INTERNATIONAL JOURNAL OF COMPUTER ENGINEERING & TECHNOLOGY (IJCET)

ISSN 0976 – 6367(Print)
ISSN 0976 – 6375(Online)
Volume 4, Issue 4, July-August (2013), pp. 462-466
© IAEME: www.iaeme.com/ijcet.asp
Journal Impact Factor (2013): 6.1302 (Calculated by GISI)
www.jifactor.com

IJCET

©IAEME

OPTIMAL APPROACH FOR TEXT SUMMARIZATION

Madhuri K. Gawali[1], Prof. M. S. Bewoor[2], Dr. S. H. Patil[3]

[1](M. Tech. Student, Bharati Vidyapeeth Deemed University College of Engineering, Pune, India)
[2](Associate Prof. Bharati Vidyapeeth Deemed University College of Engineering, Pune, India)
[3](Head, Department of Computer Engineering, Bharati Vidyapeeth Deemed University College of Engineering, Pune, India)

ABSTRACT

Large amount of unstructured information is available on the internet. Retrieving relevant documents containing the required information is difficult, because of huge amount of data. The query-specific document summarization has become an important problem. It is difficult task for the user to go through all these documents, as the number of documents available on particular topic will be more [1, 4]. It will be helpful for the user, if query specific document summery is generated. Various clustering algorithms will be evaluated which provide better results for summarization. Single query is act as input with various clustering algorithms and it will generate summary of document for each algorithm [1, 2, 3].

Evaluation of algorithms will be performing on the basis of parameters like precision, recall, time, space complexity, and quality of summary. After evaluating these algorithms suggest better algorithm for summarization. So it will help to find the better query dependent clustering algorithm for text document summarization.

Keywords: Clustering, Precision, Recall Summarization.

1. INTRODUCTION

Current document clustering methods basically represent documents in terms of document matrix and perform clustering algorithm on it. These clustering methods can group the documents satisfactorily. But it is difficult for people to capture the meanings of the documents, since there is no satisfactory analysis for each document [2, 3, 4].

Single query is taken as input to system and then different clustering algorithms like Hierarchical clustering algorithm, Query based summarization, Graph theoretic clustering algorithm, Fuzzy C-means clustering and DB Scan clustering applies on it. Different results will be generated for each algorithm. These results will be evaluating with each other in terms of precision, recall, time, space complexity, and quality of summary.

Depending on performance parameters algorithm is better for summarization will be suggested. So it will help to find the better query dependent clustering algorithm for text document summarization. Fig. 1 shows, document clustering system consists of five different clustering algorithms to get optimal solution or best document clustering algorithm.

