International Journal of Computer Theory and Engineering, Vol. 1, No. 5, December, 2009
1793-8201

Studies on Graph based Approaches for Single and Multi Document Summarizations

Shanmugasundaram Hariharan and Rengaramanujam Srinivasan

***Abstract*—This paper focuses attention on summarizing news articles using graph based approaches. The foundation for the graphical techniques is the adjacency matrix evaluated based on a suitable similarity measure between the sentences of the document. Two techniques, cumulative sum proposed by us and the degree of centrality method proposed by Erkan et al. are investigated. We also propose a recursive method of repeatedly using the above two methods after discounting the already selected sentences. We introduce two metric Effectivenes1 and Effectiveness2 to evaluate the summaries prepared by the system in comparison to the 'golden standard' summary prepared by the human judges. Comprehensive investigations with single and multi document summaries show that the discounting methods are superior to their basic counterparts and provide promise and scope for further improvements.**

***Index Terms*—Single/multi document summarization; similarity measure; degree centrality; Effectiveness; evaluation**

I. INTRODUCTION

Graph based approaches have received considerable attention in the area of text summarization [1-5]. Erkan and Radev [4] have introduced three new measures for measuring the centrality or importance of the sentences. These are centrality degree, LexRank and Continuous LexRank. These methods are inspired from the prestige of social networks.

In all these methods a sentence in a document or in a cluster of documents is represented by a vertex node. The similarities between sentences - based upon a suitable similarity measure, are represented as links, with link weights corresponding to similarity values. Overlap and cosine measures are close competitors for the similarity measure. We have shown earlier [6] that cosine similarity measure is superior and we have adopted the same through out this paper. We have used cosine similarity measure without and with the incorporation of Inverse Document Frequency (IDF) factor. Thus the representation of a document or a set of documents will be by its symmetric adjacency matrix.

The paper is organized as follows. Section II details the graph based approaches used in this paper, while Section III discusses the evaluation of summaries. Section IV describes the experimental setup and in Section V results of our summarization study are analyzed. Section VI discusses the related research work and finally Section VII lists the conclusions.

II. GRAPH BASED APPROACHES

In this section we discuss four graph based methods. They are (i) Cumulative similarity proposed by us (ii) Degree of Centrality proposed by Erkan et al.[4] . (iii) and (iv) Recursive methods of discounting already selected sentences corresponding to methods (i) and (ii).

Let us illustrate the above four methods using a pseudo document having 4 sentences. Fig. 1 presents a graphical representation of the document while Table I gives the adjacency matrix. The entries in the matrix correspond to the similarities between sentences. Thus sentences 1 and 2 have a similarity of 0.4. Let us explain the working of each of the four methods using the data presented in Fig. 1 and Table I.

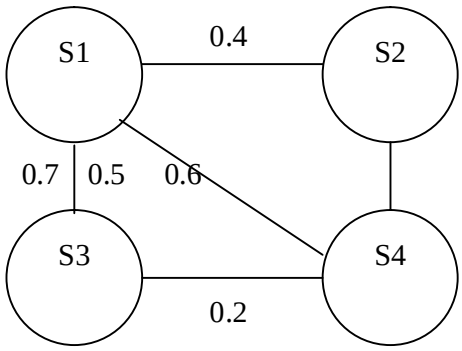

Fig 1: Representation of pseudo document by a graph

TABLE I: ADJACENCY MATRIX CORRESPONDING TO FIG 1

				Sum	Degree
1.00	0.40	0.70	0.50	2.6	4
0.40	1.00	0.00	0.60	2.0	3
0.70	0.00	1.00	0.20	1.7	2
0.50	0.60	0.20	1.00	2.3	3

Shanmuasundaram Hariharan is with Department of Information Technolgy, B.S.Abdur Rahman University,Chennai,Tamilnadu,India.(Phone :04422751347, Mobile: 9884204036, He is working as Assistant Professor and currently pursuing his doctoral studies in the area of Information Retrieval.

