Single Document Summarization with Document Expansion

Xiaojun Wan and Jianwu Yang

Institute of Computer Science and Technology
Peking University, Beijing 100871, China
{wanxiaojun, yangjianwu}@icst.pku.edu.cn

Abstract

Existing methods for single document summarization usually make use of only the information contained in the specified document. This paper proposes the technique of document expansion to provide more knowledge to help single document summarization. A specified document is expanded to a small document set by adding a few neighbor documents close to the document, and then the graph-ranking based algorithm is applied on the expanded document set for extracting sentences from the single document, by making use of both the within-document relationships between sentences of the specified document and the cross-document relationships between sentences of all documents in the document set. The experimental results on the DUC2002 dataset demonstrate the effectiveness of the proposed approach based on document expansion. The cross-document relationships between sentences in the expanded document set are validated to be very important for single document summarization.

Introduction

Document summarization aims to automatically creating a concise representation of a given document that delivers the main topic of the document. Automatic document summarization has drawn much attention for a long time because it becomes more and more important in many text applications. For example, existing search engines usually provide a short summary for each retrieved document so as to facilitate users to browse the results and improve users' search experience. News agents usually provide concise headline news describing hot news and they also produce weekly news review for users, which saves users' time and improves service quality.

Document summary can be either query-relevant or generic. Query-relevant summary should be closely related to the given query. Generic summary should reflect the main topic of the document without any additional clues and prior knowledge. In this paper, we focus on generic single document summarization.

Existing methods conduct the summarization task using only the information contained in the specified document to be summarized. Very often, a few documents topically close to the specified document can be retrieved from a large corpus through search engines, and these neighbor documents are deemed beneficial to evaluate and extract summary sentences from the document. The underlying assumption is that the topic-related documents can provide more knowledge and clues for single summarization of the specified document. From human's perception, users would better understand a document if they read more topic-related documents. This study proposes to use the document expansion technique to build an appropriate knowledge context of a small document set by adding a few neighbor documents close to the specified document. The enlarged knowledge within the context can be used in the summarization process and help to extract better summary from the document.

This study employs the graph-ranking based algorithm for single summarization of the specified document by making use of both the cross-document relationships and the within-document relationships between sentences in the expanded document set, where the within-document relationships reflect the local information existing in the specified document and the cross-document relationships reflect the global information existing in the expanded document set.

We perform experiments on the DUC2002 dataset and the results demonstrate the good effectiveness of the proposed approach. The use of the cross-document relationships between sentences can improve the performance of single document summarization. We also investigate how the size of the expanded document set influences the summarization performance and it is encouraging that a small number of neighbor documents can improve the summarization performance.

Related Work

Generally speaking, single document summarization methods can be either extraction-based or abstraction-based and we focus on extraction-based methods in this paper.

Extraction-based methods usually assign each sentence a saliency score and then rank the sentences in the document. The score is usually assigned based on a combination of statistical and linguistic features, including term frequency, sentence position, cue words, stigma words, topic signature (Hovy & Lin, 1997; Lin & Hovy, 2000), etc. Machine learning methods have also been employed to extract sentences, including unsupervised methods (Nomoto & Matsumoto, 2001) and supervised methods (Kupiec et al., 1995; Conroy & O'Leary, 2001; Amini & Gallinari, 2002;

Shen et al., 2007). Other methods include maximal marginal relevance (MMR) (Carbonell & Goldstein, 1998), latent semantic analysis (LSA) (Gong & Liu, 2001). In Zha (2002), the mutual reinforcement principle is employed to iteratively extract key phrases and sentences from a document.

More recently, the graph-ranking based methods, including TextRank (Mihalcea & Tarau, 2004, 2005) and LexPageRank (ErKan & Radev, 2004) have been proposed for document summarization. Similar to Kleinberg's HITS algorithm (Kleinberg, 1999) or Google's PageRank (Brin & Page, 1998), these methods first build a graph based on the similarity between sentences in a document and then the importance of a sentence is determined by taking into account global information on the graph recursively, rather than relying only on local sentence-specific information.

