Applying Two-Level Reinforcement Ranking in Query-Oriented Multidocument Summarization

Furu Wei
Department of Computing, Hong Kong Polytechnic University, Hong Kong, and Department of Computer Science and Technology, Wuhan University, P.R. China. E-mail: csfwei@comp.polyu.edu.hk; frwei@whu.edu.cn

Wenjie Li and Qin Lu
Department of Computing, Hong Kong Polytechnic University, Hong Kong. E-mail: {cswjli, csluqin}@comp.polyu.edu.hk

Yanxiang He
Department of Computer Science and Technology, Wuhan University, P.R. China. E-mail: yxhe@whu.edu.cn

Sentence ranking is the issue of most concern in document summarization today. While traditional feature-based approaches evaluate sentence significance and rank the sentences relying on the features that are particularly designed to characterize the different aspects of the individual sentences, the newly emerging graph-based ranking algorithms (such as the PageRank-like algorithms) recursively compute sentence significance using the global information in a text graph that links sentences together. In general, the existing PageRank-like algorithms can model well the phenomena that a sentence is important if it is linked by many other important sentences. Or they are capable of modeling the mutual reinforcement among the sentences in the text graph. However, when dealing with multidocument summarization these algorithms often assemble a set of documents into one large file. The document dimension is totally ignored. In this article we present a framework to model the two-level mutual reinforcement among sentences as well as documents. Under this framework we design and develop a novel ranking algorithm such that the document reinforcement is taken into account in the process of sentence ranking. The convergence issue is examined. We also explore an interesting and important property of the proposed algorithm. When evaluated on the DUC 2005 and 2006 query-oriented multidocument summarization datasets, significant results are achieved.

Introduction

The explosion of the World Wide Web has brought with it a vast amount of information. It has become virtually impossible for anyone to read and understand large numbers of individual documents that are abundantly available. Automatic document summarization provides an effective means to manage such an exponentially increased collection of information and to support information seeking and condensing goals.

The main evaluation forum providing benchmarks for the researchers who work on document summarization (Mani & Maybury, 1999; Jones, 2007) to exchange their ideas and experiences is the Document Understanding Conferences (DUC). The goals of the DUC are to enable researchers to participate in large-scale experiments on the standard benchmark and to increase the availability of appropriate evaluation techniques. Over the past years the DUC evaluations have gradually evolved from single-document summarization to multidocument summarization and from generic summarization to query-oriented summarization (Over, Dang, & Harman, 2007). Query-oriented multidocument summarization initiated by the DUC in 2005 aimed to produce a short and concise summary for a set of relevant documents according to a given query that describes a user’s information need.

Up to the present, the dominant approaches in document summarization regardless of the nature and the goals of the tasks have still been built on the sentence extraction framework. Under this framework, sentence ranking is the issue of most concern. It computes sentence significance based on certain criteria and ranks the sentences according to significance. Most previous work in the literature addresses the ranking issue by examining the features of each individual sentence, such as its content, grammatical structure, etc. Recently, the relationship or the association among the sentences is taken into account in the graph-based models that represent
a document or a set of documents as a text graph. The text graph is constructed by taking a text unit, such as a term or a sentence, as a node, and the similarity, the relevance, or the association between text units as links. The significance of a node in the graph is then estimated by the graph-based ranking algorithms that normally take into account the global information recursively computed from the entire graph rather than merely relying on the local information. So far, the most popular graph-based ranking algorithms applied in document summarization are Google’s PageRank (Brin & Page, 1998) and its variations. LexRank (Erkan & Radev, 2004), developed for generic summarization, is one example of the PageRank-like algorithms. LexRank has also been extended to its topic-sensitive version (Otterbacher, Erkan, & Radev, 2005) to accommodate the new challenge of query-oriented summarization.

The existing PageRank-like ranking algorithms for document summarization are capable of modeling the mutual reinforcement among the sentences, as they are in a position to capture the fact that a sentence is important when it is linked by many other important sentences. However, when dealing with the task of multidocument summarization, they often assemble a set of documents into one large file and ignore the intrinsic difference between the tasks of single document summarization and multidocument summarization. Or rather, the information carried by the document dimension is totally ignored during the reinforcement ranking. In this article we study how to model the reinforcement of the documents in the sentence ranking process. The main contributions of the paper are three-fold:

1. A formal framework is presented to model both the internal and the external reinforcement (i.e., two-level mutual reinforcement) among two sets of relevant objects (e.g., documents and sentences in this article).
2. A solid mathematical description for the document-sentence mutual reinforcement (D-S MR for short) ranking is provided. The convergence of the algorithm is guaranteed. More important, an interesting and important property has been observed when the symmetric weight matrix is designed.
3. The effectiveness of the proposed two-level mutual reinforcement ranking algorithm is examined in the context of query-oriented multidocument summarization.

The remainder of this article is organized as follows. The section Related Work reviews existing feature-based and graph-based ranking approaches applied in document summarization. Document-Sentence Mutual Reinforcement Framework and Ranking Algorithm introduces the proposed two-level reinforcement framework and explains the D-S MR algorithm and its extension Qs-MR algorithm. Query-Oriented Multidocument Summarization: An Application presents the application of the Qs-MR in query-oriented summarization, especially focusing on sentence ranking. Then, Experiment and Evaluation presents experiments and evaluations, and Discussion: DS-MR vs. PR presents discussions and a case study. The final section concludes the article.
that a term should have a high salience score if it appears in many sentences with high salience scores while a sentence should have a high salience score if it contains many terms with high salience scores. This mutual reinforcement principle was reduced to a solution for the singular vectors of the transition matrix \( \text{of the bipartite graph} \). In fact, as early in 1998 a similar idea was used in HITS algorithm (Kleinberg, 1999) to identify hub and authority Web pages in a small subset of the Web graph. Zha’s work was later advanced by Wan et al. (2007), who additionally calculated the links among the sentences and the links among the terms. Zha’s and Wan et al.’s works are the ones most relevant to our studies presented in this article. But they all concentrated on single-document generic summarization.

The use of the PageRank family was also very popular in event-based summarization approaches (Leskovec, Grobelnik, & Milic-Frayling, 2004; Vanderwende, Banko, & Menezes, 2004; Yoshioka & Haraguchi, 2004; Li, Wu, Lu, Xu, & Yuan, 2006). In contrast to conventional sentence-based approaches, event-based approaches took event terms, such as verbs and action nouns and their associated named entities as graph nodes, and connected nodes according to their co-occurrence information or semantic dependency relations. They were able to provide finer text representation and thus could be in favor of sentence compression that was targeted to include more informative contents in a fixed-length summary. Nevertheless, these advantages largely relied on appropriately defining and selecting event terms.

Document-Sentence Mutual Reinforcement Framework and Ranking Algorithm

In many text processing applications, such as information retrieval, question answering, and document summarization, the text people often manipulate and evaluate two different granularities: document and sentence. While document ranking is indispensable to information retrieval, sentence ranking is one of the most fundamental issues in document summarization. Comparatively speaking, sentence ranking is more challenging than document ranking since a sentence carries much less information than a document for measuring the similarity of text. However, the sentence does not stand alone in the text without the context.

