Enhanced Graph Based Approach for Multi Document Summarization

Shanmugasundaram Hariharan[1], Thirunavukarasu Ramkumar [2], and Rengaramanujam Srinivasan[3]
[1]Pavendar Bharathidasan College of Engineering and Technology, India
[2]A.V.C College of Engineering, India
[3]Bangladeshi Students Association University, India

Abstract: *Summarizing documents catering the needs of an user is tricky and challenging. Though there are varieties of approaches, graphical methods have been quite popularly investigated for summarizing document contents. This paper focus its attention on two graphical methods namely – LexRank (threshold) and LexRank (Continuous) proposed by Erkan and Radev. This paper proposes two enhancements to the above work investigated earlier by adding two more features to the existing one. Firstly, discounting approach was introduced to form a summary which ensures less redundancy among sentences. Secondly, position weight mechanism has been adopted to preserve importance based on the position they occupy. Intrinsic evaluation has been done with two data sets. Data set 1 has been created manually from the news paper documents collected by us for experiments. Data set 2 is from DUC 2002 data which is commercially available and distributed or accessed through National Institute of Standards Technology (NIST). We have shown that the based upon precision and recall parameters were comprehensively better as compared to the earlier algorithms.*

Keywords: *Page rank, lexical rank, damping, threshold, summarization.*

Received July 13, 2011; accepted December 29, 2012

1. Introduction

Automatic text summarization sets its goal as condensing the given text to its essential contents, based upon user's choice of brevity. The basic foundation for summarization was laid five decades ago [2, 15] and since then numerous techniques have been extensively studied. Automatic text summarization is a multi facetted endeavor that typically branches out in several dimensions, which can be grouped into several over lapping categories [8]. For the study chosen here the data set remains to be a clustered sequence and it is not need to use any feature selection methods for categorization [31]. Based on the methodology or technique used summarization approaches can be divided into two broad groupings - as extraction and abstraction schemes. Abstraction involves reformulation of contents, while in extraction method the important sentences of the original document are picked up in toto for summary generation. Speed, simplicity, non requirement of background knowledge, and domain independency are some of the features that favour extraction, where as abstraction is domain dependent in nature and requires human knowledge and is goal oriented [1].

Investigations on summarization using data mining and other related tasks spreads across multiple disciplines like softwares [24], scientific papers[19] and others. Of all such approaches text based approaches using graphical approaches are well investigated using graph based techniques in multi document scenarios [3, 5, 17]. These methods are modelled under two types of social networks. Let us consider the real world situation to define these two types to realize their importance. A person with extensive contacts or communications 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 of 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.

[3] have presented in their excellent paper, three graph based methods of summarization; Centrality Degree based on the democratic popularity approach of social network and prestige based approaches of LexRank and Continuous LexRank. However, these methods have certain drawbacks in sentence selection namely not eliminating the redundancy among the chosen sentences and not incorporating position weight schemes. Also the recent approaches (discussed in literature review section) focus only on some additional dimensions like time, query etc., we propose

enhancements to the above presitage based PageRank type methods of [3] and show that with these proposed enhancements the summarizer performance is vastly improved.

The rest of the paper is organized as follows. Section 2 describes the review works carried out in graph based summarization, while section 3 focuses on the working of LexRank and Continuous Lexical Rank approaches developed earlier [3]. Section 4 briefs the proposed enhancements. Section 5 deals with experimental investigations and finally Section 6 list the conclusions and future enhancements.

2. Literature Review

Li *et al*., [9] have proposed event based summarization approach that would select sentences for summary by making use of inter and intra relevance information of events or sub events that the sentences describes. The authors found that events have their own internal structure, and often relates to other events semantically, temporally, spatially, causally or conditionally. PageRank ranking algorithm is then applied to estimate the significance of an event for inclusion in a summary from the event relevance derived.

Litvak and Last [13] introduced and compared two novel approaches namely supervised and unsupervised methods, for identifying the keywords to be used in extractive summarization of text documents. Both these approaches are based on the graph-based syntactic representation in form of text documents, which enhances the traditional vector-space model by taking into account some structural document features like word co occurrence, size of the co occurrence window are considered . In supervised approach, the training phase was done with the help of classification algorithm by using a summarized collection of documents. In unsupervised approach, HITS algorithm was run on the document graphs under the assumption that the top-ranked nodes should represent the document keywords.