All the above methods make use of only the information contained in the specified document. The use of neighbor documents to improve single document summarization has not been investigated yet.

Proposed Approach

Overview

Given a specified document d_0 to be summarized, the proposed approach first finds a few neighbor documents for the document. The neighbor documents are topically close to the specified document and build a knowledge context for the specified document. In other words, document d_0 is expanded to a small document set D which provides more knowledge and clues for single summarization of document d_0. Given the expanded document set, the proposed approach adopts the graph-ranking based algorithm to incorporate both the within-document relationships (local information) and the cross-document relationships (global information) between sentences to summarize the specified document within the context. Figure 1 sketches the framework of the proposed approach.

1. **Document Expansion**: *Expand the specified document d_0 to a small document set $D=\{d_0, d_1, d_2, ..., d_k\}$ by adding k neighbor documents. The neighbor documents $d_1, d_2, ..., d_k$ can be obtained by document similarity techniques;*
2. **Document Summarization**: *Given document d_0 and the expanded document set D, perform the following steps to produce the summary for d_0:*
   a) **Affinity Graph Building**: *Build a global affinity graph G based on all sentences of the documents in D; Let $S=\{s_1, s_2, ... , s_n\}$ denotes the sentence set for the document set.*
   b) **Informativeness Score Computation**: *Based on the global affinity graph G, the graph-ranking based algorithm is employed to compute the informativeness score $IFScore(s_i)$ for each sentence s_i, where $IFScore(s_i)$ quantifies the informativeness of the sentence s_i.*
   c) **Redundancy Removing**: *The greedy algorithm is employed to remove redundancy for the informative sentences in d_0. Finally, the sentences of d_0 which are both informative and novel are chosen into the summary.*

Figure 1: The framework of the proposed approach

For the first step in the above framework, different similarity search techniques can be adopted to obtain neighbor documents close to the specified document. The number k of the expanded documents influences the summarization performance and will be investigated in the experiments.

For the second step in the above framework, step a) aims to build a global affinity graph reflecting the relationships among all sentences in the expanded document set of k+1 documents. Step b) aims to compute the informativeness score of each sentence based on the global affinity graph. The informativeness of a sentence indicates how much information about the main topic the sentence contains. Step c) aims to remove redundant information in the summary and keep the sentences in the summary as novel as possible. A summary is expected to include the sentences with high informativeness and minimum redundancy.

Document Expansion

Given a specified document, document expansion aims to find a few nearest neighbors for the document from a text corpus or on the Web. The k neighbor documents $d_1, d_2,..., d_k$ and the specified document d_0 construct the expanded document set $D=\{d_0, d_1, d_2,..., d_k\}$ for d_0, which can be considered as the enlarged knowledge context for document d_0. Figure 2 shows the document expansion for document d_0.

