Query-Sensitive Mutual Reinforcement Chain and Its Application in Query-Oriented Multi-Document Summarization

Furu Wei[1,2], Wenjie Li[1], Qin Lu[1], Yanxiang He[2]

[1]Department of Computing
The Hong Kong Polytechnic University, Hong Kong
{csfwei, cswjli, csluqin}@comp.polyu.edu.hk

[2]Department of Computer Science and Technology
Wuhan University, China
{frwei, yxhe}@whu.edu.cn

ABSTRACT

Sentence ranking is the issue of most concern in document summarization. Early researchers have presented the mutual reinforcement principle (MR) between sentence and term for simultaneous key phrase and salient sentence extraction in generic single-document summarization. In this work, we extend the MR to the mutual reinforcement chain (MRC) of three different text granularities, i.e., document, sentence and terms. The aim is to provide a general reinforcement framework and a formal mathematical modeling for the MRC. Going one step further, we incorporate the query influence into the MRC to cope with the need for query-oriented multi-document summarization. While the previous summarization approaches often calculate the similarity regardless of the query, we develop a query-sensitive similarity to measure the affinity between the pair of texts. When evaluated on the DUC 2005 dataset, the experimental results suggest that the proposed query-sensitive MRC (Qs-MRC) is a promising approach for summarization.

Categories and Subject Descriptors

I.7 [**Document and Text Processing**]: Miscellaneous

General Terms: Algorithms, Language

Keywords

Mutual Reinforcement Chain, Query-Sensitive Similarity, Query-Oriented Summarization, Ranking Algorithms

1. INTRODUCTION

The explosion of the WWW has brought with it a vast hoard of information. It has become virtually impossible for anyone to read and understand large numbers of individual documents that are 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 researchers working on document summarization [10] [13] to exchange their ideas and experiences is the Document Understanding Conferences (DUC [4]). The goals of DUC are to enable researchers to participate in large-scale experiments upon the standard benchmark and to increase the availability of appropriate evaluation techniques. Over the past years, the DUC evaluations have evolved gradually from single-document summarization to multi-document summarization and from generic summarization to query-oriented summarization [17]. Query-oriented multi-document summarization initiated by the DUC in 2005 aims to produce a short and concise summary for a cluster 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 upon the sentence extraction framework. Under this framework, sentence ranking is the issue of most concern. Most previous work in the literature addressed the ranking issue by merely examining characteristic of sentence, such as its content, its grammatical structure, the relationship or association of each other and etc. It is a remarkable advance in our understanding when Zha [22] proposed the following mutual reinforcement (MR) principle:

> "A term should have a high saliency score if it appears in many sentences with high saliency scores while a sentence should have a high saliency score if it contains many terms with high saliency scores."

Based on this MR principle, Zha developed a generic summarization model by representing the documents as a weighted undirected bipartite text graph and linking a term and sentences containing that term together. As they assembled a cluster of documents to a single larger document, the model in essence is a single-document summarization model. The advantage of this model is that at the end of iterative reinforcement both significant sentences and key phrases could be obtained simultaneously. Intuitively, Zha's MR principle is sound and applicable. Since a document is always structured into meaningful text units such as paragraph, sentence, phrase and word in turn, the affiliation relation or the affinity relation between sentence and term (i.e., text of different granularity) can effectively mutually reinforce the importance of each other. This allows for more genuine and reliable judgment.

In this study, query-oriented multi-document summarization, a more practical yet challenging task, is of particular interest to us. However, though starting from the MR principle, our goal is not only simply to adapt it to a different summarization task, but also to establish a general framework that can be extensible to other

SIGIR'08, July 20–24, 2008, Singapore.
Copyright 2008 ACM 978-1-60558-164-4/08/07...$5.00.

applications, such as document retrieval. To achieve this goal, two issues are essential to us. They are how to introduce the document information and how to incorporate the query information (which can be deemed as the external context outside the documents) in the course of mutual reinforcement.

We extend the MR principle to the document, sentence and term mutual reinforcement chain (D-S-T MRC or MRC for short) framework, upon which three interrelated iterative ranking algorithms are developed. Although ranking sentences is the primary goal of document summarization, eventually both documents and terms are also ranked and these by-products can be of potential advantage for other text processing purposes. Notice that the MRC's capability of exploiting the relation among documents extremely facilitates multi-document summarization which is required to handle the sentences coming either from the same document or from the different documents. To cope with query-oriented summarization, we further advance the MRC to the query-sensitive MRC (Qs-MRC). As in previous work, we consider the relevant of a text unit to the query. But different from them, we also count the query impact on the relation of text units. A query-sensitive similarity measure is particularly devised for this purpose to judge the affinity of a pair of text units with respect to a given query. We would like to point out that although the MRC and the Qs-MRC based ranking algorithms are developed for the application of query-oriented multi-document summarization in this paper, the unified framework is general enough to be applied to the other text ranking tasks. Another contribution of this study is trying to provide a formal mathematical modeling for the MRC.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the mutual reinforcement chain (MRC) framework and explains how the general MRC can be extended to the query-sensitive MRC (Qs-MRC). Section 4 proposes the query-sensitive similarity measure. Section 5 suggests a summarization model based on Qs-MRC. Section 6 reports experiments and evaluation results. Section 7 concludes the paper.

2. RELATED WORK

Sentence ranking is the issue of most concern under the framework of extractive summarization. Traditional feature-based approaches evaluated sentence significance and ranked sentences depending on the features that were designed to characterize the different aspects of the sentences. A variety of statistical and linguistic features such as term frequency (distribution), sentence dependency structure, sentence position and query relevance etc. have been extensively investigated in the past. Among them, centroid [18] and signature term [8] are most remarkable. The features were often linearly combined and the weights of them were either experimentally tuned or automatically derived by applying a certain learning-based mechanism [16].