Yeh *et a*l., [30] proposed a novel graph-based ranking method called iSpreadRank that extracts sentences and presents summary to user. iSpreadRank exploits the concept of spreading activation theory to formulate a general concept from social network analysis by taking into consideration, the importance of its connected nodes also. The algorithm recursively reweighs the importance of sentences by spreading their sentence-specific feature scores throughout the network and adjusts the importance of other sentences.

Patil and Brazdil [22] presented a graph theoretic technique called SumGraph for automatic text summarization to produce extractive summaries for single documents. The authors have adopted the concept of Pathfinder Network Scaling (PFnet) technique to compute importance of a sentence in the text. Each text is represented as a graph with sentences as nodes while weights on the links represent intra-sentence dissimilarity. Experiments using Latent Semantic Indexing was also performed. The system is empirically evaluated on DUC2001 and DUC2002 datasets using ROUGE measure.

Liu *et al*., [14] presented a novel multi-document summarization approach based on Personalized PageRank (PPRSum). The algorithm trains each sentences by making use of the global features provided by the corresponding sentence using Naive Bayes Model. Then a relevance model for each corpus utilizing the query is generated, followed by calculation of probability for each sentence in the corpus utilizing the salience model. Based on the probability value it obtains Personalized PageRank ranking process is performed depending on the relationships among all the other sentences. Additionally, the redundancy penalty is imposed on each sentence. Finally summary sentences are chosen based on information richness with high information novelty.

Wan [29] exploited graph-based ranking algorithm for multi-document summarization under the assumption that all the sentences in the graph model are indistinguishable. The algorithm also focus on two different aspects namely taking into account the relationship of sentences with each others in the documents as well the document information to globally reflect the importance the theme of the multi document cluster.

TextRank demonstrated [18] is a system for unsupervised extractive summarization that relies on the application of iterative graph based ranking algorithms to graphs encoding the cohesive structure of a text. The distinguishing characteristics of the proposed system is that it does not rely on any language-specific knowledge resources or any manually constructed training data, and thus it is highly portable to new languages or domains. It is shown by the author that iterative graph-based ranking algorithms work well on the task of extractive summarization since they do not only rely on the local context of a text unit (vertex), however it takes the information recursively drawn from the entire text (graph) into account.

Wan [28] proposed TimedTextRank algorithm for multi document summarization that lies on the foundation of graph based ranking algorithm namely TextRank. The proposed algorithm overcomes the problems in earlier approaches by introducing temporal dimension. From the preliminary study carried out to measure the effectiveness of the proposed TimedTextRank algorithm, it is seen that use of temporal information of documents based on the graph-ranking for dynamic multi-document summarization leads to results that are promising.

3. Lexrank and Continuous Lexrank Approaches

In this section we discuss LexRank and Continuous LexRank methods which are developed based on modification of the most popular page ranking algorithms designed for web link analysis [21]. Such ranking models have been successfully exploited for multi document summarization by making use of the link relationships between sentences in the document set, under the assumption that all the sentences are indistinguishable from each other. A link between two sentences is considered as a vote cast from one sentence to the other sentence. The score of a sentence is determined by the votes that are cast for it, and the scores of the sentences casting these votes [17].

In sentence extraction process all the words in a sentence cannot be treated as equal importance, hence we perform necessary preprocessing like removal of stop words and stemming [23]. It is also found from our previous work that IDF would definitely improve the performance of the system [5]. Equations 1 and 2 give the LexRank and Continuous LexRank for the given document as proposed by Erkan and Radev [3].

$$LexRank[i] = \frac{d}{N} + (1-d) * \sum_{j \in S[i]} \frac{LexRank[j]}{deg[j]} \quad (1)$$

$$\text{Continuous LexRank}[i] = \frac{d}{N} + (1-d) * \sum_{j \in S[i]} \frac{idf\_modified\_Cosine(i,j) * PR[j]}{\sum_{k \in S[j]} idf\_modified\_Cosine(j,k)} \quad (2)$$

Where N is the total number of sentences in the document, d is the damping factor which is typically chosen in the interval [0 to 1], PR(j) represents the centrality of node j, S[i] denotes the set of nodes that are adjacent to 'u' and deg(j) is the degree of the node j.

