INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 5 (2016) PAGES 1178 - 1186
RECEIVED : 30/08/2016. PUBLISHED : 04/09/2016
September 4, 2016

MULTI DOCUMENT SUMMARIZATION BY USING GRAPH BASED TEXT MINING TECHNIQUES

[1]Tamizharasi.V, [2]Hariharan.P
[1]M.Phil Scholar, [2]Assistant Professor
Department of Computer Science & Applications, Adhiparasakthi College of Arts & Science
G.B.Nagar, Kalavai 632506, Vellore District, Tamilnadu.

ABSTRACT

The World Wide Web has become one of the largest information and knowledge repositories in the world. In spite of its easy access, it is virtually impossible for any user to browse or read a large number of such individual documents available online. Text summarization fulfils such information-seeking goals by providing a method for the user to quickly view the highlights or relevant portions of document collection. With tons of information uploaded on the web on a daily basis, the task of summarizing becomes a necessity. Also, locating and browsing information quickly from a collection of documents within a short span of time becomes possible with the help of summarization. This has led to large-scale research efforts in text summarization. The issues discussed above necessitate the need for an automated summarization system. The objective of this paper is to find enhancements to existing graph-based methods for summarizing single documents and multi-document clusters. The objective of automated text summarization is to condense the given text to its essential contents, based upon the user's choice of brevity. The summarization techniques are broadly categorized into two schemes, extraction and abstraction. Extraction involves picking up the most important sentences from a document using statistical approaches. Abstraction, on the other hand, involves the reformulation of content depending upon the type of summary. This technique involves more adoptable linguistic processing tools. Though abstraction leads to better summaries, extraction is the preferred approach and is widely adopted by the research community. The process of text summarization using either a single document or multiple documents is quite tricky and challenging, with multi-document summarization facing additional challenges. As this paper focuses on multi-document summarization, the first task is to cluster the documents based on their contents. To measure the similarity among the documents, several choices are available like cosine, dice, and Jaccard.

Keywords – Graph-based, summaries, Jaccard, cluster.

1. INTRODUCTION

Automated text summarization has drawn a lot of interest among the Natural Language Processing and Information Retrieval communities in recent years. The initial interest for automated text summarization started during the late 1960s in American research libraries, where a large number of scientific papers and books were to be digitally stored and made searchable. Before the invention of personal computers and the emergence of the World Wide Web (WWW) as a global digital library, locating text materials of relevance was a strenuous task. After the advent of WWW the form and function has been altered, where in people, academicians, researchers or lay end users get huge benefits by browsing the contents online. Though this has reduced the burden of information gathering, the task of acquiring the relevant information in a concise manner is still a challenge.

Text summarization is the solution to address this issue. Summarization is a technique in which a computer automatically creates an abstract or summary of one or more documents. Automated text summarization is the process of automatically constructing summaries for a text depending on the user's needs. A summary is a precise representation of information depending on the specified target compression ratio. Systems summarizing single documents are called single document summarization

systems, while systems which perform the same task with multiple related sets of documents are called multi-document summarization systems.

Research on automated summarizing has been reviewed in several dimensions over the last decades. There has been quite extensive work on summarizations adopting several methods like semantic graphs, combination of several extraction features, probabilistic approaches, fuzzy logic, budgeted median problem, contextual information, Sim with First methods, significant words textual association networks using sentences, words and paragraphs, graph-based approaches and complex network based approaches. Current sentence extraction based approaches are dependent not only on the similarity measures but also adopt the sentence clustering approach. Identifying sentences for a summary with a focus on reducing similarity among the sentences is a challenging task. The objective of my research paper is to extend the existing graph based methods for finding extractive summaries for single as well as multiple news documents. Contributions have been made in three distinct aspects of this objective. Firstly, for multi document summarization, clusters of documents are required to be formed, based on a suitable similarity measure. Investigations have been made on the choice of similarity measures and the various parameters that influence the content-based similarity of the document sets. A discriminator measure to focus on the sharpness of the categorization has been proposed.

