Empirical Analysis of Single and Multi Document Summarization using Clustering Algorithms

Mrunal S. Bewoor
Department of Computer Engineering
Bharati Vidyapeeth (Deemed to be) University College of Engineering
Pune, India
msbewoor@bvucoep.edu.in

Suhas H. Patil
Department of Computer Engineering
Bharati Vidyapeeth (Deemed to be) University College of Engineering
Pune, India
shpatil@bvucoep.edu.in

***Abstract*—The availability of various digital sources has created a demand for text mining mechanisms. Effective summary generation mechanisms are needed in order to utilize relevant information from often overwhelming digital data sources. In this view, this paper conducts a survey of various single as well as multi-document text summarization techniques. It also provides analysis of treating a query sentence as a common one, segmented from documents for text summarization. Experimental results show the degree of effectiveness in text summarization over different clustering algorithms.**

Keywords-text mining; text summarization; clustering

I. INTRODUCTION

The extensive use of Internet has caused a vast growth in the usage of digital information. People use online information services, like social media, every day resulting to the availability of a huge amount of unstructured digital information. This information is directly accessible to a large number of end-users [1-2]. The user accesses this information through queries, but the improvement of precision and speed is always an issue. The information retrieval (IR) systems have resolved this to some extent. This information overload problem is more sensitive when there is a need of taking a decision or of deep understanding of a problem The IR systems solve this through user issued queries The obtained result most of the times overwhelms users with too many answers, and provided documents that may not be relevant to the topic asked. The multi document summarization has an ability to summarize a complete document set. Ideally it is a process of query shared information extraction through a set of multiple text documents. The techniques used in single-document summarization can also be used in multi-document summarization [3]. The comparison of single and multi-document summarization is presented in Table I.

Web information retrieval relevant to the issued query is a tedious task. Information retrieval tools can be used for retrieval relevant to the topic specified by the query. The results obtained sometimes may not preserve the required content. Summary generation or automatic text summarization is the creation of abstracts or summaries, with the help of a computer program, from one or more documents. There are specifically two types of text summarization techniques, generic and query specific [4]. It becomes a difficult task for the user to go through a large number of retrieved documents [5]. This difficulty can be resolved with the use of query specific document summary generation. The generated summary or abstract must preserve the semantics and central idea of an input text [6]. Below we present the main existing approaches to multi document summarization:

A. Feature Based Method

The extractive type summarization approach identifies the most related sentences from the original text and place them together to generate a concise summary. The process identifies relevant sentences based on features like sentence length, word frequency title word, sentence position, cue word, proper noun etc.

B. Cluster Based Method

The initiative of clustering is to group similar objects into classes. In case of multidocument summarization, these objects refer to the sentences and the classes represent the cluster each sentence belongs to. Considering the type of documents that concentrate on different subjects or topics, some of the researchers try to integrate the clustering concept based on the sentence similarity. The most common similarity measure is cosine similarity. The sentence selection is performed by selecting sentences from each cluster on the basis of ranking tf-idf in that cluster.

C. Graph Based Method

This method uses the basic concept behind the graph to represent the relationship elements. Related elements in the graph are linked. In case of text, the element relationship is the similarity between the sentences. It is represented as an ordered pair graph G=(V, E), where V is set of elements representing sentences and E is set of edges representing the association among the sentences. The strongly connected sentences are considered in the summary. Many graph based approaches use cosine similarity to identify the association.

D. Knowledge based Method

The documents are organized with the text content related to a specific topic belonging to a particular domain. Every domain has a common knowledge structure. The researchers have common background knowledge structure (i.e. ontology) to improve the summary results. There have been efforts to utilize the background knowledge. Many applications have tailored their model to be ontology-driven [4]. Ontology can be useful for domain specific documents where key concepts corresponding to the domain can be identified. The technique is implemented as query specific related to the respective domain by identifying keywords.

E. Our Approach

This paper presents a combined approach by using topic queries or important keywords corresponding to the document set and the fundamental concept of clustering as well as language features to extract the relevant sentences from the original document set. The features of clustering algorithms and NLP based retrieval can be useful in preserving the context of the information in the retrieval process [7].

