Evaluator and Comparator : Document Summary Generation based on Quantitative and Qualitative Metrics for International Journal of Scientific & Engineering Research

Roma V J, M S Bewoor, S H Patil

Abstract— The increased use of World Wide Web has overloaded the internet with huge amount of information. The information available on the internet is unstructured. To retrieve relevant and significant information manually from the World Wide Web becomes difficult process for the end user. This problem can be resolved by using query specific document summarization. Document summarization is the process of condensing the input document into shorter or abstract version by preserving its original content and semantics of the given document .The various clustering techniques can be used for document summarization. The proposed system will get input as one query and retrieve all the documents relevant to query and on these documents different clustering techniques will be used for summary generation. The quality of summary generated from various clustering techniques will be compared by using various metrics such as Compression Ratio, Retention Ratio, Recall and Precision.

INDEX TERMS—Clustering, Compression ratio, Hierarchical ,Precision ,Recall,retention ratio .

1 INTRODUCTION

The automatic text summarization is an approach through which computer program generates the summary or abstract of an input text. The generated summary or retrieved abstract should preserve the semantics and central idea of an input text. Typically there are main two approaches for automatic text summarization.

a) Generic – In which important sentences from given document are extracted and the extracted sentences are arranged in an appropriate order.

b) Query Based text summarization - which will be used to generate query dependent summary, the sentences are scored based on the query given by user. It is one of the statistical methods to make a summary by extracting relevant sentences from a document [1].

The criterion for extraction is given as a query .The sentences in the summary result are considered as most important information related to the given document. Most of the existing methodologies are query independent. The exponential growth of information on the internet has resulted in the extensive use of web search engines. Thus web is supported by Information Retrieval tools, which has given rise the need of query specific document retrieval.

- *Roma V J is currently pursuing masters degree program in computer engineering in Bharati University College ofEngineering,India,PH-9881051462.E-mail:rvpanjwani@bvucoep.edu.in*
- *M S Bewoor is currently pursuing doctorate degree program in computer engineering in Bharati University College Of Engineering,India.*
- *S.H.Patil has completed doctorate in computer Engineering and is currently working as Professor in the department of Computer Engineering in Bharati University College Of Engineering.*

The number of documents retrieved by information retrieval system need to be summarized. The Natural language processing is an program which deals with the interpretation of text .In Natural language processing the input text has to work out through different phases where sentences can be split, which goes through parsing, tokenization and finally chunking [7].The input text under processing is considered as a single node and each node will be compared with other node. The Word Net can be used for calculating weight. Then it is represented in the form of document graph. The various clustering techniques will be applied before summary generation such as DBSCAN, Hierarchical, and Fuzzy C-means. The summary generated from the clustering techniques will be compared to yield better query specific document summarization algorithm. Fig (a) shows the proposed system architecture.

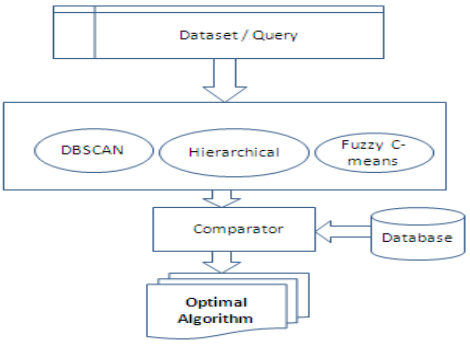
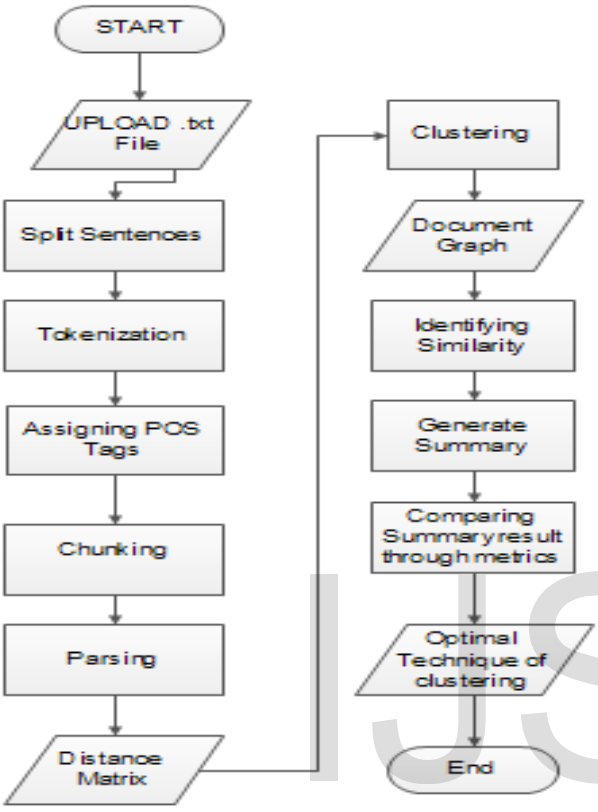


