Data Cleansing for Web Information Retrieval Using Query Independent Features

Yiqun Liu, Min Zhang, and Rongwei Cen
State Key Lab of Intelligent Technology & Systems, Tsinghua University, Beijing, People's Republic of China. E-mail: liuyiqun03@mails.tsinghua.edu.cn

Liyun Ru
R&D center, Sohu Corporation, Beijing, People's Republic of China. E-mail: ruliyun@sohu-rd.com

Shaoping Ma
State Key Lab of Intelligent Technology & Systems, Tsinghua University, Beijing, People's Republic of China. E-mail: msp@tsinghua.edu.cn

Understanding what kinds of Web pages are the most useful for Web search engine users is a critical task in Web information retrieval (IR). Most previous works used hyperlink analysis algorithms to solve this problem. However, little research has been focused on query-independent Web data cleansing for Web IR. In this paper, we first provide analysis of the differences between retrieval target pages and ordinary ones based on more than 30 million Web pages obtained from both the Text Retrieval Conference (TREC) and a widely used Chinese search engine, SOGOU (www.sogou.com). We further propose a learning-based data cleansing algorithm for reducing Web pages that are unlikely to be useful for user requests. We found that there exists a large proportion of low-quality Web pages in both the English and the Chinese Web page corpus, and retrieval target pages can be identified using query-independent features and cleansing algorithms. The experimental results showed that our algorithm is effective in reducing a large portion of Web pages with a small loss in retrieval target pages. It makes it possible for Web IR tools to meet a large fraction of users' needs with only a small part of pages on the Web. These results may help Web search engines make better use of their limited storage and computation resources to improve search performance.

Introduction

The explosive growth of data on the Web makes information management and knowledge discovery increasingly difficult. The size of the Web document collection has become one of the main obstacles for most Web-based information management technologies, such as Web information retrieval (IR) and Web data mining. The number of pages indexed by Web information retrieval tools (or search engines) has been increasing at a high speed. For example, Google (http://www.google.com/) indexed over 8 billion pages in December 2004: about 20 times as many as what it indexed in the year of 2000 (Sullivan, 2005). However, even such a huge amount of pages still cannot cover the whole size of the Web, which already contained more than 20 billion surface Web pages and 130 billion deep Web pages in February 2003, according to the How Much Info project (Lyman & Varian, 2003).

In September 2005, Google dropped from its home page the famous count of pages in its index, which indicated the end of the index size war between commercial search engines (Hedger, 2005). This event also suggests that Web IR tools currently pay more attention to data quality than data quantity. It is well known that the Web is filled with noisy, unreliable, low-quality, and sometimes contradictory data. Therefore, a data cleansing process is necessary before retrieval is performed. Moreover, the cleansing process should be query-independent because the cleansed data set is supposed to meet all kinds of Web search requests. Here emerges the problem: the goal of quality estimation for Web pages is closely related to users' information needs, which are reflected by their queries; however, the process of quality estimation has to be independent of users' queries. This is why Web page quality estimation is considered one of the greatest challenges for search engines by Henzinger, Motwani, and Silverstein (2003) from Google.

Although it is a difficult task to carry out Web data cleansing query-independently, several previous works may help.

For example, the success of PageRank (Brin & Page, 1998) and other hyperlink analysis algorithms such as HITS (Kleinberg, 1999) proves that it is possible to evaluate the importance of a Web page without query information. However, those hyperlink analysis algorithms estimate the quality of a Web page by its probability of being visited instead of its usefulness for a search engine user.

In order to cleanse Web data according to its importance for Web search users, we propose a novel data cleansing method: First, we try to find the differences between retrieval target pages (pages that can be answers for certain Web search queries, see Features of Retrieval Target Pages) and ordinary pages. The differences are located through an analysis of more than 37 million Web pages from bothanEnglish corpus (.GOV corpus adopted in TREC) and a Chinese corpus (collected by Sogou.com). According to statistical comparison, several query-independent features are found to be able to tell the differences between these two kinds of pages. Then a learning-based algorithm based on these features is designed to cleanse Web data using retrieval target page classification.

The main contributions of our work are the following:

1. A study of the query-independent feature of Web pages is conducted to draw a clear picture of the differences be-tween retrieval target pages and ordinary Web pages.
2. A learning-based method is proposed to locate high-quality Web pages automatically according to the chances of becoming retrieval target pages instead of the probabilities of being visited.
3. The possibility of achieving better retrieval performance with a cleansed page set is discussed.

The remaining part of the paper is organized as follows: Related Work gives a brief review of related work in Web data cleansing and Web page classification. Features of Retrieval Target Pages compares the differences between retrieval target pages and ordinary pages with query-independent feature analysis (with both commonly used features and newly proposed ones). Learning-Based Web Data Cleansing Algorithm describes the details of the data cleansing algorithm, including the query-independent features and the learning method. The experimental results are presented in Experimental Results and Discussions to assess the performance of our algorithm. Section 6 gives the conclusion of the paper and some possible future research issues.

Related Work

Data Cleansing for Web Information Retrieval

Search engine designers have realized the importance of data cleansing or Web page quality estimation for quite a long time. In 2003, Henzinger and associates (2003) from Google proposed that it would be extremely helpful for search engines to be able to identify the quality of Web pages independently of a given user's requests. However, because of lack of a large-scale Web page corpus, few researchers, aside from those in enterprises, paid much attention to related issues.

According to the work of Yi, Liu, and Li (2003), most studies in IR-oriented Web page data cleansing can be grouped into two categories: local-scale data cleansing and global-scale data cleansing. Local-scale cleansing is used to eliminate noise within an individual page, while global-scale cleansing deals with low-quality data with a size no smaller than that of a Web page.

There have been several works in local-scale Web data cleansing to meet the needs of Web data mining researchers. Even before the World Wide Web (WWW) appeared, there had already existed several studies in data cleansing for half-structured document formats such as Bibtex. Currently, this kind of cleansing is often described as a two-stage process: partitioning a Web page into blocks, and then locating the useful blocks or discarding the useless ones. For Web page partitioning, many researchers have considered using the tag information and dividing a page on the basis of the type of tags (Buyukkokten, Garcia-Molina, & Paepcke, 2001; Kaasinen, Aaltonen, Kolari, Melakoski, & Laakko, 2000; Wong & Fu, 2000). Some techniques (Buttler, Liu, & Pu, 2001; Embley, Jiang, & Ng, 1999; Wong & Lam, 2004) are based on the analysis of both the layouts and the actual content of Web pages. Cai, Yu, Wen, and Ma (2003) used the layout structures to build the visual structure of a Web page and fulfill the partitioning task in terms of the visual structure. Besides using information inside a Web page, researchers tried to find the common style of noisy data inside a Web site (called Site Style Tree by Yi, Liu, & Li [2003]). With that method they partitioned the Web pages inside a site into main content blocks and noisy blocks. After a Web page is partitioned into several blocks, algorithms based on learning mechanisms (Song, Liu, Wen, & Ma, 2004) or based on hyperlink structure (Cai, He, Wen, & Ma, 2004) can be performed to locate the important blocks or cleanse the unimportant ones.