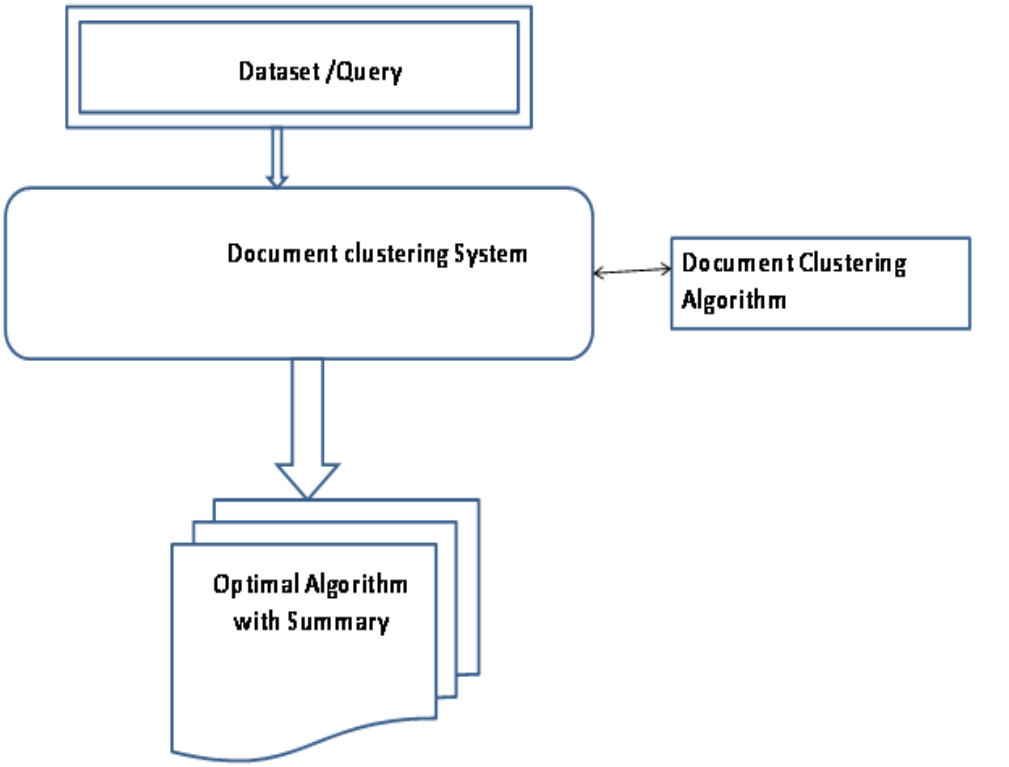

Fig. 1: Overall System Architecture

2. RELATED WORK

The paper [1] focused on creation of query specific summaries by identifying query relevant fragment and combining them using the semantic associations within the document. The best summaries are computed by calculating the top spanning trees on the document graphs. The paper [2] focused on document summarization graph method. For creating document graph each paragraph is assumed as one individual node. Node score and edge score are calculated. The paper [3] focused on the concept of Open NLP tool for natural language processing of text for word matching. And in order to extract meaningful and query dependent information from large set of offline documents, data mining document clustering algorithm are adopted.

The paper[4] describes a system the phases of natural language processing that is splitting, tokenization, part of speech tagging, chunking and parsing. Implement Expectation Maximization Clustering Algorithm to find out sentence similarity. By using the value of sentences similarity, easily summarize text. The paper [5] describes detail explanation of implementation for ROCK (Robust Clustering using links) clustering algorithm. Novel concept of links measures the similarity between a pair of data points**.**

3. PERFORMANCE PARAMETER

Two of the most common measures of system performance are;

1) Time 2) Space

The shorter the response time and smaller the space used then system considered as better. Major Performance parameters are precision, recall, F- Measure, Compression Ratio, and Retention ratio [1, 6, 12].

3.1 Precision

Precision is the fraction of retrieved document instances that are relevant document

3.2 Recall

Recall is the fraction of relevant document instances that are retrieved document.

3.3 F- Measure

F- Measure combines precision and recall. It is also known as harmonic mean. The function F assumes values in the interval [0, 1]. It is 0 when no relevant documents have been retrieved and is 1 when all ranked documents are relevant. Also the value of harmonic mean is high when both recall and precision are high [1, 12].

3.4 Compression Ratio

It is the fraction of number of terms in summary to number of total terms in data.

3.5 Retention Ratio

It is the fraction of Number relevant query words in summary to number query terms in data [1, 11, 12].

3.6 Time complexity

The time complexity of an algorithm is commonly expressed using big (O) notation, which excludes coefficients and lower order terms. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, where an elementary operation takes a fixed amount of time to perform. Thus the amount of time taken and the number of elementary operations performed by the algorithm differ by a constant factor. The better the time complexity of an algorithm is the faster the algorithm will carry out its work in practice.

3.7 Space complexity

Space complexity is one of the important parameter for performance of system. The number of memory cells which an algorithm needs.

4. SYSTEM IMPLEMENTATION

The system consists of three modules described as follows;

4.1 Document Matrix Generation using NLP

The system accepts the text to be summarized as .txt file. The input text is processed using open NLP tool for which a text has to go through the various phases of natural language processing like sentence detection, tokenization, parse tree generation, parsing, chunking, pos tagging [2, 3, 4, 5].

4.2 Document Graphs Generation using Different Clustering Algorithms

The document graph is generated using Query specific document summarization, Graph based algorithm, Expectation Maximization, DBSCAN Clustering, Fuzzy C-means clustering and hierarchical clustering techniques. From these documents graphs summaries generated are evaluated and compared by third module of the system [2, 3, 4, 5].