It is an unarguable fact that the text is always organized and structured in a certain way so that the core information would be easily identified. The assumption that document and sentence are independent of each other in delivering meanings is untenable. Therefore, even when the sentence ranking result is the only concern in summarization, the mutual constraints and the influences between document and sentence could not be ignored. In this section, we propose a new sentence ranking algorithm based on the Mutual Reinforcement (MR) of Document (D) and Sentence (S). We define the reinforcement between document and sentence as the external reinforcement.

In addition to the external reinforcement, the proposed sentence ranking algorithm also supports the calculation of the internal reinforcement within a set of documents or a set of sentences, i.e., the document-level reinforcement and the sentence-level reinforcement. The existing PageRank-like algorithms employed in summarization can be viewed as sentence-level reinforcement instances. In the past the importance of sentence relations have been stressed in graph-based summarization models and their contribution to the performance improvement has been recognized (Erkan & Radev, 2004). We put them forward to the relations at both the document level and the sentence level and move toward a more unified reinforcement model. To sum up, the external and the internal reinforcement together form a complete two-level document and sentence mutual reinforcement (D-S MR or MR for short) framework, as illustrated in Figure 1.

**Formal Description of D-S MR**

The mutual reinforcement framework is developed with an attempt to capture the following intuitions:

1. A document is important if (1) it correlates to important sentences; (2) it associates with other important documents;
2. A sentence is important if (1) it correlates to important documents; (2) it associates with other important sentences.

Then the ranking of documents and sentences can be iteratively derived from the D-S MR. Let \( R_D \) and \( R_S \) denote the ranking scores of the document set \( D \) and the sentence set \( S \), respectively, the iterative ranking can be formulated as follows:

\[
\begin{align*}
R_D^{(k+1)} &= \alpha_1 \cdot D_D \cdot R_D^{(k)} + \beta_1 \cdot D_S \cdot R_S^{(k)} \\
R_S^{(k+1)} &= \beta_2 \cdot S_D \cdot R_D^{(k)} + \alpha_2 \cdot S_S \cdot R_S^{(k)}
\end{align*}
\]

where \( D_D \) denotes the D-D affinity matrix, \( D_S \) denotes the D-S affinity matrix, and so on. The calculation of the four affinity matrices in Equation (1) will be detailed later in Query-Oriented Multidocument Summarization: An Application. \( W = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix} \) is the weight matrix used to balance the relative weights of document and sentence in D-S MR.
The coefficients in Equation (1) corresponds to a block matrix.

\[ M = \begin{bmatrix} \alpha_1 D_D & \beta_1 D_S \\ \beta_2 S_D & \alpha_2 S_S \end{bmatrix} \]  

(2)

Let \( R = \begin{bmatrix} R_D \\ R_S \end{bmatrix} \) then \( R \) can be computed as the dominant eigenvector of \( M \), i.e.:

\[ M \cdot R = \lambda \cdot R \]  

(3)

Given that the corresponding graph of \( M \) is not bipartite, we must force \( M \) stochastic, irreducible, and primitive\(^1\) in order to guarantee a unique solution of \( R \). On this account, the necessary matrix transformation explained below must be performed. We will prove to readers that the new transformed \( M \) is stochastic, irreducible, and more strictly, primitive for certain.

A sufficient condition for a stochastic \( M \) is to make the four affinity block matrices in \( M \) column stochastic. For the sake of simplicity, we let \( X \) be either of the two diagonal block matrices (i.e., \( D_D \) and \( S_D \)) and \( Y \) be either of the remaining two block matrices (i.e., \( S_S \) and \( D_S \)).

We first delete the rows and the columns that do not contain any nonzero element in \( X \). This manipulation is analogous to the strategy used in PageRank to cope with the dangling pages in the Web graph that do not have outgoing links. Since \( X \) is symmetric, if the out-degree of a document or a sentence node is zero, its in-degree must be zero as well. Such a node is actually an isolated node in a text graph. Therefore, the ranking results will not be influenced when the isolated nodes are removed. On the other hand, it is noted that there are no zero columns in \( Y \). Let us take \( S_P \), for example. The affinity of the sentence \( s \) and the document \( d \) is at least greater than zero if \( d \) contains \( s \). Now, we are ready to normalize both \( X \) and \( Y \) by columns to their column stochastic versions \( \bar{X} \) and \( \bar{Y} \). We replace \( X \) and \( Y \) with \( \bar{X} \) and \( \bar{Y} \) in \( M \), and denote the new matrix as \( \bar{M} \).

Next, we manage to make \( \bar{M} \) irreducible. Let \( \bar{X} \) denote either of the two new diagonal block matrices in \( \bar{M} \). Similar to the treatment used in PageRank calculation, we make the graph corresponding to \( \bar{X} \) strongly connected by adding (artificial) links for every pair of nodes with a probability vector \( \bar{p} \). After such an adjustment, the revised \( \bar{X} \) becomes:

\[ \bar{X} = d \cdot \bar{X} + (1 - d)E \] 

where \( 0 < d < 1 \), \( d \) is usually set to 0.85 according to PageRank. \( k \) is the order of \( \bar{X} \). The probability vector \( \bar{p} \) can be defined in many different ways. A typical definition is to assume a uniform distribution over all elements, i.e., \( \bar{p} = [1/k]_{1 \times k} \). By doing so, \( \bar{X} \) becomes both stochastic and irreducible. We finally replace \( X \) with \( \bar{X} \) in \( M \), and let \( \bar{M} \) denote the latest matrix.

After the above-mentioned transformations on the matrices, we now can prove that the final \( \bar{M} \) is column stochastic, irreducible, and primitive. For the sake of simplicity, we rewrite \( \bar{M} \) as \( \bar{P} = \begin{bmatrix} \alpha_1 \cdot P_{11} \beta_1 \cdot P_{12} \\ \beta_2 \cdot P_{21} \alpha_2 \cdot P_{22} \end{bmatrix} \) and let \( P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \) and \( W = \begin{bmatrix} \alpha_1 & \beta_1 \\ \beta_2 & \alpha_2 \end{bmatrix} \). From the previous analysis, we have:

(1) \( P_{11}(m \times m) > 0 \), \( P_{22}(m \times m) > 0 \), \( P_{21}(n \times n) > 0 \), \( P_{21}(n \times m) > 0 \);
(2) \( P_{11} \), \( P_{12} \), \( P_{21} \) and \( P_{22} \) are column stochastic;
(3) \( \forall i \in [1, n] \) and \( \exists j \in [1, m] \) such that \( P_{21}(i, j) > 0 \);
(4) \( P_{12} \) and \( P_{21} \) satisfy \( P_{21}(i, j) > 0 \iff P_{21}(j, i) > 0 \) and \( P_{21}(i, j) = 0 \iff P_{21}(j, i) = 0 \); and
(5) It is easy to ensure \( W > 0 \) and make \( W \) column stochastic.