A document can be considered as a network of sentences that are related to each other. The similarity between the two pairs of sentences x and y is determined is done after pre-processing. Though there exist several choices of measures to measure the similarity, cosine is superior (Hariharan and Srinivasan, 2008) and is preferred to measure the relevance between the two sentence vectors as modified by the inverse document frequency given by equation 3.

$$idf\_modified\_Cosine(x,y) = \frac{\sum_{w \in x,y} tf_{w,x} * tf_{w,y} * (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} * idf_{x_i})^2} * \sqrt{\sum_{y_i \in y} (tf_{y_i,y} * idf_{y_i})^2}} \quad (3)$$

Where $tf_{w,s}$ represent the number of occurrences of word 'w' in sentence 'S'. A cluster of 'n' sentences in the document can thus be represented by an n x n symmetric cosine-similarity matrix.

4. Proposed Enhancements

Our enhancements rest on the foundations of graph based approaches already developed by Erkan and Radev [3] namely LexRank and Continuous LexRank. We have introduced two enhancements to the above schemes namely discounting technique and incorporation of position weight factor.

4.1. Discounting

Discounting technique envisages that once a sentence is selected by any one of the methods (listed below in section 4.2), then the corresponding row and column values of the matrix are set to zero. The next sentence is selected based on the contributions made by the remaining 'n-1' sentences only. Thus when we use discounting technique to any of the methods proposed, the sentences were picked up desired on the target ratio, provided the adjacency matrix is modified as stipulated. The idea behind discounting technique is that once the sentence is selected, the chance for repetition of information in the succeeding sentences is minimized. The information will not be duplicated and the summary will be cohesive and meaningful in nature.

4.2. Position Weight

The location of a sentence in a document plays a significant part in determining the importance of a sentence. In the graph based approach, for multi document summarization, importance to position of the sentence can be given in a way by giving preference to sentences that occurs earlier out of the two documents considered. Consider an example to illustrate the situation clearly. For instance if document1 has 10 sentences and document2 has 5 sentences and if there is tie in selecting the first sentence, then we select sentence1 from document 1(since it gets a weight of 1/10 =0.1) rather than sentence1 from document 2 (which gets a weight of 1/5 =0.20). Position weight factor is given by

$$P_{f_i} = gama * Beta^{i-\alpha-1} \quad (4)$$

Where gama and beta are design parameters. $\alpha = 0$ for the sentences of the first document, $\alpha = n1$ for the sentences of the second document and $\alpha = n1+n2$ for the sentences of the third document etc., n_i being the number of sentences in the document. Thus position weight of any sentence is allocated based on its relative position in the document in which it is present.

In order to clearly distinguish between various methods, we call LexRank methods with the incorporation of discounting and position weight as Sentence Rank (SR) methods. The equations for Sentence Rank with threshold and Continuous Sentence Rank are given in equations 5 and 6.

$$SR_{thres}[i] = \frac{d}{N} + gama * beta^{i-\alpha-1} + (1-d) * \sum_{j \in S[i]} \frac{SR_{thres}[j]}{deg[j]} \quad (5)$$

$$SR_{Cont}[i] = \frac{d}{N} + gama * beta^{i-\alpha-1} + (1-d) * \sum_{j \in S[i]} \frac{idf-modified-Cosine(i,j) * SR_{Cont}[j]}{\sum_{\kappa \in S[j]} idf-modified-Cosine(j,k)} \quad (6)$$

In equations 5 and 6, gama and beta are parameters which affect the position influence. Thus, with no discounting:

a. when gama= 0 ; methods become LexRank
b. when gama= high value ; the summarizer is purely lead based
c. when gama= intermediate value; we have a mix of (a) and (b).

