In: Proceedings of the 2005 ACM SIGIR Conference, Salvador, Brazil, 2005. Pages 114 -- 121

Web-Page Summarization Using Clickthrough Data*

Jian-Tao Sun
Dept. of Computer Science
TsingHua University
Beijing 100084, China
sjt@mails.tsinghua.edu.cn

Dou Shen
Hong Kong University of
Science and Technology
Clearwater Bay, Kowloon, HK
dshen@ust.hk

Hua-Jun Zeng
Microsoft Research Asia
5F, Sigma Center, 49 Zhichun
Road, Beijing 100080, China
hjzeng@microsoft.com

Qiang Yang
Hong Kong University of
Science and Technology
Clearwater Bay, Kowloon, HK
qyang@cs.ust.hk

Yuchang Lu
Dept. of Computer Science
TsingHua University
Beijing 100084, China
lyc@tsinghua.edu.cn

Zheng Chen
Microsoft Research Asia
5F, Sigma Center, 49 Zhichun
Road, Beijing 100080, China
zhengc@microsoft.com

ABSTRACT

Most previous Web-page summarization methods treat a Web page as plain text. However, such methods fail to uncover the full knowledge associated with a Web page to build a high-quality summary, because the Web contains many hidden relationships that are not used in these methods. Uncovering the inherent knowledge is important to building good Web-page summarizers. In this paper, we extract the extra knowledge from the clickthrough data of a Web search engine to improve Web-page summarization. We first analyze the feasibility to utilize clickthrough data in text summarization, and then propose two adapted summarization methods that take advantage of the relationships discovered from the clickthrough data. For those pages not covered by the clickthrough data, we put forward a thematic lexicon approach to generate implicit knowledge for them. Our methods are evaluated on a relatively small dataset consisting of manually annotated pages as well as a large dataset that is crawled from the Open Directory Project website. The experimental results indicate that significant improvements can be achieved through our proposed summarizer as compared with summarizers without using the clickthrough data.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; I.5.4 [**Pattern Recognition**]: Applications—*Text processing*

General Terms

Algorithms, Experimentation, Verification

Keywords

Generic Web-page Summarization, Clickthrough Data, Latent Semantic Analysis, Thematic Lexicon

*This work was conducted when the first author was visiting Microsoft Research Asia, Beijing, China.

SIGIR'05, August 15–19, 2005, Salvador, Brazil.

1. INTRODUCTION

With the rapid growth of the WWW, there is an increasing need to succinctly summarize Web pages for the Web users. For example, the Web sites such as Yahoo! and LookSmart organize Web pages into a hierarchical structure and provide a short summary for each page in order to facilitate Web users' quick and accurate browsing and search. On a search engine such as Google, after a query is issued, a list of URL's is returned accompanied with a snippet for each URL, which gives a brief summary of the target page's content. Web-page summarization is also useful when users access Internet via small-display devices such as the personal digital assistants (PDA's) since it is hard to hold the original content of a Web page into the small screen [3]. However, it is quite expensive to generate the summaries manually. Therefore, it is urgent to provide effective automatic Web-page summarization techniques for the above applications.

Web-page summaries can be abstracts or extracts. An extract-summary consists of sentences extracted from the Web page while an abstract-summary may contain words and phrases which do not exist in the original document [16]. Web-page summary can also be either generic or query-dependent. A query-dependent summary presents the information which is most relevant with the initial search query, while a generic summary gives an overall sense of the document's content. A generic summary should meet two conditions: maintain wide coverage of the page's topics and keep low redundancy at the same time [7]. In this paper, we focus on extract-based generic Web-page summarization. Although a lot of works has been done on this kind of summarization, the result is often not satisfying [8]. One reason is that the textual information of a Web page may be scarce, and in some cases, the content contained in a Web page is

very diverse in topics. For the traditional summarization methods which focus on local contents of a document, it is difficult to capture the true meaning of a Web page. What's more, Web pages usually contain a lot of noise, which is hard to be removed simply by statistical methods. Some extra knowledge will be very helpful to distinguish the real content of a Web page from the noise.

The objective of this research is to utilize extra knowledge to improve Web-page summarization. One feasible knowledge source for generating summaries is the clickthrough data. This is because a clickthrough dataset contains users' knowledge on Web pages' content. A user's query words often reflect the true meaning of target Web page's content. Therefore, it would be helpful if the inherent knowledge contained in the clickthrough data can be extracted to complement the Web page contents. What's more, the clickthrough data is collected from millions of users all over the world. For a Web page with multiple topics, Web users may submit different queries to find the topic specific to his/her own information need. Thus the query word set of a page may cover the multiple topics of the target Web page and it could help produce a generic summary to meet the needs of general Web users, in case of biasing towards a specific user. In this paper, we develop a novel Web-page summarization method by using the clickthrough data. To be sure, this is a challenging task. First, Web pages may have no associated query words since they are not visited by web users through search engine. Especially for new emerging pages, no clickthrough data has been collected for them. Second, the clickthrough data are noisy. Web users may click on pages which are not relevant to their issued queries and the click will be recorded by search engine.

