# An Exploration of Document Impact on Graph-Based Multi-Document Summarization

**Xiaojun Wan**
Institute of Compute Science and Technology
Peking University
Beijing 100871, China
wanxiaojun@icst.pku.edu.cn

## Abstract

The graph-based ranking algorithm has been recently exploited 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 aims to explore document impact on summarization performance. We propose a document-based graph model to incorporate the document-level information and the sentence-to-document relationship into the graph-based ranking process. Various methods are employed to evaluate the two factors. Experimental results on the DUC2001 and DUC2002 datasets demonstrate that the good effectiveness of the proposed model. Moreover, the results show the robustness of the proposed model.

## 1 Introduction

Multi-document summarization aims to produce a summary describing the main topic in a document set, without any prior knowledge. Multi-document summary can be used to facilitate users to quickly understand a document cluster. For example, a number of news services (e.g. NewsInEssence[1]) have been developed to group news articles into news topics, and then produce a short summary for each news topic. Users can easily understand the topic they have interest in by taking a look at the short summary, without looking into each individual article within the topic cluster.

Automated multi-document summarization has drawn much attention in recent years. In the communities of natural language processing and information retrieval, a series of workshops and conferences on automatic text summarization (e.g. NTCIR, DUC), special topic sessions in ACL, COLING, and SIGIR have advanced the summarization techniques and produced a couple of experimental online systems.

A particular challenge for multi-document summarization is that a document set might contain diverse information, which is either related or unrelated to the main topic, and hence we need effective summarization methods to analyze the information stored in different documents and extract the globally important information to reflect the main topic. In recent years, both unsupervised and supervised methods have been proposed to analyze the information contained in a document set and extract highly salient sentences into the summary, based on syntactic or statistical features.

Most recently, the graph-based models have been successfully applied for multi-document summarization by making use of the "voting" or "recommendations" between sentences in the documents (Erkan and Radev, 2004; Mihalcea and Tarau, 2005; Wan and Yang, 2006). The model first constructs a directed or undirected graph to reflect the relationships between the sentences and then applies the graph-based ranking algorithm to compute the rank scores for the sentences. The sentences with large rank scores are chosen into the summary. However, the model makes uniform use of the sentences in different documents, i.e. all the sentences are ranked without considering the document-level information and the sentence-to-document relationship. Actually, given a document set, different documents are not equally important. For example, the documents close to the main topics of the document set are usually more important than the documents far away from the main topics

[1] http://lada.si.umich.edu:8080/clair/nie1/nie.cgi

*Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 755–762,
Honolulu, October 2008. ©2008 Association for Computational Linguistics

of the document set. This document-level information is deemed to have great impact on the sentence ranking process. Moreover, the sentences in the same document cannot be treated uniformly, because some sentences in the document are more important than other sentences because of their different positions in the document or different distances to the document's centroid. In brief, neither the document-level information nor the sentence-to-document relationship has been taken into account in the previous graph-based model.

In order to overcome the limitations of the previous graph-based model, this study proposes the document-based graph model to explore document impact on the graph-based summarization, by incorporating both the document-level information and the sentence-to-document relationship in the graph-based ranking process. We develop various methods to evaluate the document-level information and the sentence-to-document relationship. Experiments on the DUC2001 and DUC2002 datasets have been performed and the results demonstrate the good effectiveness of the proposed model, i.e., the incorporation of document impact can much improve the performance of the graph-based summarization. Moreover, the proposed model is robust with respect to most incorporation schemes.

The rest of this paper is organized as follows. We first introduce the related work in Section 2. The basic graph-based summarization model and the proposed document-based graph model are described in detail in Sections 3 and 4, respectively. We show the experiments and results in Section 5 and finally we conclude this paper in Section 6.

## 2 Related Work

Generally speaking, summarization methods can be abstractive summarization or extractive summarization. Extractive summarization is a simple but robust method for text summarization and it involves assigning saliency scores to some units (e.g. sentences, paragraphs) of the documents and extracting those with highest scores, while abstraction summarization usually needs information fusion (Barzilay et al., 1999), sentence compression (Knight and  Marcu, 2002) and reformulation (McKeown et al., 1999). In this study, we focus on extractive summarization.