In all we are considering six methods as listed below. Of the six methods, Methods I and II were proposed earlier by Erkan and Radev [3], while the rest of the methods are proposed in this paper. Methods III and IV adopts the discounting technique to the basic methods I and II, while Methods V and VI combines position weight and discounting technique together with the basic schemes proposed by Erkan and Radev [3].

- Method I -- LexRank (threshold)
- Method II -- Continuous LexRank
- Method III --Discounted LexRank (threshold)
- Method IV --Discounted Continuous LexRank
- Method V --Sentence Rank (threshold)
- Method VI --Sentence Rank (continuous)

For methods I to VI several investigations were made relating to threshold, damping factor, direction of graph and impact of self weight. While it is recommended to adopt a damping factor in the interval 0.1 to 0.2 [21], we have adopted an optimal damping factor of 0.10 [5]. We have adopted undirected graphs and threshold of 0.10 for threshold methods [6].

5. Experimental Investigations

5.1. Corpus Description

Experiments were carried out using two different data sets as shown below.

5.1.1. Data Set 1

The corpus for data set 1 was collected from news documents that are readily available from news service providers* like google, yahoo, rediff, hindu, Indianexpress and cnn. In order to obtain a target set of ideal results, the document sets were distributed to different set of judges who were appropriate to judge the quality of the summary.

*www.google.com/news , www.rediffnews.com, www.yahoonews.com, www.hindu.com, www.indianexpress.co.in

Each judge in the set was chosen and they are requested to rank the sentences according to their importance. In all there were sixteen judges chosen from the faculties of engineering, sciences and humanities as volunteers. Their age groups vary from 30 to 60 and all of them are post graduates, many of them holding doctoral degrees. For multi document experiments a cluster of 50 document set pairs were collected. All such documents pertain to news reports that are recent ones.

Study results for the methods investigated, using data set 1 is presented in Table 1. The results are based on an average of 50 document pairs. Evaluation has been done based upon precision/ recall metrics as well as Effectiveness (E1/E2) defined by equation 9. Since for data set 1 compression ratio ‘r’ has been calculated based on the number of sentences selected, both precision and recall have same values.

5.1.2. Data Set 2

Data set 2 comprises of DUC 2002 dataset extracts provided for multi document summarization. Table 2 presents the details of the corpus used. Altogether there are 4 categories of the document, with each category having 15 clusters. We have chosen 10 clusters randomly and the results are shown in Tables 3, 4 and 5. For each set, two summaries were created by NIST human assessors having approximately 200 words, and the other with 400 words. Again for each category there are two summarizer models. Evaluation has been done using Precision and Recall metrics. Since target is given as number of words, the number of sentences selected by the judges and summarizer can vary. Therefore Precision and Recall will have varying values. We have not focused on any other datasets as the recent years have concentrated on summaries that were not of pure extracts.

Table 1. Comparison of methods for Data Set 1.

Compression Ratio	Evaluation Measure	M- I	M- II	M- III	M- IV	M- V	M-VI
10%	E1	0.560	0.573	0.570	0.587	0.612	0.634
	E2	0.482	0.520	0.512	0.537	0.554	0.615
	Precision/Recall	0.384	0.394	0.401	0.422	0.452	0.476
20%	E1	0.585	0.594	0.611	0.624	0.639	0.657
	E2	0.552	0.566	0.578	0.617	0.608	0.627
	Precision/Recall	0.445	0.481	0.463	0.477	0.532	0.554
30%	E1	0.634	0. 648	0.652	0.665	0.710	0.732
	E2	0.620	0.642	0.645	0.662	0.702	0.728
	Precision/Recall	0.488	0.535	0.492	0.561	0.620	0.646

Table 2. Statistics of 2002 DUC data set.

