques are applied to information retrieval. One of the NLP process called, text summarization [5-10] is used for document retrieval by Zhou and Li [1] who 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 used text summarization for generating the summary from the input document and the query words are matched with summary for information extraction. This work lacks effective document feature and similarity measure for matching with summary and query. This has been taken into forward in this paper and utilized three different document features and similarity measure for document retrieval.

III. PROPOSED METHODOLOGY

This section presents the proposed method of document retrieval using text summarization which is the well-known method in natural language processing (NLP) for getting the summary from the input document. The ultimate aim of this work is to get the most relevant

information stored as document in the database for the user in an effective way. Here, the intention of using text summarization method for this information retrieval system is to strengthen the retrieval effectiveness by storing the most important sentences which can signify the input document more precisely as compared with other keyword-based indexing techniques. The block diagram of the proposed method of document retrieval shown in Figure 1 contains three important steps such as, preprocessing, summarization and matching.

A. Preprocessing

At first, input database of having ‗m' documents is given as input to the proposed system. Every document is then undergone preprocessing steps which transform the raw document as useful document using sentence segmentation, stop word removal and stemming.

The input document read out from the input database is directly applied to the sentence segmentation where delimiter (.) is used to split the document into a set of sentences and a separate ID is given for every sentence. After that process, stop words are removed from every sentences to make sentence having only useful keywords. Commonly, ‗a', an', ‗then', can', ‗could' are some of the stop words identified.

The removal of these words can be done from the sentence through the direct comparison of the stop words stored in a list. Once we remove the stop words from the sentences, rooting of all the derived words to its original form should be done to avoid confusion of the same two words which are in two different formats.

B. Document features-enabled text summarization system

Summarization of a text document reduces the document to create a summary which contains only the important points from the original document. The summary must be short and should preserve all important points from the document. Differentiating the most informative parts from less informative parts of the document is the main challenge in document summarization. From the summary, user can get the main points from the results without going deep into the content of the documents.

The next step of the proposed system is to generate the summarized report of every input document. The reason of doing summarization in this step is to shorten the input documents in a concise way and at the same time to preserve the main core of the document. Also, the matching of query with the direct document lead to computational overhead due to the complexity of large number of matching. This summarization avoids these computational overhead without compensating the main core of the document.

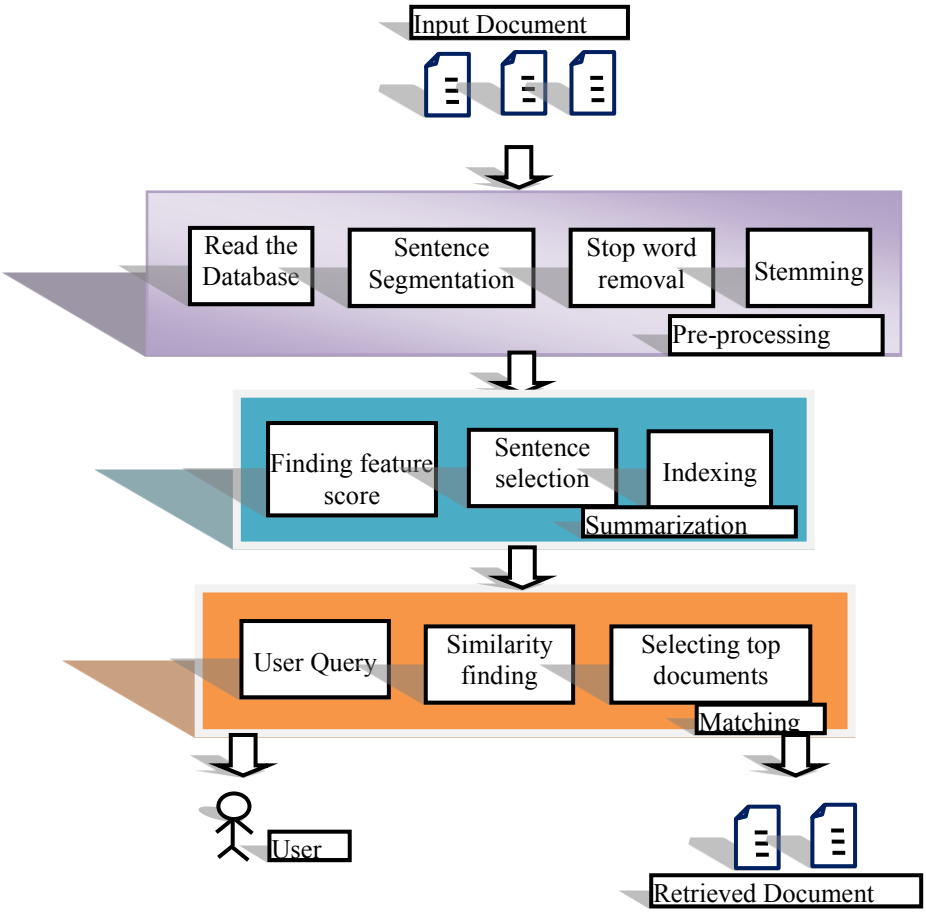

Figure.1. Block diagram of Document features-enabled Text summarization system for information Retrieval

