COMPARISON BETWEEN GRAPH BASED DOCUMENT SUMMARIZATION METHOD AND CLUSTERING METHOD

Prashant D.Joshi[1], S.G.Joshi[2], M.S.Bewoor[3], S.H.Patil[4]

[1, 3, 4]Department of Computer Engineering, Bharati Vidyapeeth University, CoE, Pune, India
[2]Department of Computer Engineering, A.I.S.S.M.S. CoE, Pune, India

ABSTRACT

Document summarization and clustering are two techniques which can be used while accessing text files within short period of time from the computer. In document summarization graph method, document graph of each text file is generated. For creating document graph each paragraph is assumed as one individual node. Node score and Edge score are calculated using mathematical formulas. Input query is applied on the document and according to that summary from the Text file is generated. Clustering ROCK algorithm can also be used for doing the summarization. Here each paragraph is considered as individual cluster and link score between two paragraphs are calculated and on that basis two clusters are merged. Here Input query is applied on the merged clusters as well as individual cluster and accordingly summary is generated. Various results are taken in to consideration and we conclude that Rock algorithm requires less time as compared to other method for document summarization. Clustering ROCK algorithm can be used with standalone machine, LAN, Internet for retrieving text documents with small amount of retrieval time.

KEYWORDS: *Input Query, Document summarization, Document Graph, Clustering, Link, Robust Hierarchical Clustering Algorithm*

I. INTRODUCTION

Today every human with basic computer knowledge is connected with the world by using an internet. WWW provides features like communication, chatting, Information Retrieval. Huge amount of data is available on N number of servers in the form of the files like text files, document files. Text Summarization is the process of identifying the most salient information in a document or text file. In existing days Query Summarization was done through the BOW (Bag of Words) approach, in which both the query and sentences were represented with word vectors. But this approach has drawback where it merely considers lexical elements (words) in the documents, and ignores semantic relations among sentences. [6]

Graph method is very important and crucial in document summarization which provides effective way to study local, system level properties at a component level. Following examples shows the importance of graphs. In the application of Biological network a protein interaction network is represented by a graph with the protein as vertices and edge is exist between two vertices if the proteins are known to interact based on two hybrid analysis and other biological experiments[3]. In stock market graph vertices are represented by stocks and edge between two point exist if they are positively correlated over some threshold value based on the calculations.[3] in Internet application an Internet graph has vertices representing IP addresses while a web graph has vertices representing websites.[3]. In this paper we are comparing clustering ROCK algorithm with graph based document summarization algorithm for generating summary from the text file.

Even though there is an increasing interest in the use of clustering methods in pattern recognition [Anderberg1973], image processing [Jain and Flynn 1996] and information retrieval [Rasmussen 1992; Salton 1991], clustering has a rich history in other disciplines [Jain and Dubes 1988] such as biology, psychiatry, psychology, archaeology, geology, geography, and marketing..[4]

Currently, clustering algorithms can be categorized into partition-based, hierarchical, density-based, grid-based and model-based. [7] In clustering, related document should contain same or similar terms. One can expect a good document cluster to contain large number of matching terms. In reality when a document cluster is large, there is no single term that occurs in all the documents of the cluster. In contrast when a cluster is small one can expect certain term to occur in its all documents [8].

Clustering and Data summarization are two techniques which are present in data mining. Data Mining is the notion of all methods and techniques, which allow to analyze very large data sets to extract and discover previously unknown structures and relations out of such huge heaps of details These information is filtered, prepared and classified so that it will be a valuable aid for decisions and strategies.[5]

II. Related Work for Document Graph Method

2.1 Document Summarization

Query-oriented summarization is primarily concerned with synthesizing an informative and well-organized summary from a collection of text document by applying an input query. Today most successful multi-document summarization systems refer the extractive summarization framework. These systems first rank all the sentences in the original document set and then select the most silent sentences to compose summaries for a good coverage of the concepts. For the purpose of creating more concise and fluent summaries, some intensive post-processing approaches are also appended on the extracted sentences.