4.3 Evaluating Quality Generated Summary through Performance Parameter

The summaries generated from second module is compared with each other using qualitative and quantitative performance parameters such as Precision, recall, F-measure, Compression ratio, retention ratio and CPU processing time.

The formulas for Performance Parameter are as below;

Precision is the probability that a retrieved document is relevant.
Precision = Number different terms in summary / Number of different terms in Query

Recall is the probability that a relevant document is retrieved in a search.
Recall = Number of correct matching sentences in summary / Numbers of relevant sentences in all data.

F-measure is the harmonic mean of Precision and recall; both have been given equal importance.
F-measure = 2*(Precision*Recall) / (Precision + Recall)

Compression Ratio = Number terms in summary / Number total terms in data
Retention ratio = Number relevant query words in summary / Number query terms in data.

5. CONCLUSION

In this article compare various performance parameters with different clustering algorithm is done and optimal algorithm among different clustering algorithms will be evaluated. Further this system can be improved to work on Doc file as well as PDF file which contain huge textual data.

REFERENCES

[1] Ramakrishna Varadarajan, Vangelis Hristidis, “A system for Query Specific Document Summarization”.
[2] Prashant D. Joshi, S. G. Joshi, M. S. Bewoor & Dr. S. H. Patil, “Comparison between graphs based document Summarization method and clustering method”, International Journal of Advances in Engineering & Technology (IJAET), 2011, Vol. 1, Issue 5, pp. 118-125.
[3] Harshal J. Jain, M. S. Bewoor, Dr. S. H. Patil, “Context Sensitive Text Summari**zat**ion Using K Means Clustering Algorithm”, International Journal of Soft Computing and Engineering (IJSCE), 2012, Vol. 2, Issue 2, pp. 301-304.
[4] Ms. Meghana N. Ingole, Prof. M. S. Bewoor, Dr. S. H. Patil, “Text Summarization using Expectation Maximization Clustering Algorithm”, International Journal of Engineering Research and Application (IJERA), 2012, Vol. 2, Issue 4, pp. 168-171.
[5] Ms. Laxmi S. Patil, Prof. M. S. Bewoor, and Dr. S. H. Patil, “Query Specific ROCK Clustering Algorithm for Text Summarization”, International Journal of Engineering Research and Application (IJERA), 2012, Vol. 2, Issue 3, pp. 2617-2620.

[6] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries", in Proceedings of the ACL-04 Workshop: Text Summarization Branches Out, Barcelona, Spain 2004, pp.74–81.
[7] "Count Data Modeling and Classification Using Finite Mixtures of Distributions", IEEE Transaction on Neural Networks.Vol.22, No.2, February 2011.
[8] "Clustering Sentence-Level Text using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions on Knowledge and Data Engineering 2011.
[9] Software Engineering: A Practitioner's Approach (Sixth Edition) - by Roger S. Pressman.
[10] The complete Reference of .NET, by Matthew, Tata McGraw Hill Publication Edition 2003.
[11] The complete Reference of modern information retrieval Ricardo baeza-yates.
[12] Sunita R Patil and Sunita M.Mahajan, "Document Summarization Using Extractive Approach", International Journal of computer applications, 2011.
[13] Roma V J, M S Bewoor and Dr.S.H.Patil, "Automation Tool for Evaluation of the Quality of NLP Based Text Summary Generated Through Summarization and Clustering Techniques by Quantitative and Qualitative Metrics", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 3, 2013, pp. 77 - 85, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
[14] V.Sujatha, K.Sriraman, K. Ganapathi Babu and B.V.R.R.Nagrajuna, "Testing and Test Case Generation by using Fuzzy Logic and N.L.P Techniques", International Journal of Computer Engineering & Technology (IJCET), Volume 4, Issue 3, 2013, pp. 531 - 538, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.
[15] Meghana. N.Ingole, M.S.Bewoor and S.H.Patil,, "Context Sensitive Text Summarization Using Hierarchical Clustering Algorithm", International Journal of Computer Engineering & Technology (IJCET), Volume 3, Issue 1, 2012, pp. 322 - 329, ISSN Print: 0976 – 6367, ISSN Online: 0976 – 6375.