The centroid-based method (Radev et al., 2004) is one of the most popular extractive summarization methods. MEAD[2] is an implementation of the centroid-based method that scores sentences based on sentence-level and inter-sentence features, including cluster centroids, position, TFIDF, etc. NeATS (Lin and Hovy, 2002) is a project on multi-document summarization at ISI based on the single-document summarizer-SUMMARIST. Sentence position, term frequency, topic signature and term clustering are used to select important content. MMR (Goldstein et al., 1999) is used to remove redundancy and stigma word filters and time stamps are used to improve cohesion and coherence. To further explore user interface issues, iNeATS (Leuski et al., 2003) is developed based on NeATS. XDoX (Hardy et al., 1998) is a cross document summarizer designed specifically to summarize large document sets. It identifies the most salient themes within the set by passage clustering and then composes an extraction summary, which reflects these main themes. Much other work also explores to find topic themes in the documents for summarization, e.g. Harabagiu and Lacatusu (2005) investigate five different topic representations and introduce a novel representation of topics based on topic themes. In addition, Marcu (2001) selects important sentences based on the discourse structure of the text. TNO's system (Kraaij et al., 2001) scores sentences by combining a unigram language model approach with a Bayesian classifier based on surface features. Nenkova and Louis (2008) investigate how summary length and the characteristics of the input influence the summary quality in multi-document summarization.

Graph-based models have been proposed to rank sentences or passages based on the PageRank algorithm (Page et al., 1998) or its variants. Websumm (Mani and Bloedorn, 2000) uses a graph-connectivity model and operates under the assumption that nodes which are connected to many other nodes are likely to carry salient information. LexPageRank (Erkan and Radev, 2004) is an approach for computing sentence importance based on the concept of eigenvector centrality. It constructs a sentence connectivity matrix and compute sentence importance based on an algorithm similar to PageRank. Mihalcea and Tarau (2005) also propose a similar algorithm based on PageRank to compute sentence importance for document summarization. Wan and Yang (2006) improve the ranking algo-

[2] http://www.summarization.com/mead/

rithm by differentiating intra-document links and inter-document links between sentences. All these methods make use of the relationships between sentences and select sentences according to the "votes" or "recommendations" from their neighboring sentences, which is similar to PageRank.

Other related work includes topic-focused multi-document summarization (Daumé. and Marcu, 2006; Gupta et al., 2007; Wan et al., 2007), which aims to produce summary biased to a given topic or query. It is noteworthy that our proposed approach is inspired by (Liu and Ma, 2005), which proposes the Conditional Markov Random Walk Model based on two-layer web graph in the tasks of web page retrieval.

## 3 The Basic Graph-Based Model (GM)

The basic graph-based model is essentially a way of deciding the importance of a vertex within a graph based on global information recursively drawn from the entire graph. The basic idea is that of "voting" or "recommendation" between the vertices. A link between two vertices is considered as a vote cast from one vertex to the other vertex. The score associated with a vertex is determined by the votes that are cast for it, and the score of the vertices casting these votes.

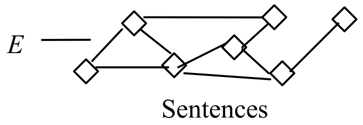

Figure 1. One-layer link graph

Formally, given a document set *D*, let *G*=(*V*, *E*) be an undirected graph to reflect the relationships between sentences in the document set, as shown in Figure 1. *V* is the set of vertices and each vertex $v_i$ in *V* is a sentence in the document set. *E* is the set of edges. Each edge $e_{ij}$ in *E* is associated with an affinity weight $f(v_i, v_j)$ between sentences $v_i$ and $v_j$ ($i \neq j$). The weight is computed using the standard cosine measure between the two sentences.

$$f(v_i, v_j) = sim_{\cos ine}(v_i, v_j) = \frac{\vec{v}_i \cdot \vec{v}_j}{|\vec{v}_i| \times |\vec{v}_j|} \qquad (1)$$

where $\vec{v}_i$ and $\vec{v}_j$ are the corresponding term vectors of $v_i$ and $v_j$. Here, we have $f(v_i, v_j)$=f($v_j$, $v_i$). Two vertices are connected if their affinity weight is larger than 0 and we let $f(v_i, v_i)$=0 to avoid self transition.