Both the above problems could be solved by our proposed thematic lexicon. By using the annotated hierarchical taxonomy of Web pages such as the one provided by ODP website (http://dmoz.org/), we build a thematic lexicon. For each category, the lexicon contains terms submitted by all Web users to browse the pages of this category, as well as the weights of these terms. In conjunction with the proposed summarization methods, the thematic lexicon can be used to complement the scarcity of Web-page content even no clickthrough data was collected associated with these pages. In addition, our method can help filter out noises contained in query words for an individual Web page through the use of statistics over all Web pages of this category. Because the category-specific thematic lexicon provides term distribution on category level and reflects users' knowledge on term usage when they locate pages of this category. Some noisy terms which may be relatively frequent in one page's query words will be given a low weight by our approach.

To leverage the extracted knowledge, we adapt two text-summarization methods to summarize Web pages. The first approach is based on significant-word selection adapted from Luhn's method [14]. The second method is based on Latent Semantic Analysis (LSA) [8]. The experimental results show that both approaches achieve improvements compared with the pure-text-based summarization.

The rest of this paper is organized as follows. In Section 2, we present the related works on Web-page summarization and clickthrough data analysis. Our proposed summarization algorithms are discussed in Section 3. In Section 4, the experimental results are given as well as some discussions. Finally, we conclude our work in Section 5.

2. RELATED WORK

Web-page summarization techniques have been widely applied in many applications [3, 5, 11, 17]. As discussed previously, it would be helpful if extra information can be used to assist in Web-page summarization process. A few works utilize the context information constructed by hyperlinks among Web pages [1, 6]. With the InCommonSense system, Amitay and Paris first extract the text segments containing a link to a Web page. Then the most accurate sentence is chosen from the text segments as the snippet of the target page [1]. Since the context information may be related to the target Web page but contains no clues for summarization, Delort et al. proposed two enhanced Web-page summarization methods using hyperlinks [6]. In their work, the authors studied the characteristics of context information of a Web page as well as the relations between the context information and the target Web page content. Our work differs from these previous ones mainly in the approaches to construct the extra knowledge. These methods try to get the knowledge from hyperlinks which may be sparse and noisy while we get them from the search-engine clickthrough data. The search queries provide a more accurate semantic labeling for the subsequently clicked Web pages, and are found to perform better for Web-page summarization. In addition, in the previous works, Web-page summaries are extracted from the context text segments, while we extract sentences from a Web page, with help of the clickthrough data.

Much research has also been done on clickthrough data analysis to improve search performance, Web-page classification or metadata extraction. Sun et al. proposed a CubeSVD approach to utilize clickthrough data for personalized Web search [18]. Liu et al. proposed a technique for categorizing Web query terms from the clickthrough logs into pre-defined subject taxonomy based on their popular search interests [13]. To the best of our knowledge, few reports in the literature have focused on using the clickthrough data for Web-page summarization. There are several works on automatic metadata extraction which are related with our work. Hulth et al. proposed to extract keywords using domain knowledge [10]. In [9], classification based methods were proposed in order to extract metadata inherent in Web pages and to build relationships among the metadata and Web page categories. In [4], the authors construct taxonomy of queries by retrieving Web pages from search engine to help represent query terms. In this paper, we also construct a thematic hierarchy of query terms. However, we use the clickthrough data instead of Web pages solely. In addition, the research works aforementioned did not concern how to use the extracted knowledge data for Web-page summarization, which is the focus of this paper.

3. SUMMARIZE WEB PAGES USING CLICKTHROUGH DATA

In this section, we first give several observations on clickthrough data, which validate our assumption to leverage it for Web-page summarization. Then we describe two proposed summarization methods which leverage the knowledge data extracted from clickthrough data. At last, we propose an approach to build a thematic lexicon as a knowledge source for summarizing Web pages which are not covered by the clickthrough data.

3.1 Empirical Study on Clickthrough Data

Consider the typical search scenario: a user (u) submits a query (q) to search engine, the search engine returns a ranked list of Web pages. Then the user clicks on the pages (p) of interest. After a period, the server side will accumulate a collection of clickthrough data, which can be represented by a set of triples $< u,\ q,\ p >$. Thus the clickthrough data records how Web users find information through queries. From a statistical point of view, the query word set corresponding with a Web page contains human's knowledge on how the pages are related with their issued queries. Some users even refine their queries in order to find the desired information. Therefore, the collection of queries is supposed to well reflect the topics of the target Web page.