Rengaramanujam Srinivasan is with B.S.Abdur Rahman University,Chennai,Tamilnadu,India. Currently he is working as Professor in Department of Computer Science and Engineering)

A. Method 1: Cumulative Sum Method

For a moment let us assume that Table I does not represent a document but present data about a set of four players and their preferences for captaincy. We assume a perfect democratic system, where each player can recommend any number of players with varying degrees of recommendation*. Thus the captain is to be determined by a preferential list supplied by each member of the team. Each player can cast a vote, varying from 0 to 1 based upon her preference. She also has the freedom to assign zero to any of the player. Let us assume all the players uniformly name themselves for the captaincy (vote: 1) or totally exclude themselves from consideration (vote: 0). Given the above scenario, which player will be selected as captain? Clearly 1 and 4 are contenders and 1 wins by a margin of 0.3. We adopt the very same approach in picking up the most salient sentences. Thus in Method 1, any sentence weight is obtained by adding all the entries in the similarity matrix, corresponding to the sentence, either row wise or column wise. Since similarity matrix is symmetric row or column addition will yield the same result. The link weight can be considered as recommendation of one sentence by another and thus importance of a sentence is given by summation of link weights. For the four sentence pseudo document case with r = 50%, sentences 1 and 4 will thus be picked up.

B. Method II: Degree of Centrality Method

Let us now consider degree of centrality method with a specified threshold proposed by Erkan et al. Here "centrality degree" of any node is the number of edges incident on the vertex, with link weight greater than or equal to specified threshold. The idea behind this approach is to eliminate link weights which have too low values – possibly noisy signals. If we choose a too high threshold the graph is not at all connected and becomes a set of islands. For example if we choose threshold > 0.7, we get an unconnected graph with 4 vertices.

If we choose a threshold value of 0.3, the centrality degree of 4 sentences are (4, 3, 2 and 3). Again sentence 1 is the top notcher and gets automatically selected. For a compression ratio of 50% sentences 2 or 4 can be selected. The tie between two sentences can be resolve by slightly increasing or decreasing the threshold value and based on the revised centrality degree obtained for the tied sentences. For example if we increase the threshold to 0.5, the degree of sentence 2 deceases to 2, while that of sentence 4 remains as 3. Hence sentence 4 can be selected.

Alternatively the tie between the two sentences can be resolved based on the position occupied by the sentence in the document. Since we are investigating news paper documents we have adopted this approach and has given preference to sentences that appear earlier in the document.

C. Method Iii: Discounted Cumulative Sum Method

Method III is similar to Method I. We form the cumulative sum, select the sentences with the highest score. There after, we remove the sentence from further consideration, by striking out rows and columns corresponding to the selected sentences. Thus after selection of sentence 1 and striking out the corresponding row and column, the modified adjacency matrix is shown in Table II. The cumulative sum values for sentences 2, 3 and 4 are (1.6, 1.2, and 1.8). Hence sentence 4 will be selected and so on.

TABLE II: MODIFIED ADJACENCY MATRIX

				Sum	Degree
1.00	0.40	0.70	0.50	-	-
0.40	1.00	0.00	0.60	1.6	2
0.70	0.00	1.00	0.20	1.2	1
0.50	0.60	0.20	1.00	1.8	2

The idea behind the discounting technique is that, once a sentence is selected, we need not select sentences which are very close to the selected sentence. Thus we ensure that the information in the selected sentence is less likely to repeat. However we would like to point out that this is unlike the real world example of captain selection by a consensus list, where once a captain is chosen, definitely she will have a stronger say in the selection of vice captain. We have found that the selection of sentences by the discounting techniques agrees more closely with the selection by the panel of judges, as compared to the basic method.

D. Method IV: Discounted Degree Centrality Method

This method is a modification of the degree centrality method. Thus with a threshold of 0.3, the degree centrality of the four sentences are (4, 3, 2, 3). After selecting the first sentence, we remove the corresponding row and column. The revised adjacency matrix is given in Table II. With the remaining elements of the matrix, the degree centrality of the sentences are (2, 1, 2). The tie between the sentences 2 & 4 can be resolved suitably as already explained.

