

Learning Query-Biased Web Page Summarization¹

Changhu Wang
Department of EES, University of Science and
Technology of China
Hefei 230027, China
(86)13581984028
wch@ustc.edu

Feng Jing^{*}, Lei Zhang^{*}, Hong-Jiang Zhang[†]
^{*}Microsoft Research Asia
[†]Microsoft Research Advanced Technology Center
49 Zhichun Road, Beijing 100080, China
(86-10)5896{3609, 3197, 5991}
{fengjing, leizhang, hjzhang}@microsoft.com

ABSTRACT

Query-biased Web page summarization is the summarization of a Web page reflecting the relevance of it to a specific query. It plays an important role in search results representation of Web search engines. In this paper, we propose a learning-based query-biased Web page summarization method. The summarization problem is solved within the typical sentence selection framework. Different from existing Web page summarization methods that use page content or link context alone, both of them are considered as the sources of sentences in this work. Most of existing learning-based summarization methods treat summarization as a sentence classification problem and train a classifier to discriminate between extracted sentences and non-extracted sentences of *all* training documents. The basic assumption of these methods is that sentences from different documents are comparable with respect to the class information. In contrast to the classification scheme, a ranking scheme is introduced to rank extracted sentences higher than non-extracted sentences of *each* training document. The underlying assumption that sentences within a document are comparable is weaker and more reasonable than the assumption of classification-based scheme. Extensive results using intrinsic evaluation metrics gauge many aspects of the proposed method.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information filtering

General Terms

Algorithms, Experimentation, Performance

Keywords

Query-biased Web page summarization, Ranking, Classification, Support vector machines

1. INTRODUCTION

Automatic text summarization is a long-standing research topic dating back to 1950s [23]. It is a multi-faceted endeavor that typically branches out in several dimensions as follows:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'07, November 6--8, 2007, Lisboa, Portugal.
Copyright 2007 ACM 978-1-59593-803-9/07/0011...\$5.00.

- General vs. query-biased: A general summarization [2][3][8][11][17][24][28] gives an overall sense of a document's content, while a query-biased summarization [4][7][15][29] presents the information that is most relevant to a search query.
- Web document vs. non-Web document: Before the naissance of Internet, most of the summarization algorithms dealt with non-Web documents, such as news documents. These algorithms could be adapted to Web document summarization [6][30]. However, compared with non-Web documents, Web documents contain textual information that is scarcer, noisier, and more diverse. Therefore, traditional summarization methods that focus on local contents of a document, is difficult to capture the true meaning of a Web page. As a result, additional knowledge will be very helpful in distinguishing the real content of a Web page from noise. On the one hand, clickthrough data that contains many users' knowledge on Web pages' contents has been used in [28] to improve Web page summarization. On the other hand, Web documents contain link context information that generic documents do not have. The link contexts that may contain human-made summaries of the documents have been shown to be beneficial to both general [11] and query-biased [1] Web page summarization. For example, with the InCommonSense system [1], Amitay and Paris first extracted the text segments containing a link to a Web page. Then, the most accurate sentence was chosen from the text segments as the query-biased summarization of the target page. Considering that the context information may be related to the target Web page but contains no clues for summarization, Delort et al. [11] proposed two enhanced summarization methods for general Web pages. Both the characteristics of context information of a Web page and the relations between the context information and the target Web page content were studied [11].
- Extract vs. abstract: An extract summary consists of text-spans extracted from a text document while an abstract summary may contain words and phrases which do not exist in the original document. Abstracting a document requires the ability to manage various hard problems such as discourse understanding and natural language processing. Thus, most existing summarization algorithms use extract instead of abstract. Furthermore, abstraction which needs detailed parsing and digesting usually takes a considerable amount of time. As a result, abstraction may be suitable for

¹ This work was performed at Microsoft Research Asia.

general summarization which could be performed offline [3]. However, for query-biased summarization which is a real time process, abstraction-based summarization will be inappropriate. Although summary by extraction is not guaranteed to have narrative coherence, it is informative enough for general summarization and rapid relevance assessment. Moreover, it could be efficient enough to be performed online.

In this paper, we focus on query-biased Web page summarization. The extraction-based scheme is adopted to meet the real time requirement. Sentences are generally used as text-span units for extraction, although paragraphs have also been considered [25]. Since the length of query-biased summarization is often limited, we use sentences as the extraction units instead of the lengthy paragraphs.

Since the main purpose of query-biased summarization is to assist users to judge the relevance of the search results, it should in some sense reflect the searching process. We argue that the key features such as anchor text [13] and extracted title [18] used in relevance ranking should also be considered and employed in snippet generation.

Most of existing works on extraction-based Web page summarization extract text-spans either from Web page contents [28][30] or link contexts [1][11] *alone*. As analyzed in [11], if the Web page content is not too few, i.e. contains more than three sentences, combining content with context will be better than using context alone. The sentences of both content and context could be extracted in our algorithm.