Newly emerged graph-based approaches like LexRank [5] and TextRank [11] [12] modeled a document or a set of documents as a weighed text graph. Different from feature-based approaches, graph-based approaches took into account global information and recursively calculated sentence significance from the entire text graph rather than only replying on single sentences. These approaches were actually inspired by PageRank [2], which has been successfully used for ranking web pages in the Web graph. The effectiveness of PageRank-like approaches came from the advantage of making use of the link structure information. It further promoted the use of topic-sensitive PageRank [6], i.e., an extension of PageRank, for query-oriented summarization [15].

While those PageRank-like approaches normally considered the similarity or the association of the sentences, Zha [23], in contrast, proposed a mutual reinforcement principle that is capable of extracting significant sentences and key phrases at the same time. In his work, a weighted bipartite document graph was built by linking together the sentences in a document and the terms appearing in those sentences. Zha argued that a term should have a high salience score if it appears in many sentences with high salience scores while a sentence should have high salience scores 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 of the bipartite graph. In fact, as early in 1998, the similar idea has been used in HITS algorithm [14] 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 [21] who additionally calculated the links among the sentences and the links among the terms. Zha's and Wan's work are the ones most relevant to our studies presented in this paper. But they all concentrated on single-document generic summarization.

Measuring similarities between two text units such as documents, sentences or terms has been one of the most fundamental issues in information retrieval and other related domains. While a great deal of previous work was found in the literature, few of them addressed the issue of measuring similarity with respect to a particular external context such as a query from a user. Tombros and Rijsbergen [20] pioneered the development of query-sensitive similarity functions. They combined the traditional cosine similarity between the pair of documents with the collective similarity of the two documents and the query together, which was defined as the cosine similarity between the centroid of the two documents and the query. Zhou and Dai [23] also proposed a query-sensitive similarity measure for content-based image retrieval based on Euclidean distance which was widely used in image processing.

3. MUTUAL REINFORCEMENT CHAIN

3.1 Mutual Reinforcement Chain (MRC) of Document, Sentence and Term

Without doubt, the most critical issue in document summarization is sentence ranking. However, the sentences could 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 identifiable. In text processing applications, people often work with the text of three different granularities, i.e., document (D), sentence (S) and term (T). They are actually not independent of each other in delivering meanings. A sentence is the component of a document and meanwhile it is the composition of a set of terms. Therefore, the constraints or the influences among documents, sentences and terms could not be ignored in sentence (document, or term) ranking although they have not been well studied before. In this paper we call the reinforcement among D, S and T the *external reinforcement*.

Meanwhile, we also consider the *internal reinforcement* within D, S or T, i.e., the reinforcement among documents (sentences or terms). In the past, the relations of sentences have been emphasized in graph-based summarization models and their contribution to the performance improvement has been recognized [21]. We put them forward to the relations at the document level as well as at the term level to make a more unified modeling. Finally, the external and the internal reinforcements together form a complete document, sentence and term mutual reinforcement chain (D-S-T MRC or MRC for short) framework, as illustrated in Figure 1.

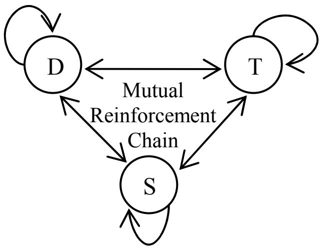

Figure 1. The document, sentence and term mutual reinforcement chain (D-S-T MRC)

This MRC framework is developed with an attempt to capture the following intuition:

1. A document is important if (1) it includes the important sentences; (2) it includes the important terms; and (3) it associates to the other important documents.
2. A sentence is important if (1) it appears in the important documents; (2) it includes the important terms; and (3) it associates to the other important sentences. Similarly,
3. A term is important if (1) it appears in the important documents; (2) it appears in the important sentences; and (3) it associates to the other important terms.

Then, the ranking of documents, sentences and terms can be iteratively derived from the MRC. Let R_D, R_S and R_T denote the ranking scores of *D*, *S* and *T*, respectively, the MRC-based ranking can be formulated as follows:

$$\begin{cases} R_D^{(k+1)} = \alpha_1 \cdot D_D \cdot R_D^{(k)} + \beta_1 \cdot D_S \cdot R_S^{(k)} + \gamma_1 \cdot D_T \cdot R_T^{(k)} \\ R_S^{(k+1)} = \alpha_2 \cdot S_D \cdot R_D^{(k)} + \beta_2 \cdot S_S \cdot R_S^{(k)} + \gamma_2 \cdot S_T \cdot R_T^{(k)} \\ R_T^{(k+1)} = \alpha_3 \cdot T_D \cdot R_D^{(k)} + \beta_3 \cdot T_S \cdot R_S^{(k)} + \gamma_3 \cdot T_T \cdot R_T^{(k)} \end{cases} \quad (1)$$

where D_D denotes the D-D affinity matrix, D_S denotes the D-S affinity matrix, D_T denotes the D-T affinity matrix, and so on. The calculation of the nine matrices in Equation (1) will be detailed in Section 5.1. $W = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 \\ \alpha_2 & \beta_2 & \gamma_2 \\ \alpha_3 & \beta_3 & \gamma_3 \end{bmatrix}$ is the weight matrix used to balance the relative weight among D-S-T in the MRC. Equation (1) corresponds to a block matrix,

$$M = \begin{bmatrix} \alpha_1 D_D & \beta_1 D_S & \gamma_1 D_T \\ \alpha_2 S_D & \beta_2 S_S & \gamma_2 S_T \\ \alpha_3 T_D & \beta_3 T_S & \gamma_3 T_T \end{bmatrix} \quad (2)$$

Let $R = \begin{bmatrix} R_D \\ R_S \\ R_T \end{bmatrix}$, then *R* can be computed as the dominant eigenvector of *M*.

$$M \cdot R = \lambda \cdot R \quad (3)$$

Given that the corresponding graph of *M* is not bipartite, in order to guarantee a unique *R*, we must force *M stochastic* and *irreducible* [7]. To this end, the necessary matrix transformation must be performed. We can prove that the new transformed *M* is stochastic and irreducible.