III. EVALUATION OF SUMMARY

Precision has long been used as a metric for evaluating of extraction summaries. If S_{sum} denotes the sentences selected by the summarizer and S_{judges} denotes the sentences selected by the judges Precision is defined as:

$$Precision = \frac{|S_{sum} \cap S_{judges}|}{Sentences\ \ extracted} \quad (1)$$

where $|\ |$ denotes a count measure. Precision is rather pessimistic approach that does not take into account the information contribution of the sentences that are selected by the summarizer, but not found in the list provided by the panel of judges.

In order to overcome the difficulty we had suggested earlier [7] a method of preparation of 'golden standard summary' and evaluating on the basis of effectiveness factor defined by us. In this approach the judges are asked to rank the sentences in the document in the order of perceived importance. For a 10 sentence document case the ranking given by the judges are presented in Table III.

We calculate the weight W_{ij} of any sentence 'i' as evaluated by the judge 'j' as:

$$wij = \frac{n+1-R_{i,j}}{n} \qquad (2)$$

where $R_{i,j}$ is the rank of the i[th] sentence under consideration, given by judge'j'. The cumulative weight of the sentences is obtained as:

$$W_i = \sum_j W_{ij} \qquad (3)$$

For a compression ratio 'r', n*r sentences with top cumulative scores are picked up. Thus with r =0.3, a perusal of Table III shows that sentences 3, 1 & 4 are picked up in that order.

If a summarizer has done the worst selection and has picked up sentences 8, 9 & 10 for the summary, the effectiveness for the selection will be (8+9+4)/ (27+25+22) = 21/74 = 28.3 %. We find that the above method of allocating weights to sentences presents too optimistic effectiveness factors.

Therefore we propose an alternative formula for the allotment of weights to sentences. Under the new proposal, w_i is given by

$$w_{ij} = [\alpha]^{R_{i,j}-1} \qquad (4)$$

$$W_i = \sum_j w_{ij} \qquad \text{(as before)}$$

Here α is a user chosen parameter, the value of which ranges between 0 and 1 (0 < α < 1). We had experimented with various values of α and have found that α =0.5 is satisfactory. With α =0.5, the first ranked sentence will have a weight of 4 times that of the third sentence and twice the weight of the second sentence. This approach is also commensurate with the real world situation, where the first prize amount often is twice the second prize amount and so on. Again using 30% compression, adopting expression (4) sentences 1, 3 and 5 will be selected. The effectiveness for a pick up of sentences 8, 9 & 10 will be (0.0082+0.0243+0.0243)/ (2.032+1.5+1.0786) = 0.0568 / 4.6106 = 1.2 %. Thus poor performance is put on a proper perspective.

There is also another benefit arising out of expression (4). When we asked the judges to rank all the sentences of the document, there was a close agreement in ranking the top sentences. However lower down the order, the judges found it difficult to rank the sentences. Therefore for quiet number of studies we had requested them to rank the specified percentage of sentences only. Adopting this approach in Table III we have given the cumulative weights of the sentences (in column marked as M3). The sentence picked up by this scheme is again 1, 3 & 4. The Effectiveness factor for a summarizer which picks up sentences 8, 9 & 10 will be 0/ 4.5 = 0. Thus the expression (4) appears to be elegant and consistent. In Table III, cumulative score M1 and M2 are calculated using expression (2) & (4) respectively. M3 again uses expression (4) but ignores ranking beyond the required number of sentences.

In order to distinguish between two effectiveness factors, we call effectiveness calculated using expression (2) as Effectiveness1 (E1) and using expression (4) as Effectiveness2 (E2). We have presented results corresponding to both effectiveness factors in the paper.