**Lemma 1.** \( \bar{P} \) is also column stochastic if the weight matrix \( W \) is column stochastic.

**Proof.** Let \( A \) and \( B \) denote the two block matrices in any column of \( \bar{P} \) under concern, \( \alpha \) and \( \beta \) the corresponding weight coefficient with respect to \( A \) and \( B \), then \( \sum_i \bar{P}_{ij} = \alpha \sum_i A_{ij} + \beta \sum_i B_{ij} = \alpha + \beta = 1. \)

**Lemma 2.** \( \bar{P} \) is irreducible.

**Proof.** Since the two graphs corresponding to the two diagonal block matrices in \( \bar{P} \) are strongly connected (i.e., they are irreducible) and the links connecting the two graphs are bidirectional, obviously the graph corresponding to \( \bar{P} \) is also strongly connected. Thus, \( \bar{P} \) must be irreducible.

Now the matrix \( \bar{P} \) is both stochastic and irreducible. More strictly, we have

**Lemma 3.** \( \bar{P} \) is primitive.

**Proof.** Considering

\[ \bar{P}^2 = \begin{bmatrix} \alpha_1 \cdot P_{11} \beta_1 \cdot P_{12} \\ \beta_2 \cdot P_{21} \alpha_2 \cdot P_{22} \end{bmatrix} \cdot \begin{bmatrix} \alpha_1 \cdot P_{11} \beta_1 \cdot P_{12} \\ \beta_2 \cdot P_{21} \alpha_2 \cdot P_{22} \end{bmatrix} = \begin{bmatrix} \alpha_1^2 \cdot P_{11}^2 + \beta_1^2 \cdot P_{12} \cdot P_{21} + \alpha_2^2 \cdot P_{22} \\ \beta_2^2 \cdot P_{21} \cdot P_{11} + \beta_1 \cdot \beta_2 \cdot P_{12} \cdot P_{22} + \alpha_1 \cdot \alpha_2 \cdot P_{11} \cdot P_{22} \end{bmatrix} \]

we have:

(1) \( \alpha_1^2 \cdot P_{11}^2 + \beta_1^2 \cdot P_{12} \cdot P_{21} > 0 \) \( \quad (P_{11}^2 > 0) \)
(2) \( \alpha_2 \cdot P_{21} \cdot P_{11} + \alpha_1 \cdot \beta_2 \cdot P_{12} \cdot P_{22} > 0 \) \( \quad (P_{11} \cdot P_{22} > 0) \)
(3) \( \beta_1 \cdot P_{11} \cdot P_{12} + \alpha_2 \cdot \beta_1 \cdot P_{21} > 0 \) \( \quad (P_{12} > 0) \)
(4) \( \alpha_1 \cdot \beta_2 \cdot P_{21} \cdot P_{11} + \beta_1 \cdot \beta_2 \cdot P_{22} \cdot P_{21} > 0 \) \( \quad (P_{22} \cdot P_{21} > 0) \)

It is easy to deduce that \( \bar{P}^2 > 0 \) and \( \bar{P} \) is primitive.

\(^1\)A matrix is irreducible if its graph shows that every node is reachable from every other node. A nonnegative, irreducible matrix is primitive if it has one eigenvalue on its spectral circle. An irreducible Markov chain with a primitive transition matrix is called an aperiodic chain. Please refer to Langville & Meyer (2004) for more details.

\(^2\)For each sentence, there exists at least one document that contains that sentence such that the element in the affinity matrix is a positive value because the affinity between them is positive, and vice versa.
The above proof can be understood from the perspective of graph. Let $G_1$ denote the graph corresponding to the matrix $P_{11}$, $G_2$ denote the graph corresponding to the matrix $P_{22}$, and $G$ denote the graph corresponding to the matrix $P$. $P_{12}$ and $P_{21}$ can be viewed as the links connecting the nodes between $G_1$ and $G_2$. Notice that any two nodes in $G_1$ or $G_2$ are connected and there is at least one link from the nodes in $G_1$ to $G_2$, and vice versa. The nodes in $G$ have been divided into two sets, i.e., $G_1$ and $G_2$. There is no question that any two nodes in the same set (i.e., within $G_1$ or within $G_2$) are able to reach each other in exactly two steps. Then we may conclude that the paths are reversible in $G$. In conclusion, any two nodes in $G$ are able to reach each other in exactly two steps, which means the matrix $P$ is reversible in $G$.

As a result, we can compute the unique dominant eigen-vector (with 1 as the eigenvalue) of $P$. It is well known that the power method applied to $P$ will converge to $R$.

$$M \cdot R = \lambda \cdot R \tag{5}$$

Eventually, we can develop an iterative algorithm to solve Equation (1).

**Algorithm 1: Rank($P_{11}$, $P_{12}$, $P_{21}$, $P_{22}$, $W$, $R^{(0)}$)**

**Require:** $R^{(0)}$ can be a random probability vector such that $|R^{(0)}| = 1$

1: $k \leftarrow 0$, $\gamma$;
2: while ($\gamma \leq 3^k$) do
3: \begin{align*}
& R_{1}^{(k+1)} = \alpha_1 \cdot P_{11} \cdot R_{1}^{(k)} + \beta_1 \cdot P_{12} \cdot R_{2}^{(k)}; \\
& R_{2}^{(k+1)} = \beta_2 \cdot P_{21} \cdot R_{1}^{(k)} + \alpha_2 \cdot P_{22} \cdot R_{2}^{(k)};
\end{align*}
4: $\gamma \leftarrow \max \left( \frac{\|R_{1}^{(k+1)} - R_{1}^{(k)}\|_1}{\|R_{2}^{(k+1)} - R_{2}^{(k)}\|_1} \right)$;
5: $k \leftarrow k + 1$;
6: End of while;
7: $R^{(k)} \leftarrow [R_{1}^{(k)} ~ R_{2}^{(k)}]$;
8: Return $R^{(k)}$.

**Weight Matrix Design**

A critical issue in implementing Equation (1) is to design the appropriate weight matrix $W$. Essentially, a positive column stochastic matrix is expected. We design a symmetric weight matrix, i.e., $W = \begin{bmatrix} \alpha & \beta \\ \beta & \alpha \end{bmatrix}$. Although $W$ is necessary to be column stochastic and positive in our previous analysis, we use the weight matrix before it is normalized to be column stochastic for ease of illustration and explanation. Generally speaking, $\alpha$ (set to 1 as reference) indicates the weight of the internal reinforcement and $\beta$ (set to $\alpha$) the external reinforcement. The motivation of this design is straightforward. It is reasonable to assume that the internal reinforcement is more important than the external reinforcement. But it seems unnecessary to further distinguish the weights of different kinds of internal reinforcement (i.e., documents-to-document and sentence-to-sentence reinforcement) or external reinforcement (i.e., documents-to-sentence and sentence-to-document reinforcement).