We use an affinity matrix *M* to describe *G* with each entry corresponding to the weight of an edge in the graph. $M = (M_{i,j})_{|V|\times|V|}$ is defined as follows:

$$M_{i,j} = \begin{cases} f(v_i, v_j), & \text{if } v_i \text{ and } v_j \text{ is connected} \\ & \text{and } \ i \neq j; \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

Then *M* is normalized to $\widetilde{M}$ as follows to make the sum of each row equal to 1:

$$\widetilde{M}_{i,j} = \begin{cases} M_{i,j} \Big/ \sum_{j=1}^{|V|} M_{i,j}, & \text{if } \sum_{j=1}^{|V|} M_{i,j} \neq 0 \\ 0 & , \text{ otherwise} \end{cases} \qquad (3)$$

Based on matrix $\widetilde{M}$, the saliency score *SenScore*($v_i$) for sentence $v_i$ can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form as in the PageRank algorithm:

$$SenScore(v_i) = \mu \cdot \sum_{all\, j \neq i} SenScore(v_j) \cdot \widetilde{M}_{j,i} + \frac{(1-\mu)}{|V|} \qquad (4)$$

And the matrix form is:

$$\vec{\lambda} = \mu \widetilde{M}^T \vec{\lambda} + \frac{(1-\mu)}{|V|} \vec{e} \qquad (5)$$

where $\vec{\lambda} = [SenScore(v_i)]_{|V|\times 1}$ is the vector of sentence saliency scores. $\vec{e}$ is a vector with all elements equaling to 1. $\mu$ is the damping factor usually set to 0.85, as in the PageRank algorithm.

The above process can be considered as a Markov chain by taking the sentences as the states and the corresponding transition matrix is given by $A = \mu \widetilde{M}^T + \frac{(1-\mu)}{|V|} \vec{e}\vec{e}^T$. The stationary probability distribution of each state is obtained by the principal eigenvector of the transition matrix.

For implementation, the initial scores of all sentences are set to 1 and the iteration algorithm in Equation (4) is adopted to compute the new scores of the sentences. Usually the convergence of the iteration algorithm is achieved when the difference between the scores computed at two successive iterations for any sentences falls below a given threshold (0.0001 in this study).

We can see that the basic graph-based model is built on the single-layer sentence graph and the transition probability between two sentences in the Markov chain depends only on the sentences themselves, not taking into account the document-level information and the sentence-to-document relationship.

# 4 The Document-Based Graph Model (DGM)

## 4.1 Overview

As we mentioned in previous section, there may be many factors that can have impact on the importance analysis of the sentences. This study aims to examine the document impact by incorporating the document importance and the sentence-to-document correlation into the sentence ranking process. Our assumption is that the sentences, which belong to an important document and are highly correlated with the document, will be more likely to be chosen into the summary.

In order to incorporate the document-level information and the sentence-to-document relationship, the document-based graph model is proposed based on the two-layer link graph including both sentences and documents. The novel representation is shown in Figure 2. As can be seen, the lower layer is just the traditional link graph between sentences that has been well studied in previous work. And the upper layer represents the documents. The dashed lines between these two layers indicate the conditional influence between the sentences and the documents.

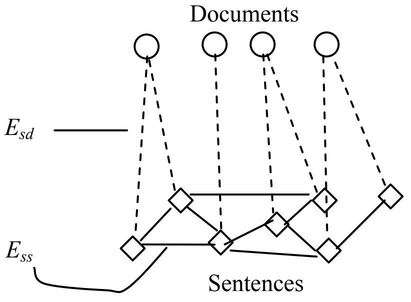

Figure 2. Two-layer link graph

Formally, the new representation for the two-layer graph is denoted as $G^*$=<$V_s$, $V_d$, $E_{ss}$, $E_{sd}$>, where $V_s$=$V$={$v_i$} is the set of sentences and $V_d$=$D$={$d_j$} is the set of documents; $E_{ss}$=$E$={$e_{ij}$|$v_i$, $v_j \in V_s$} includes all possible links between sentences and $E_{sd}$={$e_{ij}$|$v_i \in V_s$, $d_j \in V_d$ and $d_j$=$doc(v_i)$} includes the correlation link between any sentence and its belonging document. Here, we use $doc(v_i)$ to denote the document containing sentence $v_i$. For further discussions, we let $\pi(doc(v_i)) \in [0,1]$ denote the importance of document $doc(v_i)$ in the document set, and let $\omega(v_i, doc(v_i)) \in [0,1]$ denote the strength of the correlation between sentence $v_i$ and its document $doc(v_i)$.

The two factors are incorporated into the affinity weight between sentences and the new sentence-to-sentence affinity weight is denoted as $f(v_i, v_j|doc(v_i), doc(v_j))$, which is conditioned on the two documents containing the two sentences. The new conditional affinity weight is computed by linearly combining the affinity weight conditioned on the first document (i.e. $f(v_i,v_j|doc(v_i))$) and the affinity weight conditioned on the second document (i.e. $f(v_i,v_j|doc(v_j))$).

