International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 7 4929 - 4934

Keyword Merging Based Multi Document Enhanced Summarization

Ms. Ajita Patil
Depatment of Computer Engineering
DCOER, Savitribai Phule Pune University,
Pune, India
Email: ajitapatil89@gmail.com

Prof .Mane P.M.
Depatment of Computer Engineering
DCOER, Savitribai Phule Pune University,
Pune, India
Email: prashant.mane@zealeducation.com

Abstract:- Automatic text summarization is a wide research area. There are several ways in which one can characterize different approaches to text summarization: extractive and abstractive from single document or multi document. Summary is text that is produced from one or more text. Document summarization is a procedure that building coated version of document that gives respected data to the client, and multi-document summarization is to produce a summary conveying the larger part of data substance from a set of documents about an implicit or explicit primary point.This paper describes a system for the summarization of multiple documents. The system produces multi-document summaries using data merging techniques. For combining multiple document on same thing the system uses Bisecting k-means algorithm which works better than basic K-means algorithm.Our System uses Enhanced Summarization algorithm to summarize multiple document.The Enhanced algorithm is applied separately on each cluster. According to results this system gives better results as compared to NEWSUM algorithm.

Keywords—*Document Summarization; K-means; Bisecting Algorithm, Clustering*

*****

I. INTRODUCTION

The World Wide Web surrounds the information and billions of documents.It is impossible to anybody to read all the related documents once. so there is need to provide high-quality summaries in order to allow the user to quickly locate the desired information. It associated to start improvement of overall summary of system. The current system produce brief summary of system and most vital information by captivating single article, a cluster of article, a broadcast news.The improvement of numerous summarization application for news, email strings, lay and proficient medicinal information, exploratory articles, spontaneous dialogs, voice message, broadcast news and feature, and gathering recordings. The typical issue of these applications which is the numerous document that covers comparative information as on account of numerous news stories about an events or a planning of events[9]. The capacity to outline the likeness and discrepancy in information content which is delicate significance to user is the main challenge of content summarization. Multidocument summarization (MDS) involves developing a short summary from a couple of documents which focuses on a single topic . Recently, most researchers for automatic text summarization have transferred their efforts from single documents to multiple documents but they have to aware with the issues of redundancy, sentence ordering, collocation, etc. Extractive MDS uses past work for the issue of sentence determination and requesting individually for that joint model are important to produced coherent summary. The constitution is forthright: Acceptance ordering of the sentence is not exit if the given sentence is correctly chosen without unambiguous [10].Summarization required information over-burden and same information is defined by several online document. Summary offers to the user by positioning the normal information document and highlight distinct document.client takes single events through a few news wires it is very challenging to summary. Automatic combination of information over multiple documents utilizing language creation is brief summary which is presented in this paper [9]. Most of the research is on single document summarization for domain sovereign task and delivering summary it uses sentence extraction[10]. The information of first article is conflict in multi document summarization of article about the same events, when we extracting any comparative sentence then result will be repetitive.

The rest of this work is organized as follows. Section 2 presents some related work. Section 3 describes the implementation details and procedures used. Results and disscussion are presented in section 4,and finally, conclusions are given in section 5.

II. RELATED WORK

The paper[1] presents CATS system (Cats is an Answering Text Summarizer) uses the extraction of sentences to create a 250-word summary of the cluster.They present CATS, a system for summarizing multiple documents concerning a given topic at a level of granularity specified in a user profile. The system first performs a thematic analysis of the documents and then matches these themes with the ones identified in the question. Once CATS has identified a list of thematic segments containing interesting aspects related to the subject, they are sorted to find the most promising ones.Problem with this system is that there is need to improve sentence compression.

In paper [2] proposes two improvements in above work.They present a sentence-compression-based framework and design a series of learning-based compression models built on parse trees. An innovative beam search decoder is proposed to efficiently find highly probable compressions.This is query focused query-focused MDS framework which consisting of three steps:Sentence Ranking, Sentence Compression and Post-processing.This

system improves sentence compression but the sentence ordering is not done.
The approach in paper [3]summarizes a text according to the relevance of the sentences within the text is derived by Simplified Lesk algorithm and WordNet, which is an online dictionary. This approach is not only independent of the format of the text and position of a sentence in a text, as the sentences are arranged at first according to their relevance before the summarization process, the percentage of summarization can be varied according to needs. This approach is based on the semantic information of the extracts in a text. So, different parameters like formats, positions of different units in the text are not taken into account.