Fig (a) System Architecture

2. SYSTEM IMPLEMENTATION

The System implementation plays an important role but considered as more crucial stage where actual working system will be developed. The workflow is divided into following modules. Fig (b) shows flow throughout the system.



Fig(b)Flowchart for the Proposed System

A. Input text file to the system for document matrix generation –

The system accepts the input text file. The file is read and stored into stream of strings. Each and every sentence is represented through a node. Identifying the end of sentence is an important task. The symbols such as '.','!''?' can be used as separators. After splitting the sentences the input text is divided into separate tokens. Punctuation marks, spaces and terminators can be used as breaking characters. After tokenization POS (Part of Speech Tagger) is applied for grammatical semantics. The output of POS tagger is a single best POS tag for each word such as noun, verb, coordinating conjunction, verb and past tense. After assigning POS tags chunker is used to form the groups of words like noun group, verb group. Parsing is used to convert the input sentence into hierarchical structure which corresponds to the items of meaning in the sentence [8]. By carrying out lexical semantics of the words the distance matrix is prepared from the parse tree generated. This distance matrix is given as input to the next module for clustering.

B. Document Graph Generation -

The generated distance matrix is given as input to the clustering. Clustering is one of the data mining technique, Clustering is also called as unsupervised learning which is used to group the elements or objects possessing similar attributes. The proposed system aims at the generation of document graph using clustering techniques –

i) DBSCAN,

ii) Hierarchical Clustering,

iii) Fuzzy C-means,

The document graph generated through clustering is send as input to the third module for similarity identification of sentences.

i) DBSCAN :Density Based Clustering Method Based on Connected Regions

These methods have been developed based on the notion of Density .The general idea behind it is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold that is for each data point within a given cluster; the neighbourhood of a given radius has to contain at least a minimum number of points. The algorithm grows the regions with sufficiently high density into clusters of arbitrary shape in spatial databases with noise. The basic idea of density based clustering algorithm involves –

The neighbourhood within a radius ε of a given object is called the ε-neighbourhood of the object. If the neighbourhood of an object contains at least a minimum number Midpoints, of objects is called as core object.

It can be expressed as follows –

$$N\varepsilon\ (u) = \{u\ \varepsilon\ X\ /\ dist\ (u) \leq (\varepsilon\text{-neighbourhood})\}$$

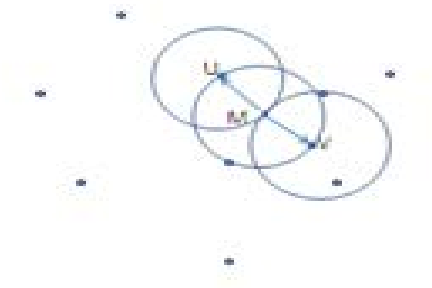
Directly Density Reachable –

Suppose X is a given set of objects, an object u is directly density reachable from object v if u is within the ε-neighbourhood of v and v is core object.

