AMDS: Sentence Extraction Based Proficient Framework For Multi-Document Summarization

C.Balasubramanian[1] ***K.G.Srinivasagan***[2] ***K.Duraiswamy***[3]

1.Professor/CSE,P.S.R.Rengasamy College of Engineering For Women,Sivakasi,India -626140
2.Professor/CSE,National Engineering College,Kovilpatti,India
3.Dean/CSE,K.S.R.College of Technology,India Email:rc.balasubramanian@gmail.com

Abstract: - Rapid improvement of electronic documents in World Wide Web has made overload to the users in accessing the information. Therefore, abstracting the primary content from numerous documents related to same topic is highly essential. Summarization of multiple documents helps in valuable decision-making in less time. This paper proposed a framework named Adept Multi-Document Summarization (AMDS) for efficient summarization of document, which achieves the aforementioned requirement. Here, the documents are preprocessed initially to remove the information that is less important. Summary of each preprocessed document is obtained through the sentence extraction process. Single document summarization is carried out based on graph model. A ranking method named Ingenious Ranking (IR) is proposed to rank and order the extracted single document summaries. It ranks the sentences in the generated summaries of each document and incorporates the individual summaries to generate a concise summary. Empirical results presented in this paper demonstrate the efficiency of the proposed AMDS framework.

Keywords: - Sentence extraction, multi-document summarization, sentence ingenious ranking, similarity measure

I. INTRODUCTION

The number of online document is enormously increased during recent years due to the outgrowth of large-capacity, low price storage dives, and with the Internet. This growth led to information overload, which implicitly made the process of searching information. Since many documents retrieved for a user query may have the same information, while differing in certain aspects. This situation is overcome by various techniques such as Information Retrieval (InR), Question-Answering (QA), Information Extraction (IE) and automatic summarization. Among all the aforementioned techniques, automatic summarization has been focused by a lot of researchers.

Summarization is an art of abstracting the key concept from single or multiple documents. Individual document summarizes would help, but they are likely to provide very similar contents. Therefore, multi-document summarization is used to provide a summary of either summary of single document summaries that are generated earlier or for set of documents. Summary of documents provide an overview, which are easier and helpful for the user to browse and to take decisions. Summarization process can be differed based on whether they are abstracting or extracting. However, both methods have the following two steps.

- Indentify what is important or salient in the given document
- Determine the way to reduce it Apart from the similarity, multi-document summarization differs from the single document summarization in the following four different ways.
- Degree of information redundancy presented inside a set of related documents is much greater on comparing with the degree of redundancy in a single document
- The news report documents may have information of temporal dimension regarding an unfolding event. Here, there exists a possibility of overriding the later information with the former information leads to incomplete sentences/accounts
- Third difference is based on compression ratio, for multi-document summarization the compression ratio is higher than the single document summarization
- Multi-document summarization faces a greater challenge of co-reference problem during summarization than single document summarization

The multi-document summarization is useful in dealing with dissimilar documents and to access the landscape information from the set of documents. This paper discusses a method that is related to the multi-document summarization in order to generate an extractive summary of set documents that are topically related to each other. The process of extractive summarization is portrayed in Figure 1.

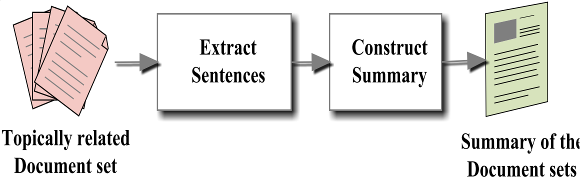


Figure 1. Process of Extractive Summary

From the Figure 1 it is easily understood that the process of extractive summary has two steps. (1) Sentence Extraction: Sentences are extracted from a set of documents that contain similar content, and (2) Summary construction: Based on the extracted sentence, summary is constructed.

A framework named AMDS is proposed for constructing the summary from a set of document, which addresses the following three issues: (1) Detect the similarities among the given set of documents, (2) removes the redundancy, and (3) Ensures the coherence summary. Figure 2 represents the steps in the proposed method.

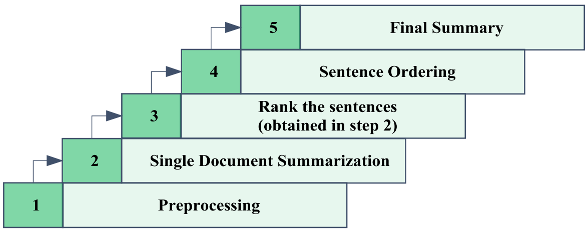


Figure 2. Steps in AMDS framework

Initially, the given documents are preprocessed to remove the figures and other noises in the document. In addition, the sentences in the documents are separated. This process helps to accelerate the summarization process. Each document that participates in the summarization process is preprocessed. The preprocessed documents then summarized individually using a graph model, which extracts the important sentences. These are grouped together to form a summary of corresponding documents. These summaries are given as input to the IR ranking technique, which ranks the sentences and order it to produce a concise summary. The concise summary represents the summary of the topically related documents that are given as input.

Rest of this paper is structured as below: In section 2, works that are related to the summary extraction is discussed briefly. The detailed explanation of the proposed AMDS framework is given in section 3. Empirical study is reported in the section 4 to prove the efficiency and accuracy of the proposed framework. Finally, section 5 concludes this paper along with the points providing directions for future work.

II. PREVIOUS WORKS

This section summarizes various research works that were carried over the past decade. Brief explanations about twenty research works were given in this section.