We first conduct an experiment to investigate whether the query words are related with the topics of the Web page. For experimental purpose, we have crawled a set of Web pages from ODP directory. One month's clickthrough data collected by MSN search engine is also available for our experiments. In the clickthrough data, among the 260,763 web pages accessed by users during this month, 109,694 of them contain "KEYWORD" metadata, which is an important indicator of the content of Web pages. According to the statistics, 45.5% of keywords occurs in the query words and 13.1% of query words appear as keywords. This result supports our hypothesis that query words are indicative of Web pages' contents, which motivates us to leverage the clickthrough data to improve Web-page summarization.

In order to give more evidence that clickthrough data is helpful to summarizing Web pages, we conducted a second experiment. We collect a small dataset consisting of 90 pages which are covered by the clickthrough data. For evaluation purposes, we asked three human evaluators to conduct a manual summarization task on these Web pages, without knowing the queries. Each evaluator was asked to extract the sentences which are important for a Web page. There is no constraint on the number of sentences extracted. According to the statistics, about 58% of the sentences in the original Web page contain query words and each sentence contains 1.48 query words on average. However, for manually created summaries, the percentage of sentences containing queries becomes 71.3% and the average query word length in each sentence becomes 2.0. It indicates that the human evaluators tend to extract the sentences with query words as Web-page summaries. Therefore, it is natural for us to pay high attention to the query words when summarizing the Web pages automatically.

From the above empirical study, it is safe for us to come to the conclusion that clickthrough data is promising to help Web-page summarization.

3.2 Adapted Web-page Summarization Methods

Suppose that we have a set of query terms for each page now, we propose two adapted summarization methods to leverage the clickthrough data.

3.2.1 Adapted Significant Word (ASW) Method

Our first summarization method is adapted from Luhn's algorithm, which is a classical algorithm designed for text summarization [14]. In Luhn's method, each sentence is assigned a significance factor and the sentences with high significance factors are selected to form the summary. In order to compute the significance factor of each sentence, a set of significant words are constructed first. In Luhn's algorithm, significant words are selected according to word frequency in a document. That is, those words with frequency between high-frequency cutoff and low-frequency cutoff are selected as significant words. After this is done, the significant factor of a sentence can be computed as follows: (1) Set a limit L for the distance at which any two significant words could be considered as being significantly related. (2) Find out a portion in the sentence that is bracketed by significant words not more than L non-significant words apart. (3) Count the number of significant words contained in the portion and divide the square of this number by the total number of words within the portion. The result is the significant factor of each sentence.

In order to customize this procedure to leverage query terms for Web-page summarization, we modify the significant word selection method. The basic idea is to use both the local contents of a Web page and query terms collected from the clickthrough data to decide whether a word is significant or not. Each candidate word is assigned with a significance factor w_i given in Equation 1.

$$w_i = (1-\alpha)\cdot\frac{tf_i^p}{\max\left(tf_i^p\right)} + \alpha\cdot\frac{tf_i^q}{\max\left(tf_i^q\right)} \tag{1}$$

In Equation 1, tf_i^p and tf_i^q denote frequencies of the i-th term in the local text content of a Web page and in the query set respectively. The significance factor w_i is measured in a weighted combination: α is a trade-off parameter when combining the two significance measurements. After the significance factors for all words are calculated, we rank them and select the top $N\%$ as significant words (N is the number of distinct terms in the Web page). Then we employ Luhn's algorithm to compute the significant factor of each sentence.

3.2.2 Adapted Latent Semantic Analysis (ALSA) Method

An LSA based text summarization method was proposed in [8]. Suppose that there are m distinct terms in a n documents collection. The corpus can be represented by a term-document matrix $X \in R^{m\times n}$, whose component x_{ij} is the weight of term t_i in document d_j. The Singular Value Decomposition (SVD) of X is given by:

$$X = U\Sigma V^T \tag{2}$$

In Equation 2, U and V are the matrices of the left and right singular vectors. Σ is the diagonal matrix of singular values. LSA approximates X with a rank-k matrix:

$$X_k = U_k\Sigma_k V_k^T \tag{3}$$

by setting the smallest $r-k$ singular values to zero (r is rank of X). That is, the documents are represented in the k dimensional space spanned by column vectors of U_k [2].

The advantage of LSA derives from its ability to capture the latent relations between terms. In the dimension-reduced space, each singular vector corresponds to a latent concept, with the corresponding singular value measuring the importance of the concept. In [8], Gong et al. proposed an extraction based summarization algorithm. Firstly, a term-sentence matrix is constructed from the original text document. Next, LSA analysis is conducted on the matrix. In the singular vector space, the i-th sentence is represented by the column vectors $\varphi_i = [v_{i1}, v_{i2}, \cdots, v_{ir}]^T$ of V^T.