The working principle of summarization system is to assign score values based on the importance model and to select the top sentences based on aggregation of these score values. By taking this principle, three different feature models are used here to weight the sentence. The first one is frequency based score which can score the sentence based on the assumption that frequent words carry the most important information. The second one is title based score which gives more score value for a sentence if title keywords are presented in the sentence. The third one is position based score which can weigh the sentence based on the location placed in the document. These features model are formulated as a mathematical formula and is given below.

3.2.1. Frequency based: It is based on the frequency of the appearance of a particular word in the document. The most important words appear more frequently in the document.

$$F_s(s) = \frac{1}{n} \sum_{i=1}^{n} \frac{f_i}{Ns} \quad (1)$$

3.2.3. Title relevancy: Here, the sentence weight is computed as a sum of all the content words appearing in the title of the document.

$$T_s(s) = \frac{1}{m} \sum_{i=1}^{m} \frac{f_i}{Ns} \quad (2)$$

3.2.3. Position method: Position method is based on the assumption that the sentences occurring in the initial positions of the test and paragraphs are more relevant.

$$P_s(s) = 1 - \frac{P_i}{S_T} \quad (3)$$

where, n is the number of words presented in the document, f_i is the frequency of the word, Ns is the highest frequency, m is the number of title words, S_T is the total number of sentences. The final score of each sentence is determined by the following equation.

$$S(s) = \frac{1}{3} * \left(F_s(s) + T_s(s) + P_s(s)\right) \quad (4)$$

All the sentences are ranked in the order of their importance and the sentences with high ranking are selected to form the summary. After finding three score values such as frequency based score, title based score and position based score for every sentences, the final score value is the average of these score values which signify the importance of that sentence in the particular document taken for summarization. Once we generate final score value for every sentences, the sentence selection can be done through the compression ratio given by the user. Finally, input document is converted into a summary which is stored in the database for matching.

C. Retrieval using Matching

In the retrieval process, user query Q is matched with the summaries stored in the indexed database using the following equation.

$$D_S = \frac{1}{Q_w} \sum_{i=1}^{S_w} \left(\frac{M_{Qw}}{S_w} \right) \quad (5)$$

where, Q_w is the number of query words and M_{Qw} equal to one when the query word is matched with the summary. S_w is the number of words presented on the summary. This score value is found out for all the summaries stored in the database with respect to the query words. Then, documents are sorted based on the score value and documents are given for the user in a ranked way as an output.

IV. RESULTS AND DISCUSSION

This section presents experimental results and discussion of the proposed document feature based document retrieval.

A. Experimental setup

The proposed algorithm is implemented with 100 web 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.

4.1.1. Precision (P): Precision (P) is the number of relevant documents that are retrieved divided by the total number of documents retrieved.

$$Precision = \frac{relevant\ items\ retrieved}{retrieved\ items} = \frac{relevant\ documents \cap retrieved\ documents}{retrieved\ documents}$$

4.1.2. Recall (R): Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. Recall (R) is the number of relevant documents that are retrieved divided by the total number of existing relevant documents.

$$Recall = \frac{relevent\ items\ retrieved}{relevent\ items} = \frac{relevant\ documents \cap retrieved\ documents}{relevant\ documents}$$

the recall based measure determines how well a system-generated summary covers the content present in the original documents.

4.1.3. F-measure: F-measure is the harmonic mean of precision and recall.

$$F = \frac{2PR}{P + R}$$

B. Performance analysis

Figure 2 plot the precision of the proposed and existing method. From the graph, the proposed method achieved the highest precision value. Figure 3 plot the recall of the proposed and existing method. From the results, for a query ‗sports', the proposed method achieved the highest value of 80% as compared with existing algorithm. F-measure plot is given in figure 4. In F-measure also, the proposed method outperformed the existing method.