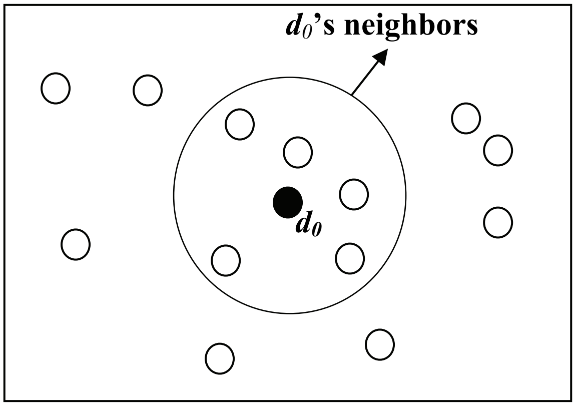

Figure 2: Document expansion for d_0

The neighbor documents can be obtained using the technique of document similarity search. Document similarity search is to find documents similar to a query document in a text corpus and return a ranked list of similar documents to users. The effectiveness of document similarity search relies on the function for evaluating the similarity between two documents. Usually, the standard Cosine measure (Baeza-Yates & Ribeiro-Neto, 1999) is considered as one of the best functions and thus widely used for document similarity search. In the vector space model (VSM), a document d_i is represented by a vector with each dimension referring to a unique term and the weight associated with term t is calculated by the $tf_{d_i,t}.idf_t$ formula, where $tf_{d_i,t}$ is the number of occurrences of term t in document d_i and $idf_t=1+log(N/n_t)$ is the inverse document frequency, where N is the total number of documents in the collection and n_t is the number of documents containing term t. The similarity $sim_{doc}(d_i, d_j)$, between documents d_i and d_j, can be defined as the normalized inner product of the two vectors $\vec{d}_i$ and $\vec{d}_j$:

$$sim_{\text{doc}}(d_i,d_j)=\frac{\vec{d}_i \cdot \vec{d}_j}{\left\|\vec{d}_i\right\| \times \left\|\vec{d}_j\right\|} \qquad (1)$$

The efficiency of document similarity search can be significantly improved by adopting some index structure in the implemented system, such as K-D-B tree, R-tree, SS-tree, SR-tree and X-tree (Böhm & Berchtold, 2001).

In the experiments, we simply use the Cosine measure to compute pairwise similarity value between the specified document d_0 and the documents in the corpus, and then choose k documents (different from d_0) with the largest similarity values as the nearest neighbors for d_0. Finally, there are totally k+1 documents in the expanded document set. For the document set $D=\{d_0, d_1, d_2, \ldots, d_k\}$, the pairwise Cosine similarity values between documents are calculated and recorded for later use.

The use of neighborhood information is worth more discussion. Because neighbor documents might not be sampled from the same generative model as the specified document, we probably do not want to trust them so much as the specified document. Thus a confidence value is associated with every document in the expanded document set, which reflects out belief that the document is sampled from the same underlying model as the specified document. When a document is close to the specified one, the confidence value is high, but when it is farther apart, the confidence value will be reduced. Heuristically, we use the Cosine similarity between a document and the specified document as the confidence value. The confidence values of the neighbor documents will be incorporated in the summarization algorithm.

Document Summarization

Affinity Graph Building: Given the sentence collection $S=\{s_i \mid 1 \le i \le n\}$ of the expanded document set, the affinity weight $sim_{sen}(s_i, s_j)$ between a sentence pair of s_i and s_j is calculated using the Cosine measure. The weight associated with term t in sentence s_i is calculated with the $tf_{s_i,t}.isf_t$ formula, where $tf_{s_i,t}$ is the frequency of term t in the sentence and isf_t is the inverse sentence frequency of term t, i.e. $1+log(N'/n_t')$, where N' is the total number of sentences and n_t' is the number of sentences containing term t.

If sentences are considered as nodes, the sentence collection can be modeled as an undirected graph by generating a link between two sentences if their affinity weight exceeds 0; otherwise no link is constructed. The links (edges) between sentences in the graph can be categorized into two classes: within-document link and cross-document link. Given a link between a sentence pair of s_i and s_j, if s_i and s_j come from the same document, the link is a within-document link; and if s_i and s_j come from different documents, the link is a cross-document link. Actually, the within-document link reflects the local information in a document, while the cross-document link reflects the global information in the expanded document set, which delivers mutual influences between documents in the set. The within-document link and the cross-document link correspond to different confidence values and the weight associated with each link is determined by both the corresponding sentence similarity value and the confidence value. Thus, we construct a weighted graph G reflecting the relationships between sentences in the expanded set. The graph G contains both kinds of links between sentences and is called as *Global Affinity Graph*. We use an adjacency (affinity) matrix **M** to describe G with each entry corresponding to the weight of a link in the graph. $\mathbf{M} = (M_{i,j})_{n\times n}$ is defined as follows:

$$M_{i,j}=\begin{cases}\lambda \times sim_{sen}(s_i,s_j), & \text{if } i \ne j \\ 0, & \text{otherwise}\end{cases} \qquad (2)$$

where λ specifies the confidence value of the sentence relationship. If the link between s_i and s_j is a within-document link, let λ=1; if the link between s_i and s_j is a cross-document link, i.e., s_i and s_j come from different document d_k and d_l, let $\lambda=sim_{doc}(d_k, d_l)$.

