Journal of Engineering and Applied Sciences 12 (21): 5533-5536, 2017
ISSN: 1816-949X

Analysis of Cluster Based Document Condensation Techniques

Mrunal S. Bewoor and Suhas H. Patil
Department of Computer Engineering,
College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India

Abstract: Availability of huge amount of text data and increase of organizational spread over has arises the need to control their data corpora, especially with the availability of big data platforms. People does not have sufficient time to read and understand each document to make decisions based on document content. This has resulted in a great demand to summarize text documents to provide the end user a representative substitute for the original text input. This arises a need to identify techniques that performs precised summary retrieval through search queries against input documents. The user expects this process in a optimum way. To improve this process of querying against the full spectrum of original documents several generic algorithmms for text summarization have been developed, each with its own advantages and disadvantages. The study conducts a survey and analysis of the cluster based summary techniques obtained through expectation maximization, DBSCAN, graph based method, hierarchical and fuzzy C-means clustering algorithms. The results of the summaries obtained using these algorithms are evaluated with the parameters precision, recall, F-measure, compression ratio and retention ratio. The study aims at the analysis, investigation, design and development of various metrics which may help the end user regarding the selection of optimal query based technique.

Key words: Text summarization, unstructured data, text mining, document clustering, regarding, optimal query

INTRODUCTION

In today's world internet usage is increasing exponentially with the advent of internet there is a vast growth in the usage of digital information. The usage of online information services, social media has become day to day need for a human being. This has caused the availability of information in a digital form. It shows that huge amount of information instantaneously available and directly accessible to a large number of end-users. User expects faster and precise results which resulted in the maximum use of information on web. The information retrieval systems are developed to tackle this problem. On the other hand; human being has limited ability to understand and organize this huge number of documents.

This information overload problem is most sensitive when there is a necessity to take any decision or requirement of deep study for a certain problem. The IR systems enables this through user supplied queries The result obtained through these overwhelms users with too many results and provide documents that may not be relevant to the topic being asked by the user. For example, if the user is searching using some keyword and the search engine finds it somewhere inside a document that document will be a "search hit" even if the document is not really relevant to the keyword query entered by the user.

The task of retrieving the information from the web relevant to the user query has become tedious. Various information retrieval tools can be used for retrieval of relevant information. The results obtained sometimes may not preserve the required contents. Summary generation or automatic text summarization is the technique where a computer program automatically creates an abstract, or summary of one or more texts i.e., creation of shortened version of text. There are specifically two types of Text Summarization techniques viz. Generic and Query Specific (Kaushik and Naithani, 2016).

In generic text summarization, the most important sentences from the given input text are extracted (independent of any query) and considered as a condensed text. In query specific text summarization, the sentences containing the words matching with the keywords of query are retrieved, scored and added into the summary. It becomes a difficult task for the end user to go through large number of relevant retrieved documents. The solution to this problem will be the use of

Corresponding Author: Mrunal S. Bewoor, Department of Computer Engineering, College of Engineering, Bharati Vidyapeeth Deemed University, Pune, India

query specific document summary generation. The generated summary or retrieved abstract should preserve the semantics and central idea of an input text.

The features of clustering algorithms and NLP based retrieval can be useful in retaining the gist of the information in the retrieval process.

MATERIALS AND METHODS

System architecture: Text summarization is a dual task, first is to identify the most important portions of text and second is to generate the coherent summaries. Information Retrieval process is used for searching documents on internet. Since, huge amount of amorphous data is available on the internet. The use of IR tools has given rise to the requirement of query dependent document condensation. The system has demoralized the Natural Language Processing techniques to support a range of Natural Language queries. This type of query processing for text summarization may result in imprecise summary and user may not view the correct or consistent results.

The system accepts input in the text form. The pre-processing step consists of phases of natural language processing such as sentence detection, tokenization, part-of-speech tagging; chunking and parsing. The OpenNLP tool is used for implementation of natural language processing of text for word matching. The output of a pre-processing step is represented in a document matrix form. This document matrix is given as an input to various clustering algorithms such as EM, graph based method, fuzzy C-means, DBSCAN and hierarchical clustering algorithms and a query specific summarization.

