IACSIT International Journal of Engineering and Technology Vol. 2, No.1, February, 2010
ISSN: 1793-8236

Enhancements to Graph based methods for Single Document Summarization

Shanmugasundaram Hariharan and Rengaramanujam Srinivasan

***Abstract*— This paper focuses its attention on extractive summarization using popular graph based approaches. Graph based methods can be broadly classified into two categories: non- PageRank type and PageRank type methods. Of the methods already proposed - the Centrality Degree method belongs to the former category while LexRank and Continuous LexRank methods belong to later category. The paper goes on to suggest two enhancements to both PageRank type and non-PageRank type methods. The first modification is that of recursively discounting the selected sentences, i.e. if a sentence is selected it is removed from further consideration and the next sentence is selected based upon the contributions of the remaining sentences only. Next the paper suggests a method of incorporating position weight to these schemes. Thus we have experimented with 12 methods –six of non- PageRank type and six of PageRank type. To clearly distinguish between various schemes, we call the methods of incorporating discounting and position weight enhancements over Lexical Rank schemes as Sentence Rank (SR) methods. Intrinsic evaluation of all the 12 graph based methods were done using conventional Precision metric and metrics earlier proposed by us - Effectiveness1 (E1) and Effectiveness2 (E2). Experimental study brings out that the proposed SR methods are superior to all the other methods.**

***Index Terms*—Page rank, lexical rank, sentence rank, recommendation, degree, damping, threshold, effectiveness, discounting**

I. Introduction

Text summarization has been an important and challenging area studied almost over the past 5 decades [10, 12] and has continued to be a steady subject of research. 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 back ground knowledge, and domain independency are some of the features that favour extraction, where as abstraction, which is domain dependent in nature, requires human knowledge and is goal oriented [9].

Extractive Summarization requires ranking sentences according to their importance. The traditional method of determining sentence importance is based upon product of term frequency and inverse document frequency (tf * idf), position weight and other parameters [14]. Graph based methods for summarization modeled on the basis of social network have been proposed [7, 11] and successfully implemented. This paper focuses its attention on graph based methods.

Graph oriented summarization methods are modeled on 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 his contacts are highly placed. 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. Erkan and Radev [7] have presented in their excellent paper 3 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. We propose enhancements to the above methods and show that with enhancements summarizer performance is vastly improved.

The rest of the paper is organized as follows. Section II describes the Centrality Degree and Lexical Rank methods already developed and the proposed enhancements. In Section III an example is worked out to illustrate all the methods. Section IV deals with experimental investigations while Section V discusses related work. Finally Section VI gives the conclusions.

II. Proposed Enhancements

Our enhancements rest on the foundations of graph based approaches already developed by Erkan and Radev [7]. They have proposed a basic method called Degree Centrality and two Page Rank type methods called LexRank and continuous LexRank. A brief description about these methods is in order.

A document can be considered as a network of sentences

Manuscript received August 29, 2009. Accepted November 24, 2009.

Shanmuasundaram Hariharan is with the Department of Information Technology, B.S.Abdur Rahman University, Chennai, Tamilnadu, India.(Phone :04422751347, Mobile: +91-9884204036, E-mail : mailtos.hariharan@gmail.com) Currently he is working as Assistant Professor and pursuing his doctoral programme 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 (E-mail: drrsrs@yahoo.com)

that are related to each other. They hypothesize that the sentences that are similar to many of the other sentences in the document are more important. The similarity between the two pairs of sentences x and y is determined by the cosine between the two sentence vectors as modified by the inverse document frequency. Though there exists several measures to evaluate the strength of relationship among the sentences, cosine metric is found to be popular and more superior than others [13] as given by expression (1).

$$idf\_\mathrm{mod}ified\_Co\sin e(x,y)=\frac{\sum_{w\in x,y} tf_{w,x}*tf_{w,y}*(\mathrm{idf}_w)^2}{\sqrt{\sum_{x_i\in x}(tf_{x_i,x}*\mathrm{idf}_{x_i})^2}*\sqrt{\sum_{y_i\in y}(tf_{y_i,y}*\mathrm{idf}_{y_i})^2}} \quad (1)$$

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. This matrix is the basis for all the methods. Also the matrix is formed after removal of stop words and stemming the terms [17, 18].

The degree centrality of a sentence is simply, the number of links connected to the sentence, with link weight above a specified threshold. The higher the degree the more important the sentence will be and the method corresponds to popularity based democratic approach.