By designing such a symmetric weight matrix, we come up with the following interesting and important significance conservation property of the ranking solution in Algorithm 1.

**Proposition 1.** The significance is redistributed within the scope of document set or sentence set in each iteration. However, the sum of the scores in each set remains the same during the ranking iterations. In other words, sentences (or documents) compete with one for a higher significance, but they will not jump across the set boundary to grab the significance from documents (or sentences). Formally, let $R^{(0)} = [R^{(0)}_1 ~ R^{(0)}_2]$, we have $|R^{(n)}_1| = |R^{(n)}_2| = \gamma$ if $|R^{(0)}_1| = |R^{(0)}_2| = \gamma$. $\gamma$ can be any arbitrary positive value. Later in the experiments, we set it to 1/2 so as to ensure $|R^{(0)}| = 1$.

**Proof.** We complete the proof by mathematical induction.

1. Given that, $|R^{(0)}_1| = \gamma$ and $|R^{(0)}_2| = \gamma$;
2. Assume $|R^{(k)}_1| = \gamma$ and $|R^{(k)}_2| = \gamma$ when $n = k$, then,
\begin{align*}
|R^{(k+1)}_1| &= |\alpha \cdot P_{11} \cdot R^{(k)}_1 + \beta \cdot P_{12} \cdot R^{(k)}_2| = (\alpha + \beta) \cdot \gamma = \gamma, \\
|R^{(k+1)}_2| &= |\beta \cdot P_{21} \cdot R^{(k)}_1 + \alpha \cdot P_{22} \cdot R^{(k)}_2| = (\alpha + \beta) \cdot \gamma = \gamma.
\end{align*}

This means that the conservation of total significance in a document set or a sentence set also holds at $n = k + 1$ if it holds at $n = k$. Therefore, Proposition 1 is true.

This proposition is meaningful in the context. Given the initial individual significance of a sentence (or a document) and the accumulated total significance of all the sentences (or the documents), the ranking Algorithm 1 can be viewed as iteratively redistributing the total significance among the sentences (or the documents) by the mutual reinforcement of document and sentence (including both external and internal) according to the link structure (i.e., the affinity graph) of them. We believe that the documents influence the ranking of sentences and vice versa. In other words, the external reinforcement from the documents provides useful hints to guide the sentence internal rank competition, and the other way around. However, this does not mean that the total significance of the sentences (or the documents) would change during ranking iterations. The redistribution of the total significance should not cross over the set boundary of the sentences (or the documents). In short, document and sentence are interactive during ranking iterations but they still have certain independence. It is meaningless for the text with different granularities to compete with each other for a higher significance. Significance of a document and a sentence are not comparable.
Another advantage of using the symmetric weight matrix is that we only need to tune and fix one parameter when we design the weight matrix. In this context, we only need to determine the proportion of the internal-reinforcement and the external-reinforcement weight. We will discuss the parameter issues later in the section The Weight Matrix.

Query-Sensitive D-S MR (Qs-MR)

In the previously introduced D-S MR framework, the reinforcement of document and sentence is query-unaware. That is, only the content of the text is concerned. However, for tasks like query-oriented summarization, how the reinforcement is biased to an external context (such as a user’s query) is often of great interest.

A general way to incorporate the query information into the general D-S MR framework is to impose the influence of a user’s query on each text unit (document or sentence) such that it works in the internal reinforcement. This can be viewed somewhat as a topic-sensitive PageRank (Haveliwala, 2003) at each level of text granularity. The key to make ranking biased toward the query rests with the definition of the query-sensitive probability vector \( \mathbf{p} \). A simple yet effective solution is to define \( \mathbf{p} \) as:

\[
\mathbf{p}_i = \begin{cases} 
    \text{rel}(t_i|q) & \text{if } \text{rel}(t_i|q) \neq 0 \\
    \theta & \text{otherwise}
\end{cases} \quad (6)
\]

where \( t_i \) can be either a document or a sentence, \( \text{rel}(t_i|q) \) denotes the relevance of \( t_i \) to \( q \) and can be calculated by cosine similarity. \( \theta \) is an extremely small real number to avoid zero elements in \( \mathbf{p} \). \( \mathbf{p} \) is further normalized to 1 in order for it to be a probability vector.

Query-Oriented Multidocument Summarization: An Application

Task Definition of DUC Query-Oriented Multidocument Summarization

The query-oriented multidocument summarization task defined in the DUC evaluations requires generating a concise and well-organized summary for a set of the relevant documents according to a given query that simulates a user’s information need. The query usually consists of one or more interrogative and/or narrative sentences. Here is a query example from the DUC 2005 document set “d331f”:

According to the task definitions, system-generated summaries are strictly limited to 250 words in length.

Existing query-oriented summarization approaches basically follow the same processes: (1) first calculate the significance of the sentences with reference to the given query from different perspectives with/without using some sorts of sentence relations; (2) then rank the sentences according to certain criteria and measures; (3) finally extract the top-ranked but nonredundant sentences from the original documents to produce a summary. Under this extractive framework, undoubtedly the two critical processes involved are sentence ranking and sentence selection.

Sentence Ranking

In Document-Sentence Mutual Reinforcement Framework and Ranking Algorithm (above), we introduced the general MR and the extended Qs-MR frameworks that can prescribe the reinforcement-based procedure for ranking the text of different granularities (document and sentence) simultaneously. The frameworks themselves are applicable to both document retrieval and sentence retrieval. In this section, however, we manage to take the advantage of the Qs-MR framework to deal with the sentence ranking problem in the query-oriented multidocument summarization task.

We now come to design the affinity matrices (i.e., \( D_D \), \( S_S \), \( D_S \), and \( S_D \)) among the documents and sentences in Equation (1) and the query-sensitive probability vector \( \mathbf{p} \) in Equation (6). In this work, the affinity between any two units is measured by their cosine similarity, which is the most popular and widely used similarity measure in information retrieval and text-mining applications (Baeza-Yates & Ribeiro-Neto, 1999, Chapter 2.5.3). Similarly, the query-sensitive probability of a sentence (or document) is defined as the relevance between a sentence (or a document) and the query. It is calculated in the same way as the affinity. The documents, the sentences, as well as the query are all represented by the vectors of terms. We will describe the term weighting scheme in detail in Experiment Set-up (below).

Let \( \mathbf{a} \) and \( \mathbf{b} \) be two text vectors, the cosine similarity between them is calculated by \( \cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \). Then, the above mentioned affinity matrices and probability vector are defined as,

**Notation**

- \( D_D(d_i, d_j) \)
- \( S_S(s_i, s_j) \)
- \( D_S(d_i, s_j) \)
- \( S_D(s_i, d_j) \)
- \( p_i(d, q) \)
- \( p_i(s, q) \)

**Definition**

- \( \cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \)

We develop the following iterative procedure based on Qs-MR to rank the sentences for query-oriented multidocument...
summarization. Sentences are ranked according to their significance eventually converged in \( R_S \).