Extraction-based summarization has been cast in the framework of supervised learning for the first time in the seminal work of [22]. It formulates summarization as a statistical classification problem. Given a training set of documents with manually selected extracts, a classifier could be learned to separate the extracted sentences from the non-extracted sentences of all training documents. New extracts can then be generated by ranking sentences according to the output of the classifier and selecting the top ranked ones. Based on several discrete features, a naïve Bayesian classifier was used in [22]. Within the classification framework, several classifiers have been used such as decision tree [8][24], neural network [8], and logistic regression [2]. Under the classification framework, one assumes that all sentences from different documents are comparable with respect to the class information. This assumption may hold for scientific articles that contain similar contents. However, for Web documents whose contents are heterogeneous, the assumption is inappropriate. To address this issue, we treat summarization as a ranking problem. The goal of the training is to rank extracted sentences higher than non-extracted sentences for each training document. For a new document, all of its sentences are ranked and top ranked ones are selected as the summary. The underlying assumption of this ranking framework is that sentences *within* a document are comparable. This weaker assumption is more reasonable for Web documents and therefore could lead to better summarization results.

The rest of the paper is organized as follows. The proposed query-biased Web page summarization algorithm is discussed in Section 2. In Section 3, extensive experimental results are given as well as some discussions. Finally, we conclude in Section 4.

2. QUERY-BIASED WEB PAGE SUMMARIZATION

2.1 Content vs. Context

Both content and context of Web pages could be used as the source for sentence selection, although most of existing Web page summarization algorithms only use one of them. We assume that the Web pages discussed here are HTML documents. For a Web page, all sentences are extracted from the body of the corresponding HTML document. The extracted sentences are called “content sentences”. For each sentence, additional information such as location and format information are also extracted for further feature extraction. All the sentences containing a link to the current page are extracted and called “context sentences.” If more than 80% words of two context sentences are identical, the shorter sentence will be removed.

2.2 Features

According to [4], two issues should be considered for a query-biased summarization: relevance and fidelity. The former shows the relevance of the summary with the query, while the later indicates the correspondence of the summary with the original document. A good summary should keep both high relevance and high fidelity.

2.2.1 Relevance

Based on the belief that the larger the number of query terms in a sentence, the more likely that sentence conveys a significant amount of the information need expressed in the query, [29] used the following formula to calculate a “query score” for a sentence s :

$$Score_{query}(s) = 2 * n^2 / q \quad (1)$$

where n is the number of query terms in sentence s and q is the total number of query terms.

2.2.2 Fidelity

To characterize the fidelity of a sentence to a document, the properties of the sentence such as location [12] and format [30] have been used for the content sentences. Considering that such properties are inapplicable for context sentences, the “gist” of the document such as title and anchor text phrases are used for both content and context sentences.

2.2.2.1 Term Occurrence

Based on the assumptions that high frequent non-stop words are “significant” and sentences with dense cluster of “significant” words are important, [23] proposed a keyword-based clustering method to measure the importance of sentences. More specifically, a word is deemed as “significant” if it is not a stop word and its term frequency is larger than a threshold T . Similar to [29], we define T as: $T = 7 + I * 0.1 * |L - n|$, where n is the number of sentences in the document. L is 25 for $n < 25$ and 40 for $n > 40$. I is 0 for $25 < n <= 40$ and 1 otherwise.

Clusters of significant words are created such that significant words are separated by not more than 4 insignificant words within the sentence. If in that way a sentence contains two or more clusters, the one with the highest significance factors is taken as the measure of that sentence. The significance score for a cluster (or the sentence that cluster represented) is:

$$Score_{TO}(s) = SW^2 / TW \quad (2)$$

where SW is the number of significant words in the cluster (both cluster boundaries are significant words), and TW is total number of words in the cluster.

2.2.2.2 Title & Extracted Title

As the titles of news articles that tend to reveal the major subject of the article, titles of Web pages serve as previews of the whole pages. Therefore, sentences containing more title words could possibly be more important. We use the following formula to calculate the title score of a sentence s :

$$Score_{title}(s) = i / n \quad (3)$$

where n is the number of title words and i is the number of title words appearing in the sentence s .

As analyzed in [18], 33.5% of HTML documents in the TREC data set have bogus titles. Therefore, besides the titles in the title field of HTML documents, we extracted titles from the body of HTML documents using the algorithm of [18]. We define an extracted title-based (E-title-based) score of sentence s as follows:

$$Score_{E-title}(s) = i / n \quad (4)$$

where n is the number of extracted title words and i is the number of extracted title words appearing in the sentence s .