The two methods- LexRank with threshold and Continuous LexRank - are based on PageRank type algorithms and are given by the expressions (2) and (3).

$$\text{LexRank[i]} = \frac{d}{N} + (1\text{-d})*\sum_{j\in S[i]}\frac{LexRank[j]}{\deg[j]} \quad (2)$$

$$\text{Continuous LexRank[i]} = \frac{d}{N} + (1\text{-d})*\sum_{j\in S[i]}\frac{idf\_\mathrm{mod}ified\_Co\sin e(i,j)*\mathrm{PR}[j]}{\sum_{\kappa\in S[j]} idf\_\mathrm{mod}ified\_Co\sin e(j,k)} \quad (3)$$

where 'N' is the total number of sentences in the document and 'd' is the damping factor.

They have shown that, compared to Degree Centrality method, LexRank with threshold and Continuous LexRank methods fare well and outperform other centroid based methods[7].

We propose two enhancements to the above methods. The first method is discounting technique while the second method incorporates position weight to the above expressions. Let us discuss each one of these enhancements.

A. Discounting Method:

Discounting technique envisages that once a sentence is selected by any one of the methods, immediately corresponding row and column values of the matrix are set to zero. Thus the next sentence is selected from contributions made by the remaining (n-1) sentences only. The algorithm for the discounting method is given in Appendix – I. Discounting methods are applicable to both non-PageRank type as well as PageRank schemes. Thus when we use discounting technique to LexRank methods, we use expression (2) and (3) but each time we pick up only the top ranking sentence and modify the adjacency matrix 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.

B. Position Weight:

The location of a sentence in a document plays a significant part in determining the importance of a sentence. Therefore all earlier methods have incorporated position weight of the sentence in calculating the overall sentence weight [8, 14]. In the graph based approach, importance to position of the sentence can be given in a way, by using directed graphs instead of undirected graphs. Thus forward directed graph gives preference to earlier sentences. The experimental results corresponding to the observations are presented in Section IV-A.

However we wanted to incorporate the position weight separately, so that its parameters can be changed to suit the characteristics of the document. Thus the position factor (P_f) of any sentence 'i' is given by

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

Gama and beta are design parameters and beta lies between 0 to 1. The above expression gives importance to sentences that appear earlier in the document. If we want to give preference to end sentences we can use expression (5).

$$P_{f_i} = gama * Beta^{n-i} \quad (5)$$

A combination of expressions (4) and (5) can be used when we want to confer importance to sentences that appear in the first and last part of the document.

C. Sentence Rank Methods:

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 expressions for Sentence Rank with threshold is given in expression (6).

$$SR[i] = \frac{d}{N} + gama * beta^{i-1} + (1\text{-}d) * \sum_{j \in S[i]} \frac{SR[j]}{deg[j]} \quad (6)$$

Similarly the expression for continuous Sentence Rank is given by expression (7).

$$SR[i] = \frac{d}{N} + gama * beta^{i-1} + (1\text{-}d) * \sum_{j \in S[i]} \frac{idf - modified - Cosine(i,j) * SR[j]}{\sum_{\kappa \in S[j]} idf - modified - Cosine(j,k)} \quad (7)$$

In expressions (6) and (7), gama and beta are parameters which affect the position influence. Thus, with no discounting:

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

Table I shows relative effect of gama for a typical 9-sentence document shown in Table II. For the document sets illustrated in experimental section, beta =0.9 and gama = 0.2 were found to be satisfactory. We have adopted these two values throughout the rest of the paper wherever position weight feature is incorporated.

TABLE I: EFFECT OF GAMA FOR 9-SENTENCE DOCUMENT AT 30% COMPRESSION RATIO

gama ⟶	**beta $^{i-1}$**					**beta $^{n-i}$**			
	0.0	**0.1**	**0.2**	**0.3**	**1.0**	**0.1**	**0.2**	**0.3**	**1.0**
Method –I	1,8,3	1,8,3	1,3,8	1,3,8	1,3,2	9,3,1	9,3,8	9,3,8	9,8,7
Method –II	1,8,3	1,8,3	1,3,8	1,3,8	1,3,2	9,3,1	9,3,8	9,3,8	9,8,7

In all we are considering 12 methods –six of non-PageRank type and six of PageRank type. We list the methods as follows:

Non-PageRank type methods:

i. Cumulative Sum
ii. Degree centrality
iii. Discounted Cumulative Sum
iv. Discounted Degree centrality
v. Discounted Cumulative Sum with position
vi. Discounted Degree centrality with position

PageRank type methods:

vii. LexRank (threshold)
viii. Continuous LexRank
ix. Discounted LexRank (threshold)
x. Discounted Continuous LexRank
xi. Sentence Rank (threshold)
xii. Continuous Sentence Rank

Of the 12 methods, methods II, VII and VIII were proposed by Radev et al [7]. Methods I, III and IV were proposed by us earlier [19]. Methods V, VI, IX, X, XI and XII are being proposed by us now. We go on show that Method VI is the best in non- PageRank type methods while SR methods proposed by us are the best of all methods and Continuous SR method XII being superior to threshold based SR method XI.