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

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

Similar to the above process, another two affinity graphs G_{intra} and G_{inter} are also built: the within-document affinity graph G_{intra} is to include only within-document links between sentences (the entries of cross-document links are set to 0); the cross-document affinity graph G_{inter} is to include only cross-document links between sentences (the entries of within-document links are set to 0). The corresponding adjacency (affinity) matrices of G_{intra} and G_{inter} are denoted by $\mathbf{M}_{\text{intra}}$ and $\mathbf{M}_{\text{inter}}$ respectively. $\mathbf{M}_{\text{intra}}$ and $\mathbf{M}_{\text{inter}}$ can be extracted from **M** and we have $\mathbf{M}=\mathbf{M}_{\text{intra}}+\mathbf{M}_{\text{inter}}$. Similar to Equation (3), $\mathbf{M}_{\text{intra}}$ and $\mathbf{M}_{\text{inter}}$

are respectively normalized to $\widetilde{\mathbf{M}}_{\text{intra}}$ and $\widetilde{\mathbf{M}}_{\text{inter}}$ to make the sum of each row equal to 1.

Informativeness Score Computation: Based on the global affinity graph *G*, the informativeness score $IFScore_{all}(s_i)$ for sentence s_i can be deduced from those of all other sentences linked with it and it can be formulated in a recursive form as follows:

$$IFScore_{all}(s_i) = d \cdot \sum_{all\, j \neq i} IFScore_{all}(s_j) \cdot \widetilde{M}_{j,i} + \frac{(1-d)}{n} \quad (4)$$

And the matrix form is:

$$\vec{\lambda} = d\widetilde{\mathbf{M}}^T \vec{\lambda} + \frac{(1-d)}{n}\vec{e} \quad (5)$$

where $\vec{\lambda} = [IFScore_{all}(s_i)]_{n \times 1}$ is the vector of informativeness scores. $\vec{e}$ is a unit vector with all elements equaling to 1. *d* is the damping factor usually set to 0.85.

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

Similarly, the informativeness score of sentence s_i can be deduced based on either the within-document affinity graph G_{intra} or the cross-document affinity graph G_{inter} as follows:

$$IFScore_{intra}(s_i) = d \cdot \sum_{all\, j \neq i} IFScore_{intra}(s_j) \cdot (\widetilde{M}_{intra})_{j,i} + \frac{(1-d)}{n} \quad (6)$$

$$IFScore_{inter}(s_i) = d \cdot \sum_{all\, j \neq i} IFScore_{inter}(s_j) \cdot (\widetilde{M}_{inter})_{j,i} + \frac{(1-d)}{n} \quad (7)$$

The final informativeness score $IFScore(s_i)$ of sentence s_i can be either $IFScore_{all}(s_i)$, $IFScore_{intra}(s_i)$ or $IFScore_{inter}(s_i)$. With different scenarios for computing informativeness score, three summarization methods are defined as follows:

UniformLink: $IFScore(s_i)$ is equal to $IFScore_{all}(s_i)$, considering both the within-document relationships and the cross-document relationships.

InterLink: $IFScore(s_i)$ is equal to $IFScore_{inter}(s_i)$, considering only the cross-document relationships.

IntraLink: $IFScore(s_i)$ is equal to $IFScore_{intra}(s_i)$, considering only the within-document relationships.

We will investigate all the above summarization methods. Note that all previous graph-ranking based methods do not consider the cross-document links and have $IFScore(s_i)= IFScore_{intra}(s_i)$.