The condition for directly density reachable is as follows-

$$U\ \varepsilon\ N_{\varepsilon}(v) \text{ and}$$

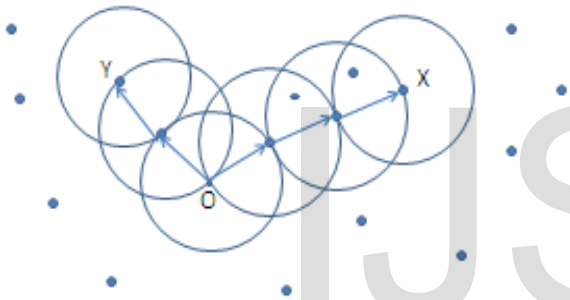
$$|N_{\varepsilon}(v)| \geq \text{Midpoints (core objects)}$$



Fig(c) (i) Density Reachability

An object u is density-reachable from object v with respect to ε and Midpoints in a set of objects, X, if there is a chain of objects ,u_1,u_2…….,u_n=v and u_n=u such that u_{p+1} is directly density reachable from u_p, with respect to ε and Midpoints,

for $1 \le p \le n$, $u_p \varepsilon X$[9].



Fig(c) (ii) Density Connectivity in density based Clustering

An object u is density Connected to object v with respect to ε and Midpoints in a set of objects, X, if there is an object z ε X such that both U and V are density reachable from o with respect to ε and MinPoints. In fig(c) M,V,O and X are core objects since each is in an ε-neighbourhood containing atleast three points .U is directly density reachable from M.M is directly density reachable from V and vice versa. O,X,Y are all density connected. For document summarization each and every sentence is considered as an object represented by a node as mentioned below, where the various clusters are formed and cluster which exactly contain the relevant information is returned.

ii)Hierarchical Clustering-

A hierarchical clustering method groups the data objects into a tree of clusters. Hierarchical clustering can be classified into Agglomerative and Divisive methods.

Agglomerative approach is a bottom-up strategy that starts by placing each object in its own cluster and then merges the individual clusters into larger and larger clusters, until all of the objects are in a single cluster or certain termination condition is satisfied.

Divisive approach is top-down strategy reverse of agglomerative approach. It starts with all objects in one cluster. It divides the cluster into smaller and smaller pieces, until each object forms a cluster on its own or until it satisfies certain termination condition. The user can specify the desired number of clusters as a termination condition. The measures for the distance between clusters are as follows-

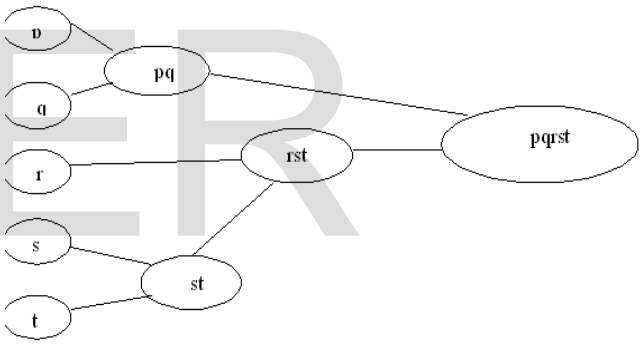
Min_dist : $d_{min}(C_i,C_j) = \min_{u \varepsilon C_i,\ u' \varepsilon C_j} |u-u'|$

Max_dist : $d_{max}(C_i,C_j) = \max_{u \varepsilon C_i,\ u' \varepsilon C_j} |u-u'|$

Mean_dist: $d_{mean}(C_i,C_j) = |m_i - m_j|$

Where $|u-u'|$ is the distance between two objects or points u and u', m_i is the mean for cluster C_i and n_i is the number of objects in C_i.

Avg_dist : $d_{avg}(C_i,C_j) = 1/n_i\, n_j \sum_{u \varepsilon C_i} \sum_{u' \varepsilon C_j} |u-u'|$



Fig(d) Hierarchical Clustering Algorithms

iii) Fuzzy C-means Clustering –

The Fuzzy C-means clustering allows the data objects to belong to more than one cluster. This technique is based on the minimization of the following objective function given below –

$W_l(a) = 1/\sum_l ((d(\text{center}(l),x)/d(\text{center}(b),x))^{2/(m-1)}$

The document graph retrieved from the clustering is traversed using shortest path spanning algorithm and obtained graph is processed using cosine similarity technique. The results obtained through above three clustering techniques are given as input to the third module to evaluate the quality of summary generated.

C. Development of quantitative and qualitative metrics-

The summary obtained from the second module from different clustering is compared with the reference summary

obtained from human natural processing through benchmarking. The parameters to be used for comparison are compression ratio, retention ratio, recall, and precision.

