# LexPageRank: Prestige in Multi-Document Text Summarization

Güneş Erkan<sup>1</sup>, Dragomir R. Radev<sup>1,2</sup> <sup>1</sup>Department of EECS, <sup>2</sup>School of Information University of Michigan {gerkan,radev}@umich.edu

#### Abstract

Multidocument extractive summarization relies on the concept of sentence centrality to identify the most important sentences in a document. Centrality is typically defined in terms of the presence of particular important words or in terms of similarity to a centroid pseudo-sentence. We are now considering an approach for computing sentence importance based on the concept of eigenvector centrality (prestige) that we call LexPageRank. In this model, a sentence connectivity matrix is constructed based on cosine similarity. If the cosine similarity between two sentences exceeds a particular predefined threshold, a corresponding edge is added to the connectivity matrix. We provide an evaluation of our method on DUC 2004 data. The results show that our approach outperforms centroid-based summarization and is quite successful compared to other summarization systems.

## **1** Introduction

Text summarization is the process of automatically creating a compressed version of a given text that provides useful information for the user. In this paper, we focus on multi-document generic text summarization, where the goal is to produce a summary of multiple documents about the same, but unspecified topic.

Our summarization approach is to assess the *centrality* of each sentence in a cluster and include the most important ones in the summary. In Section 2, we present centroid-based summarization, a well-known method for judging sentence centrality. Then we introduce two new measures for centrality, Degree and LexPageRank, inspired from the "prestige" concept in social networks and based on our new approach. We compare our new methods and centroid-based summarization using a feature-based generic summarization toolkit, MEAD, and show that new features outperform Centroid in most of the cases. Test data for our experiments is taken from Document Understanding Conferences (DUC) 2004 sum-

marization evaluation to compare our system also with other state-of-the-art summarization systems.

# 2 Sentence centrality and centroid-based summarization

Extractive summarization produces summaries by choosing a subset of the sentences in the original documents. This process can be viewed as choosing the most *central* sentences in a (multi-document) cluster that give the necessary and enough amount of information related to the main theme of the cluster. Centrality of a sentence is often defined in terms of the centrality of the words that it contains. A common way of assessing word centrality is to look at the centroid. The centroid of a cluster is a pseudodocument which consists of words that have frequency\*IDF scores above a predefined threshold. In centroid-based summarization (Radev et al., 2000). the sentences that contain more words from the centroid of the cluster are considered as central. Formally, the centroid score of a sentence is the cosine of the angle between the centroid vector of the whole cluster and the individual centroid of the sentence. This is a measure of how close the sentence is to the centroid of the cluster. Centroid-based summarization has given promising results in the past (Radev et al., 2001).

#### **3** Prestige-based sentence centrality

In this section, we propose a new method to measure sentence centrality based on *prestige* in social networks, which has also inspired many ideas in the computer networks and information retrieval.

A cluster of documents can be viewed as a network of sentences that are related to each other. Some sentences are more similar to each other while some others may share only a little information with the rest of the sentences. We hypothesize that the sentences that are similar to many of the other sentences in a cluster are more central (or *prestigious*) to the topic. There are two points to clarify in this definition of centrality. First is how to define similarity between two sentences. Second is how to compute the overall prestige of a sentence given its similarity to other sentences. For the similarity metric, we use cosine. A cluster may be represented by a cosine similarity matrix where each entry in the matrix is the similarity between the corresponding sentence pair. Figure 1 shows a subset of a cluster used in DUC 2004, and the corresponding cosine similarity matrix. Sentence ID dXsY indicates the Y th sentence in the X th document. In the following sections, we discuss two methods to compute sentence prestige using this matrix.

