# Improving the Performance of Multiple Document Summarization Using Mead Extraction

**C.V.Sheeba [1], T.Meyyappan [2], SM.Thamarai [3]**

*[1]Departmentmet of Computer Science, Alagappa University,Karaikudi, Tamilnadu, India. sheeba123raj@gmail.com[1]
*[2]Professor Department of Computer Science, Alagappa University, Karaikudi, Tamilnadu, India. meyyappant@alagappauniversity.ac.in[2]
*[3] Guest Lecturer, Alagappa Government Arts College, Karaikudi, Tamilnadu, India. lotusmeys@yahoo.com[3]

## ABSTRACT

In this paper, we present three techniques for generating extraction based summaries including a novel graph based formulation to improve on the former methods. The first method uses a sentence importance score calculator based on various semantic features and a semantic similarity score to select sentences that would be most representative of the document. It uses stack-decoder algorithm as used as a template and builds on it to produce summaries that are closer to optimal. The second approach clusters sentences based on the above semantic similarity score and picks a representative from each cluster to be included in the generated summary. The third approach is a novel graph problem based formulation where summaries are generated based on the cliques found in the constructed graph. The graph is generated by building edges between sentences which talk about similar topics but are semantically not similar.

## INTRODUCTION

### 1.1. Subject matter detection and multi-report summarization

James Allan et.al proposed a method [1] of generate a precis, one ought to first start with relevant files that one desires to summarize. The technique of figuring out all articles on an rising event is referred to as subject matter Detection and monitoring (TDT). A massive frame of studies in TDT has been created during the last years we will present an extension of our personal studies on TDT that we utilized in our summarization of multi-file clusters. the main concept we used to become aware of documents in TDT is also used to rank sentences for our summarizer.

Jaime Carbonell et.al proposed a method [2] our entry within the reputable TDT assessment, CIDR, uses changed TF_ IDF to produce clusters of information articles at the equal event (‘TF’ suggests how many times a phrase appears in a record whilst IDF measures how many of all documents in a collection include a given word). An incoming report is grouped into an current cluster, if the TF* IDF of the new report is close to the centroid of the cluster. A centroid is a set of words that statistically constitute a cluster of files. From a TDT device, an event cluster may be produced. An occasion cluster includes chronologically ordered news articles from more than one sources. those articles describe an event as it develops over the years. In our experiments, occasion clusters variety from 2 to ten documents. it's miles from those documents that summaries may be produced.

Jade Goldstein et.al proposed a method [3] evolved a new technique for multi-record summarization, known as centroid-based summarization (CBS). CBS makes use of the centroids of the clusters produced with the aid of CIDR to become aware of sentences central to the subject of the whole cluster. we've applied CBSin MEAD, our publicly to be had multi-record summarizer. A key characteristic of MEAD is its use of cluster centroids, which include phrases which might be important no longer best to one article in a cluster, however to all of the articles. at the same time as TF*IDF has been applied in single-document summarizer, ours is the first try to increase that concept to multi-file summarization.

MEAD is substantially unique from previous paintings on multi-report summarization which use techniques including graph matching, maximal marginal relevance, or language era. finally, evaluation of multi-record summaries is a hard trouble. presently, there may be no extensively normal evaluation scheme. We advise a application-primarily based evaluation scheme, which may be used to evaluate both unmarried-record and multi-file summaries. the principle contributions of this paper are: the usage of cluster-based totally relative software (CBRU) and cross-sentence informational subsumption (CSIS) for evaluation of unmarried and multi-file summaries, the improvement of a centroid-based multi-report summarizer, user research that help our findings, and an assessment of MEAD.

### 2. Informational content of sentences

### 2.1. Cluster-based relative utility (CBRU)

Hongyan Jing et.al proposed a method [5] Cluster-based totally relative application (CBRU, or relative software, RU in quick) refers back to the diploma of relevance (from zero to ten) of a specific sentence to the overall topic of the entire cluster. A application of 0 approach that the sentence isn't always applicable to the cluster and a 10 marks an important sentence. assessment structures might be constructed based on RU and accordingly provide a greater quantifiable degree of sentences

### 2.2. pass-sentence informational subsumption (CSIS)

A associated notion to RU is go-sentence informational subsumption (CSIS, or subsumption). CSIS reflects that sure sentences repeat a number of the information present in different sentences and may, consequently, be omitted at some point of summarization. If the records content of sentence a (denoted as i(a) is contained inside sentence b, then a becomes informationally redundant and the content of b is stated to subsume that of a:
$i(a) \cap i(b)$

### 2.3. Equivalence lessons of sentences

Inderjeet Mani and Eric Bloedorn et.al proposed a method [7] Sentences subsuming each other are said to belong to the same equivalence elegance. An equivalence class can also include more than two sentences within the same or specific articles. within the following instance, even though sentences (3) and (four) are not precise paraphrases of each other, they can be substituted for every other with out important loss of facts and consequently belong to the identical equivalence class. inside the consumer have a look at) we can test the manner people understand CSIS and equivalence magnificence.