TABLE II: ADJACENCY MATRIX FOR 9-SENTENCE DOCUMENT

1.000	0.000	0.217	0.067	0.081	0.060	0.087	0.095	0.223
0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.085
0.217	0.000	1.000	0.033	0.086	0.007	0.197	0.160	0.117
0.067	0.000	0.033	1.000	0.028	0.000	0.046	0.000	0.052
0.081	0.000	0.086	0.028	1.000	0.041	0.000	0.079	0.000
0.060	0.000	0.007	0.000	0.041	1.000	0.000	0.000	0.000
0.087	0.000	0.197	0.046	0.000	0.000	1.000	0.164	0.076
0.095	0.000	0.160	0.000	0.079	0.000	0.164	1.000	0.249
0.223	0.085	0.117	0.052	0.000	0.000	0.076	0.249	1.000

III. AN EXAMPLE

Let us illustrate all the 12 methods by considering a 9-sentence document. Table II presents the 9 x 9 cosine similarity matrix. This forms the basic dataset for all the methods. Tabulations for the working of the 12 methods are presented in Appendix – II. For PageRank type methods damping factor is set to 0.10. In methods where position

weight has been incorporated gama= 0.20 and beta =0.90.

A. M-I: Cumulative sum method

Consider the adjacency matrix shown in Table I. Here the weight of any sentence 'i' is obtained by adding all the entries in the similarity matrix, corresponding to the i^{th} sentence row wise or column wise. 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 9-sentence document case, with r = 30%, sentences 1, 3 and 9 will thus be picked up. (Appendix – II, Table A, Column 1).

B. M-II: Degree Centrality method

In this method "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. If we choose a threshold value of 0.10, the centrality degrees of top 3 sentences are 5, 4 and 4 (Table A, Column 2). Sentence 3 is the top notcher and gets automatically selected, followed by sentences 8 and 9. The tie, if any 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. M-III: Discounted Cumulative Sum method

Method III is similar to Method I. We form the cumulative sum, select the sentences with the highest score. Thus sentence 1 is selected (Table A, Column 1). There after, we remove the sentence from further consideration, by striking out row and column corresponding to the selected sentence and again obtain Cumulative sum. Thus sentence 8 is picked up (Table B, Column 1) with cumulative sum of 1.66. Repeating the procedure again sentence 3 is selected with cumulative sum = 1.44 (Table C, Column 1).

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. 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. M-IV: Discounted Degree Centrality method

Similar to Method II, this method picks up the first top ranked sentence, and then sets the corresponding row and column values to zero. In the next iteration, second sentence is picked up. Thus sentence 3 is picked up first. In the second round there is a tie between the sentences 8 & 9, which is resolved in favour of sentence 8 and finally sentence 1 is picked up (Tables A,B,C : Column 2).

E. M-V: Discounted Cumulative Sum with position

Discounted Cumulative Sum with position combines position with Method III. Column 3 of Tables A,B and C show that sentences 1,3 and 8 are selected.

F. M-VI: Discounted Degree centrality with position

Discounted Degree centrality with position combines position with Method IV, after converting "degrees" to relative weights. From column 4 of Tables A, B and C of Appendix- II, we find that sentences 3, 1 and 8 are picked up in that order.

G. M- VII: LexRank method

The sentence weights are calculated using the expression (2) and are presented in Column 5 of Table A. At 30% compression ratios, sentences 3, 8 and 9 are selected, with lexical scores are 1.000, 0.812 and 0.812. The lexical scores given are normalized by dividing each sentences weight with the maximum sentence weight, so that the top sentence score will be 1.

H. M-VIII: Continuous LexRank method

Continuous LexRank scores are calculated using expression (3) and the values are presented in Column 6 of Table A. From a perusal of values we find that sentences 1, 9 and 3 are picked up.

I. M- IX: Discounted LexRank method

This is similar to Method VII with discounting feature incorporated. Results are presented in Column 5 of Tables A,B and C, we find that sentences 3, 8 and 1 are picked up respectively.

J. M- X: Discounted Continuous LexRank method

This method is similar to Method VIII, except for the incorporation of discounting feature. Perusing Column 6 of Tables A, B and C we find sentences 1, 8 and 3 are selected.