Each element in φ_i measures the importance factor of this sentence on the corresponding latent concept. In the last step, a document summary is produced incrementally. For the most important concept, the sentence having the largest importance factor is selected into the summary. Then, the second sentence is selected for the next most important concept. This procedure repeated until a predefined number of sentences are selected.

Our LSA-based summarization method is a variant of Gong's method. We utilize the query-word knowledge by changing the term-sentence matrix: if a term occurs as query words, its weight is increased according to its frequency in query word collection. In this approach, we expect to extract sentences whose topics are related to the ones reflected by query words. As discussed in [2, 8], LSA is capable of capturing the latent associations among terms. If a word combination pattern is salient and recurring frequently in a document, this pattern will be captured and represented by one of the singular vectors. If we increase the weights of query terms, these terms, as well as others which frequently co-occur, will make more contributions to singular vector formulation. Thus in the sentence extraction step, the candidate sentences which are semantically related with query terms will be selected firstly.

Since the term-sentence matrix determines the SVD result, its representation may influence the summarization results. The term frequency vector of each sentence can be weighted by different weighting (global weighting and local weighting) and normalization methods. These schemes are studied in [8]. According to their experiments, the global weighting and normalization schemes lower the summarization performance, while the local weighting schemes produce similar results. Thus, in this paper, a term frequency (TF) approach without weighting or normalization is used to represent the sentences in Web pages. Terms in a sentence are augmented by query terms as follows:

$$w_i = (1-\beta)\cdot\frac{tf_i^p}{\max(tf_i^p)} + \beta\cdot\frac{tf_i^q}{\max(tf_i^q)} \qquad (4)$$

In this equation, β is a parameter used to tune the weights of query terms. Here, tf_i^p is the frequency of term i in a sentence, while tf_i^q denotes term frequency in query set.

3.2.3 Advantages of the Adapted Methods

There are several advantages for both the above modified summarization algorithms. First, the extra knowledge of query terms is utilized to help select significant words and to modify the page representation. Because of the subjective characteristic of word usage, some words may have a relatively low term frequency in the Web page, even though they are topic-related. The contributions of these words are constrained because both significant-word method and LSA are term-frequency based methods. However, these words may be used as query words by Web-search users as they are related with the topics of the Web page. Second, our approach can, to some extent, handle the noises of query words. The reason is that the contribution of each query term is proportional to its frequency in the query term collection. Thus a query term with very low frequency makes few contributions to significance factor calculation and latent concept formation respectively. Finally, for Luhn's method, the frequency-cutoff method may lead to a lot of significant words for long pages. This problem is avoided in our ASW approach by keeping the number of significant words to be $N\%$ times the number of distinct terms occurring in the Web page. Therefore, both the proposed approaches are supposed to produce better summarization results by leveraging the clickthrough data.

3.3 Summarize Web Pages Not Covered by Clickthrough Data

According to statistics on the one month's clickthrough data, only 23.1% out of the crawled ODP pages (in English) was browsed and associated with query words (the detailed numbers are given in Section 4.1). Thus for pages which are not browsed by Web users, neither summarization method proposed in Section 3.2 can be directly applied on them. In this section, we build a hierarchical lexicon using the clickthrough data and apply it to help summarize those pages.

Since all ODP Web pages have been manually organized into a hierarchical taxonomy, we combine this large knowledge source and the clickthrough data to build a thematic lexicon. For each category of the taxonomy, the lexicon contains all query terms that users have submitted to browse Web pages of this category, as well as weights of these terms, where the latter measures the likelihood that Web users will use this term to locate pages of this category. In this paper, we use $TS(c)$ to represent a set of terms associated with category c, as well as their corresponding weights. Thus the thematic lexicon is a set of TS, which correspond with categories in ODP and are organized using the ODP category structure. The lexicon is built as follows: first, TS corresponding to each category is set empty. Next, for each page covered by the clickthrough data, its query words are added into TS of categories which this page belongs to as well as all its parent categories. When a query word is added into TS, its frequency is added to its original weight in TS. If a page belongs to more than one category, its query terms will be added into all TS associated with all its categories. At last, term weight in each TS is multiplied by its Inverse Category Frequency (ICF). The ICF value of a term is the reciprocal of its frequency occurring in different categories of the hierarchical taxonomy.

After the hierarchical lexicon is built, we can use it to summarize Web pages that are not covered by the clickthrough data. For each Web page to be summarized, we first look up the lexicon for TS according to the page's category. Then the summarization methods proposed in Section 3.2 are used. Weights of the terms in TS can be used to select significant words or to update the term-sentence matrix. If a page to be summarized has multiple categories, the corresponding TS are merged together and weights are averaged. When a TS does not have sufficient terms, TS corresponding with its parent category is used.