To force *M* stochastic, we must make the nine block matrices in *M column stochastic*. For the sake of simplicity, let *X* denote any of the three diagonal block matrices (i.e., D_D, S_S, or T_T), and *Y* be any of the remaining six block matrices (i.e., S_D, T_D, D_S, T_S, or D_T, S_T). We first delete the rows and the columns that contain all zero elements[1] in *X*. Note that there are no zero columns in *Y*. Let us take S_D for example. The affinity of the sentence *s* and the document *d* is at least greater than zero, if *d* contain *s*. Then *X* and *Y* are both normalized by columns to their column stochastic versions $\overline{X}$ and $\overline{Y}$. We then replace *X* and *Y* by $\overline{X}$ and $\overline{Y}$ in *M*. Let $\overline{M}$ denote the new matrix, we can prove that:

Lemma 1. $\overline{M}$ is also column stochastic, if the weight matrix *W* is column stochastic.

Proof: Let *A*, *B* and *C* denote the three block matrices in $\overline{M}$ for any column under concern, then

$\sum_i M_{ij} = \alpha_1 \sum_i A_{ij} + \alpha_2 \sum_i B_{ij} + \alpha_3 \sum_i C_{ij} = \alpha_1 + \alpha_2 + \alpha_3 = 1$ □

Next, we manage to make $\overline{M}$ irreducible. Let $\overline{X}$ denote any of the three new diagonal block matrices in $\overline{M}$ .As used in PageRank calculation, we will make the graph corresponding to $\overline{X}$ strongly connect by adding links from one node to any other nodes with a probability vector $\vec{p}$. After such an adjustment, the revised $\overline{X}$ becomes

$$\overline{\overline{X}} = \alpha \overline{X} + (1-\alpha)E \text{ and } E = \vec{p} \times [1]_{1\times k} \quad (4)$$

where $0 < \alpha < 1$ and α is usually set to 0.85 according to PageRank, and *k* is order of $\overline{X}$. The probability vector $\vec{p}$ can be defined in many different ways. A typical definition is to assume a uniform distribution over all elements, i.e., $\vec{p} = [1/k]_{k\times 1}$. By doing so, $\overline{\overline{X}}$ is both stochastic and irreducible. We finally replace $\overline{X}$ by $\overline{\overline{X}}$ in $\overline{M}$, and let $\overline{\overline{M}}$ denote the new matrix. We can prove that:

Lemma 2: $\overline{\overline{M}}$ is irreducible.

Proof: Since the three graphs corresponding to the three diagonal block matrices in $\overline{\overline{M}}$ are strongly connected (they are irreducible) and the links connected the three graphs are *bidirectional*, the graph corresponding to $\overline{\overline{M}}$ is also obviously strongly connected. Thus, $\overline{\overline{M}}$ must be also irreducible. □

Now the matrix $\overline{\overline{M}}$ is stochastic and irreducible. Meanwhile, it is easy to check that $\overline{\overline{M}}^2 > 0$, so $\overline{\overline{M}}$ is also primitive. As a result, we can compute the unique dominant eigenvector (with 1 as the eigenvalue) of $\overline{\overline{M}}$. It is well-known that the power method applied to $\overline{\overline{M}}$ will converge to *R*.

3.2 Query-Sensitive Mutual Reinforcement Chain (Qs-MRC)

In the above introduced MRC, the reinforcements of documents, sentences and terms are *query-unaware*. It means that only the content of the text is concerned. However, for the 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.

Generally, the query information can be incorporated into the general MRC framework in two alternative ways. The first way is to impose the influence of a user's query on each text unit, (document, sentence, or term) such that it works in the internal reinforcement.

[1] This corresponds to the strategies used to cope with the dangling nodes in Web graph in PageRank. Since *X* is symmetric, so there will be no dangling nodes in *X*. Dangling nodes in Web graph correspond to isolated nodes in *X*, because if the out-degree of a node is zero, its in-degree will also be zero. As a result, we simply delete the isolated nodes, and this will not influence the ranking results.

This somewhat can be viewed as a topic-sensitive PageRank [6] at each level of text granularity. In the second way, the query effect is modeled in the affinity matrices in both the internal reinforcement and external reinforcement. By doing so, the MRC turns into a query-sensitive mutual reinforcement chain (Qs-MRC).

With regard to the first way, the key to making ranking biased towards the query rests with the definition of the query-sensitive probability vector $\vec{p}$. A simple yet effective solution is to define $\vec{p}$ as

$$\vec{p}_i = \begin{cases} rel(o_i \mid q) \\ \theta, if\ rel(o_i \mid q) = 0 \end{cases} \tag{5}$$

where, $rel(o_i \mid q)$ denote the relevance of o_i to *q*. It can be calculated by cosine similarity. θ is an extremely small real number to avoid zero elements in $\vec{p}$. $\vec{p}$ is further normalized to 1 in order for it to be a probability vector.