### 3.MEAD extraction set of rules

MEAD decides which sentences to include within the extract by means of ranking them in step with a hard and fast of parameters. The input to MEAD is a cluster of articles (e.g., extracted by using CIDR), segmented into sentences and a value for the compression fee R. The output is a series of n * r sentences from the unique files presented in the equal order as the input files. for instance, if the cluster contains a complete of 50 sentences (n = 50) and the fee of R is 20%, the output of MEAD will comprise 10 sentences. Sentences appear inside the extract inside the identical order because the authentic documents are ordered chronologically. We benefit right here from the time stamps associated with each document.We used three capabilities to compute the salience of a sentence: Centroid fee, Positional cost, and first-sentence overlap. these are described in full below.

### 3.1 Centroid value

The centroid cost Ci for sentence Si is computed as the sum of the centroid values Cw,i of all phrases inside the sentence. as an example, the sentence ‘‘President Clinton met with Vernon Jordan in January’’ could get a score of 243.34 that is the sum of the character centroid values of the words (Clinton=36.39; Vernon=47.54; Jordan=75.81; January=83.60)
$C_i=\sum_w C_{w,i}$

### 3.2 Positional value

The positional fee is computed as follows: the primary sentence in a record gets the identical score Cmax as the best-ranking sentence in the document in keeping with the centroid fee. The score for all sentences within a file is computed in keeping with the following formulation:
$P_i=\frac{n-i+1}{n}*C_{max}$

### 3.3 First-sentence overlap

The overlap price is computed because the inner product of the sentence vectors for the modern-day sentence i and the primary sentence of the file. The sentence vectors are the n-dimensional representations of the phrases in every sentence, whereby the price at role i of a sentence vector shows the quantity of occurrences of that phrase in the sentence.

### 3.4 Combining the 3 parameters

We examined numerous sentence weighting strategies using linear combinations of 3 parameters: phrases in centroid (C), sentence function (P), and phrases in title or first sentence (F ). The score of a sentence is the weighted sum of the rankings for all words in it. due to the fact we have not included getting to know the weights robotically, on this paper we used an same weight for all three parameters. thus, we use the subsequent score values to approximate cluster-based totally relative utility, wherein i is the sentence number within the cluster.Redundancy-based totally set of rules

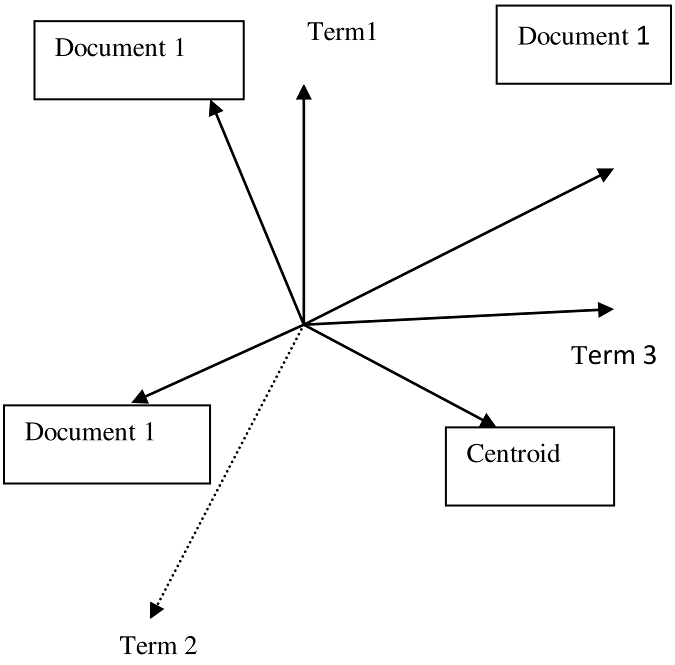


Fig 1: Conceptual centroid representation

### 4. Techniques for evaluating summaries

Inderjeet Mani and Mark Maybury et.al proposed a method [8] Summarization assessment methods can be divided into two classes: intrinsic and extrinsic. Intrinsic assessment measures the quality of summaries immediately. Extrinsic methods degree how nicely the summaries help in appearing a particular venture. Extrinsic assessment, also called assignment-based assessment, has acquired extra interest these days at the record information conference

### 4.1. single-record summaries

Kathleen McKeown et.al proposed a method [9] two techniques commonly used to degree interjudge settlement and to assess extracts are (A) precision and bear in mind, and (B) percent agreement. In both cases, an robotically generated summary is in comparison towards an ‘‘best’’ summary. To assemble the suitable summary, a set of human topics are requested to extract sentences. Then, the sentences selected by means of a majority of humans are covered in the proper summary. The precision and recall suggest the overlap between theperfect summary and the automatic precis. It suppose need to determine which of the two structures that selected precis sentences at a compression rate of 20% is higher.