The hierarchical lexicon-based Web-page-summarization method has at least two advantages. First, the category-specific TS provides a distribution of topic terms in this category, which reflects Web users' knowledge on term usage when they locate information of this category. Second, some noisy terms which may be relatively frequent in one page's query words will be given a low weight through the use of statistics over all Web pages of this category.

4. EXPERIMENTS

In this section, we will investigate whether the adapted Web-page summarization methods are superior to the ones

without using clickthrough data. We introduce the experiment data set, the evaluation metrics and the experiment results.

4.1 Data Set

The clickthrough data was collected from MSN search engine. This data set contains about 44.7 million records of 29 days from Dec 6 of 2003 to Jan 3 of 2004. As we collected the clickthrough data, a set of Web pages of the ODP directory are crawled. Among the 3,074,678 Web pages crawled, we removed those which belong to "World" and "Regional" categories, as many of them are not in English. At last we got 1,125,207 Web pages, 260,763 of which are clicked by Web users using 1,586,472 different queries.

Two different data sets were used for experiment. The first one, denoted by DAT1, consists of 90 pages which are selected from the 260,763 browsed pages. Table 1 gives the query numbers associated with each page of DAT1. Three human evaluators were employed to summarize these pages. Each evaluator was requested to extract the sentences which he/she deemed to be the most important ones for a Web page. There is no constraint on the number of sentences to be extracted. Table 2 describes the overall consistencies among the three evaluators. For example, for the Evaluator1:Evaluator2 pair, 0.45 means 45% sentences in evaluator1's summary are also included in evaluator2's summary. Each number given in this table is an average result over all Web pages. From Table 2, we can find the three evaluators have a relatively high disparity on the 90 pages. Similar observations of high disparities between human evaluators are also observed in previous works, such as [8]. Thus, in this paper, the experiments are evaluated using the annotation results of all three evaluators. The average results are also reported.

We also use a relatively large scale data set, denoted by DAT2, to evaluate our summarization methods. We preprocess the 260,763 pages contained in the clickthrough data using a layout analysis algorithm. After the content body of each page is extracted, the textual content is segmented into sentences by a sentence segmenting program implemented in our group. In addition, descriptions of each page ("DESCRIPTION" metadata) are also extracted. We keep the Web pages with a description of over 200 characters and containing at least 10 sentences, from which 10,000 pages are randomly selected and constitutes DAT2 data set. Since the description is provided by the page editor to give a general description of this page, we use it as the ideal summary.

Table 1: Number of Queries Associated with Annotated Pages

Number of Queries	Number of Pages
<5	44
$5 \sim 10$	20
$10 \sim 50$	13
> 50	23
Total	90

4.2 Performance Evaluation

Both intrinsic and extrinsic methodes are proposed for automatic summarization evaluation [12, 15]. In this paper, we employ two intrinsic evaluation approaches to evaluate the proposed approaches.

Table 2: Disparities among the Human Evaluators

	Evaluator1	Evaluator2	Evaluator3
Evaluator1		0.45	0.50
Evaluator2	0.54		0.50
Evaluator3	0.55	0.46	

4.2.1 Precision, Recall and F_1

Precision, recall and F_1 are straightforward measures widely used in summarization evaluation. For each document, the manually extracted sentences are considered as the reference summary. This approach compares the candidate summary with the reference summary and computes the precision, recall and F_1 values:

$$P = \frac{\mid S_{ref} \cap S_{cand} \mid}{\mid S_{cand} \mid}; R = \frac{\mid S_{ref} \cap S_{cand} \mid}{\mid S_{ref} \mid}; F_1 = \frac{2PR}{P+R} \quad (5)$$

where S_{cand} and S_{ref} denotes the sentences contained in the candidate summary and the reference summary respectively.

4.2.2 ROUGE Evaluation

ROUGE[1] is a software package adopted by DUC[2] for automatic summarization evaluation [12]. It measures summarization 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 is an n-gram recall measure which is defined as follows:

$$ROUGE-N = \frac{\sum_{S \in \{ref\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ref\}} \sum_{gram_n \in S} Count(gram_n)} \quad (6)$$

In Equation 6, N stands for the length of the n-gram, $Count_{match}(gram_n)$ is the maximum number of n-grams co-occurring in the candidate summary and the reference summary, $Count(gram_n)$ is the number of n-grams in the candidate summary.

According to [12], among the evaluation methods implemented in ROUGE, ROUGE-N (N=1, 2) is relatively simple and works well in both single document summarization evaluation tasks and evaluation of very short summaries. In this paper, we evaluated our experiments using all the methods provided by the ROUGE package and only reported ROUGE-N where N=1, since the conclusions drawn from different methods are quite similar.

4.3 Experimental Results and Analysis

In this subsection, we conduct experiments to evaluate the summarization methods proposed in Section 3.2. On the DAT1 dataset, in order to evaluate our methods on summaries of different evaluators, we keep the number of sentences extracted be equal with that of the human summary. In this case, the precision, recall and F_1 measure are all equal. On both the data sets, the summarization methods are evaluated by the ROUGE software.