Category	Document category	No. of Clusters chosen	No. of documents in each cluster (separated by commas)	No. of sentences in each cluster (separated by commas)
1	Single Natural disaster, created within at most seven day window	4	6,5,6,10	146,118,121,222
2	Single vent in any domain, created within at most seven day window	3	8,6,5	246,140,112
3	Multiple distinct events of single type (no limit on time window)	1	6	115
4	Bibliographical information about a single individual	2	7,11	191,149
		10	70	1560

5.2. Evaluation

Evaluation is a crucial step for multi-document summarization and is categorized into two major categories as intrinsic and extrinsic modes of evaluation [16]. In intrinsic evaluation humans judge the quality of summary by directly analyzing it in terms of fluency, coverage or resemblance to manually constructed ideal summary. The second type of evaluation method is extrinsic, where the quality of summary is judged based on how it affects the completion of some other task. We stick on to the former method of evaluation by evaluating the automated summary with the human generated reference summary based on ranking of sentences by judges.

Precision and Recall have long used as important evaluation metrics in IR field. If "retrieved" (represented as 'A' shortly for convenience), denotes the number of sentences retrieved by the summarizer and "relevant" (represented as 'B' shortly for convenience) denotes the sentences that are relevant as compared to target set, precision and recall is computed based on equations 7 and 8.

$$\Pr ecision = \frac{A \cap B}{A} \tag{7}$$

$$\text{Re}call = \frac{A \cap B}{B} \tag{8}$$

Instead of this boolean-based method, a utility-based evaluation scheme have been suggested [25,26]. Considering the drawbacks of both these evaluation schemes, we have proposed an effectiveness based evaluation method which is an enhancement of the earlier utility based evaluation mechanism [7]. We have defined Effectiveness1 (E1) and Effectiveness2 (E2) by equation 9. Definitions for E1 and E2 are quiet similar. In case of E1, judges assign score to all the sentences in the document where as in case of E2 judges rank only the required number of sentences, corresponding to the stipulated compression ratio. In this case the score of the sentences that are not picked up by any of the judges is set to zero.

$$E_1 \text{ or } E_2 = \frac{Score\ of\ the\ selected\ sentences\ by\ the\ sumarizer}{Maximum\ possible\ score} \tag{9}$$

Tho ugh ROUGEeval [10, 11, 12] is used as a defacto standard for automated evaluation of summaries in annual Document Understanding Conferences [20], we are not considering the same, since it has been found that even poor quality summaries can also have very high ROUGE scores [27].

5.3. Experiments

Table 3 presents the results of all the six methods using data set 2. Figure 1 presents the plot of various methods using precision and recall at target ratio of 200 of 400 words, using data set 2. Tables 4 and 5 present the precision and recall results for each of the 10 sets. Each document set has a unique identifier named as 'DocSet number', category as represented in Each summary of specified length is generated by several judges (Code assigned as A to J). Past best

results of DUC participants are denoted in the column marked 'maximum' denotes the maximum among the runs (for 200 and 400 word sizes separately).

Table 3. Comparison of methods for data set 2.

Evaluation Measure	Target size	M- I	M- II	M- III	M- IV	M- V	M-VI
Precision	200	0.200	0.240	0.277	0.281	0.299	0.369
	400	0.235	0.254	0.287	0.317	0.341	0.419
Recall	200	0.149	0.187	0.230	0.271	0.305	0.321
	400	0.201	0.225	0.268	0.286	0.331	0.360

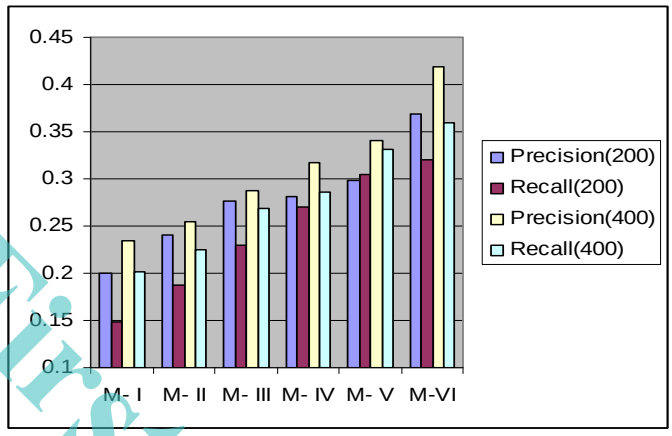

Figure 1. Methods compared using data set1.

Table 4. Comparison of SR(Continuous) Precision values with those of Best DUC results using Data set 2.