#### 3.1 Degree centrality

In a cluster of related documents, many of the sentences are expected to be somewhat similar to each other since they are all about the same topic. This can be seen in Figure 1 where the majority of the values in the similarity matrix are nonzero. Since we are interested in *significant* similarities, we can eliminate some low values in this matrix by defining a threshold so that the cluster can be viewed as an (undirected) graph, where each sentence of the cluster is a node, and significantly similar sentences are connected to each other. Figure 2 shows the graphs that correspond to the adjacency matrix derived by assuming the pair of sentences that have a similarity above 0.1, 0.2, and 0.3, respectively, in Figure 1 are similar to each other. We define *degree centrality* as the degree of each node in the similarity graph. As seen in Table 1, the choice of cosine threshold dramatically influences the interpretation of centrality. Too low thresholds may mistakenly take weak similarities into consideration while too high thresholds may lose much of the similarity relations in a cluster.

ID	Degree (0.1)	Degree (0.2)	Degree (0.3)
d1s1	4	3	1
d2s1	6	2	1
d2s2	1	0	0
d2s3	5	2	0
d3s1	4	1	0
d3s2	6	3	0
d3s3	1	1	0
d4s1	8	4	0
d5s1	4	3	1
d5s2	5	3	0
d5s3	4	1	1

Table 1: Degree centrality scores for the graphs in Figure 2. Sentence d4s1 is the most central sentence for thresholds 0.1 and 0.2.

#### 3.2 Eigenvector centrality and LexPageRank

When computing degree centrality, we have treated each edge as a *vote* to determine the overall prestige value of each node. This is a totally democratic

method where each vote counts the same. However, this may have a negative effect in the quality of the summaries in some cases where several unwanted sentences vote for each and raise their prestiges. As an extreme example, consider a noisy cluster where all the documents are related to each other, but only one of them is about a somewhat different topic. Obviously, we wouldn't want any of the sentences in the unrelated document to be included in a generic summary of the cluster. However, assume that the unrelated document contains some sentences that are very prestigious considering only the votes in that document. These sentences will get artificially high centrality scores by the local votes from a specific set of sentences. This situation can be avoided by considering where the votes come from and taking the prestige of the voting node into account in weighting each vote. Our approach is inspired by a similar idea used in computing web page prestiges.

One of the most successful applications of prestige is PageRank (Page et al., 1998), the underlying technology behind the Google search engine. PageRank is a method proposed for assigning a prestige score to each page in the Web independent of a specific query. In PageRank, the score of a page is determined depending on the number of pages that link to that page as well as the individual scores of the linking pages. More formally, the PageRank of a page A is given as follows:

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + \dots + \frac{PR(T_n)}{C(T_n)})$$
(1)

where  $T_1...T_n$  are pages that link to A,  $C(T_i)$  is the number of outgoing links from page  $T_i$ , and d is the damping factor which can be set between 0 and 1. This recursively defined value can be computed by forming the binary adjacency matrix, M, of the Web, where M(u, v) = 1 if there is a link from page u to page v, normalizing this matrix so that row sums equal to 1, and finding the principal eigenvector of the normalized matrix. PageRank for *i*th page equals to the *i*th entry in the eigenvector. Principal eigenvector of a matrix can be computed with a simple iterative power method.

This method can be directly applied to the cosine similarity graph to find the most *prestigious* sentences in a document. We use PageRank to weight each vote so that a vote that comes from a more prestigious sentence has a greater value in the centrality of a sentence. Note that unlike the original PageRank method, the graph is undirected since cosine similarity is a symmetric relation. However,