As for the second way, the remaining problem is how to design a query-sensitive similarity measure so that the query dimension can be taken into consideration when measuring the affinity between the text units (documents, sentences or terms) in the affinity matrices. In the following sections, we will first introduce the query-sensitive similarity and then explain how to apply the Qs-MRC to query-oriented multi-document summarization.

4. QUERY-SENSITIVE SIMILARITY

Existing similarity measures produce static and constant scores. However, the similarity between a pair of texts may vary with the different contexts involved. Now, we consider a more general problem: measuring the similarity between any two text objects o_i and o_j with respect to the given query *q*. An object *o* can be a document, a sentence or a term. We believe that the similarity between o_i and o_j themselves would be adjusted when *q* is involved. This is not difficult to understand. Similar to the social relationship, when the third party comes in, the connection of the two will undoubtedly change more or less.

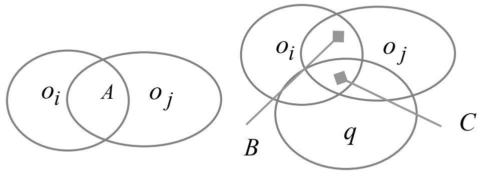

Figure 2. Illustration of Query-Sensitive Similarity from Set Overlapping Perspective

This problem can be explained from the set intersection perspective. As illustrated in Figure 2, "*A*" in the left picture corresponds to the similarity between the two objects o_1 and o_2 regardless of the query. It is what we call *query-unaware* similarity. But in the right picture, when the query *q* becomes explicit, it divides the overlapping area of o_i and o_j into two separated parts, i.e., "*B*" and "*C*". We could then deduce that a *query-sensitive* similarity measure should consist of two parts, i.e., the *query-independent* part and the *query-dependent* part. The query-independent part concerns the dedication of query-unaware similarity, while the query-dependent part further highlights the contribution of the terms in common not only in o_i and o_j but also in *q*. Formally, the query-sensitive similarity can be formulated as,

$$sim(o_i, o_j \mid q) = f\left(sim(o_i, o_j \mid\in q), sim(o_i, o_j \mid\notin q)\right) \tag{6}$$

Let $o_i = \{o_{i1}, o_{i2}, ..., o_{in}\}$, $o_j = \{o_{j1}, o_{j2}, ..., o_{jn}\}$ and $q = \{q_1, q_2, ..., q_n\}$ be the three *n*-dimensional vectors. $\mu = \max(q_k)$ and $\eta = \min(q_k)$ where $1 \le k \le n$ and $q_k \ne 0$. We then define the following weight coefficient function,

$$S(q_k) = \begin{cases} \xi, q_k = 0 \\ \xi + \left(\theta_1 + \dfrac{q_k - \eta}{\mu - \eta} \cdot (\theta_2 - \theta_1)\right), q_k \ne 0 \end{cases} \tag{7}$$

where 0< $\theta_1 < \theta_2 < \xi$ <1. $q_k = 0$ and $q_k \ne 0$ regards "*B*" and "*C*" in Figure 2. Notice that there is a special case that the above function can not cope with, i.e., $\mu = \eta$. In this case

$$S(q_k) = \begin{cases} \xi, q_k = 0 \\ \xi + \left(\dfrac{\theta_2 - \theta_1}{2}\right), q_k \ne 0 \end{cases} \tag{8}$$

Then, we define the query-sensitive similarity function as

$$sim(o_i, o_j \mid q) = \frac{\sum_{k=1}^{n} S(q_k) \cdot o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} \tag{9}$$

When we move from query-unaware similarity $sim(o_i, o_j)$ to query-sensitive similarity $sim(o_i, o_j \mid q)$, the range of the adjustment is subject to the certain constraints by Equation (8).

Proposition 1. The query-sensitive similarity defined by Equation (9) ranges from $\xi \cdot sim(o_1, o_2)$ to $(\xi + \theta_2) sim(o_1, o_2)$, more precisely from $\xi \cdot sim(o_1, o_2) + \theta_1 \cdot sim'(o_1, o_2)$ to $\xi \cdot sim(o_1, o_2) + \theta_2 \cdot sim'(o_1, o_2)$. $sim(o_1, o_2)$ is defined as the cosine similarity between the two corresponding vectors. $sim'(o_1, o_2)$ is contributed from the query- dependent part.

Proof: Let $n^q = \{k \mid q_k \ne 0\}$, then

$$sim(o_i, o_j \mid q) = \frac{\xi \cdot \sum_{k=1}^{n} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} + \frac{\left(\theta_1 + \dfrac{q_k - \eta}{\mu - \eta} \cdot (\theta_2 - \theta_1)\right) \cdot \sum_{k \in n^q} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}}$$

$$\begin{aligned} sim(o_i, o_j, q) &\ge \frac{\xi \cdot \sum_{k=1}^{n} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} + \frac{\theta_1 \cdot \sum_{k \in n^q} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} \\ &= \xi \cdot sim(o_i, o_j) + \theta_1 \cdot sim'(o_i, o_j) \ge \xi \cdot sim(o_i, o_j) \end{aligned}$$

$$\begin{aligned} sim(o_i, o_j \mid q) &\le \frac{\xi \cdot \sum_{k=1}^{n} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} + \frac{\theta_2 \cdot \sum_{k \in n^q} o_{ik} \cdot o_{jk}}{\sqrt{\sum_{k=1}^{n} o_{ik}^2} \cdot \sqrt{\sum_{k=1}^{n} o_{jk}^2}} \\ &= \xi \cdot sim(o_i, o_j) + \theta_2 \cdot sim'(o_i, o_j) \\ &\le (\xi + \theta_2) \cdot sim(o_i, o_j) \end{aligned}$$

□

There are three parameters in Proposition 1, i.e., ξ, θ_1 and θ_2. They are all meaningful and can be appropriately determined according to the practical application requirements. ξ is the contribution degree of the original query-unaware similarity $sim(o_i, o_j)$ in the query-sensitive similarity function. θ_1 and θ_2 can be viewed as the lower and upper bounds of the contribution from the query-dependent part. Query-sensitive similarity is extremely important for the study of query-oriented summarization. The given query can be treated as the external context that influences or even determines how the summary is produced.