TABLE I. COMPARISON OF SINGLE AND MULTIPLE DOCUMENT SUMMARIZATION PROCESS

Comparison features	Single Document Summarization	Multi-document Summarization
Degree of coherence	Change in sequence of sentence selection do not affect the degree of coherence.	The order of sentence selection may affect the degree of coherence.
Redundancy	Topics in a single document are related. The degree of redundancy is high.	Some information that may be seen as redundant might be important and vice versa.
Compression Ratio	Usually much smaller	Usually higher.
Cross-reference	Cross reference resolution can be easily resolved	Cross reference resolution is a greater challenge

II. SYSTEM ARCHITEXTURE

The process of automatic text summarization consists of mainly two tasks. The first one is to recognize the most significant text portions and the second is to obtain the coherent summaries. Information retrieval (IR) process is used to search documents on web. Since massive amount of vague data is available on the web [8], the use of IR tools has given rise to the necessity of query dependent document summarization. The IR system has demoralized the natural language Processing techniques to support a range of natural language queries. The type of query processing for text summarization without NLP support may result in imprecise summary and user may not view correct or reliable results [9].

The system considers original document or document set in txt form. Documents are preprocessed using basic steps of natural language processing (NLP), like sentence detection, tokenization, part-of-speech tagging, chunking and parsing. The NLP is implemented using the Open NLP tool. The NLP steps help to identify the correct word match with respect to the context within the document by removing the ambiguity if any [10]. The result obtained after pre-processing is further given to clustering algorithms along with the keywords used for summarization process. The clustering algorithms such as EM, Graph Based Method, Fuzzy C-Means, DBSCAN, and Hierarchical clustering algorithms are used to obtain the summary. The summary is also computed with a simple query specific approach [11]. The result obtained as a summary can be evaluated on the basis of qualitative and quantitative metrics. The precision, recall and F-measure are quality measuring metrics and compression retention ratio are quantity measuring metrics [12].

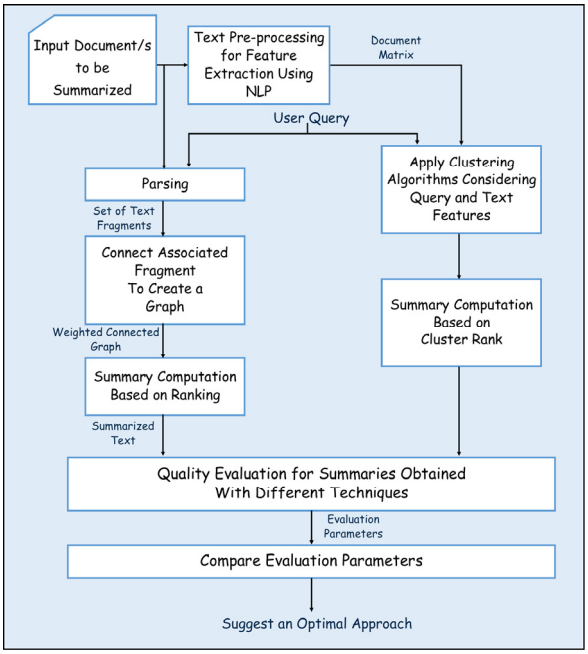

Fig. 1. System Architecture.

III. IMPACT OF CLUSTERING IN AUTO TEXT SUMMARIZATION

The concept of clustering is very helpful in the text domain as document objects as words, sentences, paragraphs to be clustered are of varying granularities. Clustering is particularly useful to put together documents to get better retrieval and support browsing. In [13], authors recognized and selected clustering algorithms for obtaining document summary. The main motive of the research was to extract in the summary those sentences that are more relevant to the original input text by using clustering algorithm features which can group the objects based on the relevancy [14]. The approach tries to combine two major approaches of summary generation: extractive and is abstractive. Results obtained with different methods are then evaluated for the summary quality.

IV. EXPERIMENTATION AND RESULT DISCUSSION

Results are tested for the inputs from existing datasets like Reuter. Reuter dataset is a popular dataset for text mining experiments. Different splits into training test and unused data have been considered. In the case of abstractive type of

summarization, the quality of the summary obtained is more important. This quality is judged on the basis of qualitative evaluation parameters. The length or the size of the summarized text is evaluated on the basis of quantitative measures compression ratio and retention ratio. Results obtained are compared on the basis of these evaluation measures. The summaries generated are also compared with existing query summarizers, Copernic summarizer and web Summarizer. These two tools are query based summarizers[15]. The values for these parameters for the corresponding document are calculated as shown below.

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

Precision=No. of different terms in summary/No. of different terms in Query.

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

Recall=No. of correct matching sentences in the summary/No. of all relevant sentences in the original document

F-Measure is the harmonic mean of precision and recall:

F-Measure= 2*((Precision*Recall)/(Precision+ Recall)

Compression Ratio= No. of sentences in summary/Total No. of sentences in original document.