[1]http://www.isi.edu/ cyl/ROUGE/
[2]http://duc.nist.gov/

4.3.1 Summarization Results on DAT1

We first conduct experiments to investigate whether the adapted summarizers can benefit from query terms associated with each page. For the ASW method, we vary α from 0 to 1 (in step 0.1) to change the influence of query terms on significant word selection. The summarization results are listed in Figure 1, measured by precision and ROUGE-1 respectively. When α is 0, the clickthrough data is ignored and only the local features of a Web page are used to select significant words. When α is 1, only query words are used. For ALSA, the parameter β is also varied from 0 to 1. The results are given in Figure 2. From the two figures, we can find that both methods achieve significant improvements when evaluated on the manual summaries of different evaluators. For ASW method, the average precision of the three evaluators got a relative improvement of 20.7%, from 0.29 (when no query terms are used) to 0.35 (when query terms are used and $\alpha = 0.5$). The average ROUGE-1 measure got a relative improvement of 11.5% (from 0.52 to 0.58). The ALSA method achieves a relative improvement of 12.9% and 11.5% when measured by precision and ROUGE-1 respectively, compared with without using the clickthrough.

On this data set, we also evaluate our summarization methods using the thematic lexicon approach. We first use the ODP directory and the clickthrough data to build a thematic lexicon. Since the clickthrough data contains only 260,763 pages, our lexicon contains 141,869 categories, which is a subset of the ODP category structure. When we summarize a page, we ignore the query words associated with it. As described in section 3.3, the thematic lexicon is used to help generate query words. We select query terms from the lexicon based on the category of the page to be summarized. If terms under this category have more than P% overlap with distinct terms in the Web page, then they are used for summarization. Otherwise, we try to use lexicon terms of its parent category. This process continues until we find a category which covers enough query terms or until we reach the root of the thematic lexicon. In this experiment, P is empirically set to 6. Evaluation results are described in Figure 3 and Figure 4. We can find that the ASW method achieves an improvement of 7.8% and 5.3% measured in precision and ROUGE-1 respectively. The ALSA method achieves an improvement of 7.6% and 6.7% when measured in precision and ROUGE-1 respectively.

4.3.2 Summarization Results on DAT2

From the above experimental results, we can find the ROUGE-1 measure and the precision measure are consistent with each other in most cases, especially when the average results of the three evaluators are averaged. On this data set, only ROUGE-1 measure is used for evaluation. The extract-summary of each page is generated by our proposed summarization algorithms and the leading characters of length equal with the description sentences are used for evaluation. The results are illustrated in Figure 5. In Figure 5, "Text" denotes summarization based on textual content of Web pages. "Query" denotes summarization using query words issued to locate Web pages. "Lexicon" denotes the page queries are ignored while the thematic lexicon is used to help summarization. Since the description length is commonly short and the ROUGE-1 measure is recall based, the summarization results are relatively poor. From the results in Figure 5, we can find that the clickthrough data can improve the Web-page summarization. Even the real queries of Web pages are dropped, the thematic lexicon-based methods can still lead to better summaries compared with local textual content based summarizers, although the performance is not so good as that of page-query based methods.

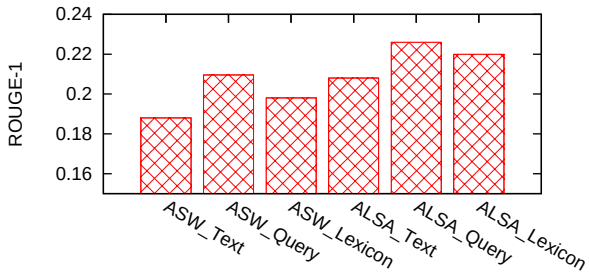


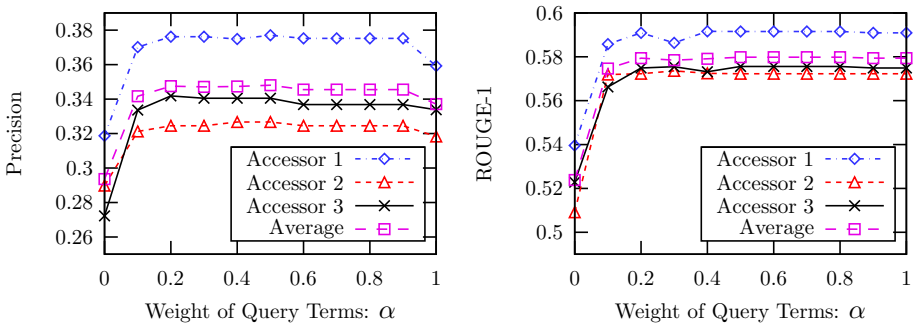
Figure 5: Summarization results on DAT2, evaluated by ROUGE-1 measure