2.2.2.3 Anchor Text

As the descriptions from people other than authors, anchor text phrases describe the target Web pages more precisely and objectively. Therefore, they could be effectively used in the evaluation of sentences' importance. We consider all the anchor text phrases. Considering that some anchor text phrases may be the same, we deem them as one anchor text with a weight proportional to their count. More specifically, assume that there are N unique anchor text phrases $\{AT_i\}$, $i = 1, \dots, N$ with the number of occurrences of AT_i being O_i . Then, the weight W_i of AT_i is defined as:

$$W_i = O_i / \sum_{j=1}^N O_j \quad (5)$$

Based on the weights, we define the anchor text-based score of a sentence s as follows:

$$Score_{anchor}(s) = \sum_{i=1}^N W_i m_i / n_i \quad (6)$$

where n_i is the number of words in AT_i and m_i is the number of words of AT_i appearing in s .

Since all the aforementioned scoring methods calculate importance score of sentences, any combination of them could be used as a summarizer by ranking and selecting sentences according to the sum of scores.

2.3 Learning Algorithms

2.3.1 Classification

As aforementioned in Section 1, extraction-based summarization could be treated as a classification problem. Given a training set of documents with manually selected extracts, a classifier could be learned to separate the extracted sentences from the non-extracted sentences. For a new document, all of its sentences are ranked

according to the output of the classifier and the top ranked ones are selected as the summary.

We choose SVM as the classifier due to its sound theoretical foundations and proven empirical successes.

In the basic form, SVM tries to find a hyperplane that separates the positive and negative training data with maximal margin [19]. More specifically, finding the optimal hyperplane is translated into the following optimization problem:

$$\text{Minimize: } \frac{1}{2} \|\bar{w}\|^2 + C \cdot \sum_{i=1}^n \xi_i \quad (7)$$

$$\text{subject to: } \forall k : y_k (\bar{w} \cdot \bar{x}_k + b) \geq 1 - \xi_k \quad (8)$$

where \bar{x}_i is a feature vector of sentence s_i , and y_i is equal to 1 if s_i is a summary sentence or -1 otherwise.

Considering that there are usually more non-extracted sentences than extracted ones, we use different penalty parameters (C) to resolve this issue of unbalanced training data. More specifically, formula (7) becomes:

$$\text{Minimize: } \frac{1}{2} \|\bar{w}\|^2 + C_+ \cdot \sum_{y_i=1} \xi_i + C_- \cdot \sum_{y_i=-1} \xi_i \quad (9)$$

To summarize a new document, a significance score $f(s)$ for each sentence s is calculated:

$$f(s) = \bar{w} \cdot \bar{x} + b \quad (10)$$

where \bar{x} is the feature vector of s . Then, the sentences with highest scores are selected as the summary.

2.3.2 Ranking

Since a document is summarized by ranking its sentences and selecting the top ones, it is more natural to treat summarization as a ranking problem. Instead of training a classifier that discriminates extracted sentences from non-extracted sentences of *all* training documents, a ranking function is learned to rank extracted sentences higher than non-extracted ones *within* each training document.

Learning to rank is an active research area in machine learning community. Several ranking algorithms have been proposed in recent years including ordinal regression [16], perception [9], RankNet [5], RankBoost [14] and Ranking SVM [20]. Some has been successfully used in information retrieval [5][20]. Considering that SVM classifier is used in the classification-based summarization, we choose Ranking SVM as the ranking algorithm:

$$\text{Minimize: } \frac{1}{2} \|\bar{w}\|^2 + C \cdot \sum_{i,j,k} \xi_{i,j,k} \quad (11)$$

$$\text{Subject to: } \forall s_i \in r_1, \forall s_j \notin r_1 : \bar{w} \cdot \bar{x}_i \geq \bar{w} \cdot \bar{x}_j + 1 - \xi_{i,j,1} \quad (12)$$

$$\forall s_i \in r_N, \forall s_j \notin r_N : \bar{w} \cdot \bar{x}_i \geq \bar{w} \cdot \bar{x}_j + 1 - \xi_{i,j,N}$$

$$\forall i \forall j \forall k : \xi_{i,j,k} \geq 0$$

where N is the number of training documents, r_k is a sentence set containing all extracted summary sentences of document k , and x_i is the feature vector of sentence s_i . New documents are

summarized in the same way as in Section 2.3.1 except that : $f(s) = \bar{w} \cdot \bar{x}$.

Both the classification and ranking problems are solved using SVM Light [19].

3. EXPERIMENTS

Extensive experiments were performed to gauge many aspects of the proposed method. In the following parts, we'll first describe the data set and evaluation measures. Then, features, sentence sources and learning algorithms are evaluated respectively.