2. LITERATURE SURVEY

Md. Mohsin Ali et al [1]. Proposed Document summarization is an emerging technique for understanding the main purpose of any kind of documents. To visualize a large text document within a short duration and small visible area like PDA screen, summarization provides a greater flexibility and convenience. Simulation results demonstrate that CPSL shows better performance for short summarization than MEAD and for remaining cases it is almost similar to MEAD. Simulation results demonstrate that LESM also shows better performance for short summarization than MEAD but for remaining cases it does not show better performance than MEAD. The main disadvantages of this method Precision and recall and relative utility based evaluation methods are very poor.

Ben Hachey [2] proposed a novel representation is introduced based on generic relation extraction (GRE), which aims to build systems for relation identification and characterization that can be transferred across domains and tasks without modification of model parameters. Results demonstrate performance that is significantly higher than a non-trivial baseline that uses tf*idf -weighted words and at least as good as a comparable but less general approach from the literature. The various representations are substituted in the interpretation phase of a multi-document summarisation task and used as the basis for extracting sentences to be placed in the summary. System summaries are compared by calculating term overlap with reference summaries created by human analysts. This approach did not give a convergence proof for the reranking style extraction algorithm. Lei Li et al [3] provides a description of the methods applied in Center for Intelligence Science and Technology (CIST) system participating ACL MultiLing 2013. Summarization is based on sentence extraction. Hierarchical Latent Dirichlet Allocation (hLDA) topic model is adopted for multilingual multi-document modeling [13]. Various features are combined to evaluate and extract candidate summary sentences. Sentences are clustered into sub-topics in a hierarchical tree. They evaluate the sentence importance in a sub-topic considering three features. 1) Sentence coverage, which means that how much a sentence could contain words appearing in more sentences for a sub-topic. 2) Word Abstractive level. hLDA constructs a hierarchy by positioning all sentences on a three level tree. Level 0 is the most abstractive one, level 2 is the most specific one, and level 1 is between them. 3) Named entity. Authors consider the number of named entities in one sentence. This time only have

time to use Stanford's named entity recognition toolkit4, which could identify English person, address and institutional names.

Yang Gao et al [4] approach proposed to combine the statistical topic modelling with pattern mining techniques to generate pattern-based topic models with the purpose of enhancing the semantic representations of the traditional word-based topic models. Utilizing the proposed pattern-based topic model, users' interests can be modelled with multiple topics and each of which is represented with semantically rich patterns. This proposed pattern-based topic model is adopted in the field of Information Filtering (IF) for representing long-term user's interests as well as in the field of Information Retrieval (IR) for representing short-term user's interests, especially for improving the accuracy of query expansion. The Pattern-based Topic Model (PBTM) and Structural Pattern-based Topic Model (StPBTM). The main distinctive features of the proposed models include, (1) user information needs are generated in terms of multiple topics; (2) document relevance ranking is determined based on topic distribution and topic related semantic patterns; (3) patterns are organized structurally based on the patterns' statistical and taxonomic features for representing user interests for each topic. (4) Significant matched patterns and maximum matched patterns are proposed based on the patterns' statistical and taxonomic features to enhance the pattern representations and document ranking.