Table 3: Top 10 Terms in Thematic Lexicon, before and after ICF Re-Weighting Approach is Applied (after Stemming)

Computer/Software		*Computer/Hardware*	
Before ICF	**After ICF**	**Before ICF**	**After ICF**
mapquest	mapquest	com	logitech
com	kazaa	www	epson
www	winmx	logitech	lexmark
free	webshot	epson	pny
kazaa	winzip	lexmark	toshiba
wallpap	winrar	toshiba	factorydirect
screensav	imesh	comput	viewson
christma	winamp	sharp	microcent
screen	spybot	tivo	msi
download	microsoft	tw	mstrox

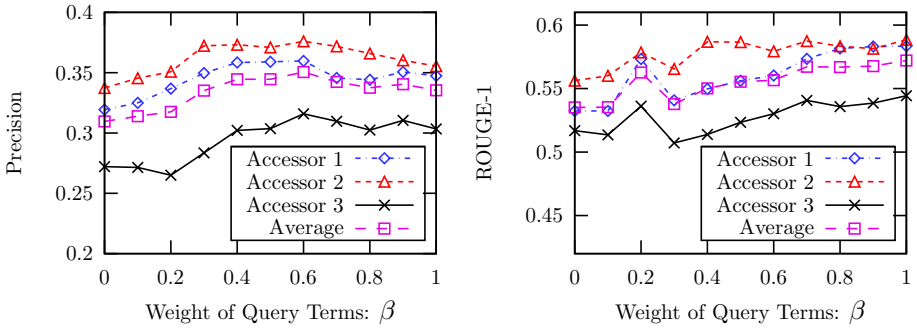
4.3.3 Discussions

All the above experiments indicate the clickthrough data are helpful for generic Web-page summarization and both our proposed methods can leverage this knowledge source well. When the thematic lexicon is used to help summarize the Web pages which are not covered by the clickthrough data, the improvements are not as significant as when queries of a page are directly used. However, the lexicon-based approach achieves better results compared with pure-text-based summarizers. This is because the thematic lexicon built from clickthrough data can discover the topic terms associated with a specific category and the ICF-based approach can effectively assign weights to terms of this category. As illustrated in Table 3, the top 10 terms of two categories are listed. From this table, we can find that ICF-based re-weighting can help discover topic terms of a specific category, while terms like "com", "www", "download" are assigned with lower weights. Although there are high disparities between different human summarizers, improvements can be achieved when the results are evaluated on summaries of each evaluator. This indicates our methods can help produce summaries which meet general Web users by leveraging clickthrough data. In most cases, our



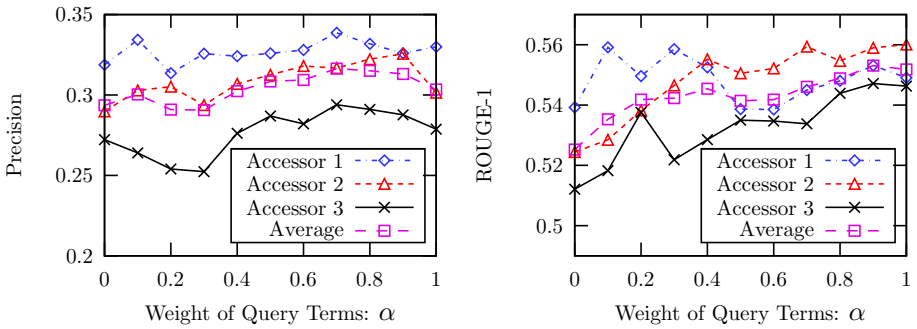
(a) Measured by Precision (b) Measured by ROUGE-1

Figure 1: Summarization results using ASW method, on DAT1 with query words



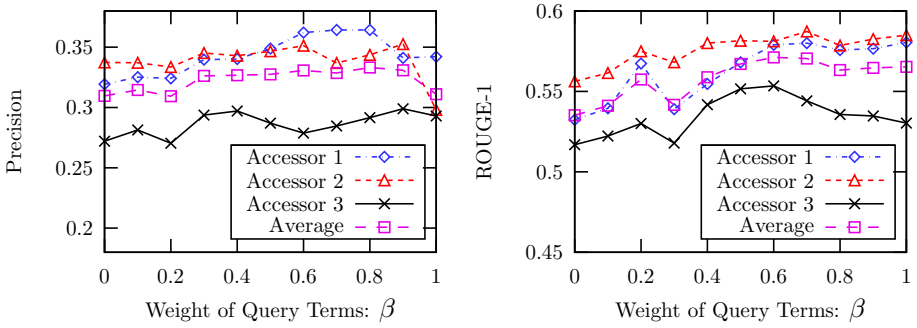
(a) Measured by Precision (b) Measured by ROUGE-1

Figure 2: Summarization results using ALSA method, on DAT1 with query words



(a) Measured by Precision (b) Measured by ROUGE-1

Figure 3: Summarization results using ASW method, on DAT1 without queries



(a) Measured by Precision (b) Measured by ROUGE-1

Figure 4: Summarization results using ALSA method, on DAT1 without queries

methods achieve optimal results when the local contents of a Web page and the clickthrough data are combined together, which verify our hypothesis that the clickthrough data can complement the textual contents of Web pages for summarization tasks.