Here input query as *q* and the collection of documents as *D*. The goal of QS is to generate a summary which best meets the information needs expressed by *q* . To do this, a Query Summarization system generally takes two steps: first, the stop words from documents as well as from input query is removed. Second sentences are selected until the length of the summary is reached.

For making document graph node weights, edge weights should be known.

Nodes are nothing but the paragraphs. Node weights are calculated after applying an input query.

Following formula is refereed for calculating the node score.[1]

$$\sum_{t\in Q,d} \ln \frac{N-df+0.5}{df+0.5} \cdot \frac{(K1+1)tf}{\left(k1(1-b)+b\frac{dl}{avdl}\right)+tf} \cdot \frac{(k3+1)qtf}{k3+qtf} \text{....(1) [1]}$$

where N is total number of text files present on the system.
df is total number of text files that contains the input term.
tf means total count of input keywords in text file.
qtf means number of times keyword occurred in input query.
k1, b, k3 are constant value. Here k1 is assumed as 1, b =0.5, k3 =2
dl is the total text file length.
avdl is average document length assume as 120.

2.2 Problem Definition for Document Summarization using Graph based Algorithm

Lets we have n document i.e.d1, d2, to dn. Size of document is total number of words. i.e. size (di).

Term frequency tf(d,w) is no of words present in documents.

Inverse document frequency is i.e.idf(w) Means inverse of documents contain word w in all Documents.

Keyword query is set of words. i.e.Q(w1,w2…wn).

The document graph G (V, E) of a document d is defined as follows:

- d is split to a set of non-overlapping text fragments t(v),each corresponding to a node v€V.
- An edge e(u,v) €E is added between nodes u,v if there is an association between t(u) and t(v) in d.

Two nodes can be connected using edges. Such edge weight is calculated by following formula. Here t (u) means first paragraph and t (v) means second paragraph. Like this edge weights between all paragraphs are calculated and stored in the database. Size t (u) shows number of keyword in first paragraph and t (v) shows number of keyword in second paragraph. Edge weight can be calculated before applying the input query because no. of text files are present on the system.[1]

$$\textbf{Escore (e)}=\frac{\sum_{w\in(t(u)\cap t(v))}((tf(t(u),w)+tf(t(v),w)).idf(w)))}{size(t(u))+size(t(v))}\ldots(2)\ [1]$$

w€ t(u) ∩ t(v) means that common word present in both paragraph. Common keyword count is assigned to w. in this fashion all edge score of all text files are calculated and they are permanently stored in the database. When new file is added then this module is run by administrator for storing the edge weights in the database.

Summary module is referring the concept of spanning tree on the document graph because multiple Nodes may have input query so which nodes will be selected? Different combinations from graph are identified and node score is generated using following formula.

$$\textbf{Score (T)} = a \sum_{edge\ e\in T} \frac{1}{Escore(e)} + b \frac{1}{\sum_{node\ v\in T} Nscore(v)} \quad \ldots.(3)\ [1]$$

Equation (3) will calculate the spanning tree score in the document graph.[1] From spanning tree table the minimum score of spanning tree is considered and that paragraph is displayed as summary.

III. CLUSTERING

Clustering can be considered as the most important unsupervised learning problem. Various techniques can be applied for making the groups. A loose definition of clustering could be "the process of organizing objects into groups whose members are similar with certain property. The similarity criterion is distance: two or more objects belong to the same cluster if they are "close" according to a given distance (in this case geometrical distance). This is called distance-based clustering.[4]

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

3.1 Example:

Clustering concept is always used with library where we have different subject's book. These books are arranged in proper structure to reduce the access time. Consider books of operating system they will be kept in operating system shelf. Shelf has also assigned numbers for managing books efficiently. Likewise all subjects' books are arranged in cluster form.

Clustering algorithms can be applied in many fields, for example

- City-planning: globally houses are arranged by considering house type, value and geographical location;
- Earthquake studies: clustering is applied while observing dangers zone.
- World Wide Web: in WWW clustering is applied for document classification and document summary generation.
- Marketing: for getting the details of the customer who purchase similar thing from huge amount of data.
- Biology: classification of plants and animals given their features;
- Libraries: organizing book in efficient order for reducing the access delay.
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds [4]

Problem definition: Assume n is no of text documents with size p number of paragraphs. Generate the summary from text files while applying the input query q. This paper follows following system architecture for implementing text file summarization using clustering as well as graph based method. Below fig.1.1 shows the system architecture for implementation of this system.