5. QS-MRC BASED QUERY-ORIENTED MULTI-DOCUMENT SUMMARIZATION

5.1 Task Definition of DUC Query-Oriented Multi-Document Summarization

The query-oriented multi-document summarization task defined in DUC evaluations requires generating a concise and well-organized summary for a cluster 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 cluster "d331f".

```
<topic>
<num> d331f </num>
<title> World Bank criticism and response </title>
<narr>
Who has criticized the World Bank and what criticisms have they made
of World Bank policies, activities or personnel. What has the Bank
done to respond to the criticisms?
</narr>
<granularity> specific </granularity>
</topic>                                              [d331f]
```

According to the task definition, 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 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 non-redundant sentences from the original documents to create a summary. Under this extractive framework, undoubtedly the two critical processes involved are sentence ranking and sentence selection.

5.2 Sentence Ranking Based on Qs-MRC

In Section 3, we introduce the general MRC and the extended Qs-MRC frameworks that can prescribe the reinforcement-based procedure for ranking the text of different granularities simultaneously. The frameworks themselves are applicable to document retrieval, sentence retrieval or key word extraction etc. In this section, however, we manage to take the advantage of the Qs-MRC framework to deal with the sentence ranking problem in the query-oriented multi-document summarization task.

To this end, the design of the affinity matrices among the documents, the sentences and the terms in Equation (1) is the first of all. In this work, we define the affinity between any two text units as their query-sensitive similarity, as proposed in Section 4. The documents, the sentences as well as the queries can be naturally represented by the vectors of the terms. However, this representation does not suit the single terms. We cannot say that it is impossible for two different terms to be relevant to each other. This is because the single terms (or even those short text snippets without any overlapping terms) themselves do not carry sufficient information for measuring the similarity of them.

In the past, many researchers have proposed different methods that attempted to capture more context of a single term or a short text snippet (such as the query posed by the user to the search engine that contains only a few words). For example, Sahami and Heliman [19] submitted each snippet as a query to the Web search engine and created a context vector of the snippet by collecting a number of returned documents that contain the words in the snippet. The context vector created in this way contains the words that tend to occur in context with the snippet. When calculating similarity, they used a context vector as a substitute for snippet. Similarly, Bollegala et al. [1] measured similarity between words or entities by making use of the information available on the Web such as page counts and text snippets returned, while Cilibrasi and Vitanyi [3] on the other side extracted Google similarity distance of words and phrases from WWW using Google page counts.

The basic idea behind these approaches is try to expand single terms or short text snippets by exploring web resources for more relevant information. The approaches using *web context vectors* are sound, but they heavily rely on the effectiveness of search engines and most important they are time-consuming. In this work, we utilize a semantic lexical resource WordNet[2] that has been widely used in the natural language processing community. WordNet groups terms into sets of synonyms called synsets and provides short, general definitions of them. A term may belong to many different synsets. We use all its descriptions in those synsets to composite the context vector, i.e., *WordNet context vector*, of that term. See Algorithm 1 below.

Algorithm 1: GetWordNetContextVector(*t*)

```
Input: The term t.
Output: The WordNet context vector C for t.
$C \leftarrow \{\ \}$;
Get all the set of part-of-speech (POS) for t in WordNet, denoted as POS;
For each POS pos in POS
    Get the set of synset for t with pos, denoted as Syn;
    For each synset syn in Syn
        Get the gloss of syn, denoted as gloss[3];
        $C \leftarrow C \cup gloss$;
    End;
End;
Return C.
```

The top 10 terms that are supposed to be close to an example term "policies" in document cluster "d331f" are given below. The similarities calculated by means of WordNet context vector are reasonable.

Cosine Similarity between WordNet Context Vector of term "policies" and Top Ten Close Terms

policy-based	0.67	planning	0.28
advise	0.35	accountable	0.27
achievement	0.33	guidelines	0.25
persuade	0.31	responsibility	0.24
policy-making	0.30	strategy	0.23

Finally, each element in affinity matrices is defined as the query-sensitive similarity between the two text units $sim(o_i, o_j|q)$ and it is calculated by Equation (10). *q* is a query vector, *o* can be a vector of a document, a sentence or a WordNet context vector of a term. The Qs-MRC based ranking algorithm formulated in Equation (1) is then implemented using the following iterative procedure. Sentences are ranked according to their ranking scores eventually converged in R_S.

[2] WordNet 2.0 is used in this work (http://wordnet.princeton.edu). The Java WordNet Library (JWNL) is used as the interface for searching a word in WordNet (http://sourceforge.net/projects/jwordnet/).

[3] We omit the examples in the gloss. Only the definitions are returned.