A system was proposed in [1] to obtain a summary for the articles of economic matter. The system designed in this paper supports only the documents that were in Turkish text. Here, the given text was converted to HTML document to give a formal structure. The system design depends on a study that concentrates on the statistical analysis of words, sentences, and paragraphs in the article according to the specific weight that were predefined. Though the predefined weight finds the skeleton of the summary, the summarization sentences were chosen to emphasize the semantic integrity. The summary of the articles depended on the summarization ration, which is provided by the user. The experimental results depict that the title and the keywords played the major role in summarization.

Following this work, a spreading activation based technique for document summarization was proposed in [11]. A network was constructed based on the important entities among the documents along with the relation between them. The method was designed for multilingual since the documents were annotated syntactically and semantically with the Global Document Annotation (GDA) tags. An example of GDA tagged sentence was shown in the Figure 3. The proposed model in [11] has the following capabilities.

- Extract the suitable documents that were related to a specific hostage incident
- An entity-relation graph was constructed, which depicts the relation between the important entities in the documents
- Important sentences were extracted and pruned
- Cross-document references were used to substitute the expressions

```
<s>
      <n> Time </n>
      <v> files</v>
      <ap>
            Like
            <n>an arrow</n>
      </ap>
</s>
```

Figure 3. GDA taggedsentence

A survey on multi-document summarization was reported in [14]. Hundred datasets were generated. For all the 100 dataset annotators have prepared the data that were as follows.

- Table style summary
- Axis
- Sentence extraction type summarization
- Free style summarization

Here, they have concentrated mainly on the axis based and the table based summaries. These styles were experimented, and their results showed that the axis based summarization works better only when the axis were determined correctly. The results depict that table style summary was very useful for the large percentage of dataset. A model was proposed in [13], which classifies each website's Top Stories whose specific news category was unknown using a supervised learning.

The papers [1, 11] discussed does not concentrate on the sentence ordering. Sentence ordering is one of the techniques to extract the most important information from the documents. It is also a successful technique, but there exists some difficulties in implementing in it. A profound technique for sentence ordering was proposed in [8]. This technique well suited for text-to-text generation since it works on the surface-level rather than on the logical form. Here, unsupervised probabilistic model was proposed to determine the text structuring, which learns the constraints in ordering the sentences from multiple documents as well as the sequence of the features that likely to co-occur.

Another technique was proposed in [12], which orders the sentences of newspaper documents coherently. Conventional techniques have used the chronological ordering for arranging the sentences that were extracted. In [12] they enhanced this method through incorporating the chronological sentence ordering and topical segmentation. This technique used the sentence's precedence relation. In addition to that in [3] a bottom-up approach was used to arrange the extracted sentence from multiple documents. The relationships between the sentences were found using the chronology, precedence, topical-closeness, and succession. The research work presented in [4] uses probabilistic parameter along with the parameters that were used in [12] to detect the relation among the sentences. A study for ordering based on corpus methodology can be found in [2]. Two different ordering strategies namely the chronological ordering and majority ordering were analyzed. The analysis shows that the majority ordering performs better only when the documents were topically related. On the other hand, the chronological ordering provides only the acceptable performance for summarization of news article. This paper aims to incorporate the best part of both the majority and chronological ordering.

The papers [8, 12] show the importance of sentence ordering. Sentence ordering or alignment was carried out using sentence similarity measures. Different sentence similarity detecting techniques were developed to aid sentence alignment. In paper [16], various sentence similarity approaches were discussed and proposed a new customized technique by combining the best parts of different techniques. The composite technique has two passes to align the extracted sentences. In the first pass, customizable number of Weighted Sentence Length (WSL) was used to generate a tentative alignment. With the obtained tentative alignment, second pass takes the following concepts and realigns the sentences.

- Word correspondence
- Numeric, Phonetic, and Cognate (NPC) matching
- Common Word Count (CWC)
- Hypernym and Synonym Intersection (HNI and SNI)

A study was carried out in [18] with an aim to differentiate between within-document and cross document relationship among the sentences. For the topic-focused summarization, this algorithm adapts the graph-ranking. Main contribution of this paper was in two-folds. (1) Graph based ranking algorithm was used to determine the sentence relationship and relative importance was explored. (2) Sentence relevancy was integrated for topic-focused multi-document summarization. Following the work of [18], [6] also introduces a graph based summarization named LDA. This technique enables the automatic findings of semantic topics from a set of documents that were to be summarized. These topics were used to build the bipartite graph, which denotes the document. Salience score for topics and sentences were computed simultaneously. This improves the scoring process. Another topic extraction method was proposed in [9] and in [19].

Clustering based summarization system named SIMBA was presented in [15]. Double clustering process namely (1) similarity clustering, and (2) keyword clustering were carried to summarize multiple documents. Ranking based summarization process was discussed in [20]. It proposed a model that enhances the manifold-ranking method. This is carried through mutual reinforcement between the theme clusters the sentences. The above discussed techniques failed to analyze semantically. Therefore, to address this need, [7] introduced a method that uses the tag cluster on the flicker.

In [17] a system was designed to summarize a single document that depends on the word frequency and local topic identification. This paper addresses the problem such as redundancy, structure and coherent those were generated using the automatic summarization method. These issues were arisen in automatic summarization since it uses the physical feature. So, the proposed system in [17] used logical structure features for successful multi-document summarization. The sentence similarity is used to cluster the document into the local topics. Following the similarity computation, word frequency was manipulated to extract the sentences from the topically related documents. With this approach, redundancy is eliminated. Following [17], paper [10] also summarizes the document using the frequency of terms in the document. Another approach for reducing the redundancy was proposed in [5] named Optimal Combinatorial Covering Algorithm for Multi-document Summarization (OCCAMS) for scoring and extracting sentences. Here, the document word's latent distribution was studied using the Latent Semantic Analysis, Budgeted Maximal Coverage, and Fully Polynomial Time Approximation Scheme (FPTAS) for knapsack to select the sentences. This technique also maximizes the combined weight of the covered terms.