TABLE III: COMPARISON OF ALTERATIVE METHODS OF ALLOCATING WEIGHTS TO SENTENCES

Sentence Number	Ranking by judges			Cumulative score			Selected sentences		
	J1	J2	J3	M1	M2	M3	M1	M2	M3
1	1	1	6	25	2.0322	2	II	I	I
2	3	3	10	17	0.5021	0.5			
3	2	2	2	27	1.5000	1.5	I	II	II
4	4	4	3	22	0.5000	0.25	III		
5	5	7	1	20	1.0786	1		III	III
6	7	6	4	16	0.1733	0			
7	6	5	5	17	0.1572	0			
8	10	10	9	4	0.0082	0			
9	8	8	8	9	0.0243	0			
10	9	9	7	8	0.0243	0			

IV. EXPERIMENTAL SETUP

This section details the corpus used and the particulars of the experiments carried out for single and multi document summarizations. We focus on summarizing the contents at 10%, 20% and 30% compression ratios.

A. Corpus Description:

In order to obtain a target set of ideal results, we distributed document sets to different judges and requested them 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

single document experiments we have chosen at random 65 documents and for multi document experiments we had 50 document set pairs. All the documents are news reports.

B. System Description:

The four methods for summarization are explained in detail using a sample news report given in Fig. 2. The corresponding inter sentence similarity matrix is given in Table IV. Though several measures are available for measuring the similarity, two measures cosine and overlap have been used widely for text summarization task. Of the two measures cosine is superior because it provides standard baselines [9]. Cosine measure reflects the degree of similarity in corresponding terms and term weights, while overlap measures the degree to which two sets overlap. Comparing the two metrics overlap measure takes the min operator and provides higher magnitude than cosine [10]. Further cosine is independent of length, but overlap measure greatly varies depending on length. We have used throughout this paper cosine measure only as given by expression (5), after eliminating stop words and stemming [8].

$$Cosine(t_i,t_j)=\sum_{h=1}^{k} t_{ih}t_{jh}/\sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2} \qquad (5)$$

Here i,j refers to the i^{th} and j^{th} sentences of the document. The above expression is without incorporation of IDF. With the incorporation of IDF, the expression for the similarity will be

$$Cosine(t_i,t_j)=\sum_{h=1}^{k} t_{ih}t_{jh}*(idf_h)^2 / \sqrt{\sum_{h=1}^{k} (t_{ih}*idf_h)^2 \sum_{h=1}^{k} (t_{jh}*idf_h)^2} \qquad (6)$$

1. New Delhi, November 29: The strike called by AIIMS doctors in protest against the passage of a bill restricting the retirement age of its Director at 65 years on Thursday saw muted response with the medicos backing institution head P.Venugopal abstaining from work while those opposing him attending OPD and other services.
2. By and large the OPDs functioned normally with only some doctors who want the present Director to continue in office staying away from work.
3. The AIIMS administration led by Medical Superintendent D K Sharma had, following the Supreme Court orders, asked the resident doctors to refrain from striking work.
4. All OPDs are functioning normally- There is no problem at all i am supervising the OPDs operations," Sharma said, adding "a few doctors have boycotted (the OPDs)".
5. The strike was called by Resident Welfare Association on Wednesday when the AIIMS bill was passed in the Rajya Sabha amid wide protests from opposition benches.
6. Anil Sharma, General Secretary, RDA, AIIMS, said, "we have boycotted the OPDs and the future course of action would be decided at a GoM later today.
7. The bill seeks to restrict the age of the AIIMS Director at 65 years, the move may pave the way for removal of Venugopal from the post- Union Health Minister A Ramdoss has been in the last two-and-a-half years locked in a turf war with Venugopal.
8. AIIMS spokesperson Shakti Kumar Gupta said that all the centres in the hospital were functioning normally.
9. The Director has appealed to all faculty members, resident doctors, nurses, officers and all the staff to continue their work and activities including teaching and research," he said, adding the functioning of OPDs has not been affected.
10. We will review the situation in the afternoon and then we will decide the further course of action," he said, adding action would taken against those staying away from work.