Retention Ratio=No. of relevant query words in summary/No. of Query terms in original data.

Precision, Recall and F-measure measure the quality of the summary, and so they, along with execution time, are called qualitative parameters [16]. Compression ratio and retention ratio measure the length or quantity of the sentences in the summary and therefore are called quantitative parameters.

Results are shown in Tables II-IV and Figures 2-7 for various methodologies in data obtained from WikiArt (data set A, single document summarization), Reuter (data set B, multi document summarization) and Wiki Internet (data set C, multi document summarization) data sets.

TABLE II. PERFORMANCE EVALUATION FOR DATA SET A

Methodology	Precision	Recall	F-measure	Compression ratio	Retention ratio	Execution Time
Expectation Maximization Clustering	1	0.1	0.18182	0.04776	0.70394	5,516
Fuzzy C-Means	1	0.07627	0.14173	0.03582	0.86227	78
DB-SCAN Clustering	1	0.03383	0.06545	0.08955	0.93577	123
Graph Theorotic Clustering	1	0.0604	0.11392	0.06567	0.94715	76892
Hierarchical Agglomerative Clustering	1	0.03435	0.06642	0.08955	0.93577	496
Query Specific	1	0.03383	0.06545	0.08955	0.93577	41
Copernic	1	0.5263	0.1	0.0806	0.63737	1,451
Web Summarizer	1	0.5263	0.1	0.0806	0.5708	786

TABLE III. PERFORMANCE EVALUATION FOR DATA SET B

Methodology	Precision	Recall	F-measure	Compression ratio	Retention ratio	Execution Time
Expectation Maximization Clustering	0.77778	0.10588	0.18639	0.02738	0.91125	20,206
Fuzzy C-Means	1	0.01415	0.02791	0.0306	0.95639	995
DB-SCAN Clustering	1	0.00904	0.01791	0.0306	0.98448	46
Graph Theorotic Clustering	0.55556	0.02206	0.04243	0.01852	0.98742	6,788
Hierarchical Agglomerative Clustering	1	0.00907	0.01798	0.0306	0.98448	8,809
Query Specific	1	0.00904	0.01791	0.8306	0.98448	20
Copernic	0.22222	0.01305	0.02466	0.00725	0.8198	3,593
Web Summarizer	0.22222	0.01004	0.01921	0.00725	0.74515	1,601

TABLE IV. PERFORMANCE EVALUATION FOR DATA SET C

Methodology	Precision	Recall	F-measure	Compression ratio	Retention ratio	Execution Time
Expectation Maximization Clustering	1	0.15254	0.26471	0.07595	0.73022	3,419
Fuzzy C-Means	1	0.07627	0.14173	0.10127	0.85801	60
DB-SCAN Clustering	1	0.03125	0.06061	0.10127	0.94967	20
Graph Theorotic Clustering	1	0.07563	0.14062	0.10127	0.95827	705
Hierarchical Agglomerative Clustering	1	0.03169	0.06143	0.10127	0.94967	662
Query Specific	1	0.03125	0.06061	0.10127	0.94967	43
Copernic	1	0.02083	0.04082	0.06329	0.6572	380
Web Summarizer	1	0.02778	0.05405	0.08228	0.69023	842