3. SIMILARITY MEASURES FOR TEXT DOCUMENTS

There are several parameters by which similarity can be evaluated. The first category of similarity evaluation is based on the document size and structure. The length of the document, the number of paragraphs, number of sentences, average number of characters per word, average number of words per sentence etc. The second category is based on "style", whether the contents have been written in the first person conversational style or in the third person and so on. Thirdly, similarity can be based on the set of words used in the document. For example the original text of the novel "A Tale of two cities" written by Charles Dickens may contain 20,000 distinct words, whereas the same novel rewritten for seventh standard students may contain only a set of 1000 words. The fourth category of similarity is "content similarity" which reflects to what extent the contents of the two documents are alike. This category is adopted throughout this thesis wherever similarity is talked of hereafter. The similarity between two documents is computed by any one of the several similarity measures based on the two corresponding feature vectors, e.g. cosine, dice, and jaccard measure. The common framework for the document clustering model starts with the representation of any document as a feature vector of the terms (words) that appear in the document collection. Let D = (D1,D2, . . .,Dn) denote the collection of documents, where 'n' is the number of documents in the collection. Let T = (T1, T2, . . .Tm) represent all the terms that occurred in the document collection 'D'. Here 'm' is the number of unique terms in the document collection. In most clustering algorithms, the dataset to be clustered is represented as a set of vectors, where each vector corresponds to a single object and is called the feature vector.

The representation of a set of documents as vectors in a common vector space is known as a Vector Space Model (VSM). A collection of 'N' documents can thus be viewed as a collection of vectors, leading to the natural view of a collection as a term-document matrix; this is an M× N matrix whose rows represent the M terms (dimensions) of the N columns, each of which corresponds to a document. The standard way of quantifying the similarity between two objects 'ti' and 'tj' is to compute the similarity of their vector representations, using the IDF in following expressions. In these formulas it is assumed that the similarity is being evaluated between two vectors ti ={ ti1,…….. tik} and tj ={ tj1,…….. tjk}, and the vector entries usually are assumed to be nonnegative numeric value.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 5 (2016) PAGES 1178 - 1186
RECEIVED : 30/08/2016. PUBLISHED : 04/09/2016
September 4, 2016

$$Cosine(ti,tj) = \sum_{h=1}^{k} (t_{ih} * t_{jh}) * (idf_h) \Big/ \sqrt{\sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2}$$

$$Dice(ti,tj) = 2 \sum_{h=1}^{k} (t_{ih} t_{jh}) * (idf_h) \Big/ \sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2$$

$$Jaccard(ti,tj) = \sum_{h=1}^{k} t_{ih} t_{jh} * (idf_h) \Big/ \sum_{h=1}^{k} t_{ih}^2 * (idf_h)^2 \sum_{h=1}^{k} t_{jh}^2 * (idf_h)^2 - \sum_{h=1}^{k} t_{ih} t_{jh} * (idf_h)$$

Fig.1. IDF in Expressions.

3.1 DISCRIMINANT FACTOR

A new factor called the Discriminant is defined as follows. The Discriminant factor is explained with an example. Let D1, D 2, D 3, D4, D5, D6, D7, D8, D9 and D10 be a set of documents. D1 and D2, D3 and D4, D5 and D6, D7 and D8 and D9 and D10 are similar documents. Then, the Discriminant for document set 1 is defined as follows:

$$Discriminant_1 = \frac{Sim(D_1, D_2) - max(Sim(D_1, D_3 \ldots D_{10}))}{min(Sim(D_1, D_2))}$$

$$Discriminant_2 = \frac{Sim(D_1, D_2) - max(Sim(D_2, D_3 \ldots D_{10}))}{min(Sim(D_1, D_2))}$$

$$Discriminant_{Cluster\,1} = min(Discriminant_1, Discriminant_2)$$

$$Discriminant = min(all\ Discriminant_{Clusters})$$

Fig.2. Discriminant Factor.

Similarly, the discriminant is calculated for all the document sets. Finally, the discriminant factor is chosen as the minimum of all discriminants in the cluster. Thus, the discriminant is a suitable measure to find whether the method adopted is sharp enough to segregate similar and dissimilar documents. The discriminant has been calculated in two ways. The first is as per the formula already given. The second is based upon the average values. To find that IDF approach yields larger discriminant factors for the three measures under consideration based on either ways. Therefore, the recommend the TF+IDF approach for the documents under consideration.