Fig 2: Sample news paper report (SDS28)

TABLE IV: COSINE SIMILARITY FOR DOCUMENT SDS28

1.000	0.157	0.189	0.042	0.173	0.049	0.255	0.063	0.118	0.051	2.097	1.94	0.049
0.157	1.000	0.139	0.310	0.000	0.072	0.053	0.092	0.347	0.222	**2.392**	-	-
0.189	0.139	1.000	0.112	0.172	0.129	0.048	0.083	0.156	0.067	2.095	1.956	1.827
0.042	0.310	0.112	1.000	0.000	0.289	0.000	0.149	0.280	0.060	2.242	1.932	1.643
0.173	0.000	0.172	0.000	1.000	0.059	0.044	0.076	0.048	0.000	1.572	1.572	1.513
0.049	0.072	0.129	0.289	0.059	1.000	0.050	0.086	0.054	0.276	2.064	**1.992**	-
0.255	0.053	0.048	0.000	0.044	0.050	1.000	0.064	0.120	0.000	1.634	1.581	1.531

0.063	0.092	0.083	0.149	0.076	0.086	0.064	1.000	0.070	0.000	1.683	1.591	1.505
0.118	0.347	0.156	0.280	0.048	0.054	0.120	0.070	1.000	0.111	2.304	1.957	**1.903**
0.051	0.222	0.067	0.060	0.000	0.276	0.000	0.000	0.111	1.000	1.787	1.565	1.289

Using Table IV the selection of sentences by the four methods can be done. For Method I the cumulative scores for sentences are given as {2.09,2.39,2.10,2.24,1.57,2.07,1.62,1.68,2.30,1.79,}. Based on these scores, we pick up sentences 2, 9 and 4. For Method III, at 30% compression ratio sentences 2, 6 and 9 with cumulative scores {2.39, 2.00, 1.91} are chosen.

For degree centrality Method II, adopting a threshold weight of 0.10, the centrality degree of the sentences are {2,4,1,4,1,3,2,1,3,3}. Hence we choose sentences 2, 4 and 6 for the summary. For discounted Method IV the choice of the sentences would be 2, 4 and 1.

V.STUDY RESULTS AND DISCUSSION

We present the results of study made by us, for both single and multi document summarizations, using the four graph based methods. Results are presented using conventional measure of precision as well as using the two metrics proposed by us Effectiveness1 (E1) and Effectiveness2 (E2). In all the cases, values presented are average values obtained as an average for the entire document set under consideration.

A. Choice of Threshold on Degree-based Selection:

For degree based methods of Method II and Method IV, we had made studies, keeping the thresholds at 0.05, 0.10, 0.20, 0.30, 0.40 and 0.50. If we keep the threshold very high, there is very little coupling between the sentences and each sentence becomes an island with no connectivity between other sentences. If we keep the threshold too low, 'noise' words may give unintentional coupling, and either case leads to poor selection of sentences for the summary. We have found a threshold of 0.1 is quite satisfactory and we are presenting the results corresponding to this threshold.

B. Effect of IDF Factor:

We present in Table V and Table VI, comparison of performances of the four methods, with the incorporation of IDF and without incorporation of IDF in the evaluation of cosine similarity. From a perusal of the Tables we clearly find, that both precision and effectiveness values are uniformly higher, with the incorporation of IDF for all compression ratios. The rest of the results are presented with the incorporation of IDF only.

TABLE V: METHODS I AND III WITH AND WITHOUT IDF

Evaluation System Adopted	Method Adopted	Approach Used	Without IDF			With IDF		
			10%	20%	30%	10%	20%	30%
Precision	Method I	Single	0.192	0.418	0.490	0.315	0.424	0.532
	Method III		0.223	0.500	0.615	0.325	0.433	0.557
	Method I	Multi	0.239	0.317	0.415	0.339	0.397	0.476
	Method III		0.264	0.354	0.424	0.354	0.435	0.494
Effectiveness (E1)	Method I	Single	0.679	0.739	0.768	0.719	0.755	0.768
	Method III		0.712	0.771	0.813	0.724	0.762	0.821
	Method I	Multi	0.396	0.482	0.543	0.500	0.579	0.619
	Method III		0.460	0.525	0.573	0.542	0.592	0.628