Most work on global-scale data cleansing has been based on the hyperlink structure of Web pages. Recently, many researchers have tried to improve existing hyperlink analysis algorithms such as PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1999). This kind of analysis is based on two basic assumptions, according to suggested by Craswell and colleagues (Craswell, Hawking, & Robertson, 2001): a recommendation assumption and a topic locality assumption. It is assumed that if two pages are connected by a hyperlink, the page linked is recommended by the page that links to it (*recommendation*) and the two pages share a similar topic (*locality*). Hyperlink analysis algorithms are used by many commercial search engines (such as the work of Page, Brin, Motwani, & Winograd, 1998) and adopted by many researchers (such as Kumar, Raghavan, Rajagopalan, & Tomkins, 2006). These algorithms rely on these two assumptions, which only hold for an ideal Web environment. However, the WWW is currently filled with spam links and advertising links so the assumptions as well as the algorithms based on them are nott working very well. A better global-scale data cleansing algorithm should use additional information from both inside a page and across pages. However, to our knowledge, there has been little research on such an algorithm.

Our data cleansing algorithm can be regarded as a global-scale one. It makes use of query-independent page features both from hyperlink analysis and from page layout analysis. These features are independent of user requests so that the cleansing process can be performed to estimate Web page quality in an offline way.

High-Quality Web Page Classification

Web page classification is adopted in our data cleansing algorithm to separate potential retrieval target pages from ordinary ones using query-independent features. The retrieval target page classification problem shares a similar difficulty with the Web page classification problem described by Yu, Han, and Chang (2004) in the lack of negative examples. Positive examples can be annotated by a number of assessors using techniques such as pooling (Hawking & Craswell, 2005). However, there may be many reasons why a Web page is not a high-quality one so a uniform sampling without bias is almost impossible.

Several learning mechanisms based on unlabeled data and a number of positive examples are proposed to accomplish the task of Web page classification. Techniques such as PEBL learning framework (Yu et al., 2004), semisupervised learning (Nigam, McCallum, Thrun, & Mitchell, 2000), single-class learning (Denis, 1998), and one class support vector machine (OSVM) (Manevitz & Yousef, 2002) have been adopted to solve the problem. These algorithms prove to be effective in dealing with traditional Web page classification problems. However, they may not be applicable for retrieval target page classification for the following reasons.

1. They are not designed for low dimensionality and high-density instance space, which are essential issues for retrieval target page classification (the number of query-independent features is usually small, but the number of instances is huge).
2. Several of these algorithms require the knowledge of positive instance proportion within the universal set, which is not available for retrieval target classification.
3. These algorithms are usually time-consuming so they are not suitable for the task of cleansing billions of pages required by search engines.

Unlike these algorithms, our data cleansing approach is based on the naive Bayes learning method (Mitchell, 1997, chap. 6), which is believed to be both effective and efficient for low-dimensional instance spaces. Besides, our method does not require prior knowledge of the original dataset. Finally, our method is the application of query-independent features from both inside a page and across pages, which involves more information than hyperlink analysis algorithms.

Features of Retrieval Target Pages

Types of Retrieval Target Page

According to Sullivan (Sullivan, 2003), commercial search engines deal with millions of user requests each day.

The huge number of user request produces a large variety in queries and corresponding retrieval target pages. However, these target pages are believed to share something in common such as their popularity and reliability. These common attributes result in the queryindependent features we present in Query-Independent Features of Retrieval Target Pages.

Before analyzing the differences between retrieval target pages and ordinary ones, we first draw a clear picture of what a retrieval target page is. On the basis of the query log analysis of Alta Vista, Broder (2002) and Rose and Levinson (2004) grouped Web search queries into three categories: navigational, informational, and transactional queries. The relationships among these three kinds of queries and their corresponding target pages are shown in Table 1.

This classification categorization is from a large-scale search engine survey. It is accepted by most Web search researchers and is adopted by the recent Text Retrieval Conference (TREC) Web tracks (Hawking & Craswell, 2002; Hawking & Craswell, 2003; Craswell & Hawking, 2004).

The major difference between navigational and informational/transactional type queries is whether the user has a fixed search target page or not. For navigational type queries, the user has a fixed search target and the search purpose is to reach a particular Web page. On the other hand, when the user submits an informational or transactional type query, he or she does not have a fixed search target page. Instead, the user has a more general search purpose and wants to get some information or service sources on a certain topic.

Navigational type query is related to two types of search tasks: home page finding (target page is a home page) and named page finding (target page is a particular page described by its name, called *named page*). The Web search task handling informational and transactional type queries is usually called *topic distillation* and its corresponding target page is called the *key resource*. To be a key resource, a page should be a good entry point to a Web site. This site should provide credible information on a certain topic. It should be principally devoted to the topic and not be part of a larger site that is also principally devoted to the topic (this definition is from Hawking & Craswell, 2003). According to this definition, most home pages belong to the category of key resources because almost all of them are entry pages that can meet the needs specified. Then we can classify the Web search target page (including home pages, named pages, and key resources)

TABLE 1.   Query types and their corresponding search tasks/target pages (based on Broder, 2002), in addition with search tasks and target page types (according to Craswell & Hawking, 2004).

Query type	Search task	Target page type	Proportion
Navigational	Named page finding Home page finding	Named page	About 20%
Informational/ transactional	Topic distillation	Key resource page	About 80%

into one of two categories, named page or key resource, as shown in Table 1.

There can be two approaches to the Web data cleansing task: reducing useless pages or locating retrieval target ones. Reducing useless pages may be difficult because a Web page can be useless for various reasons: redundancy, spam, false information, and so on. Therefore, it would be difficult to pick up all kinds of useless pages. And failing to reduce a single kind of useless pages may lead to the failure of the whole cleansing task. In this case, it will be a better idea to focus on finding high-quality pages since only retrieval target pages are regarded as important for IR-oriented Web data cleansing tasks. However, although the number of retrieval target types is much smaller than that of useless Web page types, there are also two types of retrieval target pages: navigational pages and informational pages. Should they be treated separately? What are the differences between retrieval target pages and ordinary pages? We answer these questions in terms of statistical analysis of large-scale Web page corpora.

Query-Independent Features of Retrieval Target Pages

Each Web page has several features that are independent of search engine users' requests (see Table 4 for examples). Most of these features are irrelevant to page content, such as in-link number, out-link number, PageRank value, and URL length; some features are content-correlated, such as encode information and length of in-link anchor text. All these features cannot be customized by search engine users so they are called *query-independent features*.