3.1 Data Set

We used the queries and relevance judgments of Web Tracks in TREC-2003. The queries are classified into three categories: named-page finding (NP), homepage finding (HP) and topic distillation (TD) [10]. Considering that NP and HP queries have few relevant Web pages, we only use TD queries. 10 queries were randomly chosen from the total 50 TD queries. The selected queries are: Schizophrenia, robots, arctic exploration, genealogy searches, Shipwrecks, Literacy, deafness in children, wireless communications, Counterfeit money, mining gold silver coal. The first five queries and their corresponding documents were used as training data, while other queries and documents were testing data. In order to evaluate the effectiveness of summarization algorithms on different kinds of pages, we chose three types of web pages from the .GOV data for each query: relevant pages, top ranked irrelevant pages based on certain information retrieval criterion, and randomly selected irrelevant pages excluding those top ranked ones. The retrieval method we used is BM25 [27]. We call the three kinds of pages Relevant, BM25 Irrelevant and Random Irrelevant respectively. We randomly selected ten Relevant Pages, five Random Irrelevant Pages, and five BM25 Irrelevant Pages for each query. If there are less than ten Relevant Pages, all of them will be selected. As a result, there are totally 175 pages. For each selected page, content sentences, context sentences and anchor text phrases were extracted.

Two human evaluators were asked to summarize all the 175 pages by extracting sentences according to the corresponding queries. Both content sentences and context sentences were shown to the evaluators for selection. For each TD query in TREC-2003, there is a detailed description. For example, the description of query "wireless communications" is "Information on existing and planned uses, research/technology, regulations and legislative interest." The descriptions are shown to the evaluators to help them decide the relevance of sentences. The number of sentences for a summary is recommended to be three. However, if there are less or more appropriate sentences, they could be selected in spite of the recommendation. On the one hand, 48.44% sentences of evaluator 2's summaries are also included in evaluator 1's summaries. On the other hand, 44.09% sentences of evaluator 1's summaries are also included in evaluator 2's summaries. Note that there are on average 27 sentences for each page and only about three sentences were selected. Therefore, the summarization results of two evaluators are relatively consistent.

3.2 Evaluation Measures

Summarization evaluation methods could be classified into two categories: intrinsic and extrinsic [21]. Intrinsic evaluation tests

the summarization system in itself, while extrinsic evaluation tests the summarization based on how it affects the completion of some other task.

Considering that we have the human labeled summaries, intrinsic measures were adopted. Precision, recall and F_1 are straightforward measures widely used in intrinsic summarization evaluation [26]. For each document, the manually extracted sentences are considered as the reference summary and the top three sentences ranked by certain algorithm are considered as candidate summary. The precision, recall and F_1 values are computed by comparing the candidate summary with the reference summary as follows:

$$P = \frac{|S_r \cap S_c|}{|S_c|}; R = \frac{|S_r \cap S_c|}{|S_r|}; F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (13)$$

where S_c and S_r denote the sentences contained in the candidate summary and the reference summary.

Since there are two evaluators whose extracted summaries are not totally consistent, we average the P , R and F_1 scores calculated using each evaluator's summaries as the reference summaries. The averaged F_1 score is used as the evaluation measure in all the following evaluations.

3.3 Title

Since the title of a Web page can be extracted easily from the title field of the corresponding HTML document, most of existing search engines display title and query-biased summarization separately as in Figure 1. We consider a sentence as the title sentence if all its words could be found in the title field of the Web page, after stemming and removing the stop words. Out of all the 175 Web pages, the title sentences of on average 25 pages are selected by the evaluators. This shows the importance of title in query biased Web page summarization. Considering the separation of title and summarization in current search engines, we remove all the title sentences from both candidate summaries and reference summaries.

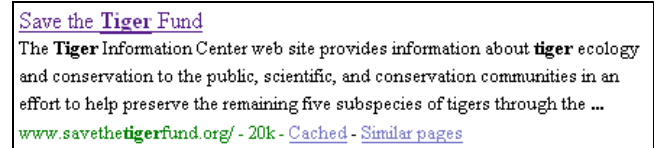


Figure 1. An item of retrieval results list of Google for query "tiger"

3.4 Features

Each of the five features introduced in Section 2.2 could be used as a summarizer that ranks sentences according to their values. Since we have three kinds of pages: relevant pages, BM25 irrelevant pages and random irrelevant pages, we first evaluate all the features on each type of pages. The F_1 scores of the five features on different page sets are shown in Figure 2. For each feature, the performances on different kinds of pages are consistent and similar. Therefore, in the following evaluations, the three page sets were mixed into one set. The performances of five features and their combination are shown in Figure 3. When used alone, title is the most effective feature, while query, title, anchor and E-title are clearly better than TO. The combination of five features is clearly better than using any single feature.