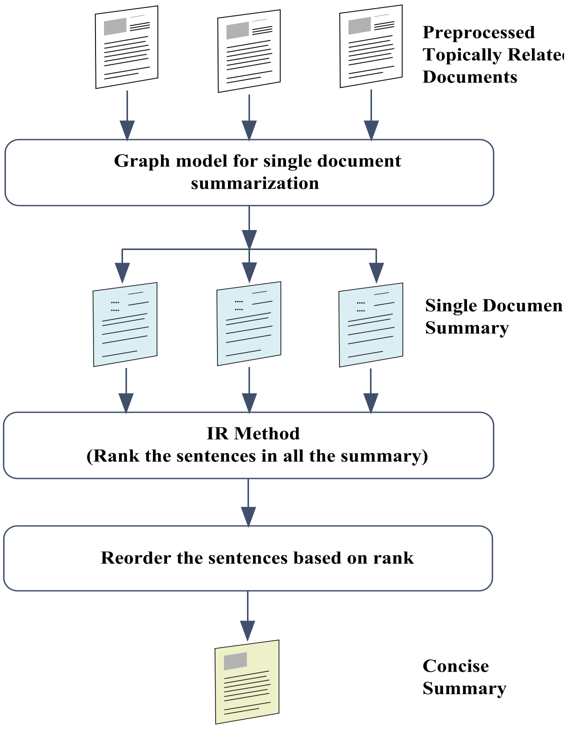

Preprocessed
Topically Relate
Documents
Graph model for single document summarization
Single Documen
Summary
IR Method
(Rank the sentences in all the summary)
Reorder the sentences based on rank
Concise
Summary

Figure 4. Framework overview

III. PROPOSED METHOD

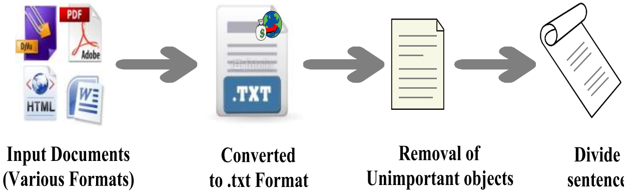
Figure 4 illustrates the overview of the proposed framework for multi-document summarization. Individual documents are given as input to the framework, which are topically related to each other. On obtaining the documents, the framework preprocesses and analyzes each document and generates summaries separately. These summaries are further processed by the IR technique to obtain a concise summary of all the documents. The summarization process can be decomposed into the following phases.

(1) Preprocessing
(2) Single document summarization
(3) IR ranking
(4) Generation of Concise summary

The detailed process is explained in the following subsequent subsections.

A. *Preprocessing*

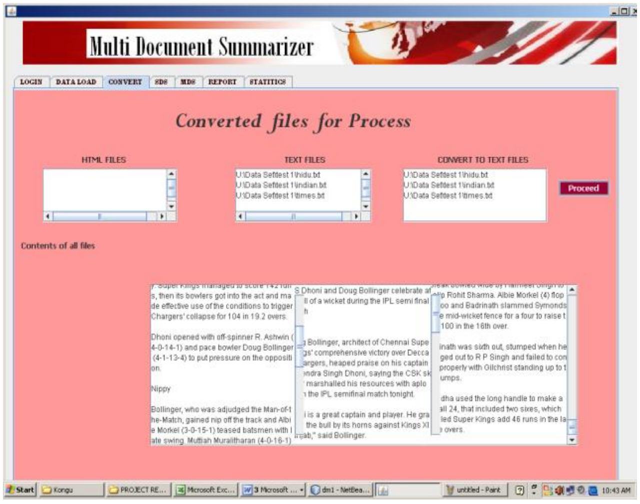
Initially, the documents are preprocessed to remove the unnecessary details. The process of preprocessing is represented in Figure 5. First step in preprocessing is to convert the given document into the text format. The input document can be of any type like pdf, html, etc. The information that is not important is removed from the documents such as headings, subheadings, tables, figures. This process reduces the overall time required for sentence extraction process.



DjVu
PDF
Adobe
HTML
.TXT
Input Documents (Various Formats)
Converted to .txt Format
Removal of Unimportant objects
Divide sentence

Figure 5. Steps in Preprocessing

After removing the unimportant objects, the document is further divided into sentences. The sentences are divided using the boundaries such as full stop/period, semicolon, exclamation mark, and question mark. Usually, the full stop is considered as the sentence boundary since it is highly ambiguous. In this work, along with the full stop following three ambiguities are also considered while dividing the sentences. (1) Periods may be present in the non standard words like email ids, website urls, etc., (2) Each sentence in the given document starts with the uppercase, and (3) Either lower case or upper case can be used for the titles and subtitles of a document. Figure 6, demonstrates the sample of the preprocessing process of a document.



TEXT FILES:
U:\Data Seftest 1\hidu.txt
U:\Data Seftest 1\indian.txt
U:\Data Seftest 1\times.txt

CONVERT TO TEXT FILES:
U:\Data Seftest 1\hidu.txt
U:\Data Seftest 1\indian.txt
U:\Data Seftest 1\times.txt

Contents of all files

s, then its bowlers got into the act and ma
de effective use of the conditions to trigger
Chargers' collapse for 104 in 19.2 overs.

Dhoni opened with off-spinner R. Ashwin (
4-0-14-1) and pace bowler Doug Bollinger
(4-1-13-4) to put pressure on the oppositi
on.