Formally, the conditional affinity weight is computed as follows to incorporate the two factors:

$$\begin{aligned}
& f(v_i,v_j \mid doc(v_i),doc(v_j)) \\
& \quad = \lambda \cdot f(v_i,v_j \mid doc(v_i)) + (1-\lambda) \cdot f(v_i,v_j \mid doc(v_j)) \\
& \quad = \lambda \cdot f(v_i,v_j) \cdot \pi(doc(v_i)) \cdot \omega(v_i,doc(v_i)) \\
& \qquad + (1-\lambda) \cdot f(v_i,v_j) \cdot \pi(doc(v_j)) \cdot \omega(v_j,doc(v_j)) \\
& \quad = f(v_i,v_j) \cdot (\lambda \cdot \pi(doc(v_i)) \cdot \omega(v_i,doc(v_i)) \\
& \qquad + (1-\lambda) \cdot \pi(doc(v_j)) \cdot \omega(v_j,doc(v_j))) \\
& \quad = sim_{\cos ine}(v_i,v_j) \cdot (\lambda \cdot \pi(doc(v_i)) \cdot \omega(v_i,doc(v_i)) \\
& \qquad + (1-\lambda) \cdot \pi(doc(v_j)) \cdot \omega(v_j,doc(v_j)))
\end{aligned} \tag{6}$$

where $\lambda \in [0,1]$ is the combination weight controlling the relative contributions from the first document and the second document. Note that usually $f(v_i, v_j|doc(v_i), doc(v_j))$ is not equal to $f(v_j, v_i|doc(v_j), doc(v_i))$, but the two scores are equal when $\lambda$ is set to 0.5. Various methods can be used to evaluate the document importance and the sentence-document correlation, which will be described in next sections.

The new affinity matrix $M^*$ is then constructed based on the above conditional sentence-to-sentence affinity weight.

$$M^*_{i,j} = \begin{cases} f(v_i,v_j \mid doc(v_i),doc(v_j)), & \text{if if } v_i \text{ and } v_j \text{ is connected} \\ & \text{and } i \neq j \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

Likewise, $M^*$ is normalized to $\tilde{M}^*$ and the iterative computation as in Equation (4) is then based on $\tilde{M}^*$. The transition matrix in the Markov chain is then denoted by $A^* = \mu \tilde{M}^{*T} + \frac{(1-\mu)}{|V|}\vec{e}\vec{e}^T$ and the sentence scores is obtained by the principle eigenvector of the new transition matrix $A^*$.

## 4.2 Evaluating Document Importance ($\pi$)

The function $\pi(doc(v_i))$ aims to evaluate the importance of document $doc(v_i)$ in the document set $D$. The following three methods are developed to evaluate the document importance.

***π*₁:** It uses the cosine similarity value between the document and the whole document set as the importance score of the document[3]:

$$\pi_1(doc(v_i)) = sim_{\cos ine}(doc(v_i), D) \tag{8}$$

***π*₂:** It uses the average similarity value between the document and any other document in the document set as the importance score of the document:

$$\pi_2(doc(v_i)) = \frac{\sum_{d' \in D \text{ and } d' \neq doc(v_i)} sim_{\cos ine}(doc(v_i), d')}{|D| - 1} \tag{9}$$

***π*₃:** It constructs a weighted graph between documents and uses the PageRank algorithm to compute the rank scores of the documents as the importance scores of the documents. The link weight between two documents is computed using the cosine measure. The equation for iterative computation is the same with Equation (4).

### 4.3 Evaluating Sentence-Document Correlation (*ω*)

The function *ω*(*v*ᵢ, *doc*(*v*ᵢ)) aims to evaluate the correlation between sentence $v_i$ and its document *doc*($v_i$). The following four methods are developed to compute the strength of the correlation. The first three methods are based on sentence position in the document, under the assumption that the first sentences in a document are usually more important than other sentences. The last method is based on the content similarity between the sentence and the document.

***ω*₁:** The correlation strength between sentence $v_i$ and its document *doc*($v_i$) is based on the position of the sentence as follows:

$$\omega_1(v_i, doc(v_i)) = \begin{cases} 1 \text{ if } pos(v_i) \leq 3 \\ 0.5 \text{ Otherwise} \end{cases} \tag{10}$$

where *pos*($v_i$) returns the position number of sentence $v_i$ in its document. For example, if $v_i$ is the first sentence in its document, *pos*($v_i$) is 1.