SNo	ID	Text
1	d1s1	Iraqi Vice President Taha Yassin Ramadan announced
	4101	today, Sunday, that Iraq refuses to back down from its
		decision to stop cooperating with disarmament
		inspectors before its demands are met.
2	d2s1	Iraqi Vice president Taha Yassin Ramadan announced
		today, Thursday, that Iraq rejects cooperating with the
		United Nations except on the issue of lifting the
		blockade imposed upon it since the year 1990.
3	d2s2	Ramadan told reporters in Baghdad that "Iraq cannot
		deal positively with whoever represents the Security
		Council unless there was a clear stance on the issue
	10.0	of lifting the blockade off of it.
4	d2s3	Baghdad had decided late last October to completely
		cease cooperating with the inspectors of the United
		Nations Special Commission (UNSCOM), in charge of disarming Iraq's weapons, and whose work became
		very limited since the fi fth of August, and announced
		it will not resume its cooperation with the Commission
		even if it were subjected to a military operation.
5	d3s1	The Russian Foreign Minister, Igor Ivanov, warned
		today, Wednesday against using force against Iraq,
		which will destroy, according to him, seven years
		of diffi cult diplomatic work and will complicate the
		regional situation in the area.
6	d3s2	Ivanov contended that carrying out air strikes against
		Iraq, who refuses to cooperate with the United
		Nations inspectors, "will end the tremendous work
		achieved by the international group during the past seven years and will complicate the situation in the
		region."
7	d3s3	Nevertheless, Ivanov stressed that Baghdad must
		resume working with the Special Commission in
		charge of disarming the Iraqi weapons of mass
		destruction (UNSCOM).
8	d4s1	The Special Representative of the United Nations
		Secretary-General in Baghdad, Prakash Shah,
		announced today, Wednesday, after meeting with the
		Iraqi Deputy Prime Minister Tariq Aziz, that Iraq
		refuses to back down from its decision to cut off cooperation with the disarmament inspectors.
9	d5s1	British Prime Minister Tony Blair said today, Sunday,
7	u351	that the crisis between the international community
		and Iraq "did not end" and that Britain is still
		"ready, prepared, and able to strike Iraq."
10	d5s2	In a gathering with the press held at the Prime
		Minister's offi ce, Blair contended that the crisis with
		Iraq "will not end until Iraq has absolutely and
		unconditionally respected its commitments" towards
		the United Nations.
11	d5s3	A spokesman for Tony Blair had indicated that the
		British Prime Minister gave permission to British Air
		Force Tornado planes stationed in Kuwait to join the aerial bombardment against Iraq.
L		nie aeriai bollibarument against fraq.
	1 2	
1		.45 0.02 0.17 0.03 0.22 0.03 0.28 0.06 0.06 0.00
2	0.45 1	.00 0.16 0.27 0.03 0.19 0.03 0.21 0.03 0.15 0.00

0.45 1.00 0.16 0.27 0.03 0.19 0.03 0.21 0.03 0.15 0.00 3  $0.02 \ 0.16 \ 1.00 \ 0.03 \ 0.00 \ 0.01 \ 0.03 \ 0.04 \ 0.00 \ 0.01 \ 0.00$ 4  $0.17 \ 0.27 \ 0.03 \ 1.00 \ 0.01 \ 0.16 \ 0.28 \ 0.17 \ 0.00 \ 0.09 \ 0.01$ 0.03 0.03 0.00 0.01 1.00 0.29 0.05 0.15 0.20 0.04 0.18 5 0 22 0 19 0 01 0 16 0 29 1 00 0 05 0 29 0 04 0 20 0 03 6 7 0.03 0.03 0.03 0.28 0.05 0.05 1.00 0.06 0.00 0.00 0.01 0.28 0.21 0.04 0.17 0.15 0.29 0.06 1.00 0.25 0.20 0.17 8 9 0.06 0.03 0.00 0.00 0.20 0.04 0.00 0.25 1.00 0.26 0.38 10  $0.06 \ 0.15 \ 0.01 \ 0.09 \ 0.04 \ 0.20 \ 0.00 \ 0.20 \ 0.26 \ 1.00 \ 0.12$  $0.00 \ 0.00 \ 0.00 \ 0.01 \ 0.18 \ 0.03 \ 0.01 \ 0.17 \ 0.38 \ 0.12 \ 1.00$ 11

Figure 1: Intra-sentence cosine similarities in a subset of cluster d1003t from DUC 2004.

this does not make any difference in the computation of the principal eigenvector. We call this new measure of sentence similarity *lexical PageRank*, or *LexPageRank*. Table 3 shows the LexPageRank scores for the graphs in Figure 2 setting the damping factor to 1. For comparison, Centroid score for each sentence is also shown in the table. All the numbers are normalized so that the highest ranked sentence gets the score 1. It is obvious from the figures that threshold choice affects the LexPageRank rankings of some sentences.