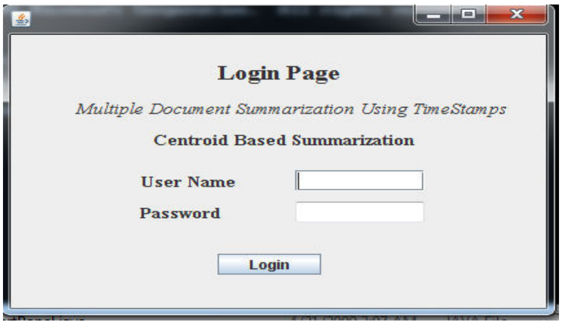
The use of precision and recall shows that the performance of systems 1 and 2 is 50% and 0%,respectively. gadget 2 seems to should worst possible overall performance, considering precision and don't forget deal with sentences S3 to S10 as equally awful. the use of percentage settlement, the overall performance is eighty% and 60%, respectively. however, percent agreement is fairly dependent on the compression charge. Dragomir et.al proposed a method [10]

### 4.2 Multiple document summaries

As opposed to P&R or percent settlement, you may measure the insurance of the suitable summary application. In the instance in desk five, the use of each evaluation methods A and B, gadget 1 achieves 50%, while system 2 achieves 0%. If we examine relative software, device 1 fits 18 out of nineteen utility points in the appropriate summary and system 2 gets 15 out of 19. In this example, the overall performance of system 2 isn't as low as while the use of strategies A and B. suggest to model each inter judge settlement and device evaluation as actual-valued vector matching and now not as boolean (methods A and B). via giving credit for ‘‘much less than best’’ sentences and distinguishing the diploma of importance between sentences, the software-based scheme is a greater herbal model to evaluate summaries. Dragomir et.al proposed a method [11]

### 5. Experimental Results

.



Login Page

*Multiple Document Summarization Using TimeStamps*

Centroid Based Summarization

.

Fig: 2 Document Images to Login Page

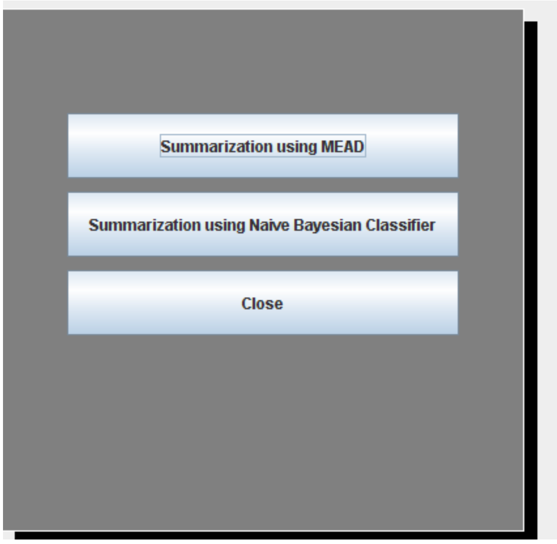


Fig: 3 Document image to Single Processing Page

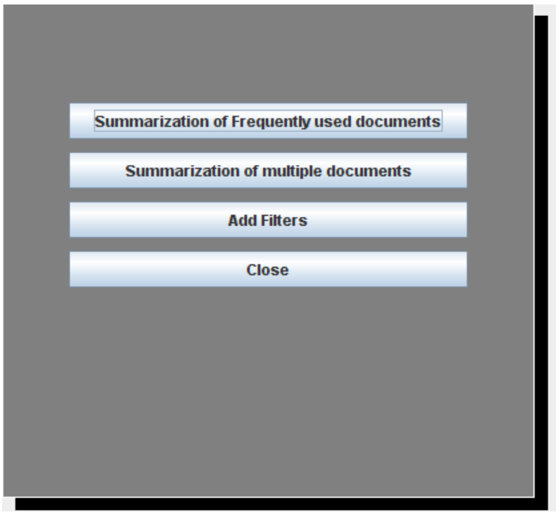
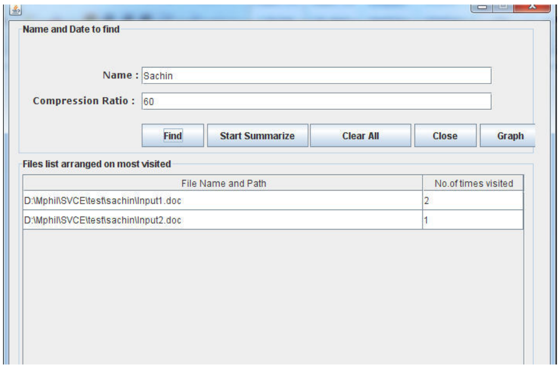