TABLE VI: METHODS II AND IV WITH AND WITHOUT IDF

Evaluation System Adopted	Method Adopted	Approach Used	Without IDF			With IDF		
			10%	20%	30%	10%	20%	30%
Precision	Method II	Single	0.462	0.526	0.591	0.562	0.613	0.647
	Method IV		0.492	0.533	0.611	0.569	0.617	0.673
	Method II	Multi	0.330	0.372	0.446	0.376	0.473	0.546
	Method IV		0.337	0.381	0.454	0.389	0.486	0.571
Effectiveness (E1)	Method II	Single	0.797	0.833	0.825	0.848	0.847	0.853
	Method IV		0.821	0.845	0.837	0.855	0.853	0.859
	Method II	Multi	0.533	0.554	0.609	0.545	0.581	0.612
	Method IV		0.542	0.563	0.613	0.556	0.601	0.642

C. Comparison of Methods:

Comparisons of performances of the four methods for 3 compression ratios are presented in Table VII and Fig. 3 for single document case and in Table VIII and Fig. 4 for multi document case. We have also investigated two baseline methods for each data set. The first scheme is picking up randomly the required number of lines from the document or document cluster corresponding to single and multi document cases. Five random runs were performed and the

average of these is given as random performance. The second scheme is lead based, i.e. with a compression ratio 'r', the first n*r sentences are picked up. For news document this method is extremely popular. From the above mentioned Tables and Figures the following conclusions can be drawn.

a. All the four methods are superior to random selection of sentences.
b. Discounted methods i.e. Method II and IV are superior to their corresponding basic methods- Method I and Method III.
c. Method II is superior to Method III and best performance is obtained with Method IV for both single and multi document cases.

TABLE VII: EVALUATION OF SINGLE DOCUMENT SUMMARIZATION

Compression Ratio	Evaluation Measure	Method I	Method II	Method III	Method IV	Random	Lead
10%	E1	0.719	0.848	0.724	0.855	0.574	0.978
	E2	0.439	0.604	0.464	0.619	0.423	0.953
	P	0.315	0.562	0.325	0.569	0.198	0.846
20%	E1	0.755	0.847	0.762	0.853	0.629	0.958
	E2	0.531	0.650	0.537	0.667	0.487	0.893
	P	0.424	0.613	0.433	0.617	0.325	0.817
30%	E1	0.768	0.853	0.821	0.859	0.669	0.947
	E2	0.629	0.724	0.654	0.743	0.521	0.896
	P	0.532	0.647	0.577	0.673	0.423	0.827

TABLE VIII: EVALUATION OF MULTI DOCUMENT SUMMARIZATION

Compression Ratio	Evaluation Measure	Method I	Method II	Method III	Method IV	Random	Lead
10%	E1	0.500	0.545	0.542	0.556	0.217	0.607
	E2	0.431	0.474	0.445	0.493	0.125	0.598
	P	0.339	0.376	0.354	0.389	0.068	0.380
20%	E1	0.579	0.581	0.592	0.601	0.270	0.643
	E2	0.526	0.564	0.533	0.575	0.207	0.621
	P	0.397	0.473	0.435	0.486	0.168	0.443
30%	E1	0.619	0.612	0.628	0.642	0.317	0.660
	E2	0.612	0.658	0.621	0.685	0.259	0.638
	P	0.476	0.546	0.494	0.571	0.247	0.533

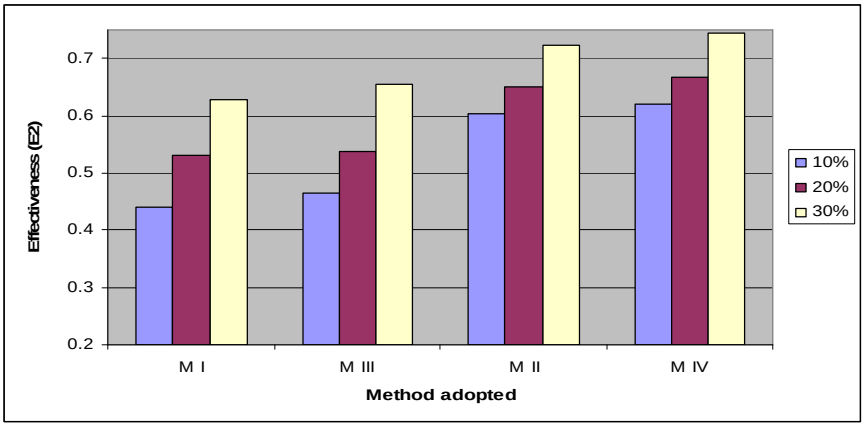

Fig 3 A Comparison of Effectiveness of the 4 methods (single document)