Redundancy Removing: For the specified document d_0 to be summarized we can extract a sub-graph G_{d_0} only containing the sentences within d_0 and the corresponding edges between them from the global affinity graph *G*. We assume document d_0 has *m* (*m*<*n*) sentences and the sentences' affinity matrix $\mathbf{M}_{d_0} = (\mathbf{M}_{d_0})_{m \times m}$ is derived from the original matrix **M** by extracting the corresponding entries. Then $\mathbf{M}_{d_0}$ is normalized to $\widetilde{\mathbf{M}}_{d_0}$ as Equation (3) to make the sum of each row equal to 1. The greedy algorithm (Zhang et al., 2005), which is actually a variant form of the MMR algorithm and thus denoted as "MMR" in next section, is used to penalize the sentences highly overlapping with other informative sentences based on $\widetilde{\mathbf{M}}_{d_0}$. The basic idea of the algorithm is to decrease the overall rank score of less informative sentences by the part conveyed from the most informative one. Finally, the overall rank score for each sentence within the document is obtained and the sentences with highest overall rank scores are both highly informative and highly novel, which are chosen into the summary for d_0 according to the summary length limit. The details of the algorithm are omitted due to page limit.

Empirical Evaluation

Evaluation Setup

We used task 1 of DUC 2002 (DUC, 2002) for evaluation. The task aimed to evaluate generic summaries with a length of approximately 100 words or less. DUC 2002 provided 567 English news articles collected from TREC-9 for single-document summarization task. The sentences in each article have been separated and the sentence information has been stored into files. The DUC2002 dataset was considered as the corpus for document expansion in this study, which could be easily expanded by adding more documents. Each specified document was expanded by adding k documents (different from the specified document) most similar to the document. The stopwords were removed and the remaining words were stemmed using Porter's stemmer (Porter, 1980).

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

$$ROUGE - N = \frac{\sum_{S \in \{\text{Ref Sum}\}} \sum_{n\text{-}gram \in S} Count_{match}(n - gram)}{\sum_{S \in \{\text{Ref Sum}\}} \sum_{n\text{-}gram \in S} Count(n - gram)} \quad (8)$$

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

ROUGE toolkit reported separate scores for 1, 2, 3 and 4-gram, and also for longest common subsequence cooccurrences. Among these different scores, unigram-based ROUGE score (ROUGE-1) has been shown to agree with human judgment most (Lin & Hovy. 2003). We showed three of the ROUGE metrics in the experimental results: ROUGE-1 (unigram-based), ROUGE-2 (bigram-based), and ROUGE-W (based on weighted longest common subsequence, weight=1.2).

In order to truncate summaries longer than length limit, we used the "-l" option in ROUGE toolkit.

Evaluation Results

The proposed approach considering neighbor documents (i.e. UniformLink) is compared with the baseline method depending only on the specified document (i.e. IntraLink). We also show the results of InterLink to demonstrate how reliable the cross-document relationships are. Table 1 shows the comparison results after removing redundancy (i.e. w/ MMR) and Table 2 shows the comparison results before removing redundancy (i.e. w/o MMR). For the methods of UniformLink and InterLink, the parameter *k* is heuristically set to 1, 5 and 10, respectively.

Table 1: Comparison results after removing redundancy

System	ROUGE-1	ROUGE-2	ROUGE-W
UniformLink (k=1)	0.46562	0.19557	0.16056
UniformLink (k=5)	0.46738	0.19618	0.16156
UniformLink (k=10)	**0.47162***	**0.20114***	**0.16314**
InterLink (k=1)	0.46641	0.19430	0.16060
InterLink (k=5)	0.46703	0.19574	0.16141
InterLink (k=10)	0.46870*	0.19800*	0.16211
IntraLink	0.46261	0.19457	0.16018