K. M-XI: Sentence Rank (Degree)

This is modification of Method IX with the addition of position weight. Looking at column 7 of Tables A,B and C we find that sentences 2, 3 and 1 are chosen.

L. M-XII: Continuous Sentence Rank

This is modification of Method X, with the incorporation of position weight From a perusal of data presented in Column 8 (Tables A,B,C) we infer that sentences 1, 3 and 2 are selected in that order .

From a perusal of selected sentences by various methods, we find that sentence 9 is a member of all the four non-discounted methods, while it cannot find a place in any of the discounted methods. In Table III we present the precision values for all the 12 methods, corresponding to 10%, 20% and 30% compression ratios. The 'golden' target chosen by the judges for the 9-sentence document is 1, 3 and 8. We find that Methods III, IV , V, VI, IX and X achieve 100% precision at 30% compression ratio, while M-V achieves a 'perfect 10' in the sense that all compression ratios precision is 100%. This is rather fortuitous and actual performance comparison has to be based on an average values obtained over a collection of document set. This is attempted in the next section.

TABLE III: PRECISION FOR ALL THE METHODS FOR 9-SENTENCE CASE SHOWN IN TABLE II

Compression Ratio	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII
10%	1.00	0.00	1.00	0.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	1.00
20%	1.00	0.50	0.50	0.50	1.00	1.00	0.50	0.50	0.50	0.50	0.50	1.00
30%	0.66	0.66	1.00	1.00	1.00	1.00	0.66	0.66	1.00	1.00	0.66	0.66

IV. EXPERIMENTAL INVESTIGATIONS

A Study was conducted in order to clearly assess the relative performance of all the 12 methods. This study results presented throughout the rest of the paper is based on the average of 30 documents. The corpus were collected from news documents that are readily available from news service providers like Google News, Hindu, Indian Express, Deccan Herald and other news services[16]. The results of the summarizer were compared with the selection by a panel of judges, using the conventional metric of precision as well as by the two metrics proposed by us Effectiveness1 (E1) and Effectiveness2 (E2). The definitions of these metrics are given in Appendix-III. Before we could compare the relative performance of the various methods, the impact of important parameters on graph based algorithms is discussed.

A. Direction of Graph

A document having 'n' sentences is an interconnected graph, where the links between two sentences have weights based on cosine similarity between them. The graph can be represented as undirected, directed forward or directed backward. Undirected graph is symmetric in nature i.e. recommendation given by sentence 'Si' and 'Sj' and vice versa is the same. Direct forward have edge weights only where i > j and direct backward have edge weights only where i < j. For all the three categories we have included self weight, i.e. a sentence voting for itself. With this the diagonal elements in all the three cases will be 1. A study was conducted for the method II and the results are presented in Table IV. We find that for the news document, directed backward approach, giving importance to earlier sentences is better. However, we prefer to go in for undirected approach, which is unbiased towards position of the document. We incorporate position weight separately. Thus handling position of sentence separately enables us to design summarizer to any type of document and gives a good flexibility in the design.

B. Effect of Self weight:

The study presented in Table IV included the self weight, i.e. diagonal elements of adjacency matrix is set to 1. We did experiments removing self weight, i.e. all the diagonal elements were set to zero. The results are presented in Table V. Comparing Tables IV and V we infer that excluding self weight affects the summary evaluation at 30% compression marginally, leaving 10% and 20% compression ratios unaltered. We have included self weight in the rest of the studies presented.

TABLE IV: EFFECT OF GRAPH DIRECTION WITH SELF WEIGHT

Compression Ratio	Evaluation Measure	Graph direction		
		Undirected	Direct forward	Direct backward
10%	E1	0.620	0.421	0.642
	E2	0.593	0.242	0.621
	P1	0.511	0.187	0.555
20%	E1	0.637	0.463	0.682
	E2	0.623	0.284	0.657
	P1	0.538	0.215	0.580
30%	E1	0.739	0.502	0.762
	E2	0.728	0.318	0.751
	P1	0.638	0.287	0.664

TABLE V: EFFECT OF GRAPH DIRECTION WITHOUT SELF WEIGHT

Compression Ratio	Evaluation Measure	Graph direction		
		Undirected	Direct forward	Direct backward
10%	E1	0.620	0.421	0.642
	E2	0.654	0.242	0.621
	P1	0.450	0.187	0.555
20%	E1	0.637	0.463	0.682
	E2	0.623	0.284	0.657
	P1	0.538	0.215	0.580
30%	E1	0.757	0.513	0.774

	E2	0.738	0.327	0.762
	P1	0.650	0.299	0.681

C. Effect of Threshold

Degree Centrality refers to the degree of a sentence corresponding to a given threshold. We have experimented with various thresholds and have found that a threshold of 0.10 is quiet satisfactory [19]. So we adopt a threshold factor of 0.10 through out for studies involving threshold.