In the paper[4] gives an approach to cluster multiple documents by using document clustering approach and to produce cluster wise summary.It is based on based on featurebased sentence extraction strategy. Related documents are grouped into same cluster using document clustering algorithm. This system is used for more different type of dataset with necessary changes to make it efficient.
In the paper[5],gives a systemization of representation of multisets and basic operations of multisets, and an overview of the applications of multisets in mathematics, computer science and related areas.

The paper[14] focuses on using data merging techniques to summarize set multiple documents.Here fb optimal merge function is used to find out set of keywords and then this system uses Enhanced summarization algorithm to generate summary.

III. IMPLEMENTATION DETAIL

A. System Overview

The following Figure 1.Shows the proposed system architecture

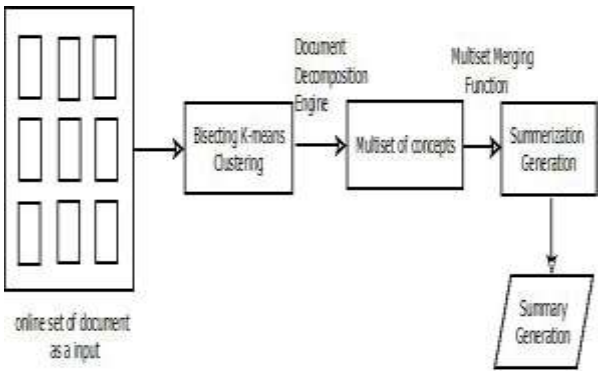

Fig.1: System Architecture

This system uses the online set of document as aninput.Online set of documents contains large no of documents of differents on different themes.So it is essential to cluster the documents on same themes. So,for this purpose I use Bisecting Kmeans algorithm. Bisecting Kmeans requires less time as compared to Kmeans algorithm.From the cluster documents the important keywords are find out, which are merge using f_β-optimal merge function.The important keyword are also called as keyconcepts. The key concept obtain after the merging technique are used as an input to Enhanced Summarization Technique which produces summary of multiple documents. The summarization process of documents is frequently separated into two main steps: the content selection step and the content presentation step. In the content selection step it is decided what is deliberated to be the information we want to present in the second step. The main focus of this paper lies with the content selection step. We introduce Enhance summarization technique that uses sentence extraction in order to produce summarizations.

B. Enhanced Summarization Technique

In current algorithm of summarization i.e. NEWSUM [15] summary comprehends the number of sentences is equal or less than size of the keyword set. As per design of NEWSUM algorithmin each iteration only one sentence is selected and keywords covered in that sentence are removed from the keyword set to reduce redundancy, but in next iteration removed keyword are notconsidered for the scoring of the sentence. So, there is possibility to miss sentences which are important than selected sentences in previousiteration.Also sentence ordering is not done properly.Summary also contains statements like for ex. etc.These statements should be removed. We proposed the new algorithm for summarization which will overcome this issue.

Algorithm works in following steps:

1) Sentences are mapped to keywords; sentence can be mapped to multiple keywords.
2) Keywords-Sentence hash table is formed in which each keyword has mapped sentences.
3) Sentence Frequency table is formed which describes number of the keywords to which sentences are mapped.
4) For each keyword in the Keyword sentence hash table all sentences are sorted into decreasing order of sentence frequency
5) For each keyword top k sentences are selected which are not in the summary and added into summary.
6)Sentence score is calculated for each selected sentence in summary. Sentence score is calculated by finding the no. of occurrences of nouns.
7)Then select such sentences which are above some specific threshold.
8)Also remove the sentences which contains words like examples,,in short or signs like curly braces etc.

C. Algorithm

Algorithm 1: K-means Clustering

1: Place K points into the space.
2: Assign each object to the group having the closest centroids.
3: Recalculate the positions of the K centroids.

4: Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated. Where μi is the mean of points in Si.

Algorithm 2: Bisecting K-means Clustering

1: Initialize the list of clusters to contain the cluster all points.
2: repeat
3: Select a cluster from the list of clusters.
4: for i=1 to number of iterations do
5: Bisect the selected cluster using basic K-means.
6: end for
7: Add the two clusters from the bisecting with the lowest SSE to the list of clusters until the list of clusters contains K clusters

Algorithm 3:Enhanced Summarization Techniques

1: For each sentence S in cluster,
2: Map sentence S to KeyConceptKc
3: End.
4: For each KeyConcept Kc
5: Select top t sentence Occurrence is Maximum
6: End.