Summarization of Frequently used documents
Summarization of multiple documents
Add Filters
Close

Fig: 4 Document image to Multiple Processing Page



Name and Date to find

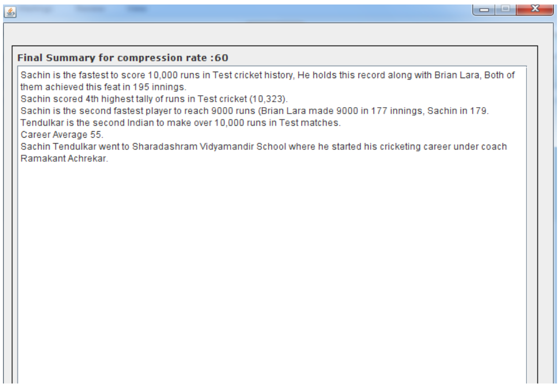
Name : Sachin

Compression Ratio : 60

Files list arranged on most visited

| File Name and Path | No.of times visited |
|---|---|
| D:\Mphil\SVCE\test\sachin\Input1.doc | 2 |
| D:\Mphil\SVCE\test\sachin\Input2.doc | 1 |

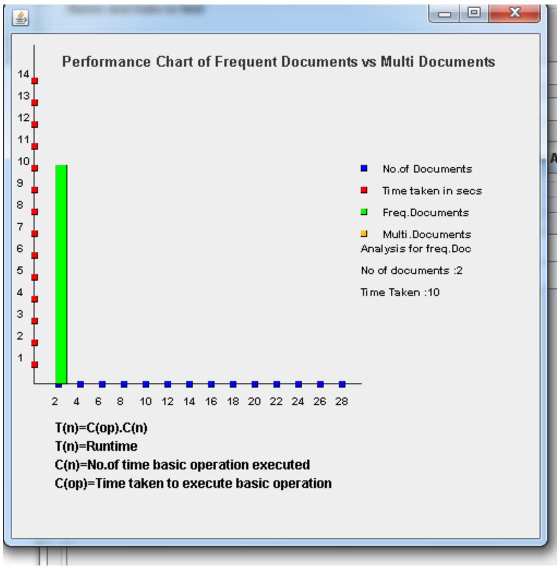
Fig: 5 Document Images to be upload file



Final Summary for compression rate :60

Sachin is the fastest to score 10,000 runs in Test cricket history, He holds this record along with Brian Lara, Both of them achieved this feat in 195 innings.
Sachin scored 4th highest tally of runs in Test cricket (10,323).
Sachin is the second fastest player to reach 9000 runs (Brian Lara made 9000 in 177 innings, Sachin in 179.
Tendulkar is the second Indian to make over 10,000 runs in Test matches.
Career Average 55.
Sachin Tendulkar went to Sharadashram Vidyamandir School where he started his cricketing career under coach Ramakant Achrekar.

**Fig:** 6 Document viewed as a output page



Analysis for freq.Doc

No of documents :2

Time Taken :10

T(n)=C(op).C(n)
T(n)=Runtime
C(n)=No.of time basic operation executed
C(op)=Time taken to execute basic operation

**Fig:** 7 Document viewed as a Performance chart

## 6. CONCLUSION

From the experiment, it is found out that if the proposed system was evaluated using ROUGE-1, the values of the average-F score value lie in an interval of 0.24 to 0.37, the highest value was achieved when the average-R = 0.54 and average-P = 0.29 if it uses the dataset of cluster D134H. Moreover, if the proposed system uses the dataset of D134H, it also achieved the highest average-F score value if it was evaluated using ROUGE-S and ROUGE-SU. However, if the proposed system evaluated using Thus, it can be concluded that the proposed system outperformed MEAD system if it was evaluated using the dataset of cluster D133C and D134H and evaluated using ROUGE-1, ROUGE-S and ROUGE SU for cluster D133C and ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L and ROUGE-W for cluster D134H. This shows that the proposed system captures the important words in the extracted summary and it generates longer sentences as longer sentence contains more material that would match the one in the reference summaries.