Frameworks for summary generation and evaluation have been proposed by Varadarajan and Hristidis (2006), Gholamrezazadeh *et al.* (2009), Teng-Kai and Chia (2010), Mean *et al.* (2010) and Walters (2011) which have defined the various metrics for summary evaluation.

RESULTS AND DISCUSSION

A significance of clustering algorithms in document summarization: Clustering can be very useful in the text domain as objects of a document as words, sentences, paragraphs to be clusters are of varying granularities. Clustering is especially useful to organize documents to improve retrieval and support browsing. Roma *et al.* (2013ab) has identified and selected these clustering algorithms for obtaining the summary of the document. The main motive of the research is to extract the sentences in the summary those are more relevant to the original input text by using features of clustering algorithms which can group the objects based on the relevancy. This is an attempt to combine two major approaches of summary generation one is extractive and the other is abstractive (Fig. 1).

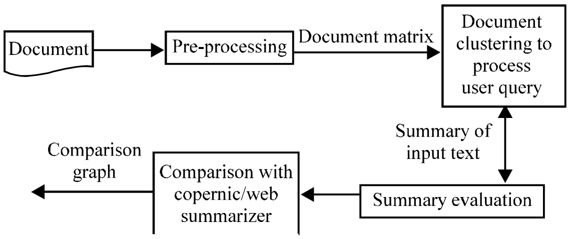

Fig. 1: System architecture

The generated summary is evaluated for the performance analysis of the used methodology. The results of these evaluation measures are compared for precision, recall, F-measure, compression ratio and retention ratio. The summaries generated are also compared with existing query summarizers copernic summarizer and web summarizer. These two tools are query based summarizers. The values for these parameters for the corresponding document are calculated using following formulas.

Precision indicates the probability at which the retrieved document is relevant in the search:

$$\text{Precision} = \frac{\text{No. of different terms in summary}}{\text{No. of different terms in query}}$$

Recall is the probability that relevant document is retrieved in the search:

$$\text{Recall} = \frac{\text{No. of correct matching sentences in the summary}}{\text{No. of all relevant sentences in the original document}}$$

F-measure is the harmonic mean of precision and recall both are of equal impertinence:

$$\text{F-measure} = \frac{2\times\left(\text{Precision}\times\text{Recall}\right)}{\left(\text{Precision}+\text{Recall}\right)}$$

$$\text{Compression ratio} = \frac{\text{No. of sentences in summary}}{\text{Total No. of sentences in original document}}$$

$$\text{Retention ratio} = \frac{\text{No. of relevant query words in summary}}{\text{No. of query terms in original data}}$$

Table 1: Performance evaluation

Methodology	Precision	Recall	F-measure	Compression ratio	Retention ratio	Execution time
Expectation maximization clustering	1	0.10000	0.18182	0.04776	0.70394	5.516
Fuzzy C-means	1	0.07627	0.14173	0.03582	0.86227	78.000
Query specific (graph based)	1	0.06040	0.11392	0.06567	0.88292	1.313
Copernic	1	0.52630	0.10000	0.08060	0.63737	1.451
Web summarizer	1	0.52630	0.10000	0.08060	0.57080	786.000