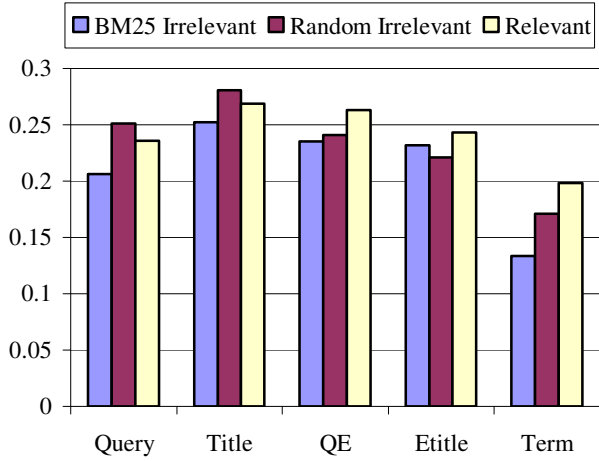


Figure 2. Performance comparison for different kinds of pages using different single features

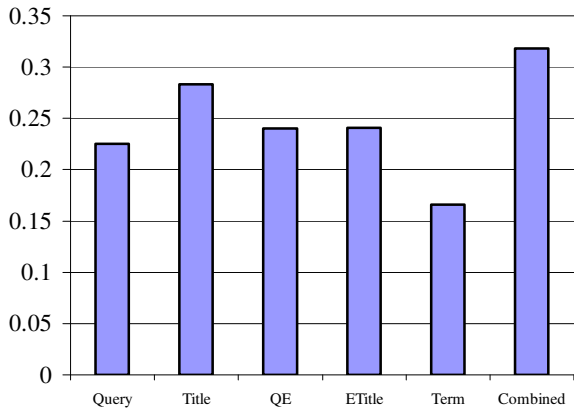


Figure 3. Performance comparison of single features and their combination

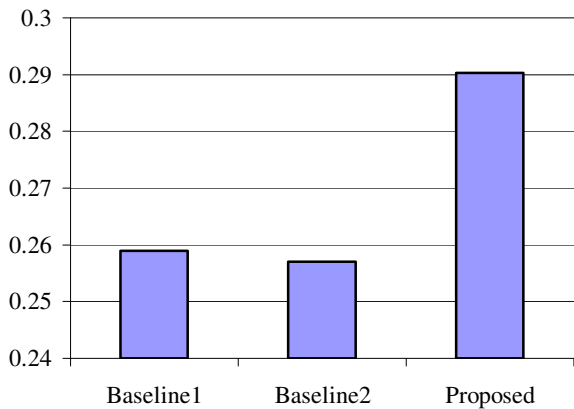


Figure 4. Performance comparison of different feature combinations using content sentences

Furthermore, to show the effectiveness of the proposed feature combination (Proposed), two combinations of [30] and [29] were used as two baselines. More specifically, [30] (Baseline 1) uses title, location, query, and format information. [29] (Baseline 2)

uses query, title, location, and term occurrence information. Since location and format are not available for context sentences, only content sentences were used in the evaluation. The results are shown in Figure 4. Proposed is better than both Baseline1 and Baseline2. This also shows the effectiveness of the two new features: anchor and E-title.

3.5 Learning Algorithms

Since the number of non-summary sentences is about five times of that of summary sentences in the training set, we set the ratio of C_+ to C_- (please refer to formula (9)) to five. For both SVM classifier and Ranking SVM, we used the Gaussian kernel:

$$k(x, y) = e^{-\|x-y\|^2/2\sigma^2} \quad (14)$$

To decide the value of σ , the training set was divided into three parts and three-fold cross-validation was performed using two parts to train and the left part to validate. σ is initially set to 0.0015625. It is multiplied by 1.5 iteratively until its value is 100. The results of Ranking SVM and SVM classifier are shown in Figure 5 and 6. The σ value that leads to best performance is chosen. The σ values for Ranking SVM and SVM classifier are listed in Table 1.

Table 1. σ value for Ranking SVM and SVM classifier on three sentence sources

	Content	Context	Both
SVM classifier	3.4638	0.3041	2.3092
Ranking SVM	0.0118	11.690	0.0901

Based on the appropriate parameters, both Ranking SVM and SVM classifier were trained on the whole training set. Then, all testing documents were summarized and evaluated. The results are shown in Figure 7. “No Learning” corresponds to the linear combination of all five features. Both SVM classifier and Ranking SVM outperformed “No Learning”, which shows the effectiveness of learning algorithms. Moreover, Ranking SVM is better than SVM classifier as expected.