If we want to conduct data cleansing independently of user queries, it is natural to make use of these features because they are independent of user queries. However, there has been no research on the differences in query-independent features between retrieval target pages and ordinary pages as far as we know. In our research, we have found out that these features can tell the differences between retrieval target pages and ordinary pages according to our analysis of large-scale Web page corpora.

We analyzed two different Web page corpora: the .gov corpus[1], which is made up of 1.2 million English Web pages, and the sogou corpus, which contains 37 million Chinese Web pages.[2] Some characteristics of these two corpora are shown in Table 2.

.GOV is a TREC (http://trec.nist.gov/) test collection, which served in TREC Web tracks between 2003 and 2004. This corpus is well organized and many Web IR researches have been performed based on it. However, it was crawled 5 years ago and only pages in .gov domain were collected. This may be a limitation of the research that uses this corpus alone for experiments.

TABLE 2.   Statistics of the .GOV and SOGOU corpora.

	Number of pages	Language	Total size	Average doc size	Domain limit	Crawling time
.GOV	1,247,753	English	18.1 G	15.2 k	.gov	Early 2002
SOGOU	37,205,218	Chinese	558.0 G	15.0 k	No limit	Nov. 2005

In order to get a more recent and representative Web test collection for our data cleansing research, we collected the SOGOU corpus by randomly sampling pages from the Web page collection indexed by Sogou.com. The size of the SOGOU collection is about 4.3% of the whole Chinese Web, which contained about 870 million Web pages at the end of 2005 according to a China Internet Network Information Center (CNNIC) report.[3] It is not limited to a specific domain and thus is more practical than the .gov data set. Although the Chinese Web collection may be seen as a specific collection in an international context, most conclusions should not change much in a multilanguage environment if the query-independent features we selected are not specific to Chinese.

Generally speaking, .GOV was collected from .gov domain only, so the overall page quality is higher than that of SOGOU, which is constitutive of pages crawled from all domains. But the latter corpus is more recent and more representative.

We build retrieval target page sample sets for both corpora. These sets serve as positive examples in the learning process. The sample set for .GOV is selected from TREC2003–2004 Web track answers (Craswell & Hawking, 2004; Hawking & Craswell, 2003) and the sample set for SOGOU corpus is labeled by three assessors using pooling technology[4] (Hawking & Craswell, 2005). We use the following steps to get a sample of the retrieval target page set:

1. Collect a number of user requests that can represent a majority of user interests.
2. Crawl search results for these requests from several popular search engines and build a result pool for each request with these results.
3. Assess whether one result can be regarded as a retrieval target page for a given request and form a retrieval target page sample set.

In this way the quality and uniformity of retrieval target page sampling depend on whether the selected user requests can represent most user interests. TREC Web track offers several hundreds of queries collected from search engine logs or designed by assessors. Further, we collected 650 queries from SOGOU search logs to represent users' search requests in several popular fields, such as film/TV stars, songs, software, movies, novels, PC/TV games, people's

---

[1] Detail information is online at http://es.csiro.au/TRECWeb/govinfo.html.

[2] A simplified version is available online at http://www.sogou.com/labs/dl/t.html.

[3] This report is available online at http://www.cnnic.cn/download/2006/20060516.pdf. However, there are no editions in other language.

[4] The pool of possible retrieval target pages is built using five well-known Chinese search engines: Baidu.com, Google.com, Yisou.com, Sogou.com, and Zhongsou.com.

names, news topics, positions, and sports. With these queries, we built a retrieval target page sample set that contains 2,631 pages for .GOV and 48,930 pages for SOGOU. We used about half of these pages for testing the effectiveness of the data cleansing algorithm (1,732 pages for .GOV, 24,927 pages for SOGOU) and the others for algorithm training.

The training and test sets for SOGOU corpus are the largest retrieval target page set ever used by Web IR researchers to our knowledge. However, it is unavoidable that these sample sets can only cover a tiny part of potential user requests proposed to Web search engines. Fortunately, search user requests can be grouped in three categories and it is possible to organize a uniform sampling of these three kinds of queries. Our training sets in both SOGOU and .GOV corpora are designed to cover all three kinds of Web search queries, informational, navigational, and transactional. The fractions of these query types are set according to practical Web search environments (see Table 1) so our page sets are believed to be the most reliable retrieval target page sample sets ever used.

Besides, we only adopted query-independent features in our data cleansing algorithm. This means although retrieval target pages are collected through queries, there is no relationship between the query content and the retrieval target page attributes we found. There are enough retrieval target pages representing all query types. Therefore, the sampling can be regarded as a simulation of the practical users' requests in our data cleansing algorithm training process.

With analysis into the corpora and corresponding retrieval target sample sets, we found that retrieval target pages have totally different query-independent features from ordinary ones. The correlation values[5] between retrieval targets and ordinary pages in several query-independent features are shown in Figure 1 for .GOV corpus.

In Figure 1, we use five query-independent features to show the differences between retrieval target pages and ordinary pages. These features are Doc Length (number of words in a certain Web page), anchor length (number of words in in-link anchor text for a certain Web page), PageRank (obtained using the algorithm described by Brin & Page [1998]), Indegree (number of in-links), and Outdegree (number of out-links). We can make the following observations from the statistics shown in Figure 1:

1. Retrieval target pages and ordinary pages have different statistical distributions in values of query-independent features. Take PageRank, for example: the correlation value between named page and ordinary page is 0.07, which represents a lack of correlation.

---

[5] Correlation value is defined as $Cov(X,Y)/\sigma_x \cdot \sigma_y$, in which Cov(X, Y) represents the covariance value of arrays X and Y. It is used to describe the relationship between two or more variables. Correlation coefficients can range from -1.00 to $+1.00$. The value of -1.00 represents a perfect negative correlation while a value of $+1.00$ represents a perfect positive correlation. A value of 0.00 represents a lack of correlation.
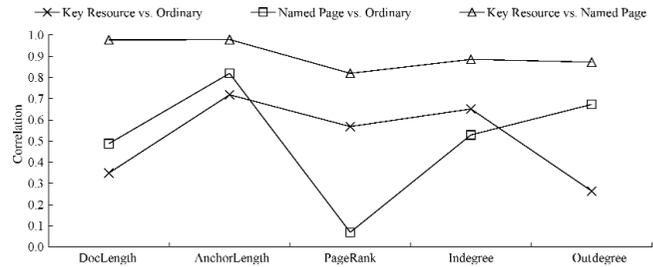
FIG. 1. Differences in qurey-independent feature distributions represented by correlation values. The category axis show query-independent features.

2. The two kinds of retrieval target pages, that is, named pages and key resource pages, are similar in these query-independent features. The correlation values of all five features between named pages and key resource pages are all above 0.8, indicating that these two kinds of pages are positively correlated. This means although these two kinds of retrieval target pages are from different search requests, we should treat retrieval target pages as a whole instead of separately in Web-IR oriented data cleansing research.