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 5 (2016) PAGES 1178 - 1186
RECEIVED : 30/08/2016. PUBLISHED : 04/09/2016
September 4, 2016

4. SINGLE DOCUMENT SUMMARIZATION

Graph-based approaches for summarizations are quite popular. These methods are modeled under two types of social networks. Let us consider a real world situation to define these two types to realize their importance. A person having extensive contacts with people in an organization is considered more important than a person with fewer contacts. Hence, the person's prominence can be simply determined in a democratic way, by the number of contacts he has. On the other hand, let us consider the case of a second person who has fewer contacts, but all his contacts are highly placed and influential persons. Clearly, in this situation, the second person may have profound influence and prestige compared to the former. The second method takes care of not only the number of supports the target person receives but also the influence or prestige of the person who is lending him support. Three graph based methods of summarization, namely, the Centrality Degree based on the democratic popularity approach of social network, and the prestige based approaches of LexRank (Threshold) and LexRank (Continuous).

In the graph-based approach, each document is represented as a graph. The entries in the matrix correspond to the similarities between sentences. Each sentence in a document or in a cluster of documents is represented by a vertex node. The similarity between sentences is based on a suitable similarity measure and is represented as links, with link weights corresponding to the similarity values. Though several measures are available for measuring the similarity, two measures cosine and overlap have been used widely for the text summarization task. Of the two measures cosine is superior because it provides standard baselines.

The similarity between the two pairs of sentences 'x' and 'y' is determined after the removal of the stop words and stemming. The cosine measure also reflects the degree of similarity in the corresponding terms and term weights, while 'overlap' measures the degree to which the two sets overlap. Comparing the two metrics, the overlap measure takes the min operator and provides a higher magnitude than cosine. Further cosine is independent of length, but the overlap measure greatly varies depending on length.

The proposed two enhancements to the already existing graph based approaches. These enhancements are applicable to all the existing methods. This approach describes the two enhancements of the discounting technique and the incorporation of the position weight in the next two subsections.

4.1 DISCOUNTING TECHNIQUE

The discounting technique envisages that once a sentence is selected by any one of the methods, immediately, the corresponding row and column values of the matrix are set to zero. Thus the next sentence is selected from the contributions made by the remaining sentences only. Discounting methods are applicable to both Non-PageRank as well as PageRank schemes. When the discounting technique is applied and sentences are chosen for summary generation based on a given compression ratio, the adjacency matrix is modified as stipulated. The idea behind the discounting technique is that once a sentence is selected, the chance for the repetition of information in the succeeding sentences is minimized. Discounting also envisages less redundancy among the chosen sentences.

The six PageRank type methods are available that are listed below.

- LexRank (Threshold)
- LexRank(Continuous)

INTERNATIONAL RESEARCH JOURNAL IN ADVANCED ENGINEERING AND TECHNOLOGY (IRJAET) E - ISSN: 2454-4752 P - ISSN : 2454-4744
VOL 2 ISSUE 5 (2016) PAGES 1178 - 1186
RECEIVED : 30/08/2016. PUBLISHED : 04/09/2016
September 4, 2016