***ω*₂:** The correlation strength between sentence $v_i$ and its document *doc*($v_i$) is based on the position of the sentence as follows:

$$\omega_2(v_i, doc(v_i)) = 1 - \frac{pos(v_i) - 1}{sen\_count(doc(v_i))} \tag{11}$$

where *sen_count*(*doc*($v_i$)) returns the total number of sentences in document *doc*($v_i$).

***ω*₃:** The correlation strength between sentence $v_i$ and its document *doc*($v_i$) is based on the position of the sentence as follows:

$$\omega_3(v_i, doc(v_i)) = 0.5 + \frac{1}{pos(v_i) + 1} \tag{12}$$

***ω*₄:** The correlation strength between sentence $v_i$ and its document *doc*($v_i$) is based on the cosine similarity between the sentence and the document:

$$\omega_4(v_i, doc(v_i)) = sim_{\cos ine}(v_i, doc(v_i)) \tag{13}$$

## 5 Empirical Evaluation

### 5.1 Dataset and Evaluation Metric

Generic multi-document summarization has been one of the fundamental tasks in DUC 2001[4] and DUC 2002[5] (i.e. task 2 in DUC 2001 and task 2 in DUC 2002), and we used the two tasks for evaluation. DUC2001 provided 30 document sets and DUC 2002 provided 59 document sets (D088 is excluded from the original 60 document sets by NIST) and generic abstracts of each document set with lengths of approximately 100 words or less were required to be created. The documents were news articles collected from TREC-9. The sentences in each article have been separated and the sentence information has been stored into files. The summary of the two datasets are shown in Table 1.

| | DUC 2001 | DUC 2002 |
|---|---|---|
| **Task** | Task 2 | Task 2 |
| **Number of documents** | 309 | 567 |
| **Number of clusters** | 30 | 59 |
| **Data source** | TREC-9 | TREC-9 |
| **Summary length** | 100 words | 100 words |

Table 1. Summary of datasets

We used the ROUGE (Lin and Hovy, 2003) toolkit (i.e. ROUGEeval-1.4.2 in this study) for evaluation, which has been widely adopted by DUC for automatic summarization evaluation. It measured summary quality by counting overlapping units such as the n-gram, word sequences and word pairs between the candidate summary and the reference summary. ROUGE-N was an n-gram recall measure computed as follows:

$$ROUGE-N = \frac{\sum_{S \in \{\text{Ref Sum}\}} \sum_{n\text{-}gram \in S} Count_{match}(n-gram)}{\sum_{S \in \{\text{Ref Sum}\}} \sum_{n\text{-}gram \in S} Count(n-gram)} \tag{14}$$

[3] A document set is treated as a single text by concatenating all the document texts in the set.

[4] http://www-nlpir.nist.gov/projects/duc/guidelines/2001.html

[5] http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

where $n$ stood for the length of the n-gram, and $Count_{match}$(*n-gram*) was the maximum number of n-grams co-occurring in a candidate summary and a set of reference summaries. *Count*(*n-gram*) was the number of n-grams in the reference summaries.

ROUGE toolkit reported separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence co-occurrences. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most (Lin and Hovy. 2003). We showed three of the ROUGE metrics in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on weighted longest common subsequence, weight=1.2). In order to truncate summaries longer than length limit, we used the "-l" option in ROUGE toolkit. We also used the "-m" option for word stemming.

## 5.2 Evaluation Results

In the experiments, the combination weight $\lambda$ for the proposed summarization model is typically set to 0.5 without tuning, i.e. the two documents for two sentences have equal influence on the summarization process. Note that after the saliency scores of sentences have been obtained, a greedy algorithm (Wan and Yang, 2006) is applied to remove redundancy and finally choose both informative and novel sentences into the summary. The algorithm is actually a variant version of the MMR algorithm (Goldstein et al., 1999).

The proposed document-based graph model (denoted as DGM) with different settings is compared with the basic graph-based Model (denoted as GM), the top three performing systems and two baseline systems on DUC2001 and DUC2002, respectively. The top three systems are the systems with highest ROUGE scores, chosen from the performing systems on each task respectively. The lead baseline and coverage baseline are two baselines employed in the generic multi-document summarization tasks of DUC2001 and DUC2002. The lead baseline takes the first sentences one by one in the last document in the collection, where documents are assumed to be ordered chronologically. And the coverage baseline takes the first sentence one by one from the first document to the last document. Tables 2 and 3 show the comparison results on DUC2001 and DUC2002, respectively. In Table 1, SystemN, SystemP and System T are the top three performing systems for DUC2001. In Table 2, System19, System26, System28 are the top three performing systems for DUC2002. The document-based graph model is configured with different settings (i.e. $\pi_1$-$\pi_3$, $\omega_1$-$\omega_4$). For example, DGM($\pi_1$+$\omega_1$) refers to the DGM model with $\pi_1$ to evaluate the document importance and $\omega_1$ to evaluate the correlation between a sentence and its document.