**Algorithm 2: RankSentence\((D, S, q)\)**

**Input:** The document set \( D \), the sentence set \( S \), and the query \( q \).

**Output:** The ranking vectors of \( R_D \) and \( R_S \).

1. Construct the affinity matrices \( D_D, D_S, S_D \) and \( S_S \);
2. Transform the four block matrices as mentioned in Section 3.1;
3. Design the symmetric weight matrix \( W \);
4. Choose (randomly) the initial non-negative vectors \( R_D^{(0)}, R_S^{(0)} \), such that \( |R_D^{(0)}| = 1/2 \) and \( |R_S^{(0)}| = 1/2 \);
5. If threshold \( \delta \) is satisfied, then stop;
6. \( J \leftarrow J + \delta \) do
7. \( \Pi \leftarrow \Pi \cup S_i \);
8. If \( (J < \xi) \) break;
9. End
10. End
11. Return \( \text{Rank}(D_D, D_S, S_D, S_S, W, R_D^{(0)}) \).

**Sentence Selection by Removing Redundancy**

In multidocument summarization, the number of the documents to be summarized can be very large. This makes the information redundancy problem appear to be more serious in multidocument summarization than in single-document summarization. Redundancy removal becomes an inevitable process. Since our focus in this study is the design of an effective (sentence) ranking algorithm, we apply the following straightforward yet effective sentence selection principle. We incrementally add into the summary the highest ranked sentence of concern if it does not significantly repeat the information already included in the summary until the word limitation of the summary is reached.

**Algorithm 3: GenerateSummary\((S, length)\)**

**Input:** sentence collection \( S \) (ranked in descending order of significance) and \( length \) (the given summary length limitation).

**Output:** The generated summary \( \Pi \)

1. \( \Pi \leftarrow \{ \}
2. \( \ell \leftarrow \text{length} \);
3. For \( i \leftarrow 0 \) to \( |S|\) do
4. \( \text{threshold} \leftarrow \max(\text{sim}(s_i, s) | s \in \Pi) \);
5. If \( \text{threshold} \leq \xi \) then
6. \( \Pi \leftarrow \Pi \cup S_i \);
7. \( \ell \leftarrow \ell - \text{len}(S_i) \);
8. If \( (\ell = 0) \) break;
9. End
10. End
11. Return \( \Pi \).

**Experiment and Evaluation**

**Experiment Set-up**

We conduct the experiments on the DUC 2005 and DUC 2006 datasets. Table 1 shows the basic statistics of the datasets. Each set of documents is accompanied with a query description representing a user’s information need.

**Parameter Setting on the DUC 2005**

The aim of the following experiments is to examine and fix the involved parameters on the DUC 2005 dataset. There are three parameters in our algorithm. They are: (1) the damping factor \( d \) used in the internal reinforcement; (2) the similarity threshold \( \delta \) for avoiding the link-by-chance phenomena; and (3) the weight matrix parameters \( \alpha \) and \( \beta \) for balancing the internal and external reinforcement.

**The Damping Factor.** We examine the damping factor \( d \) first. In these experiments, we set the similarity threshold \( d \) as \( 0.7 \) for the internal reinforcement.

**Table 1. Basic statistics of the DUC datasets.**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total number of document sets</th>
<th>Average number of documents per set</th>
<th>Average number of sentences per set</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC 2005</td>
<td>50</td>
<td>31.86</td>
<td>1002.54</td>
</tr>
<tr>
<td>DUC 2006</td>
<td>50</td>
<td>25</td>
<td>815.22</td>
</tr>
</tbody>
</table>

As for the evaluation metric, it is difficult to come up with a universally accepted method to measure the quality of machine-generated summaries. In fact, summary evaluation methods themselves are still an ongoing research in the summarization community. Much of the literature has addressed different methods for automatic evaluations other than human judges. Among them, ROUGE\(^5\) (Lin & Hovy, 2003) is supposed to produce the most reliable scores in correspondence with human evaluations. More important, it offers the advantage of being readily applied to compare the performance of different approaches on the same dataset. Given the fact that judgments by humans are time-consuming and labor-intensive and ROUGE has been officially adopted by the DUC for automatic evaluations since 2005, like other researchers we also use it as the evaluation criteria in this article.

Documents and queries are preprocessed by segmenting sentences and splitting words. Stop-words are then removed\(^6\) and the remaining words are stemmed with Porter Stemmer (18). In all the following experiments, both text units (i.e., documents or sentences) and queries are represented as the vectors of terms. Notice that the term weights are normally measured in summarization models by the TF*IDF scheme as in conventional vector space models (VSM). However, we argue that it would be more reasonable to use the sentence-level inverse sentence frequency (ISF) instead of the document-level IDF when dealing with a sentence-level text processing application. This has been verified in our early study (Wei, Li, Lu, & He, 2008). We define \( isfw = \log(N/sfw) \) where \( N \) is the total number of the sentences in the document set, and \( sfw \) is the number of the sentences where the word \( w \) appears. Then the weight of \( w \) is computed as \( tfw \cdot isfw = \theta \) in Equation (6) is assigned to 20% of the minimum value of the relevance of the documents (or the sentences) to the query in a document set.

---

\(^5\)ROUGE version 1.5.5 is used.

\(^6\)A list of 199 words is used to filter stop words.
When the value of $d$ is getting smaller, the ranking relies more on a query and vice versa. Consequently, for the task of query-oriented summarization, it is natural to find out if the results are relatively worse when $d$ is assigned with a small value or an extremely large value.

### The Similarity Threshold

The second set of experiments was conducted to examine the influence of the similarity threshold. Here we used the same weight matrix as in the previous experiments and set the damping factor $d$ to 0.75 according to the observation shown in Table 2. Intuitively, the similarity threshold is supposed to be a very small value. We run the Qs-MR algorithm by setting the range of $\delta$ to 0.00 and 0.10 and the step size to 0.01. Table 3 shows the ROUGE evaluation results.

The similarity threshold $\delta$ is used to avoid the link-by-chance phenomena when the two text units are linked together simply because they share one or two words by chance. We do not consider the links between the two sentences (or documents) if their similarity is below $\delta$. As shown in Table 3, the ranking algorithm can produce reliable results when $\delta$ is within the range of 0.02–0.05.

### The Weight Matrix

The aim of the third set of experiments is to examine the weight matrix. For simplicity of illustration, we use the weight matrix before it is normalized to be stochastic for presentation in this section. In our implementation, the corresponding normalized version is utilized. Recall that the weight matrix $W$ we design is symmetric, where parameters $\alpha$ and $\beta$ reflect the relative importance of the internal reinforcement and the external reinforcement. We set $\alpha = 1$ and then tune the values of $\beta$. In these experiments, the damping factor $d$ is set to 0.75 and similarity threshold is set to 0.03. We show the ROUGE evaluation results in Table 4(a).
Table 4(a). Experiments on weight matrix.