Table 2: Comparison results before removing redundancy

System	ROUGE-1	ROUGE-2	ROUGE-W
UniformLink (k=1)	0.46034	0.19543	0.15966
UniformLink (k=5)	0.46000	0.19478	0.15907
UniformLink (k=10)	0.46360*	0.19777*	0.16068
InterLink (k=1)	0.45925	0.19433	0.15861
InterLink (k=5)	**0.46396***	**0.19813***	**0.16084**
InterLink (k=10)	0.46345	0.19701	0.16075
IntraLink	0.45591	0.19201	0.15789

(* indicates that the improvement over the baseline "IntraLink" is statistically significant)

Seen from the tables, the proposed UniformLink always outperforms the baseline IntraLink, no matter whether the process of removing redundancy is applied, which shows that document expansion does benefit single document summarization. Moreover, we can see that the method of InterLink also always performs better than the baseline IntraLink, which demonstrates that the cross-document relationships between sentences in the expanded document set are reliable enough to evaluate and extract salient sentences from single document. Actually, the expanded document set is about the same topic with the specified document and the important information contained in the specified document would be also contained in other documents, maybe in different representations. Thus the knowledge from the expanded documents would much help to analyze and extract important information from the specified document.

In order to investigate how the size of the expanded document set influences the summarization performance, we conduct experiments with different values of *k*. Figures 3 and 4 show the ROUGE-1 and ROUGE-W values for different methods with different values of *k*. In the figures, *k* ranges from 1 to 15, indicating there are totally 2 to 16 documents in the expanded document sets. Four methods are investigated, including UniformLink, InterLink, with or without the process of removing redundancy (w/ MMR and w/o MMR).

Seen from the figures, the summarization performance first increases with *k*, however, when *k* is larger than 10, the performance tends to decrease or at least stop increasing. The figures shows that a large size of the expanded set is unnecessary, which will even deteriorate the performance because more neighbor documents run a risk of inducing more noise. Thus the size of the expanded set can be set to a small number, which will improve the computational efficiency and make the propose approach more applicable.

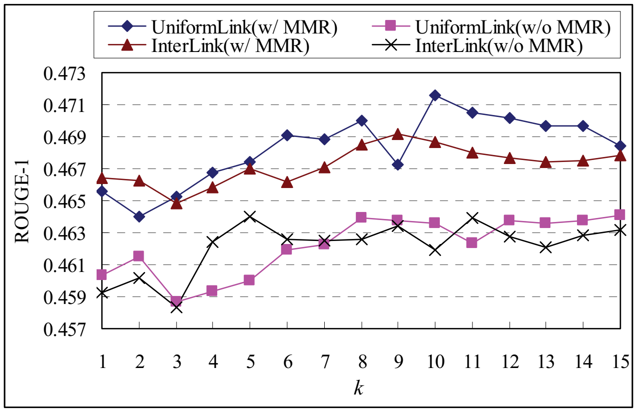

Figure 3: ROUGE-1 performance vs. *k*

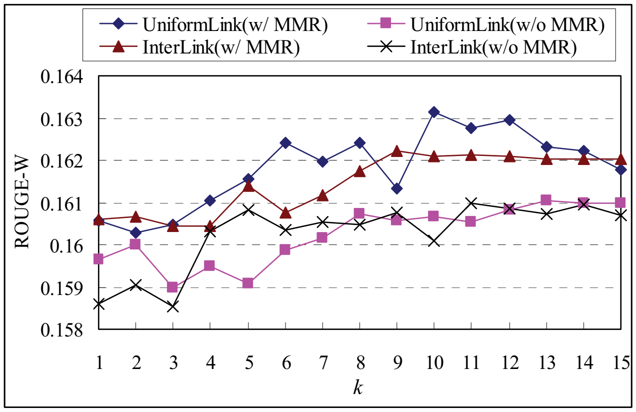

Figure 4: ROUGE-W performance vs. *k*

Frankly speaking, the proposed approach has higher computational complexity than the baseline approach because it involves more documents, and we can improve its efficiency by collaboratively conducting single document summarizations in a batch mode. Suppose there are multiple documents to be summarized separately, we can group the documents into clusters, and for each cluster, we can use all other documents as the neighbors for a specified document. Thus the mutual influences between all documents can be incorporated into the summarization algorithm and all the sentences in the documents of a cluster are evaluated collaboratively, resulting in single summarizations of all the documents in a batch mode.