| System | ROUGE-1 | ROUGE-2 | ROUGE-W |
|---|---|---|---|
| DGM($\pi_1$+$\omega_1$) | 0.35658 | 0.05926 | 0.10712 |
| DGM($\pi_1$+$\omega_2$) | 0.35945 | $0.06304^*$ | 0.10820 |
| DGM($\pi_1$+$\omega_3$) | $0.36349^*$ | $0.06472^*$ | 0.10952 |
| DGM($\pi_1$+$\omega_4$) | 0.35421 | 0.05934 | 0.10695 |
| DGM($\pi_2$+$\omega_1$) | 0.35555 | $0.06554^*$ | 0.10924 |
| DGM($\pi_2$+$\omega_2$) | $0.37228^*$ | $0.06787^*$ | $0.11295^*$ |
| DGM($\pi_2$+$\omega_3$) | $0.37347^*$ | $0.06612^*$ | $0.11352^*$ |
| DGM($\pi_2$+$\omega_4$) | 0.36340 | $0.06397^*$ | 0.11006 |
| DGM($\pi_3$+$\omega_1$) | 0.35333 | $0.06353^*$ | 0.10834 |
| DGM($\pi_3$+$\omega_2$) | $0.37082^*$ | $0.06708^*$ | 0.11235 |
| DGM($\pi_3$+$\omega_3$) | $0.37056^*$ | $0.06503^*$ | $0.11227^*$ |
| DGM($\pi_3$+$\omega_4$) | $0.36667^*$ | $0.06585^*$ | 0.11114 |
| GM | 0.35527 | 0.05608 | 0.10641 |
| SystemN | 0.33910 | 0.06853 | 0.10240 |
| SystemP | 0.33332 | 0.06651 | 0.10068 |
| SystemT | 0.33029 | 0.07862 | 0.10215 |
| Coverage | 0.33130 | 0.06898 | 0.10182 |
| Lead | 0.29419 | 0.04033 | 0.08880 |

Table 2. Comparison results on DUC2001

| System | ROUGE-1 | ROUGE-2 | ROUGE-W |
|---|---|---|---|
| DGM($\pi_1$+$\omega_1$) | 0.37891 | 0.08398 | 0.12390 |
| DGM($\pi_1$+$\omega_2$) | $0.39013^*$ | $0.08770^*$ | $0.12726^*$ |
| DGM($\pi_1$+$\omega_3$) | $0.38490^*$ | 0.08355 | 0.12570 |
| DGM($\pi_1$+$\omega_4$) | 0.38464 | 0.08371 | 0.12443 |
| DGM($\pi_2$+$\omega_1$) | 0.38296 | 0.08369 | 0.12499 |
| DGM($\pi_2$+$\omega_2$) | 0.38143 | $0.08792^*$ | 0.12506 |
| DGM($\pi_2$+$\omega_3$) | 0.38177 | $0.08624^*$ | 0.12511 |
| DGM($\pi_2$+$\omega_4$) | $0.38576^*$ | 0.08167 | 0.12611 |
| DGM($\pi_3$+$\omega_1$) | 0.38079 | 0.08391 | 0.12392 |
| DGM($\pi_3$+$\omega_2$) | 0.38103 | $0.08608^*$ | 0.12446 |
| DGM($\pi_3$+$\omega_3$) | 0.38236 | $0.08675^*$ | 0.12478 |
| DGM($\pi_3$+$\omega_4$) | $0.38719^*$ | 0.08150 | $0.12633^*$ |
| GM | 0.37595 | 0.08304 | 0.12173 |
| System26 | 0.35151 | 0.07642 | 0.11448 |
| System19 | 0.34504 | 0.07936 | 0.11332 |
| System28 | 0.34355 | 0.07521 | 0.10956 |
| Coverage | 0.32894 | 0.07148 | 0.10847 |
| Lead | 0.28684 | 0.05283 | 0.09525 |

Table 3. Comparison results on DUC2002

($^*$ indicates that the improvement over the baseline GM model is statistically significant at 95% confidence level)

Seen from the tables, the proposed document-based graph model with different settings can outperform the basic graph-based model and other baselines over almost all three metrics on both

DUC2001 and DUC2002 datasets. The results demonstrate the good effectiveness of the proposed model, i.e. the incorporation of document impact does benefit the graph-based summarization model. It is interesting that the three methods for computing document importance and the four methods for computing the sentence-document correlation are almost as effective as each other on the DUC2002 dataset. However, $\pi_1$ does not perform as well as $\pi_2$ and $\pi_3$, and $\omega_1$ and $\omega_4$ does not perform as well as $\omega_2$ and $\omega_3$ on the DUC2001 dataset.