Doc Set Number	Category	Summarizer model	Proposed approach Results		Best DUC Results	
			Target size (200 words) Max.	Target size (400 words) Max.	Target size (200 words) Max.	Target size (400 words) Max.
d061j	1	B	0.250	0.368	0.222	0.250
		I	0.444	0.462	0.429	0.450
d062j	1	A	0.200	0.462	0.167	0.308
		G	0.400	0.538	0.429	0.700
d063j	2	C	0.300	0.333	0.300	0.267
		E	0.400	0.462	0.250	0.417
d066j	4	C	0.375	0.437	0.375	0.375
		I	0.375	0.357	0.500	0.312
d067f	1	A	0.333	0.466	0.400	0.455
		I	0.429	0.500	0.500	0.583
d070f	4	G	0.500	0.353	0.500	0.357
		J	0.375	0.428	0.375	0.417
d071f	3	A	0.375	0.391	0.455	0.467
		B	0.400	0.476	0.400	0.533
d074b	2	A	0.364	0.375	0.167	0.300
		E	0.455	0.411	0.500	0.429
d097e	1	A	0.333	0.333	0.250	0.267
		J	0.400	0.384	0.333	0.385
d0113h	2	A	0.333	0.375	0.286	0.364
		I	0.300	0.461	0.286	0.455
Average			0.369	0.419	0.356	0.405

Table 5. Comparison of SR(Continuous) Recall values with those of Best DUC results using Data set 2.

Doc Set Number	Category	Summarizer model	Proposed approach Results		Best DUC Results	
			Target size (200 words) Max.	Target size (400 words) Max.	Target size (200 words) Max.	Target size (400 words) Max.
d061j	1	B	0.400	0.368	0.200	0.263
		I	0.300	0.526	0.300	0.474
d062j	1	A	0.375	0.417	0.125	0.333
		G	0.375	0.467	0.375	0.467

d063j	2	C	0.333	0.368	0.333	0.211
		E	0.429	0.214	0.143	0.357
d066j	4	C	0.333	0.350	0.333	0.300
		I	0.333	0.278	0.444	0.278
d067f	1	A	0.286	0.294	0.286	0.294
		I	0.429	0.294	0.429	0.412
d070f	4	G	0.333	0.412	0.333	0.294
		J	0.333	0.438	0.333	0.375
d071f	3	A	0.364	0.381	0.455	0.476
		B	0.444	0.300	0.444	0.450
d074b	2	A	0.200	0.333	0.100	0.200
		E	0.286	0.286	0.429	0.429
d097e	1	A	0.250	0.429	0.250	0.268
		J	0.143	0.333	0.286	0.333
d0113h	2	A	0.250	0.400	0.250	0.333
		I	0.222	0.313	0.222	0.438
Average			0.321	0.360	0.304	0.349

5.4. Study Conclusions

From a perusal of comparison of results presented in Tables 1 and 3, we find that for both data sets based on precision and recall metrics

1. Methods III and IV using discounting techniques are superior to basic LexRank (threshold) and Continuous LexRank methods (Methods I and II).
2. SR (threshold) and SR (Continuous) methods (Methods V and VI) are superior to their counter parts Methods III and IV.
3. SR (Continuous) – Method VI is superior to all the other methods.

The conclusions also hold good for Effectiveness metrics. We are unable to present E1 and E2 values for data set2 for want of data detailing actual ranking of the sentences by DUC evaluators. From the perusal of Tables 4 and 5 which present a comparison of Precision and Recall values for data set 2 we find that SR(Continuous) method are lower than best DUC results in some cases; equal to best DUC results in some cases and higher in large number of cases. On taking average for the 10 document set, we find that for 200 and 400 words summaries SR (Continuous) method emerges superior.

6. Conclusions and Future Enhancements

We have investigated in depth, two graphical methods for multi document summarization namely SentenceRank (threshold) and SentenceRank (Continuous). In each case, discounting methods proposed by us are found to be superior as compared to their basic methods and the proposed SentenceRank methods which is a combination of discounting technique along and position weight is investigated to be the best. It is brought out from the investigations presented that SentenceRank approach yields better results for both the data sets considered irrespective of evaluation measures considered. Investigations on DUC data bring out that SR (Continuous) method is superior to best DUC 2002 methods, based on the average of maximum performances. Now we focus on to measure the meaningfulness generated for the summaries by use of NLP tools.