D. Effect of Damping

The PageRank type methods M-VII to M-XII are affected by damping factor. While it is recommended to adopt a damping factor in the interval [0.1 to 0.2], we present the results for damping factor of 0.10, 0.15 and 0.20 corresponding to methods VII and VIII in Table VI. For methods involving damping, we have kept a damping factor of 0.10. Damping factor in general flattens the weights assigned to the sentences and thus smoothens the distribution.

TABLE VI: VARIATION OF DAMPING AT THRESHOLD OF 0.10

Method	**Compression Ratio**	**Damping**								
		E1			**E2**			**P**		
		0.10	0.15	0.20	0.10	0.15	0.20	0.10	0.15	0.20
VII	10%	0.806	0.798	0.791	0.598	0.593	0.586	0.517	0.514	0.507
	20%	0.815	0.804	0.793	0.602	0.595	0.589	0.541	0.533	0.528
	30%	0.850	0.842	0.833	0.621	0.617	0.610	0.558	0.550	0.542
VIII	10%	0.812	0.802	0.794	0.627	0.620	0.614	0.523	0.518	0.511
	20%	0.851	0.844	0.837	0.628	0.621	0.617	0.552	0.546	0.539
	30%	0.876	0.868	0.861	0.645	0.633	0.628	0.593	0.586	0.579

E. Effect of Discount

Tables VII and VIII bring out the importance of discounting for non-PageRank type methods and PageRank type methods respectively. Thus from Table VII we find Method III is superior to Method I and Method IV is superior to M-II. Further we find M-IV is the best of all the four non- PageRank type methods. Similarly referring to Table VIII we find discounted methods M-IX and M-X are superior to their counter parts M-VII and M-VIII and of the four LexRank methods, M-X is the best of all. By combining the study results of Tables VII and VIII, we conclude that discounting methods are superior as compared to basic methods for both non-page rank and PageRank type formulations.

F. Effect of Position

Position factor plays a major role in determining the importance of a sentence in summarization tasks. We have set gama=0.2 and beta=0.9 and the results are presented in Table IX. Comparing the M-III values and M-V values of Tables VII and IX, we find M-V values incorporating position weights are superior; so also M-VI values are superior to M-IV. Further Method VI incorporating position weight and discounting over centrality degree methods is the best of all non-PageRank type methods I to VI. Similarly we find that SR(Degree) and SR(continuous) methods XI and XII are superior to all the other methods I to X.

TABLE VII: COMPARISON OF METHODS I TO IV TO ILLUSTRATE THE EFFECT OF DISCOUNTING

Compression Ratio	**Evaluation Measure**	**Method I**	**Method II**	**Method III**	**Method IV**
10%	**E1**	0.604	0.620	0.679	0.694
	E2	0.587	0.593	0.660	0.681
	P	0.483	0.511	0.527	0.557
20%	**E1**	0.614	0.637	0.740	0.777
	E2	0.600	0.623	0.731	0.765
	P	0.486	0.538	0.552	0.577
30%	**E1**	0.694	0.739	0.757	0.816
	E2	0.677	0.728	0.744	0.808
	P	0.529	0.638	0.642	0.671

TABLE VIII: COMPARISON OF METHODS V TO VIII WITH AND WITHOUT DISCOUNTING

Compression Ratio	Evaluation Measure	Method VII	Method VIII	Method IX	Method X
10%	E1	0.806	0.812	0.823	0.837
	E2	0.598	0.627	0.636	0.664
	P	0.517	0.523	0.572	0.581
20%	E1	0.815	0.851	0.834	0.867
	E2	0.602	0.628	0.614	0.635
	P	0.541	0.552	0.590	0.595
30%	E1	0.850	0.876	0.862	0.890
	E2	0.621	0.645	0.627	0.652
	P	0.558	0.593	0.582	0.611

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 document case. 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.

Table X presents results corresponding to lead based and random selections. We find all the 12 methods are superior to random selection; SR methods are far superior to lead based selections.