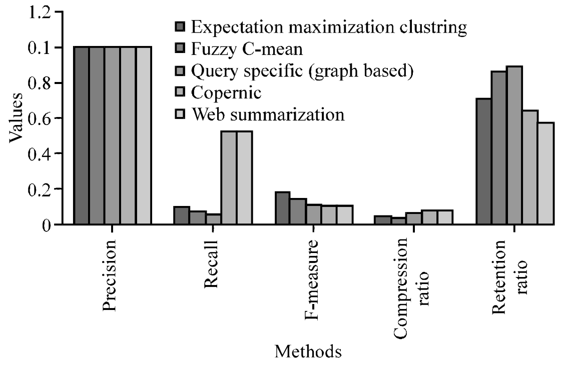

Fig. 2: Comparison graph of all the summaries obtained with various methods

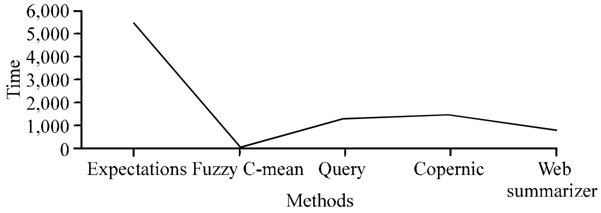

Fig. 3: Comparison graph of execution time for various methods

Precision, recall, F-measure are concerned with the quality of the summary so, these parameters are called qualitative parameters along with the execution time (Teng-Kai and Chia, 2010). The compression ratio and retention ratio measure only the length or quantity of the sentences in the summary. These parameters are called quantitative parameters (Table 1, Fig. 2 and 3).

The system is executed for various sample inputs. The analysis of the different methodologies based on the above parameters are obtained for all those sample inputs. It is observed that performance of the algorithm varies Expectation maximization algorithm gives better results for large size documents. Expectation maximization algorithm and fuzzy C-means gives better results with respect to the context within the document when file size is moderate. Precision, recall and F-measure should be accepted as better metrics than compression ratio, retention ratio and execution time as those are concerned with the gist of the original document.

CONCLUSION

The summary of the document is generated using query based approach as well as using expectation maximization, fuzzy C-means. The results generated by all these methods are evaluated considering both the quality of the result with respect to the context of the original text input as well as the length of the original input text. The goal of document condensation first focuses on the context, the length is the secondary aspect. Fuzzy C-means generates better summary. Considering both quantity and quality parameters though clustering is an unsupervised text summarization technique, it can be used as supervised technique by integrating it with a supervised approach. This may give an optimal solution for this problem The research should focus on improving the quality of clusters which directly relates with the gist of the original input document.

ACKNOWLEDGEMENTS

I wish to thank all those who have directly or indirectly contributed for the successful compilation of this study. First, I would thank Prof. Dr. Suhas H. Patil for his support and encouragement and valuable guidance. I would like to thank all the researchers whose study I have referred and they have directly or indirectly supported for completion of the research.

REFERENCES

Gholamrezazadeh, S., M.A. Salehi and B. Gholamzadeh, 2009. A comprehensive survey on text summarization systems. Proceedings of the 2nd International Conference on Computer Science and its Applications, December 10-12, 2009, IEEE, Jeju, Korea, ISBN:978-1-4244-4945-3, pp: 1-6.

Kaushik, A. and S. Naithani, 2016. A comprehensive study of text mining approach. Intl. J. Comput. Sci. Network Secur. IJCSNS., 16: 69-76.

Mean, F.O., A. Oxley and S. Suziah, 2010. Challenges and2. trends of automatic text summarization. Int. J. Inf. Telecommun. Technol., 1: 34-39.

Roma, V.J., M.S. Bewoor and D.S.H. Patil, 2013a. Automation tool for evaluation of the quality of NLP based text summary generated through summarization and clustering techniques by quantitative and qualitative metrics. Intl. J. Comput. Eng. Technol. IJCET., 4: 77-85.

Roma, V.J., M.S. Bewoor and S.H. Patil, 2013b. Evaluator and comparator: Document summary generation based on quantitative and qualitative metricsfor. Intl. J. Sci. Eng. Res., 4: 1111-1115.

Teng-Kai, F. and H.C. Chia, 2010. Exploring evolutionary technical trends from academic research papers. J. Inf. Sci. Eng., 26: 97-117.

Varadarajan, R. and V. Hristidis, 2006. A system for query-specific document summarization. Proceedings of the 15th ACM International Conference on Information and Knowledge Management, November 06-11, 2006, ACM, Arlington, Virginia, USA., ISBN:1-59593-433-2, pp: 622-631.

Walters, W.H., 2011. Comparative recall and precision of simple and expert searches in Google Scholar and eight other databases. Portal Libraries Academy, 11: 971-1006.