International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-2 Issue-4, March 2013

Review: Evaluating and Analyzer to Developing Optimized Text Summary Algorithm

Madhuri Gawali, Mrunal Bewoor, Suhas Patil

Abstract- Information available on internet is in unstructured manner, retrieving relevant documents containing the required information is difficult. Due to huge amount of data, 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. It will be helpful for the user if query specific document summery is generated. Comparing different clustering algorithms those provide better result for summarization. Based on this we provide input as one query and get all the documents related to that and on these document different clustering algorithm are used to get results of each Algorithm. Then these algorithms comparing results with each other in terms of speed, memory, and quality of summary. After comparison we can decide which algorithm is better for summarization. So it will help to find the better query dependent clustering algorithm for text document summarization.

Keywords: clustering, summarization.

I. INTRODUCTION

Large amount of information is available on web. There has been a great amount of work on query-independent summarization of document. Due to the success of Web search engines query-specific document summarization has become an important problem. [1,2]

Current document clustering methods usually represent documents as a term document matrix and perform clustering algorithm on it. Although these clustering methods can group the documents satisfactorily, it is still hard for people to capture the meanings of the documents since there is no satisfactory interpretation for each document cluster.

Use five different clustering algorithms which provide better result for summarization. Based on this we provide input as one query and get all the documents related to that and on these document we apply our five clustering algorithm (Hierarchical, Query based clustering algorithm, graph theoretic, Fuzzy C-mean and DB Scan) and get result of each Algorithm .Here we use Weka tool for clustering and getting the summary from document. Weka is a standard tool for clustering. It contains all clustering algorithm. Then we compare that result with each other in terms speed, memory, and quality of summary.

Manuscript published on 30 March 2013.
*Correspondence Author(s)
Ms. Madhuri K. Gawali, Pursuing Master in Technology (M. Tech.) in Computer Engineering from Bharati Vidyapeeth Deemed University College of Engineering, Pune (Maharastra), India.
Mr. Mrunal Bewoor, Asst. Professor, Department Computer Engineering from Bharati Vidyapeeth Deemed University College of Engineering, Pune (Maharastra), India.
Dr. Suhas Patil, Professor, Department Computer Engineering from Bharati Vidyapeeth Deemed University College of Engineering, Pune (Maharastra), India.

After comparison we can decide which algorithm is better for summarization. So it will help to find the better query dependent clustering algorithm for text document summarization. [3]

For quality of summary uses natural language parser, like Stanford NLP, Dependency Parser, and Word Net. A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb. Use Weka tool for values of precision and recall. Space complexity means amount of memory it requires to run to completion. Time complexity means amount of time it needs to run to completion. Dataset or any query act as input of integration of document clustering system. Integration of document clustering system consists of different document clustering algorithms. Finally we get output as optimal algorithm for document summarization with summary. Each algorithm generate summary, so compare summary of each algorithm in terms of quality of summary, precision and recall, space and time complexity etc. [1,4,5] Fig. 1 shows that use different datasets or queries to integration document clustering system. Integration of document clustering system consists of different five clustering algorithms. Finally we get optimal solution or best document clustering algorithm.

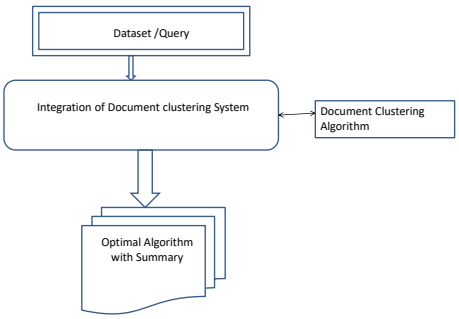

Figure 1: Overall System Architecture

II. SYSTEM IMPLEMENTATION

Implementation is very important phase; the most critical stage in achieving a successful new system so that the new system will work is effectively. The total workflow is divided into following modules:

A. Input of system in form of Text File and to Create the Document Graph

The system accepts input text file. The file is read and stored into a string. The array contains paragraphs which are further treated as nodes which creating document graph. [5]

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

B. In Document Graph Add Weighted Edges

A weighted edge is added to the document graph between two nodes if they either correspond to adjacent node or if they are semantically related, and the weight of an edge denotes the degree of the relationship. Here two nodes are considered to be related if they share common words (not stop words) and the degree of relationship is calculated by "Semantic parsing". Also notice that the edge weights are query-independent, so they can be pre-computed. [2,3,5]

C. Threshold for Edge Weights

Threshold will be created in the document graph.