In order to investigate the relative contributions from the two documents for two sentences to the summarization performance, we varies the combination weight $\lambda$ from 0 to 1 and Figures 3-6 show the ROUGE-1 and ROUGE-W curves on DUC2001 and DUC2002 respectively. The similar ROUGE-2 curves are omitted here.

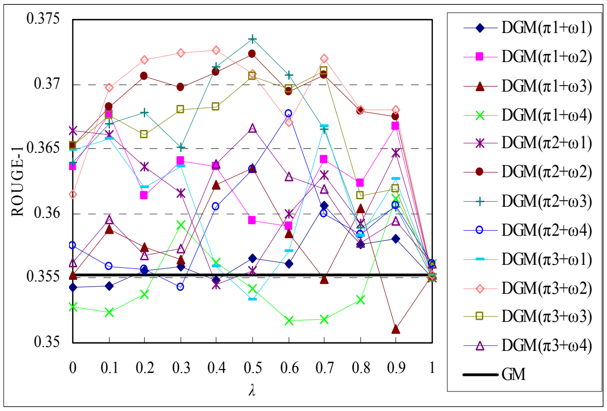

Figure 3. ROUGE-1 vs. $\lambda$ on DUC2001

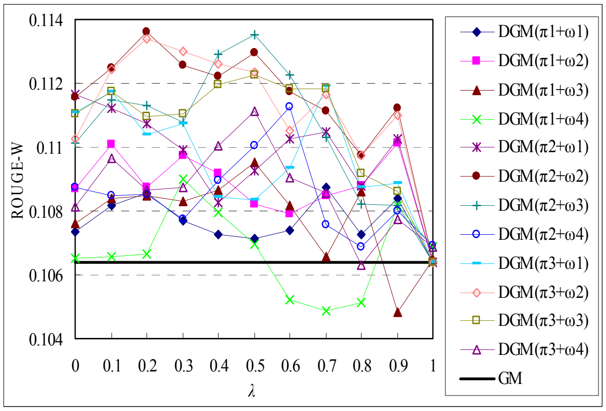

Figure 4. ROUGE-W vs. $\lambda$ on DUC2001

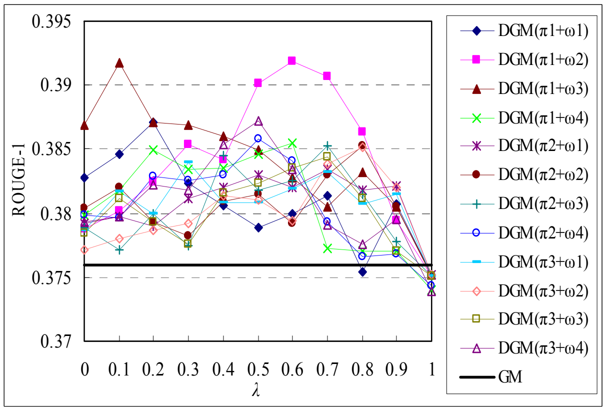

Figure 5. ROUGE-1 vs. $\lambda$ on DUC2002

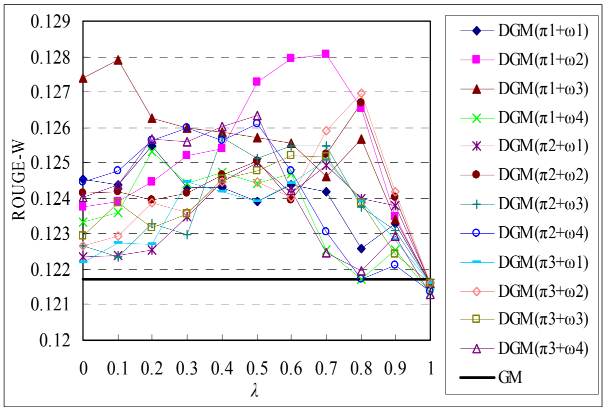

Figure 6. ROUGE-W vs. $\lambda$ on DUC2002

We can see from the figures that the proposed document-based graph model with different settings can almost always outperform the basic graph-based model, with respect to different values of $\lambda$. The results show the robustness of the proposed model. We can also see that for most settings of the propose model, very large values or very small values of $\lambda$ can deteriorate the summarization performance, i.e. both the first document and the second document in the computation of the conditional affinity weight between sentences have great impact on the summarization performance.