Figure 2: Similarity graphs that correspond to thresholds 0.1, 0.2, and 0.3, respectively, for the cluster in Figure 1.

# 3.3 Comparison with Centroid

The graph-based centrality approach we have introduced has several advantages over Centroid. First of

ID	LPR (0.1)	LPR (0.2)	LPR (0.3)	Centroid
d1s1	0.6007	0.6944	0.0909	0.7209
d2s1	0.8466	0.7317	0.0909	0.7249
d2s2	0.3491	0.6773	0.0909	0.1356
d2s3	0.7520	0.6550	0.0909	0.5694
d3s1	0.5907	0.4344	0.0909	0.6331
d3s2	0.7993	0.8718	0.0909	0.7972
d3s3	0.3548	0.4993	0.0909	0.3328
d4s1	1.0000	1.0000	0.0909	0.9414
d5s1	0.5921	0.7399	0.0909	0.9580
d5s2	0.6910	0.6967	0.0909	1.0000
d5s3	0.5921	0.4501	0.0909	0.7902

Figure 3: LexPageRank scores for the graphs in Figure 2 Sentence d4s1 is the most central sentence for thresholds 0.1 and 0.2.

all, it accounts for information subsumption among sentences. If the information content of a sentence subsumes another sentence in a cluster, it is naturally preferred to include the one that contains more information in the summary. The degree of a node in the cosine similarity graph is an indication of how much common information the sentence has with other sentences. Sentence d4s1 in Figure 1 gets the highest score since it almost subsumes the information in the first two sentences of the cluster and has some common information with others. Another advantage is that it prevents unnaturally high IDF scores from boosting up the score of a sentence that is unrelated to the topic. Although the frequency of the words are taken into account while computing the Centroid score, a sentence that contains many rare words with high IDF values may get a high Centroid score even if the words do not occur elsewhere in the cluster.

#### 4 Experiments on DUC 2004 data

#### 4.1 DUC 2004 data and ROUGE

We used DUC 2004 data in our experiments. There are 2 generic summarization tasks (Tasks 2, 4a, and 4b) in DUC 2004 which are appropriate for the purpose of testing our new feature, LexPageRank. Task 2 involves summarization of 50 TDT English clusters. The goal of Task 4 is to produce summaries of machine translation output (in English) of 24 Arabic TDT documents.

For evaluation, we used the new automatic summary evaluation metric, ROUGE<sup>1</sup>, which was used for the first time in DUC 2004. ROUGE is a recallbased metric for fixed-length summaries which is based on n-gram co-occurence. It reports separate scores for 1, 2, 3, and 4-gram, and also for longest common subsequence co-occurences. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgements most (Lin and Hovy, 2003). We show three of the ROUGE metrics in our experiment results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on longest common subsequence weighted by the length).

There are 8 different human judges for DUC 2004 Task 2, and 4 for DUC 2004 Task 4. However, a subset of exactly 4 different human judges produced model summaries for any given cluster. ROUGE requires a limit on the length of the summaries to be able to make a fair evaluation. To stick with the DUC 2004 specifications and to be able to compare our system with human summaries and as well as with other DUC participants, we produced 665-byte summaries for each cluster and computed ROUGE scores against human summaries.

#### 4.2 MEAD summarization toolkit

 $MEAD^2$  is a publicly available toolkit for extractive multi-document summarization. Although it comes as a centroid-based summarization system by default, its feature set can be extended to implement other methods.

The MEAD summarizer consists of three components. During the first step, *the feature extractor*, each sentence in the input document (or cluster of documents) is converted into a feature vector using the user-defined features. Second, the feature vector is converted to a scalar value using the *combiner*. At the last stage known as the *reranker*, the scores for sentences included in related pairs are adjusted upwards or downwards based on the type of relation between the sentences in the pair. Reranker penalizes the sentences that are similar to the sentences already included in the summary so that a better information coverage is achieved.