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

```
Input: The document set D, the sentence set S, the term set T, and the query q.
Output: The ranking vectors of $R_D$, $R_S$ and $R_T$.
Construct the affinity matrices using the query-sensitive similarity and WordNet context vector;
Transform the nine block matrices to make $M$ stochastic and irreducible as mentioned in Section 3.1;
Choose (randomly) the initial non-negative vectors $R_D^{(0)}$, $R_S^{(0)}$ and $R_T^{(0)}$, such that $\left\|R_D^{(0)}\right\|_1 = 1$, $\left\|R_S^{(0)}\right\|_1 = 1$ and $\left\|R_T^{(0)}\right\|_1 = 1$;
$i \leftarrow 0$, $\nabla \leftarrow 0$;
Repeat
    $R_D^{(i+1)} = \alpha_1 \cdot D_D \cdot R_D^{(i)} + \beta_1 \cdot D_S \cdot R_S^{(i)} + \gamma_1 \cdot D_T \cdot R_T^{(i)}$;
    $R_S^{(i+1)} = \alpha_2 \cdot S_D \cdot R_D^{(i)} + \beta_2 \cdot S_S \cdot R_S^{(i)} + \gamma_2 \cdot S_T \cdot R_T^{(i)}$;
    $R_T^{(i+1)} = \alpha_3 \cdot T_D \cdot R_D^{(i)} + \beta_3 \cdot T_S \cdot R_S^{(i)} + \gamma_3 \cdot T_T \cdot R_T^{(i)}$;
    $\nabla \leftarrow \max\left(\begin{Bmatrix} \left\|R_D^{(i+1)} - R_D^{(i)}\right\|_1 \\ \left\|R_S^{(i+1)} - R_S^{(i)}\right\|_1 \\ \left\|R_T^{(i+1)} - R_T^{(i)}\right\|_1 \end{Bmatrix}\right)$;
    $i \leftarrow i + 1$;
Until $\nabla < \zeta$[4];
$R_D \leftarrow R_D^{(i)}$, $R_S \leftarrow R_S^{(i)}$ and $R_T \leftarrow R_T^{(i)}$;
Return.
```