Nippy

Bollinger, who was adjudged the Man-of-t
he-Match, gained nip off the track and Albi
e Morkel (3-0-15-1) teased batsmen with l
ate swing. Muttiah Muralitharan (4-0-16-1)

S Dhoni and Doug Bollinger celebrate at
ll of a wicket during the IPL semi final
h

Bollinger, architect of Chennai Supe
gs' comprehensive victory over Decca
argers, heaped praise on his captain
endra Singh Dhoni, saying the CSK sk
marshalled his resources with aplo
the IPL semifinal match tonight.

is a great captain and player. He gra
the bull by its horns against Kings XI
njab," said Bollinger.

p Rohit Sharma. Albie Morkel (4) flop
oo and Badrinath slammed Symonds
e mid-wicket fence for a four to raise t
100 in the 16th over.

inath was sixth out, stumped when he
ged out to R P Singh and failed to con
properly with Gilchrist standing up to t
umps.

dha used the long handle to make a
all 24, that included two sixes, which
led Super Kings add 46 runs in the la
overs.

Figure 6. Preprocessing Sample

B. *Single Document Summarization*

The process of sentence extraction starts by constructing a graph, $G = (V, E)$. The graph based model paves a way for determining the importance of a vertex, i.e. sentence within a graph. The structure of the graph is used to determine the importance of the sentence. This paper has used the weighted graph of determining the strength of dependency between sentences in a document. The sentence in the document is modeled using the weighted word frequency vector $\overrightarrow{X_i}$. Let $W = \{w_1, w_2, w_3, \dots\dots, w_4\}$ be the word set of the given document.

The sentence vertex can be denoted as $\overrightarrow{X_i} = [X_1^i, X_2^i, X_3^i, \dots\dots, X_n^i]^W$. Here, the $X_j^i = 1$ if the word w_j present in the sentence X_i otherwise $X_j^i = 0$. The document is modeled using the graph $G = (V, E)$ where V and E represent the vertex and the edges between the vertices respectively. Sentences in the document are modeled as vertex in the graph. The connection between the two sentences is established based on the similarity measure. Similarity is the measure of content overlap. Cosine similarity is used by this paper to determine the overlap of content between sentences, which is manipulated using the equation (1).

$$S(X_i, X_j) = \frac{\overrightarrow{X_i} \cdot \overrightarrow{X_j}}{|\overrightarrow{X_i}| \times |\overrightarrow{X_j}|} \tag{1}$$

Where $\overrightarrow{X_i}$ and $\overrightarrow{X_j}$ are the vectors of the sentences X_i and X_j. The important sentences for extraction from the document are determined through the edge weight EW_{ij}, which is equal to the similarity weight. Similarly, the $EW_{ii} = 0$, which represents the weight passes to itself. The weight denotes that common words the two sentences have. This relation shows the process of recommendation: a sentence that focuses on certain concepts in the document gives a reader a suggestion to refer to another sentence in the document, which addresses the same concept. A sample graph model for 10 sentences in a document is represented in Figure 7.

From the Figure 7 it is clear that the similarity between X_1 and X_3, X_5, X_7, X_9 are greater than zero and with the remaining sentences, the similarity value is zero. The darker lines represent that the value of $S(X_1, X_3)$ and $S(X_1, X_5)$ is greater than the other similarity between other sentences. Based on the graph, sentences are ranked using the page rank algorithm employed in [21]. After ranking the sentences of the documents are ordered, and the top-ranked sentences are extracted for inclusion in the summary. Figure 8, represents the summarization of single document.

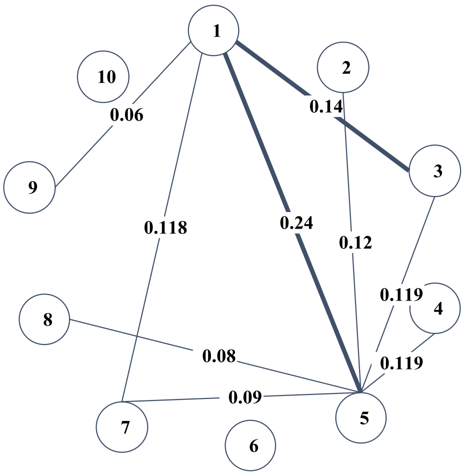
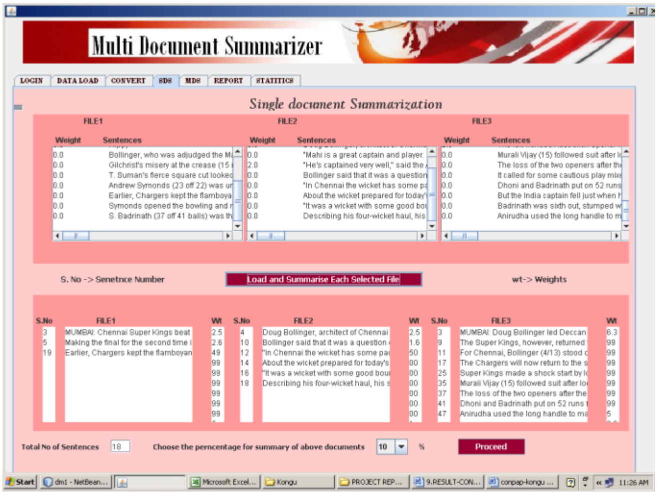


Figure 7. An example of Graph model for 10 sentences

C. IR ranking and Ordering (Multi-document Summarization)

1) Similarity Measure

Summarizes obtained in the previous subsection is given as input to the multi-document summarizer. The sentences from all the documents are grouped together, and similarities among those sentences are measured. Similar to the above section this section also uses the same graph model to determine the similarity among the sentences. This paper has introduced a similarity threshold, δ. Two sentences are connected only if they are similar to with respect δ. The value of δ is set empirically to 0.3 in implementation.