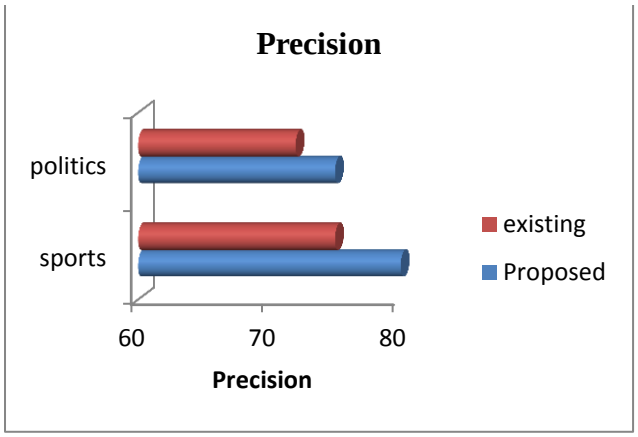

Figure 2. Precision graph

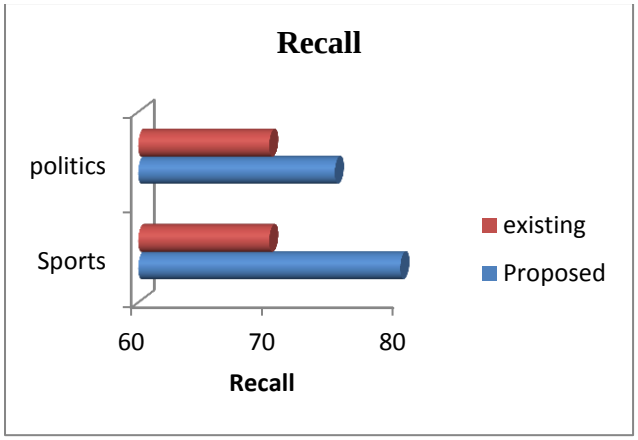

Figure 3. Recall graph

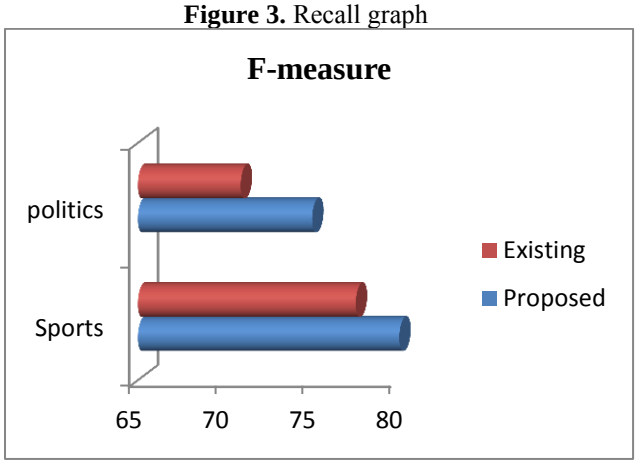

Figure 4: F-measure graphs

V. CONCLUSION

In this paper, document feature-enabled document summarization system for document retrieval method in information retrieval. Once the preprocessing was performed using stop word removal and stemming, relevant features were extracted from the sentences to generate the summary of the document. For a user query, words belonging to query is matched with summary using similarity measure to rank the documents presented in the database. The proposed document retrieval method was implemented with 100 documents having two groups, sports articles and politics' related articles. The performance of the proposed algorithm was analyzed with precision, recall and f-measure. From the experimentation evaluation, the finding is that the proposed method achieved the highest precision value. Also, for a query ‗sports', the proposed method achieved the highest value of 80% as compared with existing algorithm.

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] Pablo Castells, Miriam fernández, and David Vallet, "An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval," IEEE Transactions on Knowledge and Data Engineering,Vol.19, No.2, pp.261 - 272, 2007.

[3] Comfort T. Akinribido, Babajide S. Afolabi, Bernard I. Akhigbe and Ifiok 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.

[4] U.K.Sridevi and N. Nagaveni, ―Ontology based Similarity Measure in Document Ranking," International Journal of Computer Applications, Vol.1, No.26, pp.125–129, February 2010.

[5] Hien Nguyen, Eugene Santos and Jacob Russell,"Evaluation of the Impact of User-Cognitive Styleson the Assessment of TextSummarization", IEEE Transactions on Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 41, No. 6, pp.1038 -1051, 2011.

[6] 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.

[7] Rasim Alguliev, Ramiz Aliguliye, Makrufas Hajirahimov and Chingisz A.Mehdiyev, ―Maximum coverage and minimum redundant text summarization model", Expert systems with applications, vol. 38, no. 12, pp. 14514–14522, 2011.

[8] D.Y. Sakhare, Dr. Raj Kumar, ―Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization", International Journal of Information Technology and Computer Science (IJITCS), pp. 38-46, 2014.

[9] Rajesh Shardanand Prasad and UdayKulkarni ―Implementation and Evaluation of Evolutionary Connectionist Approaches to Automated Text Summarization", Journal of computer science, Vol. 6, no. 11, pp. 1366-1376, 2010

[10] Anjali R. Deshpande , Lobo L. M. R. J. ,"Text Summarization using Clustering Technique", International Journal of Engineering Trends and Technology(IJETT), vol. 4, no. 8, 2013.

[11] Poonam Yadav, ―Fuzzy k-mode clustering a Document Summarization System for Document Retrieval", In IJETCR, Vol. 2, No. 6, 2014.