In order to find out how the retrieval target pages behave differently than ordinary pages, we look into the in-degree (the number of in-link count for a Web page) distribution of Web pages in both corpora. The statistical distribution of the in-degree value is shown in Figure 2. In-degree is analyzed because it is an important feature for Web IR-oriented research and one of the key factors in hyperlink structure analysis algorithms.
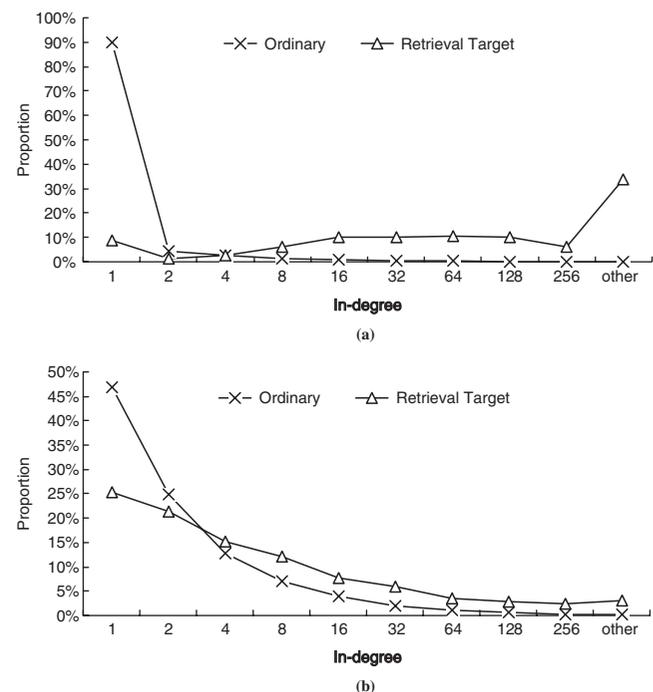

FIG. 2. Different In-degree distributions of retrieval target and ordinary pages in (a) SOGOU corpus and (b) .GOV corpus.

In Figure 2, most ordinary Web pages have a small number of in-links in both corpora. About 90% of Web pages in SOGOU and 50% of pages in .GOV have fewer than two in-links according to the statistics. However, retrieval target pages have much higher in-degrees: 90% retrieval targets in SOGOU and 75% in .GOV have more than two in-links. This can be explained by the fact that retrieval targets are welcomed not only by search users but also by other Web sites and pages. We also found that pages in .GOV generally had more in-links than those in SOGOU; therefore, .GOV pages are more popular. This difference is due to the different construction strategies of these two corpora. .GOV is designed to collect high-quality pages. Therefore, its collection is limited to .gov domain, in which Web page quality is higher than in the whole Web. Meanwhile, SOGOU is collected to build a representative Web collection for the whole Chinese Web so the loss in page quality can be foreseen.

Besides the hyperlink structure related features, we have several other sources of information to identify the differences, such as the DocLength feature, which is shown in Figure 1. Table 3 shows two other content-related features that can be used to separate retrieval target pages from ordinary ones.

According to Table 3, few retrieval targets have uniform resource locators (URLs) with question marks. It indicates that the information from dynamic Web pages (URL with a "?" always identify a dynamic page) is not as acceptable for users as the static Web pages. This may be caused by the low quality in forum or blog content, which is usually presented on dynamic pages. We can also see in Table 3 that the percentage of non-GBK (Chinese Internal Code Specification) encoded retrieval target page is very low. It may be explained by the fact that Sogou.com is a Chinese search engine and users mainly pay attention to Web pages written in simplified Chinese.

On the basis of the statistical analysis mentioned earlier, we found that retrieval target pages behave differently than ordinary ones in a number of query-independent features. These features are listed in Table 4 our learning-based data cleansing algorithm depends on them to cleanse Web data for information retrieval researches.

Learning-Based Web Data Cleansing Algorithm

There are already several studies into positive-example-based Web page classification according to the section High-Quality Web Page Classification. These works, such as the PEBL learning framework (Yu et al., 2004), improved traditional learning algorithms. Our previous work attempted to adopt these algorithms into Web data cleansing but with limited success. In 2004 (Liu, Zhang, & Ma, 2004), we succeeded in reducing 80% of .GOV pages with the ID3 decision tree algorithm. A majority of Web retrieval requests (informational/transactional type queries) also had better performance in the cleansed data set than in the whole corpus. However, decision tree learning requires knowledge about the proportion of positive examples in the corpus, which is difficult to obtain. In 2005 (Liu, Wang, Zhang, & Ma, 2005), we applied the K-means algorithm to prevent the positive proportion problem and selected about half of .GOV pages to gain retrieval performance similar to that with the whole corpus with a general purpose Web search test set. However, the K-means based algorithm suffers the problem of low time efficiency and is not suitable for practical application.

In this study, we adopt the naive Bayesian learning algorithm (Mitchell, 1997, chap. 6) to solve the retrieval target page classification problem because it is among the most practical and effective approaches for the problem of learning to classify text documents or Web pages. It can also provide explicit probabilities of whether a Web page is a retrieval target page, which can be potentially adopted in result ranking of search engines. We can also estimate the quality of a Web page according to the probabilities.

For the problem of retrieval target page classification, we consider two cases, the case when classification is based on only one feature and the case when multiple features are involved.

Case 1: Single feature analysis. If we adopt only one query-independent feature A, the probability that a Web page p with feature A is a retrieval target page can be denoted by

$$P(p \in Target\ page | p\ has\ feature\ A). \qquad (0)$$

We can use the Bayes theorem to rewrite this expression as

$$P(p \in Target\ page | p\ has\ feature\ A)$$

$$= \frac{P(p\ has\ feature\ A | p \in Target\ page)}{P(p\ has\ feature\ A)}$$

$$\times P(p \in Target\ page) \qquad (1)$$

In equation (1), $P(p \in Target\ page)$ is the proportion of retrieval target pages in the whole page set. As mentioned, this proportion is difficult to estimate in many cases, including our problem of retrieval target page classification. However, if we just compare the values of $P(p \in Target\ page | p\ has\ feature\ A)$ in a given Web page corpus, $P(p \in Target\ page)$ can be regarded as a constant value and would not affect the comparative results. So in a fixed corpus such as .GOV/SOGOU, we can rewrite equation (1) as

TABLE 3. Content-related features of retrieval target pages and ordinary pages.

	Ordinary page	Retrieval target page
URL contains a question mark ("?")	13.06%	1.87%
Encode is not GBK[a]	14.04%	1.39%

[a] GBK represents Chinese Internal Code Specification. It is widely adopted by Chinese Web sites in mainland China.

TABLE 4.   Query-independent features applied in our cleansing experiments.