Single document Summarization

FILE1		FILE2		FILE3	
Weight	Sentences	Weight	Sentences	Weight	Sentences
0.0	Bollinger, who was adjudged the M	0.0	"Mahi is a great captain and player.	0.0	Murali Vijay (15) followed suit after l
0.0	Gilchrist's misery at the crease (15	2.0	"He's captained very well," said the	0.0	The loss of the two openers after th
0.0	T. Suman's fierce square cut looked	0.0	Bollinger said that it was a question	0.0	It called for some cautious play mix
0.0	Andrew Symonds (23 off 22) was ur	0.0	"In Chennai the wicket has some pa	0.0	Dhoni and Badrinath put on 52 runs
0.0	Earlier, Chargers kept the flamboya	0.0	About the wicket prepared for today'	0.0	But the India captain fell just when h
0.0	Symonds opened the bowling and r	0.0	"It was a wicket with some good bo	0.0	Badrinath was sixth out, stumped w
0.0	S. Badrinath (37 off 41 balls) was th	0.0	Describing his four-wicket haul, his	0.0	Anirudha used the long handle to m

S. No -> Senetnce Number

wt-> Weights

S.No	FILE1	Wt	S.No	FILE2	Wt	S.No	FILE3	Wt
3	MUMBAI: Chennai Super Kings beat	2.5	4	Doug Bollinger, architect of Chennai	2.5	3	MUMBAI: Doug Bollinger led Deccan	6.3
5	Making the final for the second time i	2.6	10	Bollinger said that it was a question	1.6	9	The Super Kings, however, returned	99
19	Earlier, Chargers kept the flamboyan	49	12	"In Chennai the wicket has some pa	50	11	For Chennai, Bollinger (4/13) stood	99
		99	14	About the wicket prepared for today's	00	17	The Chargers will now return to the s	99
		99	16	"It was a wicket with some good bou	00	25	Super Kings made a shock start by l	99
		99	18	Describing his four-wicket haul, his s	00	35	Murali Vijay (15) followed suit after lo	99
		99			00	37	The loss of the two openers after the	99
		99			00	41	Dhoni and Badrinath put on 52 runs	99
		99			00	47	Anirudha used the long handle to ma	5

Total No of Sentences 18 Choose the perncentage for summary of above documents 10 %

Figure 8. Single document summarization

2) Feature Mining

To obtain the values of sentence-specific features for each sentence, a feature profile is created. Authors have used the following three different surface-level features.

a) *First-sentence overlap:* Inner-product similarity of a sentence and the first sentence in the document is measured using this feature determination.

b) *Centroid:* The relation between the centroid of the given summaries and the sentence are determined using this feature measure.

c) *Position:* The beginning of the document contains the important sentences. This feature is measured as inversely proportional to the location of a sentence from the beginning.

The score obtained for these features act as local information for every sentence. This local information of each summary is incorporated in the IR ranking, which helps to deduce global information about the sentences. The feature score is manipulated and normalized between 0 and 1.

3) IR Ranking

The similarity determined using the graph model and the profile feature act as the input to this ranking method. This ranking model has three phases that are described in the following subsections.

a) Initialization:

The graph model is constructed for the sentences present in all the summaries as in the section 3.2. This model is converted into an adjacency matrix, A. The row and column of A has an entry for each sentence and the entry value, M_{ij} determined using the equation (2).

$$M_{ij} = M_{ji} = \begin{cases} 0, & if\ i = j \\ S(X_i, X_j), & if\ i \neq j \end{cases} \qquad (2)$$

Here, $S(X_i, X_j)$ is the similarity between the pair of sentences X_i and X_j, which is measured from the equation (1). The similarity values are always greater than the similarity threshold δ.

b) Inference:

Each vertex in the graph has the important level. The IR technique recursively updates the importance of the vertex until that is stopped by the user. The iteration can be defined mathematically as linear algebra. Let Z be the n-dimensional vector that captures the importance of the vertex in the graph. The importance of the vertex at x^th^ iteration can be determined using the equation (3).

$$Z(x) = Z(0) + NZ(x-1) \qquad (3)$$

In equation 3, N is computed using the equation (4).

$$N = \alpha S^T \qquad (4)$$

From the equation (4), the value of α determines the propagation efficiency, which converts the computes the importance of a vertex with its neighbor vertex. The value of α is set as 0.65 in the experiment heuristically.

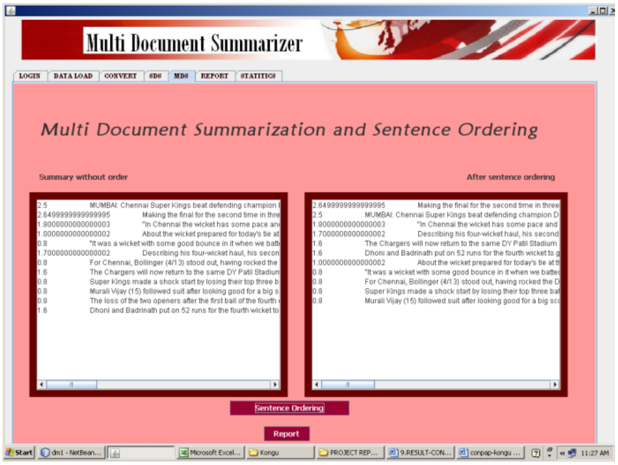
A stochastic matrix S is derived from the adjacency matrix A through the equation (5).

$$s_{ij} = \frac{M_{ij}}{\sum_k a_{ik}} \quad (5)$$

This process is continuous until a converged state is reached. In order to detect the equilibrium state, authors of this paper have introduced an equation (6).

$$\sum_i |Z_i(x) - Z_i(x-1)| \le \gamma \quad (6)$$

In the above equation $Z_i(x)$ refers to the importance of the vertex i at the step x, and the γ refers the number that has negligible value, which is set to 0.0003 by the authors for experiment. This technique stops the iteration process when the importance of every vertex in the graph is not greater than the defined number, γ. Figure 9 represents the ordering of sentences in present in the summaries generated in the section 3.2.