5. CONCLUSIONS AND FUTURE WORK

We leverage extra knowledge from clickthrough data to improve Web-page summarization. Two extract-based methods are proposed to produce generic Web-page summaries. For the pages which are not covered by the clickthrough data, we build a thematic lexicon using the clickthrough data in conjunction with an available hierarchical Web directory. The experimental results show that significant improvements are achieved compared with summarizers without using clickthrough logs.

Our experiments indicate the trade-off parameter can influence the summarization result when either proposed method is used. Therefore it would be interesting to propose a method to determine its value automatically. Besides, we will also study how to leverage other types of knowledge, such as word clusters and thesaurus, hidden in the clickthrough data to enhance Web page summarization. We also plan to evaluate our methods using extrinsic evaluation metrics and much larger data sets.

6. ACKNOWLEDGMENTS

The authors would like to thank Ya-Bin Kang for his help on organizing the Web page annotation process and trying some experiments in this work, and Xue-Mei Jiang for preparing the clickthrough data. The sudy was funded in part by Natural Science Foundation under the grant number 60473115 and 60403021. Qiang Yang and Dou Shen are supported by Hong Kong RGC HKUST6180/02E.

7. REFERENCES

[1] E. Amitay and C. Paris. Automatically summarising web sites: is there a way around it? In *Proceedings of the 9th international conference on Information and knowledge management*, pages 173–179, New York, NY, USA, 2000. ACM Press.

[2] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. *SIAM Rev.*, 37(4):573–595, 1995.

[3] O. Buyukkokten, H. Garcia-Molina, and A. Paepcke. Seeing the whole in parts: text summarization for web browsing on handheld devices. In *Proceedings of the tenth international conference on World Wide Web*, pages 652–662. ACM Press, 2001.

[4] S. Chuang and L. Chien. Enriching web taxonomies through subject categorization of query terms from search engine logs. *Decision Support Systems*, 35:113–127, 2003.

[5] W. T. Chuang and J. Yang. 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*, pages 152–159. ACM Press, 2000.

[6] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. Enhanced web document summarization using hyperlinks. In *Proceedings of the 14th ACM conference on Hypertext and hypermedia*, pages 208–215, New York, NY, USA, 2003. ACM Press.

[7] J. Goldstein, M. Kantrowitz, V. Mittal, and J. Carbonell. 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*, pages 121–128, NY, USA, 1999. ACM Press.

[8] Y. Gong and X. Liu. Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 19–25. ACM Press, 2001.

[9] C.-C. Huang, S.-L. Chuang, and L.-F. Chien. Using a web-based categorization approach to generate thematic metadata form texts. In *ACM Transactions on Asian Language Information Processing*, pages 190–212. ACM Press, 2004.

[10] A. Hulth, J. Karlgren, A. Jonsson, H. Bostrom, and L. Asker. Automatic keyword extraction using domain knowledge. *Computational Linguistics and Intelligent Text Processing*, 2004.

[11] K. Kummamuru, R. Lotlikar, S. Roy, K. Singal, and R. Krishnapuram. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the 13th international conference on World Wide Web*, pages 658–665, New York, NY, USA, 2004. ACM Press.

[12] C. Y. Lin and E. H. Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *2003 Language Technology Conference*. ACM Press, 2003.

[13] F. Liu, C. Yu, and W. Meng. Personalized web search by mapping user queries to categories. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 558–565. ACM Press, 2002.

[14] H. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[15] I. Mani, D. House, G. Klein, L. Hirschman, T. Firmin, and B. Sundheim. The tipster summac text summarization evaluation. In *Proc. of the 9th conference on European chapter of the Association for Computation Linguistics*, Bergen, Norway, 1999.

[16] I. Mani and M. T. Maybury. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, 1999.

[17] D. Shen, Z. Chen, Q. Yang, H.-J. Zeng, B. Zhang, Y. Lu, and W.-Y. Ma. Web-page classification through summarization. In *Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 242–249. ACM Press, 2004.

[18] J.-T. Sun, H.-J. Zeng, H. Liu, Y. Lu, and Z. Chen. Cubesvd: A novel approach to personalized web search. In *Proceedings of the 14th international conference on World Wide Web*, pages 652–662. ACM Press, 2005.