TABLE IX: COMPARISON OF DISCOUNTING METHODS INCORPORATING POSITION WEIGHT

Compression Ratio	Evaluation Measure	Method V	Method VI	Method XI	Method XII
10%	E1	0.722	0.737	0.915	0.921
	E2	0.699	0.710	0.896	0.909
	P	0.584	0.604	0.750	0.777
20%	E1	0.782	0.794	0.918	0.928
	E2	0.793	0.802	0.851	0.855
	P	0.658	0.670	0.687	0.710
30%	E1	0.802	0.811	0.899	0.906
	E2	0.787	0.794	0.809	0.817
	P	0.686	0.692	0.668	0.683

TABLE X: PERFORMANCE OF LEAD BASED AND RANDOM SYSTEMS

Compression Ratio	Lead			Random		
	E1	E2	P	E1	E2	P
10%	0.898	0.875	0.727	0.512	0.473	0.246
20%	0.853	0.842	0.661	0.487	0.376	0.212
30%	0.794	0.788	0.657	0.372	0.344	0.198

G. Study Conclusions

From the study the following conclusions can be drawn.

a. Directed backward graph is superior for news documents.
b. Design flexibility is obtained by adopting undirected graph and incorporating position weight separately.
c. Inclusion or exclusion of self weight of the sentence (self recommendation) affects the performance only marginally.
d. A threshold of 0.10 for degree based methods and damping of 0.10 is a good preliminary choice.
e. Discounting methods are superior to non discounting methods.
f. Incorporation of Position weight and discounting best performance under each category is obtained.
g. SR (Degree) and SR(Continuous) methods are best of all the 12 methods discussed.

V. RELATED WORK

Wenji Li et al., [1] have investigated extractive

summarization based on inter and intra relevance using information of internal association, semantic relatedness and named entity clustering. The authors found that events have their own internal structure, and meanwhile often relates to other events semantically, temporally, spatially, causally or conditionally. Then PageRank ranking algorithm is applied to estimate the significance of an event for inclusion in a summary from the event relevance derived.

Marina Litvak and Mark Last [2] 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 of text and web documents, which enhances the traditional vector-space model by taking into account some structural document features. In supervised approach, summarized collection of documents with the purpose of inducing a keyword identification model were trained using classification algorithms. In unsupervised approach, HITS algorithm was run on the document graphs under the assumption that the top-ranked nodes should represent the document keywords.

Jin Zhang et al., [15] presents a novel extractive approach based on graph-based sub-topic partition algorithm (GSPSummary), a sub-topic model based on graph representation is presented with emphasis on the implicit logic structure of the topic covered in the document collection. A new framework of MDS with sub-topic partition is also proposed by the authors. Furthermore, a novel scalable ranking criterion is adopted, in which both word based features and global features are integrated together.

Xiaojun Wan and Jianwu Yang [3] has recently exploited Markov Random Walk model 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. However, a given document set usually covers a few topic themes with each theme represented by a cluster of sentences. The topic themes are usually not equally important and the sentences in an important theme cluster are deemed more salient than the sentences in a trivial theme cluster. The work also proposes the Cluster-based Conditional Markov Random Walk Model (ClusterCMRW) and the Cluster-based HITS Model (ClusterHITS) to fully leverage the cluster-level information.

Yong Liu et al [4] presented a novel multi-document summarization approach based on Personalized PageRank (PPRSum). In this algorithm, the authors uniformly integrated various kinds of information in the corpus. At first, a salience model of sentence global features based on Naïve Bayes Model was trained. Secondly, a relevance model for each corpus utilizing the query is generated. Then the personalized prior probability for each sentence in the corpus utilizing the salience model and the relevance model are computed. With the help of personalized prior probability, a Personalized PageRank ranking process is performed depending on the relationships among all sentences in the corpus. Additionally, the redundancy penalty is imposed on each sentence. The summary is produced by choosing the sentences with both high query focused information richness and high information novelty.

Maofu Liu et al., [5] attempted to select and organize the sentences in a summary with respect to the events or the sub-events that the sentences describe. In this paper, the authors define an event as one or more event terms along with the named entities associated. Each event often relates to other events semantically, temporally, spatially, causally orconditionally. Firstly, an event relevance from external linguistic resource is derived. Then PageRank ranking algorithm was applied to estimate the significance of an event for inclusion in a summary based on the event semantic relevance derived. We make experiments on the DUC 2001 test data only using the event semantic relevance, from external linguistic resource like VerbOcean, and the results make more improvement than those based on the tf*idf.

Xiaojun Wan [6] exploited graph-based ranking algorithm for multi-document summarization by making only use of the sentence-to-sentence relationships in the documents, under the assumption that all the sentences are indistinguishable. However, given a document set to be summarized, different documents are usually not equally important, and moreover, different sentences in a specific document are usually differently important. This paper also aims to explore document impact on summarization performance. Then a document-based graph model to incorporate the document-level information and the sentence-to-document relationship into the graph-based ranking process is proposed.