- Discounted LexRank
- Discounted LexRank
- Sentence Rank
- Sentence Rank

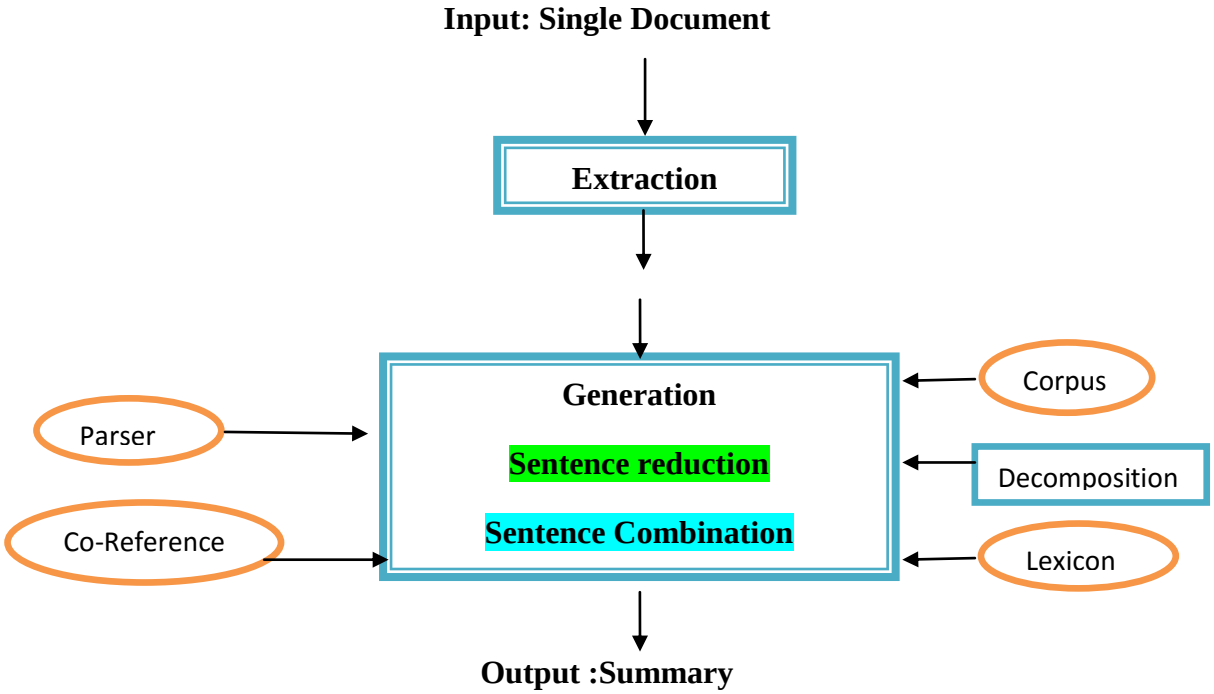

Fig.3. Single Document Summarization.

The first class corresponds to methods of the non- PageRank type, while the second group is based on the PageRank type algorithms. In each class, the discounting methods proposed in this chapter are superior to the basic methods and the proposed discounting plus position weight approach fares the best. All the twelve methods are promising in that they yield superior results as compared to random selection, based on the conventional precision metric as well as by the proposed metrics Effectiveness1 (E1) and Effectiveness2 (E2). It is brought out from the investigations presented, that based on the average performance of over a 30-document set, methods Sentence Rank (Threshold) and Sentence Rank (Continuous) – the proposed Sentence Rank (Threshold) and Sentence Rank (Continuous), yields the best results of all the 12 methods considered. The next chapter presents the investigations carried out for multi-document summarization.

5. MULTI-DOCUMENT SUMMARIZATION BY GRAPH BASED METHODS

If the first approach is used, sentences cannot be compared across documents on a common scale while trying to determine which of the sentences are important in a summary. In the second approach, a comparison of sentences within a document is possible only if the document boundary is tracked by the user. The next approach adopts suitable weights for the intradocument and cross-document characteristics of the document. This chapter follows the second approach by merging all the sentences into a single document and keeping track of the documents' identity while allocating position weights. It is assumed that the document corpus obtained is of the same time stamping. If there exists some time stamping in the document, then the summary is generated based on the descending order of time stamping consider as an example, two documents having sentences 5 and 7 sentences respectively in each. Altogether there are 12 sentences in the document. The location of a

sentence in a document plays a significant part in determining its importance. In the earlier chapter shown that a tie between two sentences can be resolved by giving preference to the sentence that appears earlier in the document. Thus, if find, that sentences 3 and 5 have equal weights, here select sentence 3, which occurs earlier in the document, to resolve the tie.