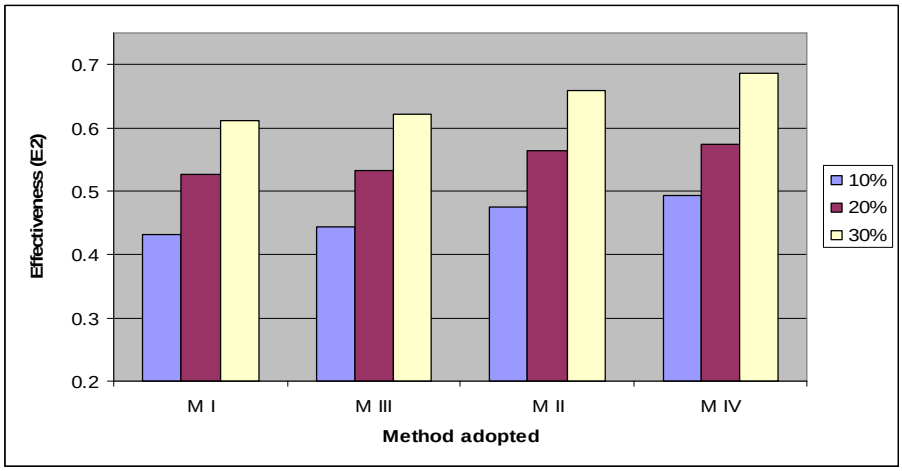

Fig 4: A Comparison of Effectiveness of the 4 methods (multi document)

d. In general, precision metric yields pessimistic values, Effectiveness1 (E1) yields optimistic values and Effectiveness2 (E2) values lie in between the two.
e. Method IV is a close competitor to lead based system, especially for multi document case.

VI. RELATED WORK

Marina and Mark (2008) introduced two novel approaches namely supervised and unsupervised methods for identifying the keywords to be used in extractive summarization of text documents. For graph-based approach, syntactic representation of text enhances the traditional vector-space model by taking into account some structural document features. In supervised approach, authors train the classification algorithms on a summarized collection of documents with the purpose of inducing a keyword identification model. In the unsupervised approach, HITS algorithm was run on document graphs under the assumption that the top-ranked nodes should represent the document keywords.

Rada et al., (2004) have demonstrated TextRank – 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. An important characteristic of the 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), but takes the information recursively drawn from the entire text (graph) into account.

Ohm Sornil et al., (2006) proposed an automatic summarization system combining content-based and graph-based features using Hopfield Network algorithm, where each node is activated in parallel and node weights were combined for each individual node. In the first stage, segments are represented by content-based feature vectors. The segment-feature matrix is then compressed into a lower dimensional matrix to uncover hidden association patterns and reduce small variations in segment characteristics by using Singular Value Decomposition (SVD). In second stage, segments are represented as nodes, and relationships between two segments whose similarity scores above a threshold are represented as edges in a document graph. The configuration that gives the best summarization performance is the undirected document graph constructed from cosine similarity.

Gunes Erkan and Dragomir (2004) introduced a stochastic graph-based method for computing relative importance of textual units for Natural Language Processing. A new approach called LexRank, for computing sentence importance based on the concept of eigenvector centrality in a graph representation of sentences. Connectivity matrix based on intra-sentence cosine similarity is used as the adjacency matrix of the graph representation of sentences in the proposed approach. Also it is found that degree-based methods (including LexRank) outperform centroid-based methods.

Xiaojun Wan et al.,(2006) describes an affinity graph based approach to multi-document summarization. Proposing an integrated framework for considering both information richness and information novelty of a sentence based on sentence affinity graph.