D. Mathematical Model

The System S is represented as: S ={I, C, M, S, N}
A. Input Online set of documents.
Let, I be the set of inputs I = (i1, i2, i3,)
Where, i1, i2, i3... are the number of online input.

B. Clustering Algorithm
Consider, C is a set for clustering and C = (c1, c2, c3,...)
Where, c1, c2, c3... are the number of clusters formed.

C. Multiset
Let, M is the set of Multiset of multisets of keyconcepts M =(m1, m2, m3,..)
Where, m1, m2, m3... are the number of multisets of documents.

E. Summarization Generator
Let, S is the set of summary S.

Multisets:

Multiset is a generalization of the notion of a set in which members are allowed to appear more than once.
A merge function over a universe U is defined by a function.
M: U →N
For each u ∈U, M(u) denotes the Multiplicity of u in M. Theset of all multisets drawn from a universe U is denoted by M(U).

Merge Function:

Functions that maps multisets of object into single object is called as merge functions. A merge function over a universe U is defined by a function:
1^{st} Order Merge Function: ϖ: M(U) →U
2^{nd} Order Merge Function: ϖ*: M(M(U)) →M(U)
Simple merge functions for multisets -Source intersection and Source Union.

Local precision and Recall:

Consider a Multiset of sources M=S1,S2,,Sn Local precision and recall are defined by functions $P^*$and r^* such that:

$$\forall u \in U : \forall j \in \mathbb{N} : p^*(u, j|M) = \frac{1}{|M|} \sum_{S \in M \wedge S(u) \geq j} M(S$$

$$\forall u \in U : \forall j \in \mathbb{N} : r^*(u, j|M) = \frac{1}{|M|} \sum_{S \in M \wedge S(u) \leq j} M(S$$

How to find out Key-concept set:

For finding out important keyword set f_β-Optimal Merge Function is used.

f_β-Optimal Merge Function:
Consider a Multiset of sources M=S1,S2 ,Sn.

$$\varpi^*(M) = \underset{\zeta \varepsilon M(U)}{argmax} f_\beta(\zeta|M)$$

$$\varpi^*(M) = \underset{\zeta \varepsilon M(U)}{argmax} \left(\frac{(1+\beta^2).p(\zeta|M).r(\zeta|M)}{\beta^2.p(\zeta|M)+r(\zeta|M)}\right)$$

The parameter _ expresses how much more weight is given to recall as opposed to precision, more specifically, recall has a weight of β_ times precision
β<1, Preference is given to precision.
β>1, Preference is given to recall.
Where, with T as triangular norm, we have that:

$$p(\zeta|M) = T_{u\varepsilon\zeta} - (\, p^*(u, \zeta(u))|M)$$
$$r(\zeta|M) = T_{u\varepsilon\zeta} - (\, r^*(u, \zeta(u))|M)$$

This is useful when we would want to select a lesser amount of sentences, knowing they give a preference to the essential data then we would want to give a preference to precision.If we want to generate a larger summarization then preference is given to recall. In this way, set f_β-Optimal Merge Function fins out the keyword wrt precision and recall. This keyword set is given as a input to Enhanced Summarization

.

E. Enhanced Summarization Technique with example.
Input-

1) Set of Keywords from each cluster C
   K= (K1, K2, K3...Kn) Where K is the set of keywords.
2) Set of Sentences from all documents in cluster C
   S= (S1, S2, S3..., Sn)

Where s is the set of sentences.

Process-

1) Sentence Mapping
For all sentences and for all keywords
{
if (Si contains keyword Ki)
Map sentence Si to keyword Ki
}
2) Hash Map(Hm) Generation
In the above table keyword K1 is found in S1, S2, S3 and S4 sentences. Thus (S1, S2, S3, and S4) be the array list of sentences map to keyword K1. Similarly keyword K2 is found in S1, S2, S4, S5 sentences. Thus (S1, S2, S4, S5) be the array list of sentences map to keyword K2.

TABLE I. HASH MAP TABLE

Keywords	Sentences
K1	S1,S2,S3,S4
K2	S1,S2,S4,S5
K3	S2,S3,S7
Kn	S1,S2,...Sn

3) Frequency count
Fi = count number of keywords to which sentence (Si) is mapped.
Where F is the Frequency count.