VI. Conclusions

We have investigated in depth, 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. We have demonstrated, that in each class discounting methods proposed by us are superior to basic methods and the proposed discounting plus weight methods fare the best. It is brought out from the investigations presented, that based on the average performance of over a 30-document set, methods XI and XII – Sentence Rank (Degree) and Sentence Rank (Continuous), proposed by us yield the best results of all the 12 methods considered. Work is in progress for the application of the SR methods to multi document summarization.

APPENDIX

APPENDIX I : DISCOUNTING ALGORITHM

```
Input: Symmetric adjacency Matrix – A;
                compression ratio r;
              Method Chosen method;
Output : Sorted list of sentences s_list< >;
begin
     s_list ← empty;          /* initially selection list is empty
     n'= n* r ;           /*  n' –number of sentences to be selected
     do while n'>0
      begin
         call chosen_method ( );  /*Calculate the sentence weight S_w by the chosen method
        SW_max= 0.0 ;
        for i=1 to n do
        if SW(i) > SW_max
                        {    SW_max = SW_i ;
             nn = i ;
           }
           s_list ← s_lsit + nn ;   /* add the sentence to the selection list
        for i= 1 to n do
                { a[i,nn] = 0; a [nn,i] = 0 } /*set the row, column value of selected sentence to 0
   n' = n - 1;
   end;
   sort  s_list < >          /* s_list is sorted in ascending order
end;
```

APPENDIX – II

(TABULATIONS OF THE WORKING FOR THE 12 METHODS)

Tables A, B and C present working details for all the methods discussed in Section II corresponding to the document whose adjacency matrix has been presented in Table I. Columns named as sum and degree reflects the aggregate sum row wise and node degree at a threshold of 0.10. Sentence weights are normalized and maximum value is set to 1. For Methods I, II, VII and VIII all sentences were picked up from Table A itself. All other methods use discounting technique and first sentence is picked up from Table A. The second and third sentences are picked up from Tables B and C. For example corresponding to M- I sentences 1, 3 and 9 will be picked up while corresponding to M- III the selection will be 1, 8 and 3.

TABLE A: SENTENCE WEIGHTS

Non- PageRank Type methods				**PageRank Type methods**			
M- I/ M–III (Sum)	**M–II/M-IV (Degree)**	**M-V**	**M-VI**	**M-VII/ M-IX**	**M-VIII/ M-X**	**M-XI**	**M-XII**
1.83	3	**1.000**	0.867	0.632	**1.000**	0.924	**1.000**
1.09	1	0.623	0.584	0.778	0.907	**1.000**	0.906
1.82	**5**	0.975	**1.000**	**1.000**	0.981	0.944	0.952
1.23	1	0.676	0.495	0.778	0.889	0.821	0.862
1.32	1	0.712	0.457	0.778	0.896	0.745	0.855
1.11	1	0.604	0.423	0.778	0.908	0.676	0.854
1.57	3	0.826	0.623	0.632	0.905	0.581	0.841
1.75	4	0.908	0.711	0.812	0.949	0.612	0.869
1.80	4	0.930	0.686	0.812	0.984	0.630	0.890

TABLE B: REVISED SENTENCE WEIGHTS AFTER FIRST SELECTION

Non- PageRank Type methods				PageRank Type methods			
M-I/ M–III (Sum)	**M–II/M-IV (Degree)**	**M-V**	**M-VI**	**M-IX**	**M-X**	**M-XI**	**M-XII**
-	2	-	**1.000**	0.698	-	0.957	-
1.09	1	0.718	0.730	0.849	0.941	-	0.969
1.60	-	**1.000**	-	-	0.996	**1.000**	**1.000**
1.16	1	0.741	0.678	0.849	0.936	0.867	0.933
1.24	1	0.775	0.632	0.849	0.941	0.786	0.924
1.05	1	0.662	0.590	0.849	0.956	0.714	0.926
1.48	2	0.902	0.765	0.698	0.949	0.611	0.909
1.66	**3**	0.992	0.944	**1.000**	**1.000**	0.653	0.943
1.58	3	0.945	0.913	1.000	0.994	0.677	0.930