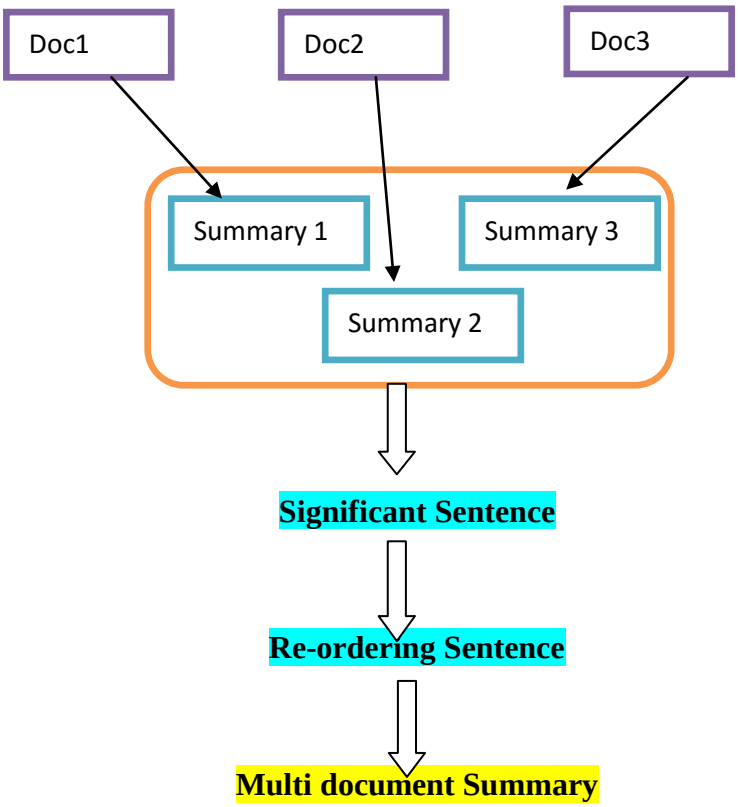

Fig.4. Multi Document Summary.

For a multi-document the strategy is to be modified as follows:

- If the tie occurs between two sentences of the same document, this approach resolved the tie based upon the single document formula. Thus, if there is a tie between sentences 3 and 5 of document 1 of a two document set, the tie is resolved in favour of sentence 3 as before.
- The tie may occur between sentences that occur in two different documents. Let us explain with an example. Let us assume that there are two documents D1 and D2; D1 has 5 sentences and D2 has 10 sentences. Let us assume that there is a tie between the first sentence of the two documents i.e., all the weights are added and the scores are equal. As before, to give importance to position, but also take care of the size of the group. Thus in D1, the first sentence is the first among five sentences, while in D2, the first sentence is the first among the ten sentences. Thus, the first sentence in D2 will be selected to resolve the tie.

6. EXPERIMENTAL RESULTS AND DISCUSSION

Experiments have been carried out for multi-document adopting approaches as proposed in the single document as well as multidocument. The existing methods for the multi-document summary generation task are categorized as Non-PageRank and PageRank type – six pertaining to the

former type and seven belonging to the latter type. For convenience, list the methods for the PageRank and Non-PageRank types as shown below. Here find all the 13 methods are superior to random selection; SR methods are far superior to Lead-based selections. The comparisons of all the 13 methods using Effectiveness1 (E1) and Effectiveness1 (E2) are presented in Figures 6.7 presents the comparison of the SentenceRank (Continuous) approach with two baseline systems, namely, Lead and Random performances.