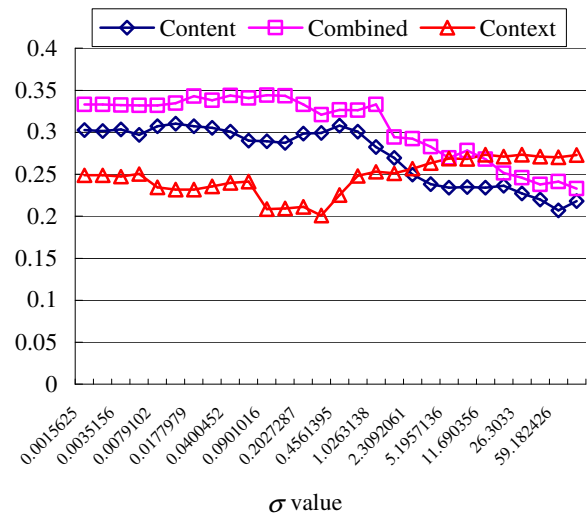


Figure 5. Performance of different σ on validation sets using Ranking SVM.

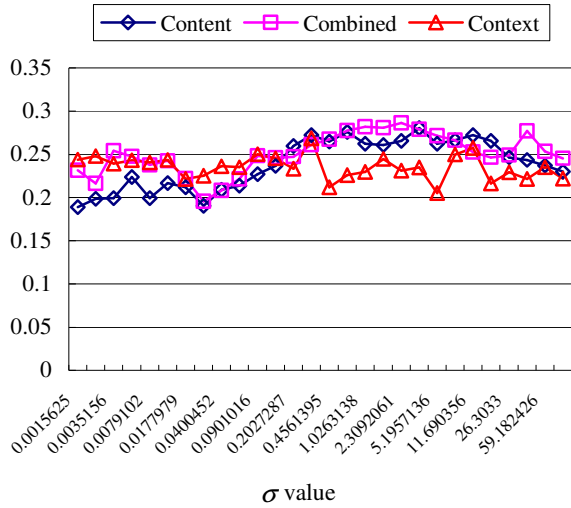


Figure 6. Performance of different σ on validation sets using SVM classifier

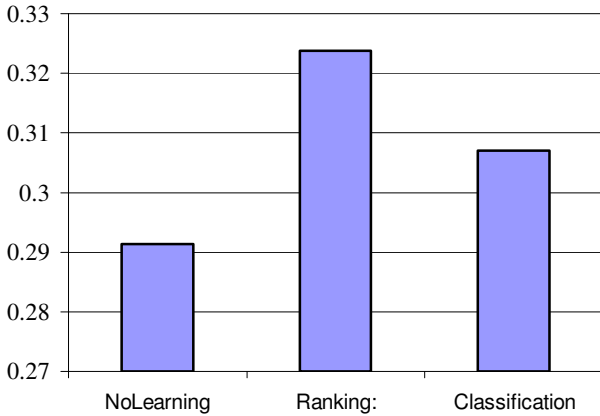


Figure 7. Performance comparison of learning algorithms using both content and context sentences

3.6 Sentence Sources

To show the effect of different sentence sources on features, the summarization results using on only content sentences (Content), only context sentences (Context), or both content and context sentences (Combined) are evaluated and shown in Figure 8. Using both content and context sentences are consistently much better than using either of them alone. Using content sentences only is better than using context sentences only. The reason is that, out of the 175 pages, there are 47 pages (27%) that have no context sentences and 90 (51%) pages that have less than 3 context sentences. For these pages, content information will be very important or even indispensable.

To show the effect of different sentence sources on learning algorithms, two classifiers and two ranking algorithms are trained based on either only content sentences or only context sentences. The parameters were tuned in the same way as in Section 3.5. The σ values for Ranking SVMs and SVM classifiers are listed in Table 1. The results of the classifiers and ranking algorithms using only content sentences, only context sentences, or both content and context sentences are shown in Figure 9 and 10. For

both SVM classifiers and Ranking SVMs, using both sentence sources is clearly better than using either of them alone. For example, the F_1 value of Ranking SVM using both sentence sources is higher than that using content (context) sentences alone by relatively 6.8% (23.2%).

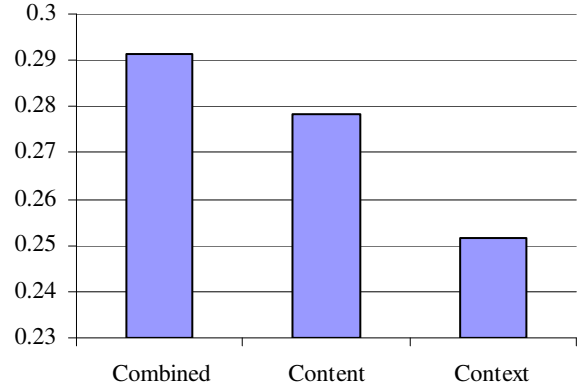


Figure 8. Performance comparison using different sentence sources without learning