Acknowledgements

The authors would like to express their gratitude to the management, staff and students of the colleges concerned for providing support and environment. The authors would also like to extend their appreciations to the anonymous reviewers for their valuable suggestions. The authors would like to express their gratitude to the well wishers who have provided valuable suggestions and feedback.

References

[1] Cretu, B., Chen Z., Uchimoto T., and Miya K., "Automatic Summarizing Based on Sentence Extraction: A Statistical Approach," *International Journal of Applied Electromagnetics and Mechanics*, IOS Press, vol. 13, no. 1-4, pp. 19-23, 2002.

[2] Edmundson H., "New Methods in Automatic Extracting," *Journal of the ACM*, vol .16, no 2, pp. 264-285, 1969.

[3] Erkan, G. and Radev D., "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp.457-479, 2004.

[4] Hariharan S. and Srinivasan,R., "A Comparison of Similarity Measures for Text Documents," *Journal of Information and Knowledge Management*, vol. 7, no. 1, pp. 1-8, 2008.

[5] Hariharan S. and Srinivasan R., "Enhancements to Graph Based Approaches for Multi Document Summarizations," *International Journal of Applied Computer Science and Mathematics*, vol. 3, no. 6, pp. 66-72, 2009.

[6] Hariharan,S. and Srinivasan.R., "Studies on Graph Based Approaches for Single and Multi Document Summarizations," *International*

Journal of Computer Theory and Engineering, vol.1, no. 5, pp. 512-519, 2009.

[7] Hariharan,S. and Srinivasan R., "Studies on Intrinsic Summary Evaluation," *International Journal of Artificial Intelligence and Soft Computing*, vol. 2, no. 1 / 2, pp.58-76, 2010.

[8] Jones K., "Automatic Summarising: The State of the Art," *Information Processing and Management*, vol. 43, no. 6, pp. 1449-1481, 2007.

[9] Li W., Wu M., Lu Q., Xu W. and Yuan C., "Extractive Summarization Using Inter- and Intra- Event Relevance," *in Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, Sydney, pp. 369-376, 2006.

[10] Lin C., "ROUGE: a Package for Automatic Evaluation of Summaries," *in Proceedings of the Workshop on Text Summarization Branches* Out *Post-conference workshop of ACL*, pp. 74-81, Spain.

[11] Lin, C.-Y. (2003) "ROUGE: Recall-Oriented Understudy for Gisting Evaluation", http://www.isi.edu/_cyl/ROUGE/.

[12] Lin C. and Hovy E., "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics," *in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language*, USA, pp. 71–78, 2003.

[13] Litvak M. and Last M., "Graph-Based Keyword Extraction for Single-Document Summarization," *in Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization Coling*, Manchester, pp. 17-24, 2008.

[14] Liu Y., Wang X., Zhang J. and Xu H., "Personalized PageRank Based Multi-document Summarization," *in Proceedings of IEEE International Workshop on Semantic Computing and Systems*, Huangshan , pp. 169-173, 2008.

[15] Luhn H., "The Automatic Creation of Literature Abstracts," *IBM Journal of Research Development*, vol. 2, no. 2, pp.159-165, 1958.

[16] Mani I. and Maybury M., *Advances in Automatic Summarization*, MIT Press, Cambridege, 1999.

[17] Mihalcea R. and Tarau P., "A Language Independent Algorithm for Single and Multiple Document Summarization," *in Proceedings of IJCNLP* 2005.

[18] Mihalcea R. and Tarau P., "TextRank: Bringing Order into Texts," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Spain, pp. 404-411, 2004.

[19] Zahri N. and F Fukumoto., "Multi-document Summarization Using Link Analysis Based on Rhetorical Relations between Sentences," *in Proceedings of international conference on Computational Linguistics and Intelligent Text Processing*, vol. 6609/2011, Berlin, pp.328-338, 2011.