	Features	Explanation
Content-related features	DocLength	Number of words in a Web page
	AnchorLength	Number of words in a Web page's in-link anchor text
	URLLength	Number of dashes "/" in a Web page's URL
	PageSize	Storage size of a Web page
	CopyNumber	Number of mirror copies of a Web page
	URLformat	Whether a URL contains a question mark
	Encode	Whether the encode of a Web page is GBK
Hyperlink structure-related features	Outdegree	Number of out-links of a Web page
	Indegree	Number of in-links of a Web page
	PageRank	PageRank value calculated according to algorithm (Brin & Page, 1998)
	In-Site-Outdegree	Number of links from a Web page to other pages in the same site

$P(p \in Target\ page | p\ has\ feature\ A)$

$$\propto \frac{P(p\ has\ feature\ A | p \in Target\ page)}{P(p\ has\ feature\ A)} \quad (2)$$

Now consider the terms in equation (2); $P(p\ has\ feature\ A | p \in Target\ page)$ can be estimated using the proportion of A-featured pages in the retrieval target page set. $P(p\ has\ feature\ A)$ equals the proportion of the pages with feature A in a given corpus. Here we obtain

$$\frac{P(p\ has\ feature\ A | p \in Target\ page)}{P(p\ has\ feature\ A)}$$

$$= \frac{\# P(p\ has\ feature\ A \cap p \in Target\ page)}{\# (Target\ page)} \bigg/ \frac{\# (p\ has\ feature\ A)}{\# (CORPUS)} \quad (3)$$

If the user query set is large enough to represent most user interests, the sampling of retrieval target page can be regarded as an approximately uniform process. Therefore, we can rewrite the numerator of (3) as

$$\frac{\# P(p\ has\ feature\ A \cap p \in Target\ page)}{\# (Target\ page)} =$$

$$\frac{\# P(p\ has\ feature\ A \cap p \in Target\ page\ sample\ set)}{\# (Target\ page\ sample\ set)} \quad (4)$$

Substituting expressions (3) and (4) into (2), we obtain

$$P(p \in Target\ page | p\ has\ feature\ A) \propto$$

$$\frac{\# P(p\ has\ feature\ A \cap p \in Target\ page\ sample\ set)}{\# (Target\ page\ sample\ set)} \bigg/ \frac{\# (p\ has\ feature\ A)}{\# (CORPUS)} \quad (5)$$

Since all terms in (5) can be obtained by statistical analysis on a Web page corpus, we can calculate the probability

of being a retrieval target for each page according to this equation.

Case 2: Multiple feature analysis. If we use more than one feature to classify retrieval target pages, the naive Bayes theorem assumes that the following equation holds:

$$P(p\ has\ feature\ A_1, A_2, \ldots, A_n | p \in Target\ page)$$

$$= \prod_{i=1}^{n} P(p\ has\ feature\ A_i | p \in target\ page) \quad (6)$$

For the problem of page classification with query-independent features, we further found that the following equation also approximately holds according to Table 5.

$$P(p\ has\ feature\ A_1, A_2, \ldots, A_n)$$

$$= \prod_{i=1}^{n} P(p\ has\ feature\ A_i) \quad (7)$$

This means for the features in Table 5, the attribute values adopted in the retrieval target page classification process are independent as well as conditionally independent given the target value.

The correlation values in Table 5 show that these features are approximately independent of one another. This may be explained by the fact that these features are obtained from different information sources and thus have little chance affecting one another. In other words, the following equations hold approximately for the retrieval target page classification task according to the naive Bayes assumption and our statistical analysis:

$$P(p \in Target\ page | p\ has\ feature\ A_1, A_2, \ldots, A_n)$$

$$= \frac{P(p\ has\ feature\ A_1, A_2, \ldots, A_n | p \in target\ page) P(p \in target\ page)}{P(p\ has\ feature\ A_1, A_2, \ldots, A_n)}$$

$$\approx \prod_{i=1}^{n} \frac{P(p\ has\ feature\ A_i | p \in target\ page) P(p \in target\ page)}{P(p\ has\ feature\ A_i)}$$

$$= \prod_{i=1}^{n} P(p \in Target\ page | p\ has\ feature\ A_i) \quad (8)$$

TABLE 5.   Correlation values between query-independent features of Web page.

	URL Format	Encode	PageRank	Cluster	DocLength	URL Length	Indegree
URLformat	1.00	0.15	0.15	0.01	0.04	0.10	0.00
Encode		1.00	0.20	0.00	0.06	0.30	0.00
PageRank			1.00	0.01	0.06	0.03	0.05
CopyNumber				1.00	0.01	0.10	0.00
DocLength					1.00	0.04	0.00
URLLength						1.00	0.02
Indegree							1.00

If we substitute (5) into (8), we can get the following equation, which is fit for multifeature cases:

$$P(\, p \in Target\ page \,|\, p\ has\ feature\ A_1, A_2, \ldots, A_n)$$

$$\propto \prod_{i=1}^{n}\left(\frac{\#\ (p\ has\ feature\ A_i \cap p \in \ Target\ page\ sample\ set)}{\neq\ (Target\ page\ sample\ set)}\middle/ \frac{\#\ (p\ has\ feature\ A_i)}{\#\ (CORPUS)}\right) \quad (9)$$

According to this equation, the probability of a Web page's being a retrieval target can be calculated with information from the Web corpus and its corresponding retrieval target sample set. As mentioned previously, we construct retrieval target sample sets for .GOV/SOGOU corpus and we select several query-independent features (shown in Table 5). Therefore, it is possible for us to use the algorithm to accomplish the classification task.

Experimental Results and Discussions

Evaluation Methods

To our knowledge, there has been little research on the evaluation of IR-oriented Web data cleansing. In our previous works (Liu et al., 2004; Liu et al., 2005), we chose the retrieval target page recall rate for a cleansed page set as the evaluation metric. If the cleansed size is significantly smaller than that of the original set while retrieval target recall is above a threshold T, we regard the cleansing method as an effective one. T is set according to practical application requirements. Although this simple method can prove the effectiveness of a certain method, it has difficulties in determining which cleansing method has better performance. There is a trade-off between cleansed corpus size and high-quality page recall. For example, one can improve recall simply by cleansing fewer pages, but doing so can increase the cleansed set size. This means we should take both factors into consideration while evaluating data cleansing methods.

In order to solve this problem of IR-oriented data cleansing evaluation, we proposed a new metric, called *high-quality page average recall* (*AR*). When Web pages in a certain corpus are ranked using a certain data cleansing method,

average recall of high-quality pages is the mean of the recall scores after each page counted.

$$AR = \sum_{i=1}^{\#(Original\ Set)} Recall(i) \middle/ \# \,(Original\ Set) \quad (10)$$

Similar to the famous IR evaluation metric average precision (*AP*), *AR* is a summary measure of a ranked page list. *AR* can also be calculated by averaging the high-quality page recall values at various points of cleansed set size because the following equation holds:

$$AR = \sum_{i=1}^{\#(Original\ Set)} Recall(i) \middle/ \# \,(Original\ Set)$$

$$= \int_{0}^{1} Recall(s)ds, \quad s = \#(Clensed\ Set)/\#(Original\ Set) \quad (11)$$

This means *AR* value can be calculated with the area under the *Size-Recall* curves. A set of such curves are shown in Figure 3. The category axis is the percentage of retained Web pages in the original set after data cleansing. The value axis is the percentage of retained high-quality pages after data cleansing.