4) Sort array list
For each keyword (Ki) in hash map
{
Sort array list of sentences according to the frequency count (Fi) in descending order.
}

5) Sentence Summary- Now each keyword (ki) in hash map (Hm) has list of sorted sentences (Li) according to frequency count (Fi). Toselect sentences, summary select top M unique sentences from list(Li) for each keyword (ki). For example. If for keyword Ki top M sentences are selected and we need to select sentences for keyword Ki+1 then we do not consider sentences which are selected for keyword Ki. If there are n keywords then (n X M) sentences are selected.
6)Sentence score-It is calculated for each selected sentence in summary. Sentence score is calculated by finding the no. of occurrences of nouns.
7)Selentence Selection-Select such sentences which are above some specific threshold.
8)Also remove the sentences which contains words like examples,,in short or signs like curly braces etc.
9)Sentence Ordering-Sentence ordering is done by arranging the sentences according to position and chronology of the original documents.

Output-

Summary of multiple documents.

F. Experimental Setup

The system is built using Java framework (version jdk1.8) on Windows platform. The Net beans (version 8.0) is used as a development tool. The system doesn't require any specific hardware to run; any standard machine is capable of running the application.

IV. RESULTS AND DISCUSSION

A. Dataset
Dataset multiple documents on different theme is input for clustering process. Clustering is done using Bisecting K-means algorithm. I use DUC2002 as input dataset.

.B. Experimental Setup
For experimental set up, I use Windows XP operating system, Intel Pentium 4 processor, 4 GB RAM, 80GB Hard disk, Net Beans IDE 8 + JDK tool. To calculate the results, Duc2002 datasets are used.

C. *Results*
Clustering Having introduced the two different summarization algorithms and their implementation we now turn to techniques of a practical study. It involves both algorithms and testing of multiple documents. The whole dataset consist of number of records. Here we use human generated summary also called as gold summary as a standard for evaluation. The Figure 9.1 shows the time comparison of K-means and Bisecting Kmeans algorithm. The time required by bisect k-means is approximately 600 ms, while the time required by the k-means clustering is 1600 ms. In the following graph in X-axis shows the k-cluster size for both algorithm while in the Y-axis shows the time in milliseconds. From the following graph it is conclude that the bisect k-means required less time than the existing algorithm which is k-means algorithm.

Summarization
we present the results of evaluating Summary of multiple documents which are coreferent. System uses Extractive summarization apporoach based on Feature Extraction Method. System uses Enhanced summarization algorithm. The performance evaluation of summarization system is held using different parameters. The Table II shows the fβ value comparison of NEWSUM and Enhanced Summarization algorithm and Expert summary. And the graph representation of Table II is shown in Figure 3. We have used fβ Optimal Merge Function, ,so we also shows graph comparion of all the three summaries depends value

of fβ. If value is less than 1, then preference is given to precision. So lesser summary is generated. If value is gerater than 1, then preference is given to recall,then larger summary is generated.If value is equal to 1 no preference is given.Results shows that our summary is highly similar to expert summary and has high recall and precision significance test than NEWSUM summary.

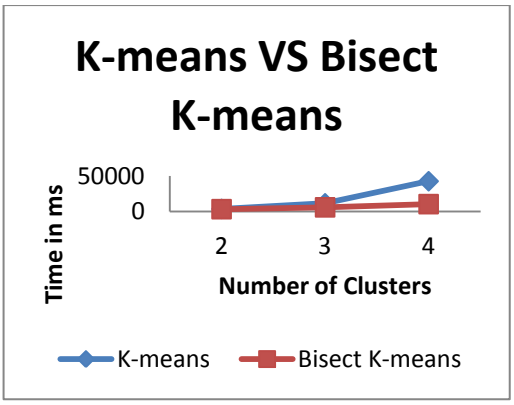

Figure.2 : Time Comparison graph between K-means and Bisect K-means clustering

TABLE II: SUMMARY COMPARISON TABLE

Value of fβ Optimal Merge Function	Existing Summary	Proposed Summary	Expert Summary
0.5	0.6	0.84	0.98
1	0.65	0.87	0.98
1.5	0.5	0.94	0.98