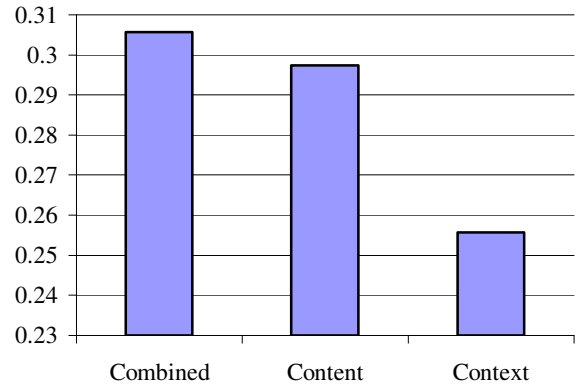


Figure 9. Performance comparison using different sentence sources with SVM classifiers

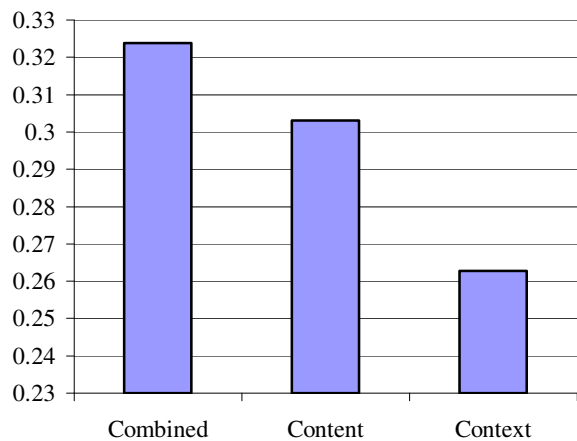


Figure 10. Performance comparison using different sentence sources with Ranking SVM

4. CONCLUSIONS

In this paper, we propose a query biased Web page summarization method based on sentence extraction and ranking. Sentences of both Web page content and link context are used. Unlike most of the learning-based summarization algorithms that treat summarization as a classification problem, we regard summarization as a ranking problem. Our assumption is that sentences within a document are comparable. It is weaker and more reasonable than the assumption of classification-based summarization, which assumes all sentences of training documents are comparable with respect to the class information. Experimental results on 175 documents of 10 queries from TREC dataset show the superiority of the proposed ranking-based summarization over a typical classification-based summarization method. Although we focused on query biased Web page summarization in this paper, the ranking scheme is also suitable for all extraction-based summarization problems. We will pursue this direction in the future.

5. REFERENCES

- [1] Amitay, E. and Paris, C. 2000. Automatically summarising Web sites: is there a way around it?. In Proceedings of the Ninth international Conference on information and Knowledge Management (McLean, Virginia, United States, November 06 - 11, 2000). CIKM '00. ACM Press, New York, NY, 173-179.
- [2] Amini, M. and Gallinari, P. 2002. The use of unlabeled data to improve supervised learning for text summarization. In Proceedings of the 25th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Tampere, Finland, August 11 - 15, 2002). SIGIR '02. ACM Press, New York, NY, 105-112.
- [3] Berger, A. L. and Mittal, V. O. 2000. OCELOT: a system for summarizing Web pages. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM Press, New York, NY, 144-151.
- [4] Berger, A. and Mittal, V. O. 2000. Query-relevant summarization using FAQs. In Proceedings of the 38th Annual Meeting on Association For Computational Linguistics (Hong Kong, October 03 - 06, 2000). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 294-301.
- [5] Burges, C., Shaked, T., Renshaw, E., Lazier, A., Deeds, M., Hamilton, N., and Hullender, G. 2005. Learning to rank using gradient descent. In Proceedings of the 22nd international Conference on Machine Learning (Bonn, Germany, August 07 - 11, 2005). ICML '05, vol. 119. ACM Press, New York, NY, 89-96.
- [6] Buyukkokten, O., Garcia-Molina, H., and Paepcke, A. 2001. Seeing the whole in parts: text summarization for web browsing on handheld devices. In Proceedings of the 10th international Conference on World Wide Web (Hong Kong, Hong Kong, May 01 - 05, 2001). WWW '01. ACM Press, New York, NY, 652-662.
- [7] Carbonell, J. and Goldstein, J. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 335-336.
- [8] Chuang, W. T. and Yang, J. 2000. Extracting sentence segments for text summarization: a machine learning approach. In Proceedings of the 23rd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Athens, Greece, July 24 - 28, 2000). SIGIR '00. ACM Press, New York, NY, 152-159.
- [9] Cramer, K. and Singer, Y. Pranking with ranking. In Proceeding of the conference on Neural Information Processing Systems (NIPS), 2001.
- [10] Craswell, N., Hawking, D., Wilkinson, R., and Wu, M. Overview of the TREC 2003 Web Track, In Proc. TREC 2003, 2003.
- [11] Delort, J., Bouchon-Meunier, B., and Rifqi, M. 2003. Enhanced web document summarization using hyperlinks. In Proceedings of the Fourteenth ACM Conference on Hypertext and Hypermedia (Nottingham, UK, August 26 - 30, 2003). HYPERTEXT '03. ACM Press, New York, NY, 208-215.
- [12] Edmundson, H. P. 1969. New Methods in Automatic Extracting. J. ACM 16, 2 (Apr. 1969), 264-285.
- [13] Eiron, N. and McCurley, K. S. 2003. Analysis of anchor text for web search. In Proceedings of the 26th Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Toronto, Canada, July 28 - August 01, 2003). SIGIR '03. ACM Press, New York, NY, 459-460.
- [14] Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. 2003. An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. 4 (Dec. 2003), 933-969.
- [15] Goldstein, J., Kantrowitz, M., Mittal, V., and Carbonell, J. 1999. Summarizing text documents: sentence selection and evaluation metrics. In Proceedings of the 22nd Annual international ACM SIGIR Conference on Research and Development in information Retrieval (Berkeley, California, United States, August 15 - 19, 1999). SIGIR '99. ACM Press, New York, NY, 121-128.
- [16] Herbrich, R., Graepel, T., and Obermayer, K. 2000. Large margin rank boundaries for ordinal regression. Advances in Large Margin Classifiers, pp. 115-132.
- [17] Hirao, T., Isozaki, H., Maeda, E., and Matsumoto, Y. 2002. Extracting important sentences with support vector machines. In Proceedings of the 19th international Conference on Computational Linguistics - Volume 1 (Taipei, Taiwan, August 24 - September 01, 2002). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1-7.
- [18] Hu, Y., Xin, G., Song, R., Hu, G., Shi, S., Cao, Y., and Li, H. 2005. Title extraction from bodies of HTML documents and its application to web page retrieval. In Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval

- (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 250-257.
- [19] Joachims, T. Making large-Scale SVM Learning Practical, in *Advances in Kernel Methods - Support Vector Learning*. B. Schölkopf et al (ed.), MIT-Press, 1999. pp. 169-184.
- [20] Joachims, T. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD international Conference on Knowledge Discovery and Data Mining* (Edmonton, Alberta, Canada, July 23 - 26, 2002). KDD '02. ACM Press, New York, NY, 133-142.
- [21] Jones, K. S., Galliers, J. R., and Galliers, J. R. 1996 *Evaluating Natural Language Processing Systems: an Analysis and Review*. Springer-Verlag New York, Inc.
- [22] Kupiec, J., Pedersen, J., and Chen, F. 1995. A trainable document summarizer. In *Proceedings of the 18th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Seattle, Washington, United States, July 09 - 13, 1995). E. A. Fox, P. Ingwersen, and R. Fidel, Eds. SIGIR '95. ACM Press, New York, NY, 68-73.
- [23] Luhn, P. H. Automatic creation of literature abstracts. *IBM Journal*, pages 159-165, 1958.
- [24] Mani, I. and Bloedorn, E. 1998. Machine learning of generic and user-focused summarization. In *Proceedings of the Fifteenth National/Tenth Conference on Artificial intelligence/innovative Applications of Artificial intelligence* (Madison, Wisconsin, United States). American Association for Artificial Intelligence, Menlo Park, CA, 820-826.
- [25] Mitra, M., Singhal, A., and Buckley, C. Automatic Text Summarization by Paragraph Extraction. In *Proceedings of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*. (1997) 31-36.
- [26] Radev, D. R., Teufel, S., Saggion, H., Lam, W., Blitzer, J., Qi, H., Çelebi, A., Liu, D., and Drabek, E. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meeting on Association For Computational Linguistics - Volume 1* (Sapporo, Japan, July 07 - 12, 2003). Annual Meeting of the ACL. Association for Computational Linguistics, Morristown, NJ, 375-382.
- [27] Robertson, S. E. and Walker, S. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the 17th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Dublin, Ireland, July 03 - 06, 1994). W. B. Croft and C. J. van Rijsbergen, Eds. Annual ACM Conference on Research and Development in Information Retrieval. Springer-Verlag New York, New York, NY, 232-241.
- [28] Sun, J., Shen, D., Zeng, H., Yang, Q., Lu, Y., and Chen, Z. 2005. Web-page summarization using clickthrough data. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Salvador, Brazil, August 15 - 19, 2005). SIGIR '05. ACM Press, New York, NY, 194-201.
- [29] Tombros, A. and Sanderson, M. 1998. Advantages of query biased summaries in information retrieval. In *Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (Melbourne, Australia, August 24 - 28, 1998). SIGIR '98. ACM Press, New York, NY, 2-10.
- [30] White, R.W., Jose, J.M., and Ruthven, I. A task-oriented study on the influencing effects of query-biased summarization in web search. *Information processing and management*, 39(5) pp 707-733, 2003.