Conclusion and Future Work

This paper proposes to summarize single document by expanding the specified document to a small document set by adding a few neighbor documents. The within-document relationships and the cross-document relationships between sentences are then incorporated in the graph-ranking based algorithm for single document summarization. The additional knowledge provided by neighbor documents is acquired through the cross-document sentence relationships. Experimental results on the DUC2002 dataset demonstrate the effectiveness of the proposed approach and the importance of the cross-document relationships between sentences.

In this study, only the graph-ranking based algorithm is adopted for document summarization, and in future work, other summarization algorithms will be integrated into the proposed framework to validate the robustness of the technique of document expansion. Furthermore, Web page summarization will be evaluated using the proposed approach to make use of the rich link information between web pages to acquire more additional knowledge.

Acknowledgements

This work was supported by the National Science Foundation of China (60642001).

References

Amini, M. R., and Gallinari, P. 2002. The use of unlabeled data to improve supervised learning for text summarization. In *Proceedings of SIGIR2002*, 105-112.

Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Modern Information Retrival*. ACM Press and Addison Wesley.

Brin, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7).

Böhm, C., and Berchtold, S. 2001. Searching in high-dimensional spaces-index structures for improving the performance of multimedia databases. *ACM Computing Surveys*, 33(3): 322-373.

Carbonell, J., and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of SIGIR-1998*, 335-336.

Conroy, J. M., and O'Leary, D. P. 2001. Text summarization via Hidden Markov Models. In *Proceedings of SIGIR2001*, 406-407.

DUC. 2002. The Document Understanding Workshop 2002. http://www-nlpir.nist.gov/projects/duc/guidelines/2002.html

ErKan, G., and Radev, D. R. 2004. LexPageRank: Prestige in multi-document text summarization. In *Proceedings of EMNLP2004*.

Gong, Y. H., and Liu, X. 2001. Generic text summarization using Relevance Measure and Latent Semantic Analysis. In *Proceedings of SIGIR2001*, 19-25.

Hovy, E., and Lin, C. Y. 1997. Automated text summarization in SUMMARIST. In *Proceeding of ACL'1997/EACL'1997 Worshop on Intelligent Scalable Text Summarization*.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Kupiec, J., Pedersen, J., and Chen, F. 1995. A.trainable document summarizer. In Proceedings of SIGIR1995, 68-73.

Lin, C. Y., and Hovy, E. 2000. The automated acquisition of topic signatures for text Summarization. In *Proceedings of ACL-2000*, 495-501.

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

Mihalcea, R., and Tarau, P. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP2004*.

Mihalcea, R., and Tarau, P. 2005. A language independent algorithm for single and multiple document summarization. In *Proceedings of IJCNLP2005*.

Nomoto, T., and Matsumoto, Y. 2001. A new approach to unsupervised text summarization. In *Proceedings of SIGIR2001*, 26-34.

Porter, M. F. 1980. An algorithm for suffix stripping. *Program*, 14(3): 130-137.

Shen, D., Sun, J.-T., Li, H., Yang, Q., and Chen, Z. 2007. Document Summarization using Conditional Random Fields. In *Proceedings of IJCAI 07*.

Zha, H. Y. 2002. Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering. In *Proceedings of SIGIR2002*, pp. 113-120.

Zhang, B., Li, H., Liu, Y., Ji, L., Xi, W., Fan, W., Chen, Z., and Ma, W.-Y. 2005. Improving web search results using affinity graph. In *Proceedings of SIGIR2005*.