Three default features that comes with the MEAD distribution are Centroid, Position and Length. Position is the normalized value of the position of a sentence in the document such that the first sentence of a document gets the maximum Position value of 1, and the last sentence gets the value 0. Length is not a real feature score, but a cutoff value that ignores the sentences shorter than the given threshold. Several rerankers are implemented in MEAD. We observed the best results with Maximal Marginal Relevance (MMR) (Carbonell and Goldstein, 1998) reranker and the default reranker of the system based on Cross-Sentence Informational Subsumption (CSIS) (Radev, 2000). All of our experiments shown in Section 4.3 use CSIS reranker.

A MEAD policy is a combination of three components: (a) the command lines for all features, (b)

<sup>&</sup>lt;sup>1</sup>http://www.isi.edu/~cyl/ROUGE

<sup>&</sup>lt;sup>2</sup>http://www.summarization.com

feature LexPageRank	LexPageRank.pl 0.2
Centroid 1 Position	1 LengthCutoff 9 LexPageRank 1
mmr-reranker-word.p2	l 0.5 MEAD-cosine enidf

Figure 4: Sample MEAD policy.

the formula for converting the feature vector to a scalar, and (c) the command line for the reranker. A sample policy might be the one shown in Figure 4. This example indicates the three default MEAD features (Centroid, Position, LengthCutoff), and our new LexPageRank feature used in our experiments. Our LexPageRank implementation requires the cosine similarity threshold, 0.2 in the example, as an argument. Each number next to a feature name shows the relative weight of that feature (except for LengthCutoff where the number 9 indicates the threshold for selecting a sentence based on the number of the words in the sentence). The reranker in the example is a word-based MMR reranker with a cosine similarity threshold, 0.5.

#### 4.3 Results and discussion

We implemented the Degree and LexPageRank methods, and integrated into the MEAD system as new features. We normalize each feature so that the sentence with the maximum score gets the value 1.

Policy	<b>ROUGE-1</b>	ROUGE-2	<b>ROUGE-W</b>
Code	(unigram)	(bigram)	(LCS)
degree0.5T0.1	0.38304	0.09204	0.13275
degree1T0.1	0.38188	0.09430	0.13284
lpr2T0.1	0.38079	0.08971	0.12984
lpr1.5T0.1	0.37873	0.09068	0.13032
lpr0.5T0.1	0.37842	0.08972	0.13121
lpr1T0.1	0.37700	0.09174	0.13096
C0.5	0.37672	0.09233	0.13230
lpr1T0.2	0.37667	0.09115	0.13234
lpr0.5T0.2	0.37482	0.09160	0.13220
C1	0.37464	0.09210	0.13071
lpr1T0.3	0.37448	0.08767	0.13302
degree0.5T0.2	0.37432	0.09124	0.13185
lpr0.5T0.3	0.37362	0.08981	0.13173
degree2T0.1	0.37338	0.08799	0.12980
degree1.5T0.1	0.37324	0.08803	0.12983
degree0.5T0.3	0.37096	0.09197	0.13236
lpr1.5T0.2	0.37058	0.08658	0.12965
C1.5	0.36885	0.08765	0.12747
lead-based	0.36859	0.08669	0.13196
lpr1.5T0.3	0.36849	0.08455	0.13111
lpr2T0.3	0.36737	0.08182	0.13040
lpr2T0.2	0.36737	0.08264	0.12891
C2	0.36710	0.08696	0.12682
degree1T0.2	0.36653	0.08572	0.13011
degree1T0.3	0.36517	0.08870	0.13046
degree1.5T0.3	0.35500	0.08014	0.12828
degree1.5T0.2	0.35200	0.07572	0.12484
degree2T0.3	0.34337	0.07576	0.12523
degree2T0.2	0.34333	0.07167	0.12302
random	0.32381	0.05285	0.11623