IV. SYSTEM ARCHITECTURE

This system is developed in network environment. The main goal of this system is to get relevant text file from the server without going through all text files. User time will be saved by just reading the summary of text file relevant to input query. Here user input query is compared with all text files and from that which text file is most relevant to input query that is generated as an output on user machine. Here User can use graphical summarization method or can use clustering algorithm for generating summary.

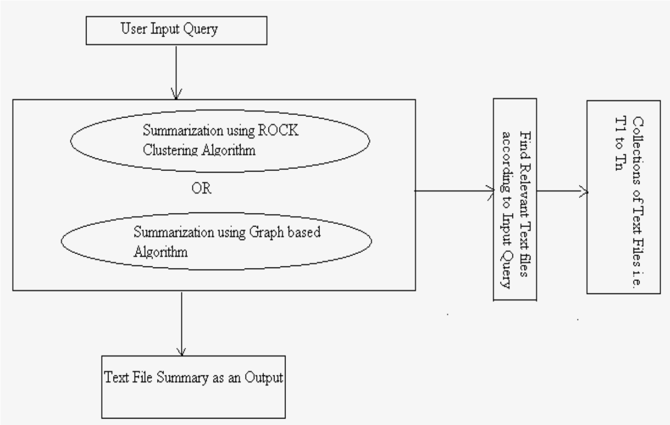
User Input Query
Summarization using ROCK Clustering Algorithm
OR
Summarization using Graph based Algorithm
Find Relevant Text files according to Input Query
Collections of Text Files i.e. T1 to Tn
Text File Summary as an Output

Fig1.1 system Architecture for Document Summarization and Clustering Method

V. ROCK ALGORITHM FOR CLUSTERING

```
Procedure cluster(S,k)Begin
Link:=compute_links(S)
For each s € S do
    q[s]:=build_local_heap(link,s)
Q:=build_gloabal_heap(S,q)
While size(Q)>k do {
    u:=extract,max(Q)
    v:=max(q[u])
    delete(Q,v)
    w:=merge(u,v)
    for each x€ q[u] U q[v] do {
        link[x,w]:= link[x,u]+ link[x,v]
        delete(q(x),u);delete(q(x),v)
        insert(q([x],w,g(x,w));insert(q[w],x,q(x,w))
        update(Q,x,q[x])
    }
    insert(Q,w,q[w])
    Deallocate(q[u];deallocate(q[v])
}
end
```

5.1 For calculating Link score here following algorithm is used.

Procedure compute_link(S)

```
Begin
1.Compute nbrlist[i] for every point i in S
2.Set link[i,j]   to be zero for all i,j.
3.for i=1 to n do {
4. N:=nbrlist[i]
5.for j:=1 to |N| -1 do
6.for l:= j+1 to |N| do
7.link[N[j],N[l]:=link[N[j],N[l]+1
8.}… [2]
```

Following Example will give the concept of clustering and how it is applied on the text file. Let's assume we have brainchip text file which contains four paragraphs.
1 Brain chip offers hope for paralyzed.
2. A team of neuroscientists have successfully implanted a chip into the brain of a quadriplegic man, allowing him to control a computer.
3. Since the insertion of the tiny device in June, the 25-year old has been able to check email an play computer games simply using thoughts. He can also turn lights on and off and control a television, all while talking and moving his head.
4. The chip, called BrainGate, is being developed by Massachusetts-based neurotechnology company Cyberkinetics, following research undertaken at Brown University, Rhode Island.
Rock algorithm is applied on above text file following thing will be done on this file and result is generated.
Count number of paragraphs in this file. Remove stop words from this file.
Assume each paragraph as individual cluster.
Above file contains 4 paragraphs. i.e.P1, P2, P3, P4.
Start with P1, compare P1 with all reaming paragraphs and find the value of link.
Link score is calculated by comparing keywords of each paragraph. The results of link score will be stored in one array.