Rada and Paul (2005) shows how a meta-summarizer relying on a layered application of graph-based techniques for single-document summarization can be turned into an effective method for multi document summarization. They represent the graph as: (a) simple undirected graph; (b) directed weighted graph with the orientation of edges set from a sentence to sentences that follow in the text (directed forward); or (c) directed weighted graph with the orientation of edges set from a sentence to previous sentences in the text (directed backward). Multi-document summaries for a document cluster are built using a "meta" summarization procedure. For each document in the cluster of documents, a single document summary is generated using one of the graph-based ranking algorithms, followed by "summary of summaries" is produced using the same or a different ranking algorithm.

VII. CONCLUSIONS

We have investigated two basic graph based methods for summarizing single and multi documents- with two

variations, with or without discounting the selected sentences. All the four methods are promising in that they yield superior results as compared to random selection, based on conventional precision metric as well as proposed metrics Effectiveness1 and Effectiveness2. The discounting methods proposed are superior to their basic counterparts. Finally Method IV of selecting sentences based up on the degree of node (with a threshold of 0.10) and with discounting emerges as a close competitor to the lead based schemes. This is true especially for multi document case. We have used cosine measure of similarity and no other additional features. There is a good scope for further improvement by combining some additional features. We also propose to investigate the variations of graphical techniques based on page rank type algorithms.

ACKNOWLEDGEMENT

The authors would like to express their thanks to Pro. Vice Chancellor Mr. Abdul Qadir A. Rahman Buhari, Vice Chancellor Dr.P.Kanniyappan, Registrar Dr. V.M.Periasamy, Former HOD/CSE Prof. Manu Natarajan and Dean & HOD/IT Dr.T.R.Rangaswamy for the environment provided.

REFERENCES

[1] Marina Litvak and Mark Last (2008) , "Graph-Based Keyword Extraction for Single-Document Summarization", Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization, pp. 17–24.

[2] Rada Mihalcea and Paul Tarau, An Algorithm for Language Independent Single and Multiple Document Summarization, in Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), Korea, October 2005

[3] Ohm Sornil and Kornnika Gree-ut (2006), "An Automatic Text Summarization Approach using Content-Based and Graph-Based Characteristics", IEEE.

[4] Gunes Erkan and Dragomir R. Radev (2004), "LexRank: Graph-based Lexical Centrality as Salience in Text Summarization", Journal of Artificial Intelligence Research,pp. 457-479.

[5] Xiaojun Wan and Jianwu Yang (2006), "Improved Affinity Graph Based Multi-Document Summarization", Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pp. 181–184.

[6] Shanmugasundaram Hariharan and Rengaramanujam Srinivasan (2008a), "A Comparison of Similarity Measures for Text Documents", Journal of Information & Knowledge Management, Vol. 7, No. 1, pp. 1–8.

[7] Shanmugasundaram Hariharan and Rengarmanujam Srinivasan (2008b), "Investigations in Single document Summarization by Extraction Method", In Proceedings of IEEE International Conference on Computing, Communication and Networking (ICCCN'08).

[8] M.F. Porter (1980), "An algorithm for suffix stripping", Program, 14(3) pp 130−137, 1980.

[9] Mani.I and Bloderm.E,(1999), "Summarizing Similarities and Differences among related documents". Information Retrieval**,** Volume 1 , Issue 1-2, pp. 35-67.

[10] Michael W.Berry, Murray Brown (2006), "Lecture notes in Data Mining", World Scientific Publishing.

[11] Rada Mihalcea and Paul Tarau, 'TextRank: Bringing Order into Texts', in Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2004), Barcelona, Spain, July 2004

Shanmugasundaram Hariharan -- born in 1981 in Tiruchirapalli, Tamilnadu, India, received his B.E degree 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 is pursuing his Ph.D degree in the area of Information Retrieval at Anna University, Chennai, India. He is presently working as Lecturer in Department of Information Technology at B.S.Abdur Rahman University, Chennai, India. His research interests include Information Retrieval and Data mining.

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. He is presently working as a Professor of Computer Science and Engineering at B.S.Abdur Rahman University, Chennai, India and is supervising doctoral projects in the areas of data mining, wireless networks, Grid Computing, Information Retrieval and Software Engineering.