## 6 Conclusion and Future Work

This paper examines the document impact on the graph-based model for multi-document summarization. The document-level information and the sentence-to-document relationship are incorporated into the graph-based ranking algorithm. The experimental results on DUC2001 and DUC2002 demonstrate the good effectiveness of the proposed model.

In this study, we directly make use of the coarse-grained document-level information. Actually, a document can be segmented into a few subtopic passages by using the TextTiling algorithm (Hearst, 1997), and we believe the subtopic passage is more fine-grained than the original document. In future work, we will exploit this kind of subtopic-level information to further improve the summarization performance.

## Acknowledgments

This work was supported by the National Science Foundation of China (No.60703064), the Research Fund for the Doctoral Program of Higher Education of China (No.20070001059) and the National High Technology Research and Development Program of China (No.2008AA01Z421). We also thank the anonymous reviewers for their useful comments.

## References

R. Barzilay, K. R. McKeown and M. Elhadad. 1999. Information fusion in the context of multi-document summarization. In Proceedings of ACL1999.

H. Daumé and D. Marcu. 2006. Bayesian query-focused summarization. 2006. In Proceedings of COLING-ACL2006.

G. Erkan and D. Radev. 2004. LexPageRank: prestige in multi-document text summarization. In Proceedings of EMNLP'04.

J. Goldstein, M. Kantrowitz, V. Mittal and J. Carbonell. 1999. Summarizing text documents: sentence selection and evaluation metrics. In Proceedings of SIGIR-99.

S. Gupta, A. Nenkova and D. Jurafsky. 2007. Measuring importance and query relevance in topic-focused multi-document summarization. In Proceedings of ACL-07.

S. Harabagiu and F. Lacatusu. 2005. Topic themes for multi-document summarization. In Proceedings of SIGIR'05.

H. Hardy, N. Shimizu, T. Strzalkowski, L. Ting, G. B. Wise. and X. Zhang. 2002. Cross-document summarization by concept classification. In Proceedings of SIGIR'02.

M. Hearst. 1997. TextTiling: segmenting text into multi-paragraph subtopic passages. Computational Linguistics, 23(1): 33-64.

K. Knight. and D. Marcu. 2002. Summarization beyond sentence extraction: a probabilistic approach to sentence compression, Artificial Intelligence, 139(1).

W. Kraaij, M. Spitters and M. van der Heijden. 2001. Combining a mixture language model and Naïve Bayes for multi-document summarization. In SIGIR 2001 Workshop on Text Summarization.

A. Leuski, C.-Y. Lin and E. Hovy. 2003. iNeATS: interactive multi-document summarization. In Proceedings of ACL2003.

C.-Y. Lin and E. H. Hovy. 2002. From single to multi-document summarization: a prototype system and its evaluation. In Proceedings of ACL-2002.

C.-Y. Lin and E. H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of HLT-NAACL2003.

T.-Y. Liu and W.-Y. Ma. 2005. Webpage importance analysis using Conditional Markov Random Walk. In Proceedings of WI2005.

I. Mani and E. Bloedorn. 2000. Summarizing similarities and differences among related documents. Information Retrieval, 1(1).

D. Marcu. Discourse-based summarization in DUC–2001. 2001. In SIGIR 2001 Workshop on Text Summarization.

K. McKeown, J. Klavans, V. Hatzivassiloglou, R. Barzilay and E. Eskin. 1999. Towards multidocument summarization by reformulation: progress and prospects, in Proceedings of AAAI1999.

R. Mihalcea and P. Tarau. 2005. A language independent algorithm for single and multiple document summarization. In Proceedings of IJCNLP'2005.

A. Nenkova and A. Louis. 2008. Can you summarize this? Identifying correlates of input difficulty for generic multi-document summarization. In Proceedings of ACL-08: HLT.

L. Page, S. Brin, R. Motwani and T. Winograd. 1998. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Libraries.

D. R. Radev, H. Y. Jing, M. Stys and D. Tam. 2004. Centroid-based summarization of multiple documents. Information Processing and Management, 40: 919-938.

X. Wan and J. Yang. 2006. Improved affinity graph based multi-document summarization. In Proceedings of HLT-NAACL2006.

X. Wan, J. Yang and J. Xiao. 2007. Manifold-ranking based topic-focused multi-document summarization. In Proceedings of IJCAI2007.