Table 1.1 Keywords of each individual paragraph

Keywords List of C1	**Keywords List of C2**	**Keywords List of C3**	**Keywords List of C4**
Brain	Team	insertion	**Chip**
Chip	neuroscientists	Tiny	BrainGate
Offers	successfully	device	Developed
Hope	implanted	June	Massachusetts_based
paralyzed	**chip**	25-year	neurotechnology
	brain	old	Company
	quadriplegic	check	Cyberkinetics
	man	Email	**Research**
	allowing	play	Undertaken
	Control	computer	Brown
	computer	games	University
		simply	Rhode
		thoughts	Island
		turn	
		Lights	
		control	
		television	
		talking	
		moving	
		head	

Table 1.2 Local heap, Link result for P1-P4

Paragraphs	Link Result	Common words
P1,p2	02	Chip, brain
P1,p3	00	Nil
P1,p4	01	Chip

Table 1.3 Local heap, Link result for P2-P4

Paragraphs	Link Result	Common words
P2,p3	02	Control, Computer
P2,p4	01	Chip

Table 1.4 Local heap, Link result for P3-P4

Paragraphs	Link Result	Common words
P3,P4	00	Nil

From Table 1.2 it can easily understand that P1_P2 Link score is maximum. So P1-P2 can be merged and one new cluster can be created. From Table 1.3 P2-P3 Link score is maximum i.e.2 so P2-P3 can be merged and one new cluster can be created. In Table 1.4 Link score of P3-P4 is zero so no need to make the cluster.

Now we have C1, C2, C3, C4 total 4 clusters. Where C1 is merged keywords of P1-P2, C2 is merged keywords of P2-P3, C3 is individual paragraph3 i.e. P3 which is not matching with any other paragraphs. Likewise C4 which is paragraph P4 having single keyword common with P1 but link score of P1-P4 is less than P1-P2. Here P4 is considered as individual cluster because input query may be present with this paragraph also. So even though two paragraphs are not matching we want to take them as separate cluster. Now apply "**Brain Chip Research"** Query on Merged cluster as well as individual cluster.

Brain chip part of Input query is present with both C1, C2 which shown with bold Letters. In C3 there is no keyword of Input "Brain Chip Research". In cluster C4 'Chip' and 'Research' keywords are present with C4. The Keyword count of Input query on cluster as well as the size of Cluster is considered while selecting final cluster as an output.

Here we are not getting "brain chip research" input query from individual cluster. So once again clustering algorithm should be applied on C1, C2, and C4. Link score between C1-C2, C1-C4, and C2-C4 is calculated and stored in database.

C1-C4, C2-C4 will give all part of input query. C1-C4 will give Keyword count of 18 where as C2-C4 will give keyword count of 24. So C1-C4 gives less count so Summary should be generated from C1, C4 clusters.

VI. EXPERIMENTAL RESULT

We have implemented above system with following Hardware and software configuration.
Pentium Processor: –IV, Hard disk: 160 Gb, RAM Capacity: 1 Gb
Software requirement for implementation of above system is:
Operating System: Windows XP, Visual Studio.NET 2008, SQL Server 2005.
We have stored 57 text files in the database, the memory capacity required for these text files were 122 kb.

Table 1.5 Clustering and Graph based Algorithm Result

Sr.No.	File Name	Input Query	Rock Algo (Time in millisecond)	Graph Algo (Time in millisecond)
1	F1.txt	eukaryotic organisms	218	234
2	F2.txt	woody plant	249	280
3	F4	Bollywood film music	296	439
4	F6	personal computers	327	592
5	F7	Taj Mahal monument	390	852
6	F8	computer programs software	468	1216

7	F13	wireless local area network	390	758
8	F15	Mobile WiMAX	780	1060
9	F16	system development	670	724
10	F22	remote procedure calls	546	1482

First query is **"eukaryotic organisms"** which is applied on the system. Rock algorithm requires 218 milliseconds where as Graph based summarization requires 234 millisecond. Second query applied is **"woody plant"** here Rock algorithm requires 249 millisecond where as Document Graph algorithm requires 280Milisecond. After observing execution time of all Input query we conclude that Clustering Rock algorithm has good performance than graph based document summarization. But when input query is not available in any of the text file then graph based summarization gives output fast as compared to Rock Algorithm.