<table>
<thead>
<tr>
<th>β</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.3805</td>
<td>0.0777</td>
<td>0.1344</td>
</tr>
<tr>
<td></td>
<td>[0.3744, 0.3863]</td>
<td>[0.0739, 0.0815]</td>
<td>[0.1309, 0.1381]</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3831</td>
<td>0.0786</td>
<td>0.1355</td>
</tr>
<tr>
<td></td>
<td>[0.3772, 0.3891]</td>
<td>[0.0748, 0.0825]</td>
<td>[0.1319, 0.1391]</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3835</td>
<td>0.0797</td>
<td>0.1361</td>
</tr>
<tr>
<td></td>
<td>[0.3772, 0.3899]</td>
<td>[0.0757, 0.0840]</td>
<td>[0.1322, 0.1400]</td>
</tr>
<tr>
<td>0.4</td>
<td>0.3840</td>
<td>0.0803</td>
<td>0.1368</td>
</tr>
<tr>
<td></td>
<td>[0.3774, 0.3902]</td>
<td>[0.0762, 0.0846]</td>
<td>[0.1327, 0.1411]</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3861</td>
<td>0.0814</td>
<td>0.1384</td>
</tr>
<tr>
<td></td>
<td>[0.3797, 0.3924]</td>
<td>[0.0774, 0.0857]</td>
<td>[0.1344, 0.1426]</td>
</tr>
<tr>
<td>0.6</td>
<td>0.3868</td>
<td>0.0806</td>
<td>0.1384</td>
</tr>
<tr>
<td></td>
<td>[0.3806, 0.3932]</td>
<td>[0.0767, 0.0848]</td>
<td>[0.1346, 0.1424]</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3860</td>
<td>0.0780</td>
<td>0.1378</td>
</tr>
<tr>
<td></td>
<td>[0.3797, 0.3925]</td>
<td>[0.0761, 0.0841]</td>
<td>[0.1339, 0.1417]</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3855</td>
<td>0.0792</td>
<td>0.1373</td>
</tr>
<tr>
<td></td>
<td>[0.3793, 0.3918]</td>
<td>[0.0758, 0.0836]</td>
<td>[0.1339, 0.1416]</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3851</td>
<td>0.0779</td>
<td>0.1329</td>
</tr>
<tr>
<td></td>
<td>[0.3788, 0.3914]</td>
<td>[0.0753, 0.0832]</td>
<td>[0.1335, 0.1413]</td>
</tr>
<tr>
<td>1.0</td>
<td>0.3859</td>
<td>0.0786</td>
<td>0.1372</td>
</tr>
<tr>
<td></td>
<td>[0.3796, 0.3923]</td>
<td>[0.0747, 0.0826]</td>
<td>[0.1335, 0.1412]</td>
</tr>
</tbody>
</table>

Table 4(b). Experiments on weight matrix.

<table>
<thead>
<tr>
<th>β</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>0.3859</td>
<td>0.0793</td>
<td>0.1370</td>
</tr>
<tr>
<td></td>
<td>[0.3797, 0.3921]</td>
<td>[0.0752, 0.0835]</td>
<td>[0.1332, 0.1411]</td>
</tr>
<tr>
<td>3.0</td>
<td>0.3817</td>
<td>0.0772</td>
<td>0.1338</td>
</tr>
<tr>
<td></td>
<td>[0.3756, 0.3877]</td>
<td>[0.0735, 0.0807]</td>
<td>[0.1302, 0.1373]</td>
</tr>
<tr>
<td>4.0</td>
<td>0.3876</td>
<td>0.0764</td>
<td>0.1329</td>
</tr>
<tr>
<td></td>
<td>[0.3736, 0.3858]</td>
<td>[0.0728, 0.0798]</td>
<td>[0.1294, 0.1363]</td>
</tr>
<tr>
<td>5.0</td>
<td>0.3787</td>
<td>0.0758</td>
<td>0.1324</td>
</tr>
<tr>
<td></td>
<td>[0.3727, 0.3846]</td>
<td>[0.0724, 0.0793]</td>
<td>[0.1290, 0.1359]</td>
</tr>
</tbody>
</table>

Table 4(c). Experiments on weight matrix.

<table>
<thead>
<tr>
<th>μ</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.3805</td>
<td>0.0777</td>
<td>0.1346</td>
</tr>
<tr>
<td></td>
<td>[0.3745, 0.3864]</td>
<td>[0.0738, 0.0816]</td>
<td>[0.1310, 0.1383]</td>
</tr>
<tr>
<td>0.2</td>
<td>0.3831</td>
<td>0.0786</td>
<td>0.1355</td>
</tr>
<tr>
<td></td>
<td>[0.3772, 0.3892]</td>
<td>[0.0749, 0.0825]</td>
<td>[0.1319, 0.1391]</td>
</tr>
<tr>
<td>0.3</td>
<td>0.3835</td>
<td>0.0798</td>
<td>0.1361</td>
</tr>
<tr>
<td></td>
<td>[0.3772, 0.3899]</td>
<td>[0.0758, 0.0840]</td>
<td>[0.1322, 0.1401]</td>
</tr>
<tr>
<td>0.4</td>
<td>0.3840</td>
<td>0.0803</td>
<td>0.1368</td>
</tr>
<tr>
<td></td>
<td>[0.3774, 0.3902]</td>
<td>[0.0762, 0.0846]</td>
<td>[0.1327, 0.1411]</td>
</tr>
<tr>
<td>0.5</td>
<td>0.3861</td>
<td>0.0814</td>
<td>0.1384</td>
</tr>
<tr>
<td></td>
<td>[0.3797, 0.3924]</td>
<td>[0.0774, 0.0857]</td>
<td>[0.1344, 0.1426]</td>
</tr>
<tr>
<td>0.6</td>
<td>0.3860</td>
<td>0.0806</td>
<td>0.1381</td>
</tr>
<tr>
<td></td>
<td>[0.3796, 0.3922]</td>
<td>[0.0767, 0.0847]</td>
<td>[0.1342, 0.1425]</td>
</tr>
<tr>
<td>0.7</td>
<td>0.3859</td>
<td>0.0800</td>
<td>0.1378</td>
</tr>
<tr>
<td></td>
<td>[0.3795, 0.3925]</td>
<td>[0.0760, 0.0841]</td>
<td>[0.1340, 0.1418]</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3856</td>
<td>0.0795</td>
<td>0.1376</td>
</tr>
<tr>
<td></td>
<td>[0.3795, 0.3920]</td>
<td>[0.0756, 0.0836]</td>
<td>[0.1338, 0.1415]</td>
</tr>
<tr>
<td>0.9</td>
<td>0.3854</td>
<td>0.0791</td>
<td>0.1375</td>
</tr>
<tr>
<td></td>
<td>[0.3791, 0.3916]</td>
<td>[0.0752, 0.0831]</td>
<td>[0.1337, 0.1414]</td>
</tr>
<tr>
<td>1.0</td>
<td>0.3859</td>
<td>0.0787</td>
<td>0.1375</td>
</tr>
<tr>
<td></td>
<td>[0.3797, 0.3922]</td>
<td>[0.0748, 0.0826]</td>
<td>[0.1337, 0.1415]</td>
</tr>
</tbody>
</table>

We can see from Table 4(a) that the ranking algorithm can produce stable and promising results in the range of 0.4–0.7 for β. We also test the cases that the external reinforcement is considered more important than the internal reinforcement (i.e., β > 1). From Table 4(b), the trend decline of the ROUGE results is observed when β gets bigger and bigger. It suggests β < 1 is a better choice than β > 1. This observation supports the common sense that the internal reinforcement should be more important than the external reinforcement in our D-S MR framework.