Multi Document Summarization and Sentence Ordering

Summary without order

```
2.5                 MUMBAI: Chennai Super Kings beat defending champion I
2.6499999999999995          Making the final for the second time in thre
1.8000000000000003          "In Chennai the wicket has some pace an
1.0000000000000002          About the wicket prepared for today's tie at
0.8                 "It was a wicket with some good bounce in it when we batt
1.7000000000000002          Describing his four-wicket haul, his secon
0.8                 For Chennai, Bollinger (4/13) stood out, having rocked the
1.6                 The Chargers will now return to the same DY Patil Stadium
0.8                 Super Kings made a shock start by losing their top three b
0.8                 Murali Vijay (15) followed suit after looking good for a big s
0.8                 The loss of the two openers after the first ball of the fourth
1.6                 Dhoni and Badrinath put on 52 runs for the fourth wicket to
```

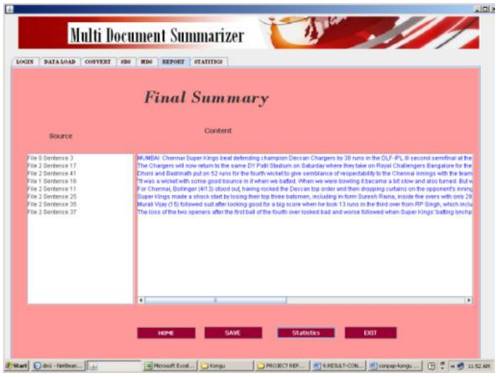
After sentence ordering

```
2.6499999999999995          Making the final for the second time in three
2.5                 MUMBAI: Chennai Super Kings beat defending champion D
1.8000000000000003          "In Chennai the wicket has some pace and
1.7000000000000002          Describing his four-wicket haul, his second
1.6                 The Chargers will now return to the same DY Patil Stadium
1.6                 Dhoni and Badrinath put on 52 runs for the fourth wicket to g
1.0000000000000002          About the wicket prepared for today's tie at t
0.8                 "It was a wicket with some good bounce in it when we batted
0.8                 For Chennai, Bollinger (4/13) stood out, having rocked the D
0.8                 Super Kings made a shock start by losing their top three bat
0.8                 Murali Vijay (15) followed suit after looking good for a big sc
```

Figure 9. Sentence Ranking and Ordering

c) *Prediction*

The state at which the IR terminates denotes the equilibrium state. The final degree of importance is given as the numeric value at this stage. The sentences are ranked depending on the importance of the entire inferred sentence. Therefore, the sentences that have more importance are extracted to generate the summary of the document. Figure 10, portrays the final summary generated from multiple documents that given as input.



Final Summary

Source	Content
File 0 Sentence 3	MUMBAI: Chennai Super Kings beat defending champion Deccan Chargers by 38 runs in the DLF-IPL II second semifinal at the
File 2 Sentence 17	The Chargers will now return to the same DY Patil Stadium on Saturday where they take on Royal Challengers Bangalore for the
File 2 Sentence 41	Dhoni and Badrinath put on 52 runs for the fourth wicket to give semblance of respectability to the Chennai innings with the team
File 1 Sentence 16	"It was a wicket with some good bounce in it when we batted. When we were bowling it became a bit slow and also turned. But w
File 2 Sentence 11	For Chennai, Bollinger (4/13) stood out, having rocked the Deccan top order and then dropping curtains on the opponent's inning
File 2 Sentence 25	Super Kings made a shock start by losing their top three batsmen, including in-form Suresh Raina, inside five overs with only 29
File 2 Sentence 35	Murali Vijay (15) followed suit after looking good for a big score when he took 13 runs in the third over from RP Singh, which inclu
File 2 Sentence 37	The loss of the two openers after the first ball of the fourth over looked bad and worse followed when Super Kings' batting lynchp

Figure 10. Concise Summary

IV. EXPERIMENTAL RESULTS

The experimental analysis is carried to determine the efficiency of the proposed framework based on the metric such as recall, precision, and accuracy. This section describes the data set used for evaluating the framework, the metric employed to measure the framework, as well as the results obtained during summarization.

A. *Data set*

The experiments where carried using 270 documents whose details are clearly given in the Table 1. 270 documents are collected from various news papers like The Hindu, Time of India, and Indian Express in the field of sports, general, business, and politics. Sports and business category has totally 75 news articles likewise, general and politics have 60 and 60 news articles respectively. These documents act as the dataset for evaluating the proposed AMDS framework.

S. No.	TYPE OF CONTENT	THE HINDU	TIMES OF INDIA	INDIAN EXPRESS
1	Sports	25	25	25
2	General	20	20	20
3	Business	25	25	25
4	Politics	25	15	20

Table 1: Data Set

B. *Results*

The AMDS framework is evaluated using the following four different metric namely, (1) accuracy, (2) Precision, (3) Recall and (4) Number of formats supported.