Policy	ROUGE-1	ROUGE-2	ROUGE-W			
Code	( <b>unigram</b> ) Task	(bigram)	(LCS)			
leg1 5TO 1			0 12427			
lpr1.5T0.1	0.39997	0.11030	0.12427			
lpr1.5T0.2	0.39970	0.11508	0.12422			
lpr2T0.2	0.39954	0.11417	0.12468			
lpr2T0.1	0.39809	0.11033	0.12357			
lpr1T0.2	0.39614	0.11266	0.12350			
degree2T0.2	0.39574	0.11590	0.12410			
degree1.5T0.2	0.39395	0.11360	0.12329			
lpr0.5T0.1	0.39369	0.10665	0.12287			
lpr1T0.1	0.39312	0.10730	0.12274			
degree1T0.2	0.39241	0.11298	0.12277			
degree2T0.1	0.39217	0.10977	0.12205			
degree0.5T0.2	0.39076	0.11026	0.12236			
degree0.5T0.1	0.39016	0.10831	0.12292			
C0.5	0.39013	0.10459	0.12202			
lpr0.5T0.2	0.38899	0.10891	0.12200			
degree1T0.1	0.38882	0.10812	0.12286			
lpr1T0.3	0.38777	0.10586	0.12157			
lpr0.5T0.3	0.38667	0.10255	0.12244			
degree1.5T0.1	0.38634	0.10882	0.12136			
degree0.5T0.3	0.38568	0.10818	0.12088			
degree1.5T0.3	0.38553	0.10683	0.12064			
degree2T0.3	0.38506	0.10910	0.12075			
degree1T0.3	0.38412	0.10568	0.11961			
lpr1.5T0.3	0.38251	0.10610	0.12039			
C1	0.38181	0.10023	0.11909			
lpr2T0.3	0.38096	0.10497	0.12001			
C1.5	0.38074	0.09922	0.11804			
C2	0.38001	0.09901	0.11772			
lead-based	0.37880	0.09942	0.12218			
random	0.35929	0.08121	0.11466			
1 4 5700 4	Task		0.10115			
lpr1.5T0.1	0.40639	0.12419	0.13445			
degree2T0.1	0.40572	0.12421	0.13293			
lpr2T0.1	0.40529	0.12530	0.13346			
C1.5	0.40344	0.12824	0.13023			
degree1.5T0.1	0.40190	0.12407	0.13314			
C2	0.39997	0.12367	0.12873			
degree2T0.3	0.39911	0.11913	0.12998			
lpr2T0.3	0.39859	0.11744	0.12924			
lpr1.5T0.3	0.39858	0.11737	0.13044			
lpr1.5T0.2	0.39819	0.12228	0.12989			
lpr2T0.2	0.39763	0.12114	0.12924			
degree2T0.2	0.39752	0.12352	0.12958			
lpr1T0.1	0.39552	0.12045	0.13304			
degree1.5T0.3	0.39538	0.11515	0.12879			
lpr1T0.2	0.39492	0.12056	0.13061			
C1	0.39388	0.12301	0.12805			
degree1.5T0.2	0.39386	0.12018	0.12945			
lpr1T0.3	0.39053	0.11500	0.13044			
degree1T0.1	0.39039	0.11918	0.13113			
degree1T0.2	0.38973	0.11722	0.12793			
degree1T0.3	0.38658	0.11452	0.12780			
lpr0.5T0.1	0.38374	0.11331	0.12954			
lpr0.5T0.2	0.38201	0.11201	0.12757			
degree0.5T0.2	0.38029	0.11335	0.12780			
degree0.5T0.1	0.38011	0.11320	0.12921			
C0.5	0.37601	0.11123	0.12605			
lpr0.5T0.3	0.37525	0.11115	0.12898			
degree0.5T0.3	0.37455	0.11307	0.12857			
	0 27220	0.09225	0.12205			
random lead-based	0.37339	0.07225	0.12203			

Table 2: Results for Task 2

# Table 3: Results for Task 4

We ran MEAD with several policies with different feature weights and combinations of features. We fixed Length cutoff at 9, and the weight of the Position feature at 1 in all of the policies. We did not try a weight higher than 2.0 for any of the features since our earlier observations on MEAD showed that too high feature weights results in poor summaries.