In Figure 3, each curve shows the performance of a certain kind of data cleansing algorithm. If we use $Area(C_i)$ to represent the area under curve C_i, then we have

$$Area(C1) > Area(C2) > Area(C3). \quad (12)$$

This means the following equation also holds according to equation 11:

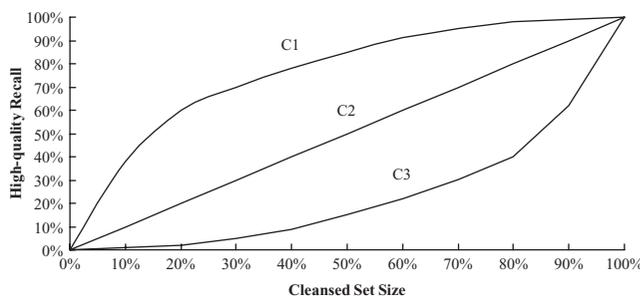$$AR(C1) > AR(C2) > AR(C3). \quad (13)$$

FIG. 3.   AR calculation with cleansed size-recall curves.

We can get the same conclusion by analyzing these three algorithms represented by the curves. For example, $C1$ corresponds to an effective data cleansing algorithm because with a small cleansed set size (such as 10%), it can cover a more than average number of high-quality pages (about 35%). $C2$ shows a random sampling method of Web pages because the high-quality page recall increases linearly with the cleansed set size and its AR value is equal to $1/2$. $C3$ can be regarded as showing a low-quality page identification method, because it retains almost no high-quality pages when cleansed set size is small (below 30%).

From equation (11) and Figure 3 we can see that AR is related to both cleansed set size and high-quality recall values. This means AR contains both cleansed-size-oriented and recall-oriented aspects and is suitable for the task of Web page data cleansing evaluation.

Data Cleansing Experimental Results

As mentioned in Query-Independent Features of Retrieval Target Pages, we keep about half of the retrieval target sample pages for testing our data cleansing algorithm. The algorithm's effectiveness is examined by the following means: First, we see whether this algorithm can pick up high-quality pages. We then compare the performance of the query-independent features applied in the algorithm to find out which one plays the most important role in data cleansing. Retrieval performance of the cleansed Web set is also examined. At last, we will check out whether this algorithm can separate low quality or even spam pages as well.

High-quality page classification. Figure shows the distribution of the probabilities of being retrieval targets for pages in .GOV corpus.

In Figure 4, the X axis is the probability of being a key resource page, which belongs to one kind of retrieval target pages; the Y axis is the probability of being a named page, which belongs to the other kind of retrieval target pages. As mentioned in Types of Retrieval Target Page, retrieval target pages can be grouped into these two kinds of pages.

We can see that pages in the retrieval target test set (marked using circle ? and triangle ?) have higher probabilities according to our 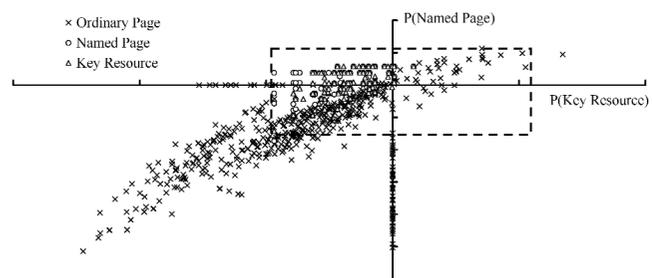algorithm (almost all these pages are located in the dashed rectangle). Meanwhile, part of ordinary pages (marked with $\times$) is not highly evaluated by our algorithm (outside the dashed rectangle). We can also find several ordinary pages mixed with retrieval target pages in the dashed rectangle. This can be explained by the fact that we cannot include all high-quality pages in our test set, and pages with high probabilities may also be high-quality pages.

Table 6 shows the cleansed corpus size and its corresponding retrieval target page recall for .GOV and SOGOU corpora.

From the statistics in Table 6, our data cleansing algorithm can retain most retrieval target pages while significantly reducing corpus size. A large proportion (95.53% in .GOV and 92.73% in SOGOU) of retrieval target pages remain in the cleansed corpus. However, there is a difference in the cleansed sizes when the algorithm is applied to different corpora. Only less than 5% pages are regarded as important by the algorithm for the SOGOU corpus while 52% pages are retained for .GOV. It may be explained by the fact that the data quality of .GOV corpus is much higher than that of SOGOU. .GOV was crawled in 2002 and its pages are limited to .gov domain, whose content is more reliable than that of the whole Web. SOGOU corpus was collected in 2005; it has many more spam and low-quality pages appearing on the Web and the crawled pages are not limited to a certain domain. It is reasonable to find a larger proportion of high-quality pages in .GOV than in SOGOU. However, compared with .GOV corpus, SOGOU corpus is closer to the practical application environment for a Web search engine.

According to the experimental results in the test set, it is possible to satisfy more than 90% of user requests with a small number of pages in the corpus, and these pages can be located query-independently using our data cleansing algorithm. Web search engines may be able to adopt hierarchy structure in their data index. The cleansed page set can be placed into a high-level, frequently used, quickly accessible index, which can meet most users' requests. The other pages can be placed into low-level indexes, because they are not so important for users and can meet the rest of search needs that cannot be handled in the high-level index.

Effectiveness of query-independent features. Figure 5 shows the *Size-Recall* curve of our data cleansing result in SOGOU corpus. According to the definition of high-quality page AR in Evaluation Methods, the AR value for our algorithm

FIG. 4. Distribution of the probabilities of being retrieval target pages in .GOV corpus.

TABLE 6. Cleansed Corpus Size and Corresponding Target Page Recall (the proportion of retained target page after data cleansing) using Data Cleansing Algorithm.

	Cleansed Corpus Size (Percentage of original set)	Retrieval Target Page recall (Training set)	Retrieval Target Page recall (Test set)
.GOV	52.00%	95.53%	93.57%
SOGOU	4.96%	92.73%	92.37%
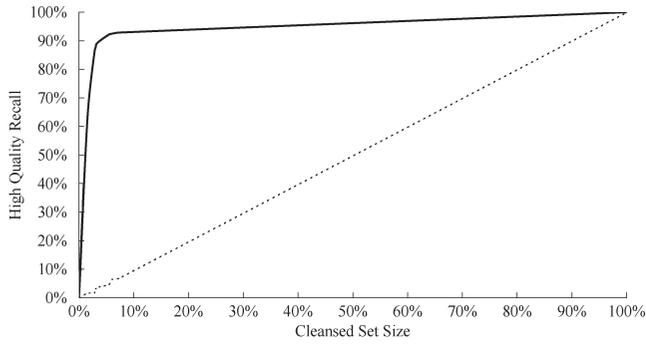
FIG. 5. Cleansed Size–Recall curve for our data cleansing algorithm in SOGOU corpus.