D. Document Clustering

Clustering is grouping of similar nodes into a group. [3] The following approaches of clustering algorithms are used:

1) Query based clustering algorithm
2) Fuzzy C-mean
3) DB Scan
4) Graph theoretic algorithm
5) Hierarchical algorithm

E. Comparison with different Parameters and Find Optimal Solution

Compare different parameters such as quality of summary, precision and recall, space and time complexity. Quality of summary: Use of natural language parser like Stanford NLP, dependency parser, word net. Precision and recall use Weka tool for values of Precision and recall. Space complexity: Amount of memory it requires to run to completion. Time complexity: Amount of time it needs to run to completion. [6,7] Stanford NLP: A natural language parser is a program that works out the grammatical structure of sentences, for instance, which groups of words go together (as "phrases") and which words are the subject or object of a verb.

Dependency Parser: Malt Parser is a system for data-driven dependency parsing, which can be used to induce a parsing model from tree bank data and to parse new data using an induced model. Word net: Word Net is a lexical database for the English language. It groups English words into sets of synonyms provides short, general definitions, and records the various semantic relations between these synonym sets. Cosine Similarity: Cosine similarity is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications and clustering. [8, 9, 10]

F. Results of the similarity measure based on the cosine and link functions

The first step of our similarity measure based on the cosine and link functions is to find the neighbors of each document.

Neighbors and link

The neighbors of a document d in a data set are those documents that are considered similar to it. Let sim d_i , d_j be a similarity function capturing the pair wise similarity between two documents, d_i and d_j and have values between 0 and 1 with a larger value indicating higher similarity. For a given threshold θ, d_i and d_j are defined as neighbors of each other if,

$$sim(d_i, d_j) \geqslant \theta, \quad \text{with } 0 \leqslant \theta \leqslant 1.$$

For two documents d_i and d_j the similarity between them can be calculated. [12,13]

Cosine similarity measure

Given two document t_a and t_b

$$SIM_C(\overrightarrow{t_a}, \overrightarrow{t_b}) = \frac{\overrightarrow{t_a} \cdot \overrightarrow{t_b}}{|\overrightarrow{t_a}| \times |\overrightarrow{t_b}|},$$

Each dimension represents a term with its weight in the document, which is non-negative. For document clustering, there are different similarity measures available. The most commonly used is the cosine function. [12,13]

III. CONCLUSIONS

In this paper we have compare the performance of different clustering algorithms and find optimal algorithm among different clustering algorithms. Further this system can be improved to work on Doc file as well as PDF file which contain huge of textual data.

REFERENCES

1. 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.
2. Anna Huang "Similarity Measures for Text Document Clustering", NZCSRSC, 2008 Christchurch, New Zealand, pp. 49-56.
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. 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.
5. Ms. Meghana N. Ingole, Mrs. M. S. Bewoor, Mr. 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.
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. B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering", proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1999.
12. Congnan Luo, Yanjun Li, Soon M. Chung "Text document clustering based on neighbors", Data & Knowledge Engineering 68, Elsevier publication, 2009, pp. 1271-1288.
13. Abdolreza Eshghi, Dominique Haughton," Identifying Groups: A Comparison of Methodologies", journal of Data Science 9, 2011, pp. 271-291.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

AUTHOR PROFILE

Ms. Madhuri K. Gawali received her Bachelor in Engineering (B.E.) degree in Computer Science & Engineering from SAVERI's College of Engineering, Pandharpur, Shivaji University Maharashtra, India, in 2006. Now she is pursuing Master in Technology (M. Tech.) in Computer Engineering from Bharati Vidyapeeth Deemed University College of Engineering, Pune, India. Her research interests include document clustering.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)