1) *Accuracy*

To determine the accuracy of summarized document, single document summarizes generated using the proposed scheme is given as input to various summarization tools like MEAD, SWESUM, MEAD 2, and the proposed ADMS. The accuracy of generated summary by all the summarization tools is measured, and the result is given as the graph in the Figure 11. The accuracy is measured in terms of percentage.

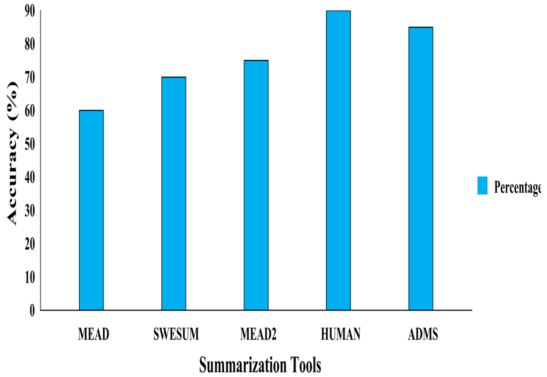


Figure 11. Accuracy

Figure 11 portrays that the proposed AMDS framework's accuracy is closely related to the human generated summary. It also represents that the proposed scheme outperforms the existing summarization tools.

2) *Precision and Recall*

The precision and recall value are determined from the equation (7) and (8) respectively where the P_s and the R_s denotes the relevant sentence and the retrieved sentences present in the document. The precision and recall results are shown as the graph in the Figure 12 and 13.

$$P = \frac{P_s \cap R_s}{R_s} \qquad (7)$$

$$R = \frac{P_s \cap R_s}{P_s} \qquad (8)$$

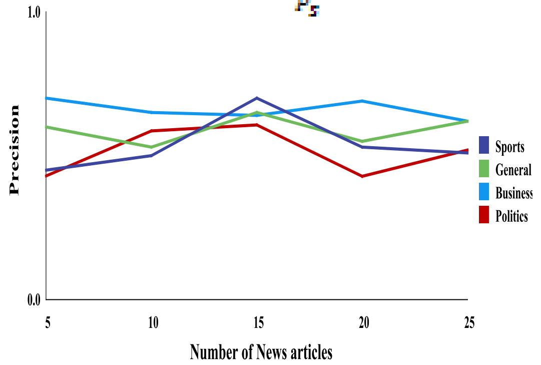


Figure 12. Precision

Figure 12 shows the precision rate summarization of ADMS framework for various categories of news articles collected from different news papers. Similarly, Figure 13 denotes the recall rate for ADMS framework.

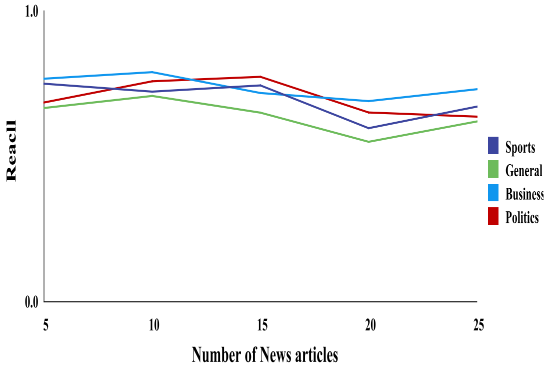


Figure 13. Recall

3) *Data format Supported*

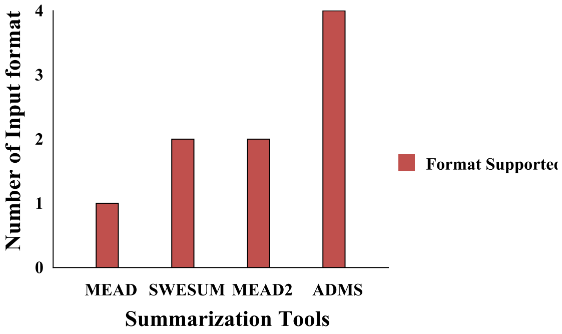


Figure 14. Number of format supported by various summarization tools

Figure 14 portrays that the proposed AMDS method supports four different file formats whereas other three existing technique supports only 1 or 2 formats. Therefore, this comparative analysis shows that the proposed scheme is versatile for various formats.

4) *Comparison with existing system*

The proposed framework is compared with an existing technique proposed in [5] based on recall, and precision. Figure 15 and 16 express the results of comparison. The comparison is carried out by varying the number of articles randomly.

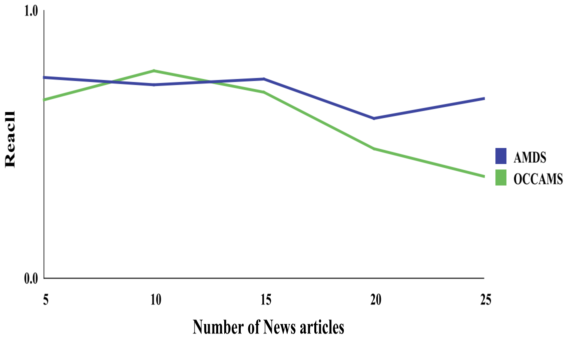


Figure 15. AMDS vs. OCCAMS Recall Comparison

Figure 15 expresses that the proposed technique has higher recall rate than the existing techniques. Similarly, Figure 16 portrays that the proposed AMDS framework has greater precision rate than the existing technique OCCAMS.

The articles that are derived for the experimental purpose are saved in various formats. The summarization tools developed in earlier days supports only limited formats. Number of supporting formats of both existing and the proposed method is shown in the figure 14.