FIG. 6. Effectiveness of PageRank and other features in data cleansing.

is 0.9064. This means that the algorithm is effective because the upper bound for *AR* value is 1.0000 and a random sampling algorithm's *AR* is 0.5000.

Furthermore, we want to find out which query-independent feature is the most important in our algorithm. We want to answer the question, Does this cleansing result arise from one or two "key" features or from a "combined" effort? If one or two features can make the algorithm work, it is not necessary to use a learning mechanism in the algorithm.

In order to answer this question, we test *AR* values of our data cleansing algorithm, each time with one single feature removed. The experimental results are shown in Table 7.

We can see from Table 7 that when PageRank or Indegree is dropped out, the *AR* value drops most. This means these two features play important roles in our data cleansing algorithm.

However, our further experimental results in Figure 6 suggest that the other features should not be treated as unimportant. We can see that the performance becomes worse when only PageRank is adopted to rank pages in the data cleansing process. A data cleansing algorithm that combines other features can gain better performance. This is in accord with the conclusion of Henzinger (Henzinger et al., 2003) that a better page quality estimation algorithm should involve other sources of information rather than using the hyperlink structure analysis alone. Although the features have different abilities in identifying high-quality pages, the cleansing performance results from a joint effort of all features instead of from one or two "key features."

Retrieval performance of the cleansed page set.   Our data cleansing algorithm aims at reducing the index size of

Table 7. Effectiveness of the query-independent features in the data cleansing algorithm.

The Feature which is dropped out	AR
URL Format	0.9037
Encode	0.9032
PageRank	0.8756
Cluster	0.9012
DocLength	0.9031
URL Length	0.8984
Indegree	0.8860
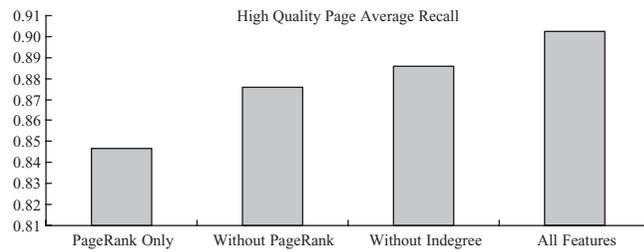
Web search engines and maintaining or improving retrieval performance at the same time. In Effectiveness of Query-Independent Features, we have shown that the cleansed set contains most high-quality pages although its size is significantly smaller than that of the original set. It is also necessary to test whether the cleansed page set really helps improve result ranking of the retrieval systems, because the effectiveness of the cleansing algorithm degrades if the algorithm reduces index size at the cost of sacrificing retrieval performance.

In order to find out the effect of data cleansing on retrieval performance, we built three cleansed page sets from .GOV corpus using our cleansing method, with about 1/4, 1/2, and 3/4 amounts of pages from the original set, respectively. TREC has developed several retrieval tasks on .GOV, so it is possible for us to find reliable test sets for retrieval experiments. We chose two search tasks from TREC 2003 and 2004 Web tracks because they are designed to be simulations of the practical Web search environment.

The TREC 2003 Web track task was focused on locating key resource pages for topic distillation queries. This query type covers about 80% of Web search requests according to Table 1. The TREC 2004 task was designed to cover all types of Web search requests; its search request set includes 1/3 home page finding queries, 1/3 named page finding queries, and 1/3 topic distillation queries. For both retrieval tasks, the retrieval performances are evaluated by mean average precision (MAP) and B-pref (Buckley & Voorhees, 2004), which are two of the most frequently used metrics in Web search research (they are also adopted by TREC as the major metrics).

Retrieval experiment results are shown in Figure 7; we tested how the retrieval performance varies with respect to the cleansed corpus size. For the TREC 2003 task, the cleansed set gets best performance while it contains about 1/4 pages of the original .GOV corpus. However, when 3/4 pages are retained, the cleansed set performs best for the TREC 2004 task. This difference can be explained by the different task settings of TREC 2003 and 2004. For TREC 2003, the topic distillation task tries to locate reliable pages (called *key resource pages*) for certain topics, so its search target pages are more likely to be left in our cleansed set. A large fraction of TREC 2004 queries are of home page or named page finding types. These kinds of pages, especially named pages, are likely to be ordinary pages (such as a certain news page) and are not highly evaluated by our cleansing algorithm. So we have to discard fewer pages to meet such navigational type search requests. This means for different
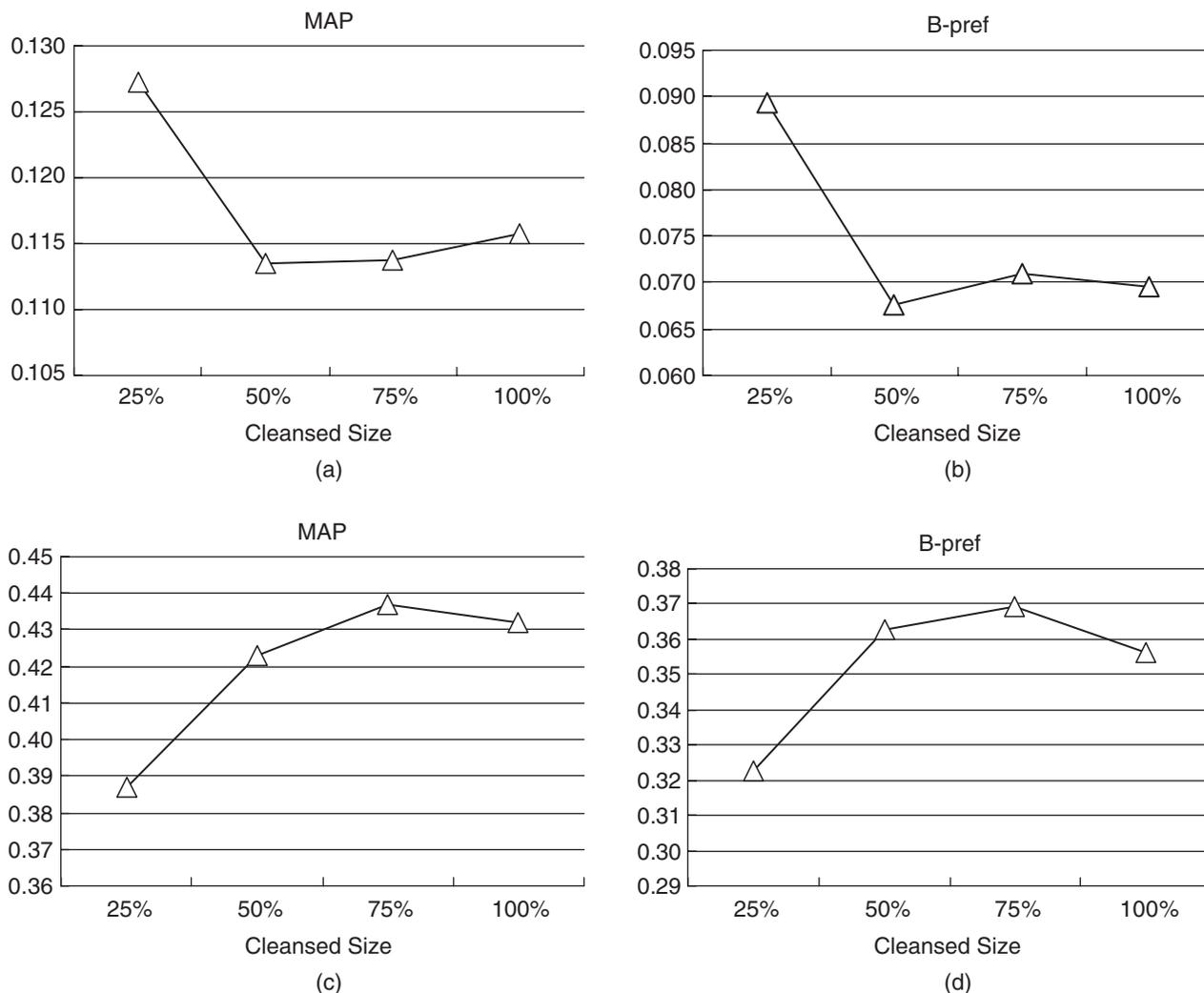
FIG. 7.   Retrieval experiment results on the cleansed .GOV corpus. (a), (b): retrieval using TREC 2003 data set; (c), (d): retrieval using TREC 2004 data set.