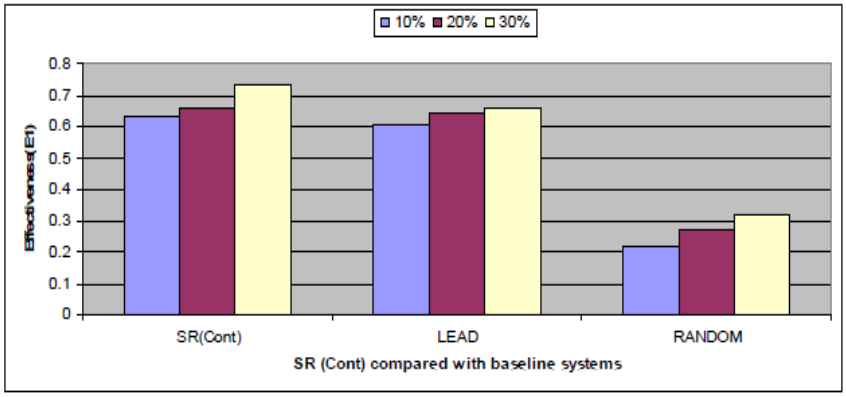

Fig.5. Comparison of the Sentence Rank (Continuous) approach with Lead and Random baseline systems.

Here find that the SentenceRank (Continuous) method results are lower than the best DUC results in some cases, equal to the best DUC results in some cases, and higher in a large number of cases. On taking the average for the 10 document set, to find that for both the 200 and 400 word summaries, the SentenceRank (Continuous) method emerges superior. The comparison of the performance of the SentenceRank (Continuous) approach with that of the best DUC 2002 results is presented using precision and recall respectively.

7. CONCLUSIONS AND FUTURE ENHANCEMENTS

The summarization of text documents has been a heavily researched area. This thesis has investigated two classes of graphical methods for text summarization. The first class corresponds to basic methods of non PageRank type, while second grouping is based on PageRank type algorithms. It is shown that in each class discounting methods proposed in this thesis is superior to basic methods and the proposed discounting technique plus position weight method fares the best. The Sentence Rank (Continuous) method is found to yield superior results as compared to the best published results no of data set. The thesis has analyzed alternative methods for the intrinsic evaluation of summaries and has proposed a new metric called ‘Effectiveness’. Further steps have been formalized for the preparation of the ‘gold standard’ reference summary. Studies done using the corpus of documents collected from commercial and research sites and DUC 2002 data set establish the superiority of the methods proposed. The Sentence Rank (Threshold) and Sentence Rank (Continuous) approaches proposed, yield better results for both the data sets, irrespective of the evaluation measures.

Generating a summary is a tricky and challenging task, especially for multi document cases. Only two aspects, namely, discounting and position weight have been considered for the study. The results obtained are not only promising but provide a good scope for further improvement using some additional features. Thus, summary generation techniques in this work do not take temporal

information into account. If such a feature is considered in future, then the summary of each document could be generated and merging can be done based on a time sequence. Linguistic processing tools may be used to analyze the semantics of the documents to improve the quality of summaries.

REFERENCES

[1]. Ali M., Ghosh M.K. and Abdullah Al Mamun (2009), 'Multidocument Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation', Proceedings of the International Conference on Future Computer and Communication, Kuala Lumpar, Malaysia, pp. 93-96.

[2]. Ben Hachey," Multi-Document Summarisation Using Generic Relation Extraction" Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 420–429, Singapore, 6-7 August 2009. c 2009 ACL and AFNLP.

[3]. Lei Li, Wei Heng, Jia Yu, Yu Liu, Shuhong Wan" Multilingual Multi-document Summarization", Proceedings of the MultiLing 2013 Workshop on Multilingual Multi-document Summarization, pages 39–44,Sofia, Bulgaria, August 9 2013.

[4]. Yang Gao, Yue Xu, and Yuefeng Li" Pattern-based Topics for Document Modelling in Information Filtering", IEEE Transactions On Knowledge And Data Engineering, VOL. 27, NO. 6, JUNE 2015.

[5]. Bollegala D., Okazaki N. and Ishizuka M. (2009), 'A Bottom-up Approach to Sentence Ordering for Multi-Document Summarization', Information Processing and Management, Vol. 46, No. 1, pp. 89-109.

[6]. Bani Ahmad S., Cakmak A., Ozsoyoglu G. and Al-Hamdani A. (2005), 'Evaluating Publication Similarity Measures', Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, Vol. 28, No. 4, pp. 21-28.