TABLE C: REVISED SENTENCE WEIGHTS AFTER SECOND SELECTION

Non- PageRank Type methods				PageRank Type methods			
M-I/ M–III (Sum)	**M- II/M-IV (Degree)**	**M-V**	**M-VI**	**M-IX**	**M-X**	**M-XI**	**M-XII**
-	**2**	-	-	**1.000**	-	**1.000**	-
1.09	1	0.797	0.806	1.000	0.947	-	**1.000**
1.44	-	-	-	-	**1.000**	-	-
1.16	1	0.801	0.747	1.000	0.945	0.906	0.972
1.16	1	0.806	0.694	1.000	0.950	0.822	0.962
1.05	1	0.730	0.647	1.000	0.956	0.746	0.956
1.32	1	0.877	0.821	1.000	0.953	0.607	0.932
-	-	**1.000**	**1.000**	-	-	0.700	0.969
1.33	2	0.975	0.748	1.000	0.974	0.769	0.958

APPENDIX III: EVALUATION METRICS

We have gone in for intrinsic evaluation. We have used 3 metrics, Precision, Effectiveness 1and Effectiveness 2. E1 and E2 have been proposed by us earlier [19]. If S_{sum} denotes the sentences picked up by the summarizer and S_{judges} denotes the sentences selected by panel of judges, precision is given as shown below:

$$Precision = \frac{|S_{sum} \cap S_{judges}|}{|Number\ of\ sentences|}$$

where $|\ |$ denotes a count measure.

$$E1\ or\ E2 = \frac{Score\ of\ the\ selected\ sentences\ by\ the\ sumarizer}{Maximum\ possible\ score}$$

where maximum possible score corresponds to the sentences selected by the judges. Definitions for E1 & E2 are similar. For E1 judges assign score to all the sentences in the document. 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 not picked up by any of the judges is set to zero.

ACKNOWLEDGMENT

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, Dean(School of Computer and Information Sciences) & HOD/CSE Dr.K.M.Mehata and Dean (Student Affairs) & HOD/IT Dr.T.R.Rangaswamy for the environment provided.

REFERENCES

[1] Wenjie Li, Mingli Wu and Qin Lu, Wei Xu and Chunfa Yuan 2006 , "Extractive Summarization using Inter- and Intra- Event Relevance", Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pp. 369–376.

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

[3] Xiaojun Wan and Jianwu Yang 2008, "Multi-Document Summarization Using Cluster-Based Link Analysis", SIGIR'08, pp. 299–306**,** Singapore.

[4] Yong Liu, Xiaolei Wang, Jin Zhang and Hongbo Xu 2008, "Personalized PageRank based Multi-document Summarization", IEEE International Workshop on Semantic Computing and Systems, pp. 169-173.

[5] Maofu Liu , Wenjie Li , Mingli Wu and Hujun Hu 2007, "Event-based Extractive Summarization using Event Semantic Relevance from External Linguistic Resource", Sixth International Conference on Advanced Language Processing and Web Information Technology, pp. 117-122.

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

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

[8] Dragomir R. Radev, Hongyan Jing, Malgorzata Stys and Daniel Tam (2004), "Centroid-based summarization of multiple documents", Information Processing and Management: Issue 40 , pp. 919-938.

[9] Bogdan Cranganu Cretu., Zhenmao Chen., Tetsuya Uchimoto. and Kenzo Miya. (2001/2002) 'Automatic Summarizing based on sentence extraction: A statistical approach', International Journal of Applied Electromagnetics and Mechanics, IOS Press, Vol 13, pp. 19-23.

[10] Luhn, H. P. (1958) 'The Automatic Creation of Literature bstracts', IBM Journal of Research Development, 2(2): 159–165.
[11] Mihalcea, R., Tarau, P 2005, "A language independent algorithm for single and multiple document summarization". In: Proceedings of IJCNLP 2005.
[12] Edmundson.H.P. (1969) 'New Methods in Automatic Extracting', Journal of the ACM, Vol .16, no 2 ,264-285.
[13] 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.
[14] 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).
[15] Jin Zhang, Xueqi Cheng, and Hongbo Xu , "GSPSummary: A Graph-Based Sub- topic Partition Algorithm for Summarization", H. Li (Eds.): AIRS 2008, LNCS 4993, pp. 321–334, 2008.
[16] www.google.com/news,www.rediffnews.com,www.yahoonews.com,www.hindu. com, www.indianexpress.co.in, www.cnn.com.
[17] M.F. Porter (1980), "An algorithm for suffix stripping", Program, 14(3) pp 130−137, 1980.
[18] http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words
[19] Shanmugasundaram Hariharan and Rengaramanujam Srinivasan, "Studies on Graph Based Approaches for Single and Multi Document Summarizations", International Journal of Computer Theory and Engineering, Vol.1 , Issue No. 5, December 2009.

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 Assitant Professor in Department of Information Technology, School of Computer and Information Sciences at B.S.Abdur Rahman University, Chennai, India. He is a member of IACSIT, ISTE and CSTA. 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.