We are also interested to know the difference between the symmetric and the asymmetric versions of the weight matrix W. Now let $W = \begin{bmatrix} \alpha & \mu \\ \beta \mu & \alpha \end{bmatrix}$ denote the weight matrix before normalization as before. We set $\alpha = 1$ and $\beta = 0.5$ according to results from the previous experiments (see Table 4(a)) and then rerun the algorithm by setting the range of $\mu$ to 0.1 and 1.0 and the step size to 0.1. We fix $\beta$ (i.e., the weight of external reinforcement from sentence to document) in these experiments because the focus here is to rank the sentences for query-oriented multidocument summarization. Table 4(c) shows the ROUGE results.

As shown, the best performance is achieved at $\mu = 0.5$ when $W$ is a symmetric matrix. The experiments here demonstrate the effectiveness of the symmetric weight matrix from the empirical perspective, while the mathematical analysis in the section Weight Matrix Design provides an important property of the symmetric weight matrix from the theoretical perspective.

Comparison of Ranking Strategies

In this section we examine the effectiveness of the proposed Qs-MR based ranking algorithm for the task of query-oriented multidocument summarization. For comparison purpose, we also implement two other widely used and well-performed ranking strategies. One is to rank the sentences according to their relevance to the query (denoted by QR). The other is the PageRank deduced iterative ranking algorithm introduced in (Otterbacher et al., 2005) (denoted by PR). In the following experiments we use the parameter setting obtained from the previous experiments, i.e., 0.75 for the damping factor $d$, 0.03 for the similarity threshold and the normalized version of $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$ for the weight matrix.

Tables 5 and 6 show the ROUGE evaluation results on DUC 2005 and DUC 2006 datasets, respectively.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Qs-MR</td>
<td>0.4012</td>
<td>0.0914</td>
<td>0.1444</td>
</tr>
<tr>
<td></td>
<td>[0.3954, 0.4069]</td>
<td>[0.0873, 0.0956]</td>
<td>[0.1409, 0.1479]</td>
</tr>
<tr>
<td>PR</td>
<td>0.3899</td>
<td>0.0856</td>
<td>0.1394</td>
</tr>
<tr>
<td></td>
<td>[0.3833, 0.3964]</td>
<td>[0.0813, 0.0899]</td>
<td>[0.1353, 0.1438]</td>
</tr>
<tr>
<td>QR</td>
<td>0.3805</td>
<td>0.0781</td>
<td>0.1326</td>
</tr>
<tr>
<td></td>
<td>[0.3751, 0.3860]</td>
<td>[0.0743, 0.0817]</td>
<td>[0.1292, 0.1359]</td>
</tr>
</tbody>
</table>

TABLE 7. Summary of improvements by Qs-MR.

<table>
<thead>
<tr>
<th></th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
<th>ROUGE-SU4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Improvements over QR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DUC 2005</td>
<td>+7.88%</td>
<td>+22.59%</td>
<td>+12.61%</td>
</tr>
<tr>
<td>DUC 2006</td>
<td>+5.44%</td>
<td>+17.03%</td>
<td>+8.90%</td>
</tr>
<tr>
<td>Improvements over PR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DUC 2005</td>
<td>+4.29%</td>
<td>+12.28%</td>
<td>+5.97%</td>
</tr>
<tr>
<td>DUC 2006</td>
<td>+2.90%</td>
<td>+6.78%</td>
<td>+3.59%</td>
</tr>
</tbody>
</table>

As seen from Tables 5 and 6, our proposed algorithm outperforms the QR algorithm significantly. Meanwhile, it can also outperform the traditional graph-based ranking algorithm (i.e., PR). We summarize the improvements in Table 7.

Comparison With DUC Systems

We then compare our results with the DUC participating systems. To provide a global picture, we present the following representative ROUGE results of (1) the worst-scoring human summary (denoted by H), which reflects the margin between the machine-generated summaries and the human summaries; (2) the top five and worst participating systems according to their ROUGE-2 scores (e.g., S15, S17, etc.); and (3) the NIST baseline, which simply selects the first sentences as summaries from the documents until the summary length is achieved. We can then easily locate the positions of our system developed based on Qs-MR among them. Notice that the ROUGE-1 scores are not officially released by the DUC.

As illustrated in Tables 8 and 9, our developed system ranks 1st out the 31 systems in the DUC 2005 and 2nd out of the 32 systems in the DUC 2006. Particularly, it is ahead of the best system in the DUC 2005, i.e., 12.28% of ROUGE-2 and 5.17% of ROUGE-SU4 over S15. These improvements are definitely very encouraging. As shown in Table 8, the best system (i.e., S15) is only 1.07% over the second-best system (i.e., S17) on ROUGE-2 and 1.46% on ROUGE-SU4. Our system is supposed to be capable of achieving more significant results when further postprocessing such as redundancy removal or sentence compression is carried out.

Discussion: DS-MR vs. PR

In the context of multidocument summarization, it is very likely that the salient information repeats in the sentences from the different documents within the same or similar topics. While multiple salient and comparable sentences are available, compared with PR, DS-MR favors the sentences that come from the important documents (or the “good” documents). Once a document receives a higher score in the ranking process, the sentences from that document will obtain relatively higher scores by the iterative mutual reinforcement between the documents and the sentences. On the other hand, the sentences from the unimportant documents (or the “bad” documents) will have fewer opportunities to get higher scores, which in turn offer them less opportunity to be included in the summaries.

However, this does not mean that all the sentences from the important documents deserve the right to be selected as the summary sentences. Both DS-MR and PR follow the same assumption, i.e., if a sentence is very important it should be associated with many other important sentences in a document set. As a result, it would obtain an incremented score from the mutual reinforcement among the sentences. Both the document-sentence reinforcement and sentence-sentence reinforcement are implemented in the DS-MR algorithm. As a matter of fact, different from PR, DS-MR obtains the document ranking along with the sentence ranking. In other words, DS-MR performs document ranking as well. The rank of a document reflects the relative importance of the document. This information can be useful for further analysis of the documents.

Going one step further, we show a case study on the document set “d632i” in the DUC 2005 dataset in the Appendix. As shown, DS-MR produces a relatively better summary than PR, although they share a certain amount of sentences. Specifically, sentences 7 and 9 in the summary produced by DS-MR are good candidate sentences for summarizing


TABLE 10. Sentence and document distribution of the generated summaries on DUC 2005 dataset.