## References

[1]James Allan, Jaime Carbonell, George Doddington,Jonathan Yamron, and Yiming Yang, *Topic detection and tracking pilot study: final report*, In Proceedings of the Broadcast News Understanding and Transcription Workshop, 1998.

[2] Jaime Carbonell and Jade Goldstein. *The use of MMR, diversity-based reranking for reordering documents and producing summaries*. In Proceedings of ACM-SIGIR'98, Melbourne, Australia, August 1998.

[3] Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell*, Summarizing Text Documents: Sentence Selection and Evaluation Metrics*, In Proceedings of ACM-SIGIR'99, Berkeley, CA,August 1999.

[4] Thérèse Hand. *A Proposal for Task-Based Evaluation of Text Summarization Systems*, in Mani, I., and Maybury, M., eds., Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization, Madrid, Spain, July 1997.

[5] Hongyan Jing, Regina Barzilay, Kathleen McKeown, and Michael Elhadad, *Summarization Evaluation Methods: Experiments and Analysis*, In Working Notes, AAAI Spring Symposium on Intelligent Text Summarization, Stanford, CA, April 1998.

[6] Inderjeet Mani and Eric Bloedorn*, Summarizing Similarities and Differences Among Related Documents*, Information Retrieval **1** (1-2), pages 35-67, June 1999.

[7] Inderjeet Mani, David House, Gary Klein, Lynette Hirschman, Leo Orbst, Thérèse Firmin, Michael Chrzanowski, and Beth Sundheim. *The TIPSTER SUMMAC text summarization evaluation*. Technical Report MTR98W0000138, MITRE, McLean, Virginia, October 1998.

[8] Inderjeet Mani and Mark Maybury. *Advances in Automatic Text Summarization*. MIT Press, 1999.

[9] Kathleen McKeown, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin, *Towards Multidocument Summarization by Reformulation: Progress and Prospects*, In Proceedings of AAAI'99, Orlando, FL, July 1999.

[10] Dragomir R. Radev and Kathleen McKeown. *Generating natural language summaries from multiple on-line sources*. Computational Linguistics, **24** (3), pages 469-500, September 1998.

[11] Dragomir R. Radev, Vasileios Hatzivassiloglou, and Kathleen R. McKeown. *A description of the CIDR system as used for TDT-2*. In DARPA Broadcast News Workshop, Herndon, VA, February 1999.

**C.V.SHEEBA**, DOB 28-06-1978 obtained B.Sc. (Computer Science)- 1995 to 1998,College – S.T.Hindu College, Nagercoil , (Manonmanium Sundaranar University, Tirunelveli) M.C.A -2007 – 2009 D.D.E – Tamil Nadu Open University, Chennai, B.Ed. (Computer Science) - 2009-2010 D.D.E – Indira Gandhi National Open University, New Delhi , 17 years Working as Computer Instructor in GHSS Rajakkamangalam,Kanyakumari District -(2001-2006), Govt.Hr.Sec.School, Sundapattivilai, Kanyakumari District from 2008 to till date. Web Auction System - for PG (Front end Tool – JSP,Back End-Microsoft Access).

**Dr.T.Meyyappan** ,M.Sc,M.Tech.,M.B.A., M.Phil,Ph.D. currently, Professor, Department of Computer Science, AlagappUniversity, Karaikudi, TamilNadu. He has organized conferences, workshops at national and international levels. He has published 90 numbers of research papers in National and International journals and conferences. He has developed Software packages for Examination, Admission Processing and official Website of Alagappa University. As a Co-Investigator, he has completed Rs.1 crore project on smart and secure environment funded by NTRO, New Delhi. As principal Investigator, he has completed Rs. 4 lakhs project on Privacy Preserving Data Mining funded by U.G.C. New Delhi. He has been honoured with Best Citizens of India Award 2012 research areas include Operational Research, Digital Image Processing, Fault Tolerant computing, Network security and Data Mining.

**SM. Thamarai** currently, guest lecturer, Alagappa Government Arts College, Karaikudi, received her Diploma in Electronics and Coomunication Engineering, Department of Technical Education, Tamilnadu in 1989 and her B.C.A. M.Sc. (University First Rank holder and Gold medalist), M.Phil. (First Rank holder) degrees in Computer Science(1998-2005) from Alagappa University. She has published 27 research papers in International, National Journals and conferences. She received her Ph.D. degree in Computer Science in 2014. Her current research interests include Operational Research and Fault Tolerant Computing.