[20] Over,P., Dang,H. and Harman,D., "DUC in Context," *Information Processing and Management,* vol. 43, no. 6, pp. 1506-1520, 2007.

[21] Page L., Brin S., Motwani R. and Winograd T., "The PageRank Citation Ranking: Bringing Order to the Web", *Technical report,* Stanford InfoLab, 1998.

[22] Patil K. and Brazdil P. "Sumgraph: Text Summarization Using Centrality In The Pathfinder Network", *IADIS International Journal on Computer Science and Information Systems*, vol. 2, no. 1, pp. 18-32, 2007.

[23] Porter M., "An Algorithm for Suffix Stripping," *Program: Electronic Library and Information Systems*, vol. 14, no. 3, pp.130-137, 1980.

[24] Quinn T., Christophe C. and Charles D., "Applications of Data Mining in Software Engineering," *International Journal of Data Analysis Techniques and Strategies*, vol. 2, no.3 pp. 243-257, 2010.

[25] Radev D. and Tam D., "Summarization Evaluation Using Relative Utility," *in Proceedings of the Twelfth International Conference on Information and knowledge Management CIKM*, USA, pp. 508-511, 2003.

[26] Radev D., Jing H., Stys M. and Tam,D., "Centroid-Based Summarization of Multiple Documents," *Information Processing and Management*, vol. 40, no. 6, pp. 919-938, 2004.

[27] Sjobergh J., "Older versions of the ROUGEeval Summarization Evaluation System were Easier to Fool," *Information Processing and Management*, vol. 43, no. 6, pp. 1500-1505, 2007.

[28] Wan X., "TimedTextRank: Adding the Temporal Dimension to Multi-Document Summarization," *in Proceedings of SIGIR*, Amsterdam, pp. 867-868, 2007.

[29] Wan,X., "An Exploration of Document Impact on Graph-Based Multi-Document Summarization," *in Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics*, Hawaii, pp. 755-762, 2008.

[30] Yeh J., Ke H., and Yang W., "ISpreadRank: Ranking Sentences for Extraction-Based Summarization Using Feature Weight Propagation in the Sentence Similarity Network," *Expert Systems with Applications*, vol. 35, no. 3, pp.1451-1462, 2008.

[31] Yaquan X. and Haibo Wang., "A New Feature Selection Method Based on Support Vector Machines for Text Categorisation," *International*

Journal of Data Analysis Techniques and Strategies, vol. 3, no.1, pp. 1 - 20, 2011.

[32] Zobia R., and Waqas A, "A Hybrid Approach for Urdu Sentence Boundary Disambiguation", *International Arab Journal of Information Technology*, vol. 9, no. 3, 2012.

Shanmugasundaram Hariharan received his B.E degree specialized in Computer Science and Engineering from Madurai Kammaraj University, Madurai, India in 2002, M.E degree specialized in the filed of Computer Science and Engineering from Anna University, Chennai, India in 2004. He holds his Ph.D degree in the area of Information Retrieval from Anna University, Chennai, India. He is a member of IAENG, IACSIT, ISTE, CSTA and has 7 years of experience in teaching. Currently he is presently working as Associate Professor in Department of Computer Science and Engineering, Trichy-621105, India. His research interests include Information Retrieval, Data mining, Opinion Mining, Web mining. He has to his credit 60 papers in referred journals and conferences. He also serves as editorial board member and as program committee member for several international conferences.

Thirunavukarasu Ramkumar is currently working as professor in the Department of Computer Applications, A.V.C.College of Engineering, Mayiladuthurai.He has received Ph.D degree in Computer Applications during the year December'2010 from Anna University, Chennai.His area of specialization includes Knowledge discovery from multiple databases and Object Computing. He is the fellow member of ISTE

Rengaramanujam Srinivasan born in 1940 in Alwartirunagari, Tamilnadu, India, received B.E. degree from the University of Madras, Chennai, India in 1962, M.E. degree from the Indian Institute of Science, Bangalore, India in 1964 and Ph.D. degree from the Indian Institute of Technology, Kharagpur, India in 1971. He is a member of the ISTE and a Fellow of Institution of Engineers, India. He has over 40 years of experience in teaching and research.