Table 2 and Table 3 show the ROUGE scores we have got in the experiments with using LexPageRank, Degree, and Centroid in Tasks 2 and 4, respectively, sorted by ROUGE-1 scores. 'lprXTY' indicates a policy in which the weight for LexPageRank is X and Y is used as threshold. 'degreeXTY' is similar except that degree of a node in the similarity graph is used instead of its LexPageRank score. Finally, 'CX' shows a policy with Centroid weight X. We also include two baselines for each data set. 'random' indicates a method where we have picked random sentences from the cluster to produce a summary. We have performed five random runs for each data set. The results in the tables are for the median runs. Second baseline, shown as 'lead-based' in the tables, is using only the Position feature without any centrality method. This is tantamount to producing lead-based summaries, which is a widely used and very challenging baseline in the text summarization community (Brandow et al., 1995).

The top scores we have got in all data sets come from our new methods. The results provide strong evidence that Degree and LexPageRank are better than Centroid in multi-document generic text summarization. However, it is hard to say that Degree and LexPageRank are significantly different from each other. This is an indication that Degree may already be a good enough measure to assess the centrality of a node in the similarity graph. Considering the relatively low complexity of degree centrality, it still serves as a plausible alternative when one needs a simple implementation. Computation of Degree can be done on the fly as a side product of Lex-PageRank just before the power method is applied on the similarity graph.

Another interesting observation in the results is the effect of threshold. Most of the top ROUGE scores belong to the runs with the threshold 0.1, and the runs with threshold 0.3 are worse than the others most of the time. This is due to the information loss in the similarity graphs as we move to higher thresholds as discussed in Section 3.

As a comparison with the other summarization systems, we present the official scores for the top five DUC 2004 participants and the human summaries in Table 4 and Table 5 for Tasks 2 and 4, respectively. Our top few results for each task are either better than or statistically indifferent from the best system in the official runs considering the 95% confidence interval.

## 5 Conclusion

We have presented a novel approach to define sentence centrality based on graph-based prestige scoring of sentences. Constructing the similarity graph of sentences provides us with a better view of *important* sentences compared to the centroid approach, which is prone to overgeneralization of the information in a document cluster. We have introduced two different methods, Degree and LexPageRank, for computing prestige in similarity graphs. The results of applying these methods on extractive summarization is quite promising. Even the simplest approach we have taken, degree centrality, is a good enough heuristic to perform better than lead-based and centroid-based summaries.

#### References

- Ron Brandow, Karl Mitze, and Lisa F. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing and Management*, 31(5):675–685.
- Jaime G. Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Research and Development in Information Retrieval*, pages 335–336.
- Chin-Yew Lin and E.H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence. In Proceedings of 2003 Language Technology Conference (HLT-NAACL 2003), Edmonton, Canada, May 27 - June 1.
- L. Page, S. Brin, R. Motwani, and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. *Technical report, Stanford University, Stanford, CA*.
- Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utilitybased evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.
- Dragomir Radev, Sasha Blair-Goldensohn, and Zhu Zhang. 2001. Experiments in single and multidocument summarization using MEAD. In *First Document Understanding Conference*, New Orleans, LA, September.
- Dragomir Radev. 2000. A common theory of information fusion from multiple text sources, step one: Cross-document structure. In *Proceedings, 1st ACL SIGDIAL Workshop on Discourse and Dialogue*, Hong Kong, October.