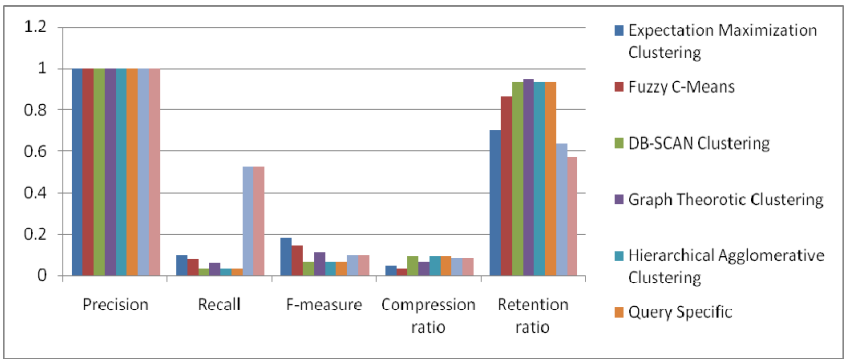

Fig. 2. Comparison graph of all summaries for data set A

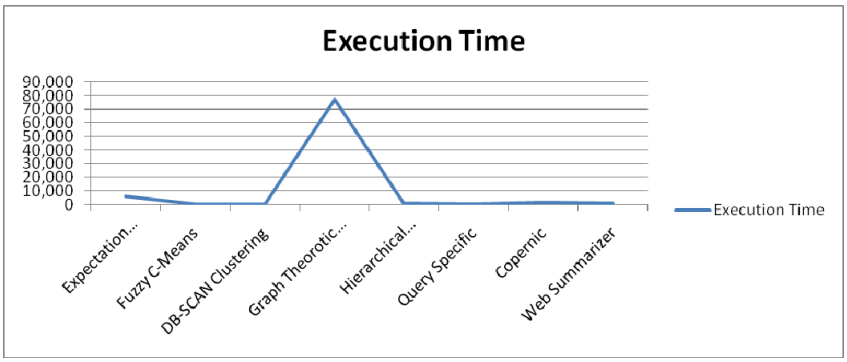

Fig. 3. Comparison graph of execution time for various summarization methods in data set A

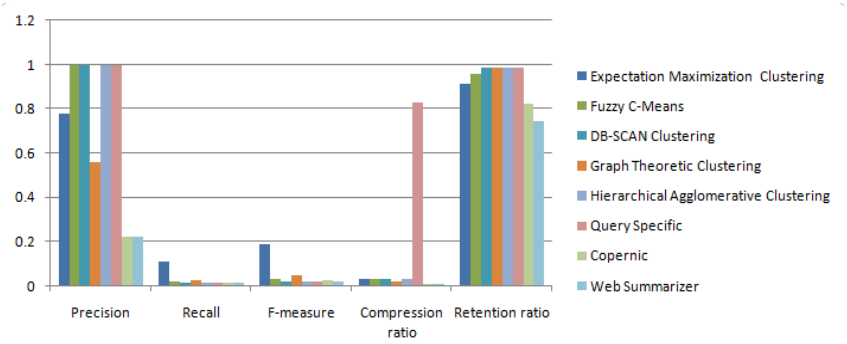

Fig. 4. Comparison graph of all summaries for data set B

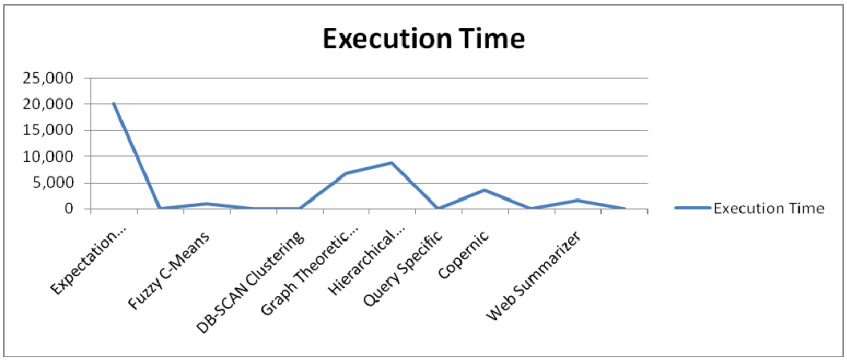

Fig. 5. Comparison graph of execution time for various summarization methods in data set B

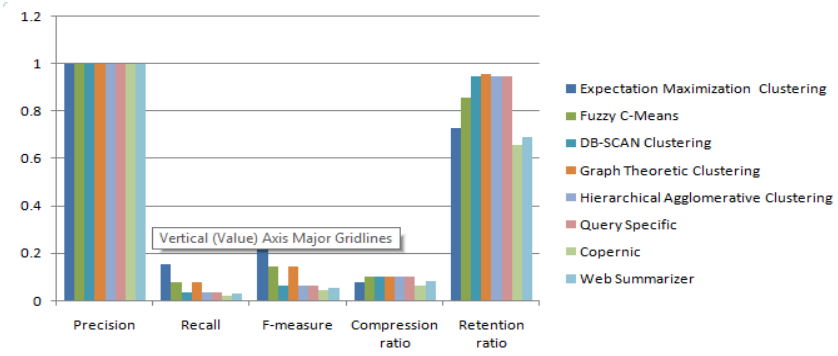

Fig. 6. Comparison graph of all summaries for data set C

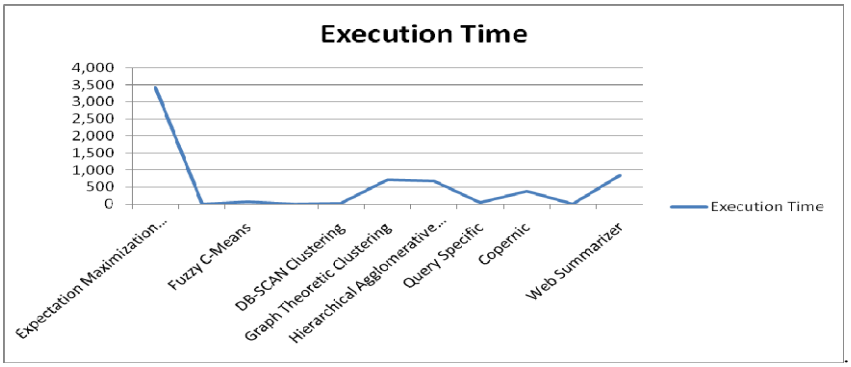

Fig. 7. Comparison graph of execution time for various summarization methods in data set C