5.3 Sentence Selection by Removing Redundancy

In multi-document summarization, the number of the documents to be summarized can be very large. This makes information redundancy problem appear to be more serious in multi-document summarization than in single-document summarization. Redundancy removal becomes an inevitable process. Since our focus in this study is the design of 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 doesn't 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 $\Pi$.
$\Pi \leftarrow \{\}$;
$l \leftarrow$ length;
For $i \leftarrow$ 0 to |S| do
    threshold $\leftarrow \max\left(sim(s_i, s) \mid s \in \Pi\right)$;
    If threshold <= 0.9 do
        $\Pi \leftarrow \Pi \cup s_i$;
        $l \leftarrow l - len(s_i)$;
        If (l<= 0) break;
    End
End
Return $\Pi$.
```

6. EXPERIMENTAL STEDIES

6.1 Experiment Set-up

Experiments are conducted on the DUC 2005 50 document clusters. Each cluster of documents is accompanied with a query description representing a user's information need. Table 1 below shows the basic statistics of the dataset. Stop-words in both documents and queries are removed[5] and the remaining words are stemmed by Porter Stemmer[6] and considered as terms.

Table 1. Basic Statistics of the DUC 2005 Dataset

Average Number of Documents per Cluster	31.86
Average Number of Sentences per cluster	1002.54
Average Number of Unique Terms per cluster	2886.54

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 this work, ROUGE[7] [8], which has been officially adopted by the DUC for automatic evaluations since 2005, is used to evaluate the system-generated summaries.

In all the following experiments, both text units 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. 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. The TF*ISF weighting scheme is applied to the WordNet context vectors as well. When a definition word does not appear in a document cluster at all, its ISF is approximated by the mean of the ISF of the terms that appears in the document cluster. Notice that the term itself is also included in its context vector. As for the weight matrix *W* in the MRC, we set it as $\begin{bmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{bmatrix}$ based on the hierarchical structure among Document-Sentence-Term. *W* is also normalized to be column stochastic. θ in Equation (6) is assigned the 20% of the minimized value of the relevance of the documents (sentences or terms) to the query in a document cluster. The three parameters in query-sensitive similarity measure presented in Equation (10) are assigned the values 0.8 for ξ, 0.1 for θ_1 and 0.2 for θ_2.

6.2 Evaluation of Mutual Reinforcement Ranking

First of all, we would like to see how the proposed Qs-MRC based ranking algorithm works for the task of query-oriented multi-document summarization. The first set of experiments is conducted for this purpose. For reference, we also implement another two widely used and well-performed ranking strategies. One is to simply rank the sentences according to their relevance to the query (denoted by QR). The other is the PageRank deduced iterative ranking algorithm (denoted by PR). The damping factor used here is set to 0.85 as the same used in Google's PageRank. To avoid the link-by-chance problem that happens when the two text units share only one or two terms by chance, we do not consider the reinforcement between them if their similarity is below a very small threshold (it is 0.05 in this work). Table 2 below shows the results of average recall scores of ROUGE-1, ROUGE-2 and ROUGE -SU4

[4] ζ is a pre-defined small real number of the convergence threshold.

[5] A list of 199 words is used to filter stop-words.

[6] http://www.tartarus.org/~martin/PorterStemmer.

[7] ROUGE version 1.5.5 is used.

along with their 95% confidence intervals within square brackets. Among them, ROUGE-2 is the primary DUC evaluation criterion.

Table 2. Qs-MRC and Two Referenced Sentence Ranking Strategies

	ROUGE-1	ROUGE-2	ROUGE-SU4
Qs-MRC	**0.3868** **[0.3810, 0.3927]**	**0.0779** **[0.0744, 0.0811]**	**0.1366** **[0.1333, 0.1399]**
PR	0.3702 [0.3672, 0.3772]	0.0725 [0.0704, 0.0766]	0.1306 [0.1274, 0.1341]
QR	0.3597 [0.3540, 0.3654]	0.0664 [0.0630, 0.0697]	0.1229 [0.1196, 0.1261]

As shown in Table 2, both Qs-MRC and PR are able to produce much better results than QR which evaluates sentence individually. Qs-MRC is above QR by 17.32% of ROUGE-2, 7.53% of ROUGE-1, and 11.15% of ROUGE-SU4. Even PR is above QR by 9.19% of ROUGE-2, 2.92% of ROUGE-1, and 6.27% of ROUGE-SU4. Qs-MRC further improves PR by increasing 7.45% of ROUGE-2, 4.48% of ROUGE-1, and 4.59% of ROUGE-SU4. The improvements by involving the MRC are promising.

6.3 Evaluation of External and Internal Reinforcement

As mentioned in previous Section 3, reinforcement can be categorized as either external or internal, and external reinforcement to sentence can come from document and/or term. The second set of experiments here is to evaluate the functions of the reinforcement in different scopes. In Table 3, the MR between document and sentence or between sentence and term are indicated by Qs-MR_DS or Qs-MR_ST. The same MRC framework is applied to Qs-MR_DS and Qs-MR_ST. The only difference is that simply the affinity and weight matrices involved are processed in calculation. Since we focus on sentence ranking in summarization, the MR between term and document is ignored in the experiments. Qs-MR_S considers internal reinforcement only.

Table 3. External and Internal Reinforcement Effects

	ROUGE-1	ROUGE-2	ROUGE-SU4
Qs-MRC	**0.3868** **[0.3810, 0.3927]**	**0.0779** **[0.0744, 0.0811]**	**0.1366** **[0.1333, 0.1399]**
Qs-MR_DS	0.3844 [0.3784, 0.3908]	0.0770 [0.0737, 0.0819]	0.1354 [0.1314, 0.1395]
Qs-MR_ST	0.3771 [0.3770, 0.3830]	0.0734 [0.0696, 0.0769]	0.1311 [0.1275, 0.1345]
Qs-MR_S	0.3730 [0.3666, 0.3789]	0.0732 [0.0708, 0.0770]	0.1307 [0.1273, 0.1343]

From the results shown in Tables 2 and 3, we come up with the following observations. First, it is obvious that all the algorithms (Qs-MRC, Qs-MR_DS, QS-MR_ST and Qs-MR_S) that take the mutual reinforcement into consideration are superior to the simple query relevance ranking algorithm (QR). Second, the external reinforcement is more useful than internal reinforcement. Qs-MRC, Qs-MR_DS and QS-MR_ST evidently outperform the Qs-MR_S. Notice that Qs-MR_S is very similar to PR but it calculates query-sensitive similarity while PR calculates normal cosine similarity. The results suggest that external reinforcement is more important than internal reinforcement. Third, for sentence ranking, the external reinforcement from document appears more significant than the same from term. Qs-MR_DS is even comparable to Qs-MRC. As recognized by other researchers, how to formulate the effect of a smaller text unit such as a term on the text unit does matter. A more comprehensive study on the term representation is expected in our future work.

6.4 Evaluation of Query Influence

Afterwards, we evaluate our modeling of query influence. Table 4 below gives the ROUGE results in terms of three different ways to incorporate the query information, i.e., to calculate text relevance to the query only (MRC_QR), query-sensitive reinforcement only (MRC_QsS) and all of them (Qs-MRC). Although absolutely the query must be taken into consideration in the query-oriented summarization task, we also present the results of MRC that ignores the query influence for reference.