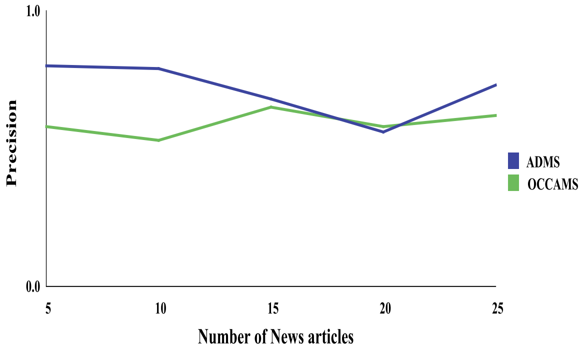


Figure 16. AMDS vs. OCCAMS Recall Comparison

V. CONCLUSION

This paper proposes a framework named AMDS for multi-document summarization. This framework initially preprocesses the document and gives it to the single document summarizer. This summarizer used graph model to compute the similarity measure between the sentences present in each document. Based on the values of similarity measure the importances of the sentences are detected. The sentences are ranked based on their importance, and the top-ranked sentences are extracted for summarization. The single document summaries are provided as input to the multi-document summarizer where the sentences from various documents are measured for its similarity, and features are extracted to generate a feature profile. The similarity measure for the sentences from various documents and the feature profile are given as input to the IR technique. This technique ranks sentences; the importance of the sentences is computed iteratively until the stable state is reached.

Experimental results show that the proposed AMDS framework summaries the given set of documents effectively than the existing techniques. The authors of this paper have planned to apply the framework to support query-oriented summarization in the future.

REFERENCES

[1] Z. Altan, "A Turkish automatic text summarization system," in Proceedings of the Artificial Intelligence and Applications, 2000.

[2] R. Barzilay and N. Elhadad, "Inferring strategies for sentence ordering in multidocument news summarization," arXiv preprint arXiv:1106.1820, 2011.

[3] D. Bollegala, N. Okazaki, and M. Ishizuka, "A bottom-up approach to sentence ordering for multi-document summarization," Information processing & management, vol. 46, pp. 89-109, 2010.

[4] D. Bollegala, N. Okazaki, and M. Ishizuka, "A preference learning approach to sentence ordering for multi-document summarization," Information Sciences, 2012.

[5] S. T. Davis, J. M. Conroy, and J. D. Schlesinger, "OCCAMS--An Optimal Combinatorial Covering Algorithm for Multi-document Summarization," in Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on, 2012, pp. 454-463.

[6] D. Gao, W. Li, Y. Ouyang, and R. Zhang, "LDA-Based Topic Formation and Topic-Sentence Reinforcement for Graph-Based Multi-document Summarization," in Information Retrieval Technology, ed: Springer, 2012, pp. 376-385.

[7] J.-U. Heu, J.-W. Jeong, I. Qasim, Y.-D. Joo, J.-M. Cho, and D.-H. Lee, "Multi-document Summarization Exploiting Semantic Analysis Based on Tag Cluster," in Advances in Multimedia Modeling. vol. 7733, S. Li, A. Saddik, M. Wang, T. Mei, N. Sebe, S. Yan, R. Hong, and C. Gurrin, Eds., ed: Springer Berlin Heidelberg, 2013, pp. 479-489.

[8] M. Lapata, "Probabilistic text structuring: experiments with sentence ordering," presented at the Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Sapporo, Japan, 2003.

[9] C. Liu, C. Zhu, T. Zhao, and D. Zheng, "Extracting main content of a topic on online social network by multi-document summarization," in Computational Intelligence and Security (CIS), 2012 Eighth International Conference on, 2012, pp. 52-55.

[10] S. Manne, Z. Shaik Mohd, and S. Sameen Fatima, "Extraction Based Automatic Text Summarization System with HMM Tagger," in Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012, 2012, pp. 421-428.

[11] U. Masao and H. Kôiti, "Multi-topic multi-document summarization," in Proceedings of the 18th conference on Computational linguistics-Volume 2, 2000, pp. 892-898.

[12] N. Okazaki, Y. Matsuo, and M. Ishizuka, "Improving chronological sentence ordering by precedence relation," presented at the Proceedings of the 20th international conference on Computational Linguistics, Geneva, Switzerland, 2004.

[13] M. W. Pope, "Automatic Classification of Online News Headlines," 2007.

[14] S. Sekine and C. Nobata, "A survey for multi-document summarization," in Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5, 2003, pp. 65-72.

[15] S. Silveira and A. Branco, "Extracting multi-document summaries with a double clustering approach," Natural Language Processing and Information Systems, pp. 70-81, 2012.

[16] A. K. Singh and S. Husain, "Exploring translation similarities for building a better sentence aligner," in Proceedings of the 3rd Indian International Conference on Artificial Intelligence, 2007, pp. 1852-1863.

[17] Z. Teng, Y. Liu, F. Ren, and S. Tsuchiya, "Single document summarization based on local topic identification and word frequency," in Artificial Intelligence, 2008. MICAI'08. Seventh Mexican International Conference on, 2008, pp. 37-41.

[18] X. Wan, "Using only cross-document relationships for both generic and topic-focused multi-document summarizations," Information Retrieval, vol. 11, pp. 25-49, 2008.

[19] H. Wang and G. Zhou, "Toward a Unified Framework for Standard and Update Multi-Document Summarization," ACM Transactions on Asian Language Information Processing (TALIP), vol. 11, p. 5, 2012.

[20] C. Xiaoyan and L. Wenjie, "Mutually Reinforced Manifold-Ranking Based Relevance Propagation Model for Query-Focused Multi-Document Summarization," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 20, pp. 1597-1607, 2012.

[21] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine," Computer networks and ISDN systems, vol. 30, pp. 107-117, 1998.