V. CONCLUSION

Results generated by all summarization methods are evaluated considering result quality with respect to the context within the input text and the length of the input text. Document summarization should focus on the context, length being the secondary aspect. Fuzzy c means generates better summary. Considering both quantity and quality parameters, clustering is an unsupervised text summarization technique, which can be used as supervised by integrating it with a supervised approach. This may give an optimal solution for this problem. The research work should focus on improving the quality of clusters which directly relates with the gist of the original input document.

REFERENCES

[1] A. Kaushik, S. Naithani, "A Comprehensive Study of Text Mining Approach", International Journal of Computer Science and Network Security, Vol. 16, No. 2, pp. 69–76, 2016

[2] R. Varadarajan, V. Hristidis, "A system for query-specific document summarization", 15th ACM international conference on Information and knowledge management, pp. 622-631, 2006

[3] J. Goldstein, V. Mittal, J. Carbonell, M. Kantrowitz, "Multi-Document Summarization By Sentence Extraction", NAACL-ANLPWorkshop on Automatic summarization, Vol. 4, pp. 40–48, 2000

[4] Y. J. Kumar, N. Salim, "Automatic multi document summarization approaches", Journal of Computer Science, Vol. 8, No. 1, pp. 133–140, 2012

[5] S. Gholamrezazadeh, M. A. Salehi, B. Gholamzadeh, "A comprehensive survey on text summarization systems", 2nd International Conference on Computer Science and its Applications, pp. 1-6, 2009

[6] M. Steinbach, G. Karypis, V. Kumar, "A Comparison of Document Clustering Techniques", KDD workshop on text mining, Vol. 400, No. 1, pp. 525-526, 2000

[7] D. Vidyadharan, A. CR "A Query Based Summerizer Based on the Context " International Journal of Science and Research, Vol. 4, No. 5, pp. 3018-3020, 2015

[8] T. K. Fan, C. H. Chang, "Exploring Evolutionary Technical Trends from Academic Research Papers", Eighth IAPR International Workshop on Document Analysis Systems, pp. 574-581, 2008

[9] D. Y. Sakhare, R. Kumar, "Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization", Internation Information Technology and Computer Science, Vol. 2014, No. 3, pp. 38–46, 2014

[10] M. N. Ingole, M. S. Bewoor, S. H. Patil, "Text Summarization using Expectation Maximization Clustering Algorithm", International Journal of Engineering Research and Applications, Vol. 2, No. 4, pp. 168–171, 2012

[11] V. J. Roma, M. S. Bewoor, S. H. Patil, "Automation Tool for Evaluation of NLP based Text Summary Generated through Summarization and Clustering Techniques by Quantitative and Qualitative Metrics", International Journal of Computer Engineering and Technology, Vol. 4, No. 3, pp. 77–85, 2013

[12] M. K. Gawali, M. S. Bewoor, S. H. Patil, "Review : Performance Evaluator of Optimized Text Summary Algorithm", International Journal of Computer Science and Technology Vol. 4, No. 1, pp. 295–296, 2013

[13] V. J. Roma, M. S. Bewoor, S. H. Patil, "Evaluator and Comparator : Document Summary Generation based on Quantitative and Qualitative Metrics for International Journal of Scientific & Engineering Research",

International Journal of Scientific & Engineering Research, Vol. 4, No. 5, pp. 1111–1115, 2013

[14] A. Nenkova, "Automatic Text Summarization of Newswire: Lessons Learned from the Document Understanding Conference", Association for the Advancement of Artificial Intelligence, Vol. 5, pp. 1436-1441, 2005

[15] M. J. A. Eugster, Benchmark Experiments—A Tool for Analyzing Statistical Learning Algorithms, PhD Thesis, Ludwig-Maximilians-Universitat, 2011.

[16] M. Hassel, Evaluation of Automatic Text Summarization, Licentiate Thesis, 2004