Peer	ROUGE-1	95% Confidence	ROUGE-2	95% Confidence	ROUGE-W	95% Confidence
Code	(unigram)	Interval	(bigram)	Interval	(LCS)	Interval
Н	0.4183	[0.4019,0.4346]	0.1050	[0.0902,0.1198]	0.1480	[0.1409,0.1551]
F	0.4125	[0.3916,0.4333]	0.0899	[0.0771,0.1028]	0.1462	[0.1388,0.1536]
Е	0.4104	[0.3882,0.4326]	0.0984	[0.0838,0.1130]	0.1435	[0.1347,0.1523]
D	0.4060	[0.3870,0.4249]	0.1065	[0.0947,0.1184]	0.1449	[0.1395,0.1503]
В	0.4043	[0.3795,0.4291]	0.0950	[0.0785,0.1114]	0.1447	[0.1347,0.1548]
А	0.3933	[0.3722,0.4143]	0.0896	[0.0792,0.1000]	0.1387	[0.1319,0.1454]
С	0.3904	[0.3715,0.4093]	0.0969	[0.0849,0.1089]	0.1381	[0.1317,0.1444]
G	0.3890	[0.3679,0.4101]	0.0860	[0.0721,0.0998]	0.1390	[0.1315,0.1465]
65	0.3822	[0.3708,0.3937]	0.0922	[0.0827,0.1016]	0.1333	[0.1290,0.1375]
104	0.3744	[0.3635,0.3854]	0.0855	[0.0770,0.0939]	0.1284	[0.1244,0.1324]
35	0.3743	[0.3615,0.3871]	0.0837	[0.0737,0.0936]	0.1338	[0.1291,0.1384]
19	0.3739	[0.3602,0.3875]	0.0803	[0.0712,0.0893]	0.1315	[0.1261,0.1368]
124	0.3706	[0.3578,0.3835]	0.0829	[0.0748,0.0909]	0.1293	[0.1252,0.1334]
2	0.3242	[0.3104,0.3380]	0.0641	[0.0545,0.0737]	0.1186	[0.1130,0.1242]
		•		•		

Table 4: Summary of official ROUGE scores for DUC 2004 Task 2. Peer codes: baseline(2), manual[A-H], and system submissions

Peer	ROUGE-1	95% Confidence	ROUGE-2	95% Confidence	ROUGE-W	95% Confidence
Code	(unigram)	Interval	(bigram)	Interval	(LCS)	Interval
Y	0.44450	[0.42298,0.46602]	0.12815	[0.10965,0.14665]	0.14348	[0.13456,0.15240]
Z	0.43263	[0.40875,0.45651]	0.11953	[0.10186,0.13720]	0.14019	[0.13056,0.14982]
Х	0.42925	[0.40680,0.45170]	0.12213	[0.10180,0.14246]	0.14147	[0.13361,0.14933]
W	0.41188	[0.38696,0.43680]	0.10609	[0.08905,0.12313]	0.13542	[0.12620,0.14464]
	•		Tasl	x 4a		
144	0.38827	[0.36261,0.41393]	0.10109	[0.08680,0.11538]	0.11140	[0.10471,0.11809]
22	0.38654	[0.36352,0.40956]	0.09063	[0.07794,0.10332]	0.11621	[0.10980,0.12262]
107	0.38615	[0.35548,0.41682]	0.09851	[0.08225,0.11477]	0.11951	[0.11004,0.12898]
68	0.38156	[0.36420,0.39892]	0.09808	[0.08686,0.10930]	0.11888	[0.11255,0.12521]
40	0.37960	[0.35809,0.40111]	0.09408	[0.08367,0.10449]	0.12240	[0.11659,0.12821]
			Tasl	c 4b		
23	0.41577	[0.39333,0.43821]	0.12828	[0.10994,0.14662]	0.13823	[0.12995,0.14651]
84	0.41012	[0.38543,0.43481]	0.12510	[0.10506,0.14514]	0.13574	[0.12638,0.14510]
145	0.40602	[0.36783,0.44421]	0.12833	[0.10375,0.15291]	0.12221	[0.11128,0.13314]
108	0.40059	[0.37002,0.43116]	0.12087	[0.10212,0.13962]	0.13011	[0.12029,0.13993]
69	0.39844	[0.37440,0.42248]	0.11395	[0.09885,0.12905]	0.12861	[0.12000,0.13722]

Table 5: Summary of official ROUGE scores for DUC 2004 Task 4. Peer codes: manual[W-Z], and system submissions