<table>
<thead>
<tr>
<th></th>
<th>Average number of sentences per summary (A)</th>
<th>Average number of summary sentences per summary (B)</th>
<th>Average number of summary sentences per summary document&lt;sup&gt;a&lt;/sup&gt; (A/B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS-MR</td>
<td>10.12</td>
<td>7.96</td>
<td>1.29</td>
</tr>
<tr>
<td>PR</td>
<td>11.02</td>
<td>8.5</td>
<td>1.27</td>
</tr>
</tbody>
</table>

<sup>a</sup>The sentence extracted from the document set and included in the summary is called a summary sentence. The document that contains at least one summary sentence is called a summary document.

“d632i,” whereas sentences 4 and 9 in the summary produced by PR are less relevant and make less sense. We also append the document ranking results produced by the DS-MR in the Appendix after two sample summaries. When taking a closer look at the above-mentioned sentences, we find that sentences 7 and 9 in the summary produced by DS-MR come from the 2nd (i.e., FT932-11184), and the 1st (FT931-345) ranked documents, respectively, while sentences 4 and 9 in the summary produced by PR both come from the 18th (i.e., LA022590-0075) ranked document out of the total of 25 documents. This demonstration illustrates the value of the document dimension in the task of sentence ranking for multidocument summarization.

If document dimension does matter, does DS-MR tend to pick out sentences from one or only a few documents? Like other readers, we are also interested in this question. Now, we take the DUC 2005 50 document set for statistics. As illustrated in Table 10, for the 250 words length summaries, there are about an average number of 10 sentences for each generated summary by DS-MR, compared to about 11 sentences by PR. Meanwhile, the sentences come from about an average number of 8 different documents by DS-MR in comparison with 8.5 by PR. These differences are not significant. It appears that the numbers of the documents from which DS-MR and PR pick up the summary sentences are similar. However, DS-MR is able to choose the sentence from those “better” documents than PR.

To summarize, on the one hand, DS-MR helps to select the “better” sentences that carry more global information and favors the sentences from the important documents based on the document-sentence mutual reinforcement. On the other hand, the final sentence scores are greatly determined by the sentences themselves, which is also highlighted in DS-MR. Notice that the internal mutual reinforcement (i.e., sentence-sentence) is supposed to be more important than the external mutual reinforcement (i.e., document-sentence), and it has been verified in previous experiments (see Table 4(a,b)).

Conclusion

In this article we introduce a two-level document and sentence mutual reinforcement framework and a ranking algorithm based on this framework that can rank both documents and sentences simultaneously in a graph linked by the affinity among them. Both the internal reinforcement (among documents or among sentences) and the external reinforcement (between documents and sentences) are well modeled. We pay more attention on the weight matrix for balancing the internal and external reinforcement. We also provide a solid mathematical description for the proposed ranking algorithm. Convergence is analyzed and an interesting and important property is explored.

We apply the proposed framework and the algorithm to the task of query-oriented multidocument summarization. Evaluation results on the DUC 2005 and 2006 datasets show that the ranking algorithm can significantly outperform the widely used and well-performed query relevance algorithm as well as the traditional graph-based PageRank-like algorithm. Meanwhile, it can outperform or be comparable to the best participating systems in the DUC 2005 and 2006.

Acknowledgments

This research was supported by the grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU5211/05E and PolyU5217/07E), a grant from National Natural Science Foundation of China (Project No. 60703008), and an internal grant from the Hong Kong Polytechnic University (Project No. A-PA6L).

References


Query description of the document set “d632i” in the DUC 2005 dataset:

<title> Southeast Asian Tin Mining </title>
<narr>
What is the status of tin mining in Southeast Asian countries? What has been the SE Asian tin industry’s history and importance to the world? What factors have affected it?
</narr>
<granularity> general </granularity>
</topic>

Summary produced by DS-MR:
1. But with prices falling, tin mines have been closing. [FT934-12248]
2. Dozens of tin mines around the world have already closed and last month Malaysia Mining Corporation, one of the world’s biggest producers, said that, after suffering three successive years of losses, it would close all its tin mines. [FT932-2680]
3. MALAYSIA’S TIN industry, already suffering its worst production slump since the second world war, has been dealt a body blow with the announcement that Malaysia Mining Corporation, one of the country’s biggest tin producers, is pulling out of tin mining. [FT932-11184]
4. For years tin mining has been the core of MMC’s activities and until recently Malaysia was the world’s largest producer of the metal. [FT932-11184]
5. The Malayan Mining Employers’ Association (MMEA), which groups Malaysia’s main tin mining and smelting companies, is considering disbanding because of the dramatic decline in the country’s tin industry. [FT942-9756]
6. Indonesia is the world’s third largest tin producer after Brazil and Malaysia and had an export quota at the end of last year of 28,376 tonnes a year under the Association of Tin Producing Countries’ market stabilisation scheme. [FT923-8983]
7. Indonesia is the world’s third largest tin producer after Brazil and Malaysia and had an export quota at the end of last year of 28,376 tonnes a year under the Association of Tin Producing Countries’ market stabilisation scheme. [FT923-8983]
8. Tin is part of the country’s history: hundreds of thousands of immigrants from China originally came to the country to work the tin mines. [FT934-12248]
9. Tin mining has been going on in one form or another in Malaysia for centuries. [FT931-345]
1. But with prices falling, tin mines have been closing. [FT934-12248]
2. MALAYSIA’S TIN industry, already suffering its worst production slump since the second world war, has been dealt a body blow with the announcement that Malaysia Mining Corporation, one of the country’s biggest tin producers, is pulling out of tin mining. [FT932-11184]
3. Dozens of tin mines around the world have already closed and last month Malaysia Mining Corporation, one of the world’s biggest producers, said that, after suffering three successive years of losses, it would close all its tin mines. [FT932-2680]
4. Rather, it has tried to fuse modern industry with history, Western business with Asian custom. [LA022590-0075]
5. Since 1857, when rich tin deposits were found here, this Southeast Asian city has progressed from a primitive boom town – not unlike California’s gold rush settlements – to a British colony, to capital of the independent nation of Malaysia. [LA022590-0075]
6. Tin is part of the country’s history: hundreds of thousands of immigrants from China originally came to the country to work the tin mines. [FT934-12248]
7. The Malayan Mining Employers’ Association (MMEA), which groups Malaysia’s main tin mining and smelting companies, is considering disbanding because of the dramatic decline in the country’s tin industry. [FT942-9756]
8. For years tin mining has been the core of MMC’s activities and until recently Malaysia was the world’s largest producer of the metal. [FT932-11184]
9. It’s a good time to visit this flower-drenched city, which is rich with Asian culture and easy to explore. [LA022590-0075]
10. In 1990 there were 141 tin mines in the country. [FT932-11184]
11. The decline has been still [FT944-10411]