Table 4. Query Influence

	ROUGE-1	ROUGE-2	ROUGE-SU4
Qs-MRC	**0.3868** **[0.3810, 0.3927]**	**0.0779** **[0.0744, 0.0811]**	**0.1366** **[0.1333, 0.1399]**
MRC_QR	0.3807 [0.3749, 0.3870]	0.0759 [0.0721, 0.0798]	0.1330 [0.1293, 0.1368]
MRC_QsS	0.3689 [0.3634, 0.3747]	0.0717 [0.0677, 0.0757]	0.1291 [0.1251, 0.1330]
MRC	0.3633 [0.3576, 0.3693]	0.0632 [0.0604, 0.0664]	0.1201 [0.1177, 0.1240]

Evidently, the query is extremely important. Qs-MRC is above MRC by 17.3% of ROUGE-2, 7.53% of ROUGE-1, and 11.1% of ROUGE-SU4. The difference is significant. Meanwhile, MRC_QR and MRC_QsS are also superior to MRC. The reason is intuitive. In query-oriented summarization, users are particularly interested in the information conveyed in the query that reflects their information needs. On the other hand, the improvements from text relevance alone are much better than query-sensitive reinforcement alone. This is reasonable. No matter how important the reinforcement is, it can not supersede the essential nature of text materials. Even so, the improvement of MRC_QsS over MRC is still very competing especially in ROUGE-2, i.e., 13.45%. These improvements are meaningful, especially when they are compared with the improvements among DUC 2005 participating systems as we will show later. Thus it is easy to conclude that the query-sensitive similarity is a direction worth further extensive study. More important, it can be applied in many applications other than query-oriented summarization.

6.5 Comparison with DUC Systems

Thirty-one systems have been submitted to DUC for evaluation in 2005. Table 5 compares the Qs-MRC with them. To provide a global picture, we present the following representative ROUGE results of (1) the worst-performed human summary (i.e., H), which reflects the margin between the machine-generated summaries and the human summaries; (2) the top five and worst participating systems according to ROUGE-2; (3) the average ROUGE scores (i.e., AVG); and (4) the NIST baseline which simply selects the first sentences in the documents. We can then easily locate the positions of the proposed models among them. Notice that the ROUGE-1 scores are not officially released by DUC.

Table 5. System Comparison

	ROUGE-1	ROUGE-2	ROUGE-SU4
H	-	0.0897	0.1510
....			
Qs-MRC	**0.3868**	**0.0779**	**0.1366**
...			
S15	-	0.0725	0.1316
S17	-	0.0717	0.1297
S10	-	0.0698	0.1253
S8	-	0.0696	0.1279
S4	-	0.0686	0.1277
...			
S23	-	0.0256	0.0557
AVG		0.0584	0.1121
NIST Baseline	-	0.0403	0.0872

It clearly shows in Table 4 that Qs-MRC outperforms the first-ranked system (i.e., S15). It is above S15 by 7.45% of ROUGE-2 and 3.80% of ROUGE-SU4. These are definitely exciting achievements since the best system (i.e., S15) is only 1.12% above the second-best system (i.e., S17) on ROUGE-2 and 1.46% on ROUGE-SU4.

7. CONCLUTION

In this paper, we propose a mutual reinforcement chain framework (MRC and Qs-MRC). Based on it, we develop an interactively reinforced ranking algorithm for the application of query-oriented multi-document summarization. The main contributions of this work are three-fold. First, we extend the mutual reinforcement principle between two objects to the mutual reinforcement chain (MRC) among three (or more than) objects and provide a formal mathematical modeling for the MRC. Second, we design a query-sensitive similarity measure and incorporate it into the MRC, i.e., the Qs-MRC. Last but not least, we exploit the effectiveness of Qs-MRC for sentence ranking in query-oriented multi-document summarization. The work suggests that it is worth further studying on more appropriate and mathematical sound query-sensitive similarity measures and more accurate term context representation.

8. ACKNOWLEDGMENTS

The work described in this paper was supported by the grants from the RGC of HK, (Project No. PolyU5211/05E and PolyU5217/07E), the grant from the NSF of China (Project No. 60703008), and the internal grant from the Hong Kong Polytechnic University (Project No. A-PA6L).

9. REFERENCES

[1] Bollegala, D., Matsuo, Y., and Ishizuka, M. 2007. Measuring Semantic Similarity between Words using Web Search Engines. In Proceedings of 16th WWW, pp 757-766.

[2] Brin, S. and Page, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 30(1-7): 107-117.

[3] Cilibrasi, R. L. and Vitanyi, P. M. B. 2007. The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No. 3, pp 370-383.

[4] DUC: http://www-nlpir.nist.gov/projects/duc/pubs.html.

[5] Erkan, G. and Radev, D. R. 2004. LexRank: Graph-based Centrality as Salience in Text Summarization, Journal of Artificial Intelligence Research 22:457-479.

[6] Haveliwala, T. H. 2003. Topic-Sensitive PageRank: A Context-Sensitive Ranking Algorithm for Web Search. IEEE Transactions on Knowledge and Data Engineering, Vol. 15, No. 4, pp 784-796.

[7] Langville, A. N. and Meyer, C. D. 2004. Deeper Inside PageRank. Journal of Internet Mathematics, 1(3): 335-380.

[8] Lin, C. Y. and Hovy, E. 2000. The Automated Acquisition of Topic Signature for Text Summarization. In Proceedings of 18th COLING, pp 495-501.

[9] Lin, C. Y. and Hovy, E. 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics, in Proceedings of HLT-NAACL, pp71-78.

[10] Mani, I. and Maybury, M. T.(Eds.). 1999. Advances in Automatic Summarization. The MIT Press.

[11] Mihalcea, R. 2004. Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summarization. In Proceedings of ACL 2004, Article No. 20.

[12] Mihalcea, R. 2005. Language Independent Extractive Summarization. In Proceedings of ACL 2005, pp 49-52.

[13] Jones, K. S. 2007. Automatic Summarising: The State of the art. Information Processing and Management 43: 1449-1481.

[14] Kleinberg, J. M. 1999. Authoritative Sources in a Hyperlinked Environment. In Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms. Pp 668-677.

[15] Otterbacher, J., Erkan, G., and Radev, D. R. 2005. Using Random Walks for Question-focused Sentence Retrieval. In Proceedings of HLT/EMNLP, pp 915-922.

[16] Ouyang, Y., Li, S. J., and Li, W. J. 2007. Developing Learning Strategies for Topic-Based Summarization. In Proceedings of the 16th CIKM, pp 79-86.

[17] Over, P., Dang, H., and Harman, D. 2007. DUC in Context, Information Processing and Management, 43(6): 1506-1520.

[18] Radev, D. R., Jing, H. Y., Stys, M., and Tam, D. 2004. Centroid-based Summarization of Multiple Documents. Information Processing and Management, 40: 919-938.

[19] Sahami, M.and Heliman, T. D. 2006. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In Proceedings of 15th WWW, pp 377-386.

[20] Tombros, A. and Rijsbergen, C. J. v. 2004. Query-Sensitive Similarity Measures for Information Retrieval. Knowledge and Information Systems (2004) 6: 617-642.

[21] Wan, X. Y., Yang, J. W., and Xiao, J. G. 2007. Towards Iterative Reinforcement Approach for Simultaneous Document Summarization and Keyword Extraction. In Proceedings of ACL.

[22] Zha, H. Y. 2002. Generic Summarization and Key Phrase Extraction using Mutual Reinforcement Principle and Sentence Clustering. In Proceedings of the 25th ACM SIGIR, pp113-120.

[23] Zhou, Z. H. and Dai, H. B. 2006. Query-Sensitive Similarity Measure for Content-Based Image Retrieval. In Proceedings of ICDM, pp1211-1215.