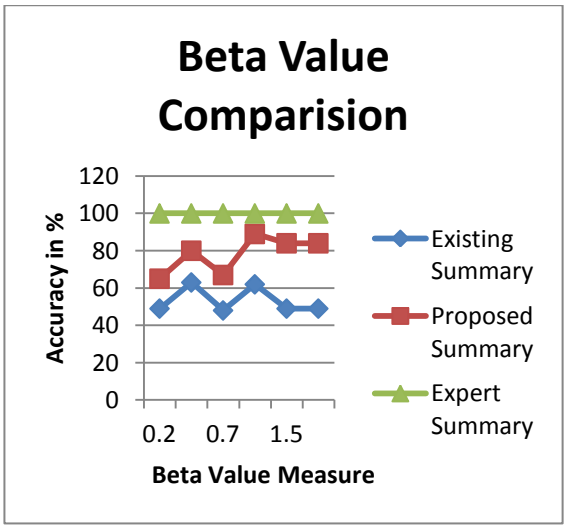

Figure 3 : Summary Compararison Graph

V. CONCLUSION

This paper explains how to generate summary of multiple Documents. Here summarization technique used is data merging technique.fβoptimal merge function is used to objectively give a preference to precision or recall the best, compared to the other merge functions. The proposed enhance summarization algorithm helps us to summarize no of documents on same theme. Multiple documents on different theme are first clustered using bisecting kmeans algorithm.Problems of NEWSUM algorithm are removed. Proposed enhanced summarization algorithm works better than NEWSUM algorithm and maximizes precision and recall result.

REFERENCES

[1] A. Farzindar, F. Rozon, and G. Lapalme, "Cats a topic-oriented multidocument summarization system," in DUC2005 Workshop, NIST. Vancouver: NIST, p. 8 pages ,oct 2005..

[2] LuWang,Hema Raghavan,Vittorio Castelli,Radu Florian,Claire Cardie , "A Sentence Compression Based Framework to Query-Focused Multi-Document Summarization",in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1384–1394, Sofia, Bulgaria, August 4-9 2013.

[3] A.Kogilavani1 and Dr.P.Balasubramani "Clustering and feature specific sentence Extraction based summarization of Multiple documents". In International journal of computer science & information Technology (IJCSIT) Vol.2, No.4,pages 13, August 2010.

[4] D. Singh, A. M. Ibrahim, T. Yohanna and J. N.Singh," An Overview Of The Applications Of Multisets," in Fuzzy Sets and Systems, Vol. 37, No. 2,pages 73-92,2007.

[5] C.-Y. Lin and E. Hovy, "From single to multi-document summarization: a prototype system and its evaluation," in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ser. ACL 02. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 457464, 2002.

[6] Marcu, D. "Discourse trees are good indicators of importance in text". In I. Mani and M. Maybury (eds), Advances in Automatic Text Summarization, 123136. MIT Press,1999.

[7] Radev, D.R. and K.R. McKeown. Generating Natural Language Summaries from Multiple On-line Sources. Computational Linguistics, 24(3):469500, 1998.

[8] J. Bleiholder and F. Naumann, "Data fusion," ACM Comput.Surv., vol. 41, no. 1, pp. 1:11:41, Jan. 2009.

[9] K. S. Jones, "Automatic summarizing: The state of the art," Inf. Process. Manage. vol. 43, no. 6, pp. 14491481, 2007.
.

[10] Marcu, D."The theory and practice of discourse parsing and summarization", Cambridge MA: MIT Press, 2000.

[11] Regina Barzilay, Kathleen R. McKeown, and Michael Elhadad, "Information fusion in the context of multi-document summarization," in ACL '99 Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, pp. 550557, 1999.

[12] Kathleen R McKeown, Judith Klavans, VasileiosHatzivassiloglou, Regina Barzilay, and

EleazarEskin. “Towards multidocument summarization by reformulation: Progress and prospects”,1999.

[13] E. Canhasi and I. Kononenko, “Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization,” Expert Systems with Applications, vol. 41, no. 2, pp. 535 543, 2014.

[14] R. M. Aliguliyev, “A new sentence similarity measure and sentence based extractive technique for automatic text summarization,” Expert Syst. Appl., vol. 36, no. 4, pp. 77647772, May 2009.

[15] Daan Van Britsom, AntoonBronselaer, Guy De Tre,“Using data merging techniques for generating multi-document summarizations”, Department of Telecommunications and Information Processing, sGhent University Sint-Pietersnieuwstraat 4, B-9000 Ghent, Belgium, 2013