The compression ratio tells how much short text summary is generated than original text. The compression ratio should be small. The another important factor to be considered for analyzing the quality of summary is retention ratio which tells how much information is retained in the summary.

The retention ratio's value should be large, which helps to identify whether summary preserves the central idea of a source text.

Fig (e) shows the relation between various comparison parameters for evaluating the quality of a summary.

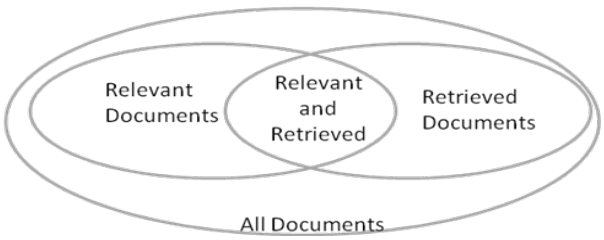
Text Summarizer	CR	RR	Precision	Recall
	Small	Large	Small	Large

Fig(e)Comparison table for relationship between various metrics.

Retention Ratio of summary = I(S) / I (T)

Where, I(S) is information retrieved in summary,

I (T) Information in a given text.



Fig(f) Relation Between Recall and Precision

Precision and recall – helps to evaluate the quality of summarizer, where precision is retrieved information that is relevant. Whereas recall is fraction of relevant data that has been retrieved.

Recall =Actual positive/(predicted(+ve) + predicted(-ve))

and

Precision = Actual (+vet) / (predicted(+ve)+Actual (-ve))

The recall should have higher value. Recall and Precision are inversely proportional to each other.

I) INTRINSIC METHODS -

It measures the system in of itself .which can be done through benchmarking where a reference summary is set and obtained summary is compared with benchmark standard. The intrinsic approach uses the coherence and informativeness as measures. Many times summary generated through cut and paste options suffer from the parts of summary being extracted out of text, which is called as coherence.To compare generated summary with the text being summarized in an effort to assess how much information from source is preserved.

II) EXTRINSIC METHODS –

It measures the efficiency (speed) and acceptability of generated summary in some context. Thus by analyzing various quantitative and Qualitative metrics and investigations methods such as intrinsic and extrinsic, we can evaluate the summary result set.

4 CONCLUSION

The proposed system aims at the development of system which will take one input file as .txt, which will be condensed into abstract version using DBSCAN, Hierarchical and Fuzzy C-means to generate query specific document summary. The generated summary from each and every clustering technique will be evaluated and compared using metrics such as compression ratio, retention ratio, recall and precision, to evolve better query specific document summarizing clustering technique.

REFERENCES

1] Ramakrishna Varadarajan "A System for Query-Specific Document Summarization" School of Computing and Information Sciences Florida International University,

2] A paper on ROCK: A Robust Clustering Algorithm for Categorical Attributes by Sudipto Guha Stanford University Stanford, CA 94305.

3] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, and Kyuseok Shim"Fast similarity search in the presence of noise, scaling, and translation in time-series databases" In Proc. of the VLDB Conference, Zurich, Switzerland, September 1995.

4]Mohammed Salem Binwahlan, Naomie Salim and Ladda,"Swarm Based Textummarization"International Association of Computer Science and Information Technology,2009,IEEE , pp.145-150.

5] Oi Mean Foong1, Alan Oxley1 and Suziah Sulaiman"Challenges and Trends of Automatic Text Summariza-

tion",International Journal of Information and Telecommunication Technology 2010.

6]Sunita R Patil and Sunita M.Mahajan"Document Summarization Using Extractive Approach" ,International Journal of computer applications, *2011*.

7]Laxmi Patil,M S Bewoor,S H Patil"Query Specific ROCK algorithm for text summarization" ,Intrenational Journal Of Engineering Research and applications vol2 Issue 3,May-June2012.pp26172620,.

8]Ms.Meghana Ingole,M S Bewoor,Dr.S.H.Patil"Text Summarization using EM clustering algorithm" ,International Journal Of Engineering Research and applications Vol. 2, Issue 4, July-August 2012, pp.168-171