VII. CONCLUSION

In this paper we have compared the performance of Graph based document summarization method with clustering method. And the performance of Rock algorithm is better than Graph based document summarization method algorithm. This system can be applied with stand alone machine, LAN, WAN for retrieving text files within short period of time. Further this system can be improved to work on Doc file as well as PDF file which contain huge number of textual data.

ACKNOWLEDGEMENT

I am thankful to Professor & H.O.D. Dr. S. H. Patil, Associate Professor M. S. Bewoor, Prof. Shweta Joshi for their continuous guidance. I am also thanks to all my friends who are directly or indirectly supported me to complete this system.

REFERENCES

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

[2]. Sudipto Guha_Stanford University Stanford, CA 94305, Rajeev Rastogi Bell Laboratories, Murray Hill, NJ 07974 Kyuseok Shim,Bell Laboratories Murray Hill, NJ 07974 Paper on "A Robust Clustering Algorithm for Categorical Attributes" .

[3]. Balabhaskar Balasundaram 4th IEEE Conference on Automation Science and Engineering Key Bridge arriott, Washington DC, USA August 23-26, 2008 " A cohesive Subgroup Model For Graph-based Text Mining".

[4]. A Review by A.K. Jain Michigan State University,M.N. Murty Indian Institute of Science AND P.J. Flynn The Ohio State University on "Data Clustering".

[5]. Johannes Grabmeier University of Applied Sciences, Deggendorf, Edlmaierstr 6+8, D-94469Deggendorf, Germany, Andreas Rudolph Universitat der Bundeswehr Munchen, Werner-Heisenberg-Weg 39, Neubiberg, Germany D-85579, on "Techniques of Cluster Algorithms in Data Mining".

[6]. Prashant D. Joshi, M. S. Bewoor,S. H. Patil on topic "System for document summarization using graphs In text mining" in "International Journal of Advances in Engineering & Technology (IJAET)".

[7]. Bao-Zhi Qiu1, Xiang-Li Li, and Jun-Yi Shen, on "Grid-Based Clustering Algorithm Based on Intersecting Partition and Density Estimation".

[8]. Jacob kogan, Department of Mathematics and statistics,Marc Teboulle, paper on "The Entropic Geometric Means Algorithm: An Approach to Building small clusters for large text datasets".

Authors

Prashant D. Joshi currently working as Assistant professor and pursuing Mtech Degree from Bharati Vidyapeeth Deemed University College of Engineering Pune. Total 5 and half years of teaching experience and six months of software development experience. He has Completed B.E. Computer science degree from Dr. Babasaheb Ambedkar University Aurangabad (MH) in year 2005 with distinction. Published 2 papers in national conferences, 2 papers in International conferences and published 1 paper in international Journal.His area of interest is Data Mining, Programming Languages, and Microprocessors.

S. G. Joshi currently working as Lecturer in A.I.S.S.M.S.College of Engineering, Pune. She is having total 2 years of teaching experience in polytechnic college. She has completed B.E. computer science engineering from Swami Ramanand Teerth Marathwada University Nanded with distinction. Her research interest is in Data Mining, Operating System, and Data Structure.

M. S. Bewoor currently working as Assistant Professor in Bharati Vidyapeeth Deemed university college of enginering,pune.she has total having 10 years of teaching experience in Engineering college and 3 years of Industry experience. She is involved in Reseacrh activity by presenting 07 papers in national conferences, 08- international conferences and 07- International journals.Her area of interest is Data Structure,Data Mining,Artificial Intelligence.

S. H. Patil working as professor and Head of Computer Department at Bharati Vidyapeeth Deemed University College of engineering Pune. Total 24 years of teaching experience. He has published more than 100 papers in National conferences, International conferences, National Journals and International Journals. His area of Interest is Operating System, Computer Network, and Database Management System.