search tasks, we should choose different parameters for cleansed set size to get the best retrieval performance. For the topic distillation task, fewer pages are needed in the cleansed set; for navigational type searches, a larger cleansed set should be used.

Although the retrieval performance varies with different retrieval tasks, it is interesting to find out that for each retrieval task the cleansed set outperforms the original set with both smaller size and better performance. For TREC 2003, the cleansed set contains only 25% of pages of the original set, while the algorithm helps improve MAP by 9.93% and B-pref by 28.3%. For TREC 2004, the cleansed set had better retrieval performance (1.25% using MAP, 3.65% using B-pref) with 25% of pages reduced.

In these experiments, data cleansing affects retrieval performance in two aspects:

1. Cleansed size. By data cleansing, it is possible to reduce low-quality pages that may be ranked higher than high-quality ones by retrieval algorithms. For example, for the topic "*The White House President Bush's cabinet*"

(TREC 2004, topic 26), the file "G45–22-1096484" in .GOV is ranked in third place by BM2500 weighting algorithm (Robertson, Walker, Hancock-Beaulieu, & Gatford, 1994) according to our experimental results. This page is from the U.S. Embassy in Tokyo and mentions Bush's cabinet, but it does not give any detailed information on that topic. In the cleansing process, this page is reduced because it is not popular and contains no important information. By this means other important pages related to this topic can be ranked higher. Therefore, MAP for this topic improves from 0.143 to 0.500 after data cleansing.

2. Target page recall. Retrieval target pages, especially those for navigational type searches, may also be reduced by our data cleansing algorithm. Such loss is not huge on average (less than 10% according to Table 6), but it may result in search failure for some particular cases if all search target pages for a certain topic are discarded by the cleansing algorithm. With techniques such as hierarchy indexing system (discussed in Retrieval Performance of the Cleansed Page Set), it is possible for our cleansing algorithm to be effective for a majority proportion of Web search requests while not affecting the others much.

The preceding two factors affect the retrieval performance at the same time. If search target pages tend to be highly evaluated by our algorithms (such as informational/transactional type target pages), the cleansed size is the key factor. Then a smaller cleansed size leads to higher performance just as occurred in TREC 2003 experiments (the smallest cleansed set has the best MAP). If search target pages are usually obtain a low evaluation by our cleansing algorithm (such as navigational type target pages), retrieval target page recall plays a more important role. In this case, a smaller size often causes lower recall and thus produces loss in performance (see TREC 2004 results). However, this does not mean a failure of our algorithm, because a relatively large cleansed set may still outperform the original set because the "cleansed size" factor also works.

If we take a closer look at the experiment results of TREC 2004 in Figure 8, we can also see how the two factors affect retrieval performance.

In Figure 8(a), informational type searches have a similar retrieval performance distribution with TREC 2003 because they share the same type of search task. Therefore, cleansed set size is the key factor that leads to high performance of the smallest cleansed set. In Figure 8(b), we obtain the highest MAP with 75% pages retained. This means retrieval target page recall plays an important role along with cleansed set size.

From the preceding experimental results we can conclude that our cleansing algorithm can improve the overall retrieval performance. Effectiveness varies with specific retrieval tasks and is heavily dependent on two key factors: cleansed size and retrieval target page recall. These two factors are also the ones we want to evaluate in our average recall (*AR*) measure, so retrieval effectiveness and data cleansing success can to a certain extent be judged together with this measure.

Reducing low-quality and spam pages.   Because the learning process is based on analysis into high-quality page samples, our data cleansing algorithm is designed to cleanse Web data by picking up high-quality pages (retrieval target pages) instead of reducing low-quality or spam pages. However, experimental results in Figure 9 show that our algorithm can also reduce a part of these harmful pages.

For SOGOU corpus, we have three assessors pick up a number of low-quality and spam pages according to the following rules: low-quality pages are pages that offer little or biased information and are not as useful as retrieval target pages, while spam pages are those making use of tricks in hyperlink structure or page content to get higher rankings than they should have in Web search results; 3,272 low-quality pages are annotated together with 120 spam pages by the assessors.

According to Figure 9, using PageRank can reduce more spam pages than using Indegree, while the latter feature is more effective in separating low-quality pages. However, the data cleansing algorithm that makes use of both features can reduce 30% of spam pages as well as 15.26% of low-quality pages. This means the algorithm can fully exploit the advantages of both PageRank and Indegree in reducing spam and low-quality pages.

Possibilities of Applying Data Cleansing Method in Web Search Engines

As shown in Data Cleansing Experimental Results, our data cleansing method is effective for SOGOU and .GOV corpus in significantly reducing collection size as well as retaining a large proportion of high-quality pages. Cleansed .GOV had better retrieval performance than the original corpus while only about 50% of pages needed to be indexed. However, there are several major differences between these corpus-based experiments and practical Web search applications, and each of them may result in failure of the lab-oriented methods.

First, the Web page corpus only covers a small (usually tiny) part of the World Wide Web, so it is not always true that the methods that work well in corpora are also effective for
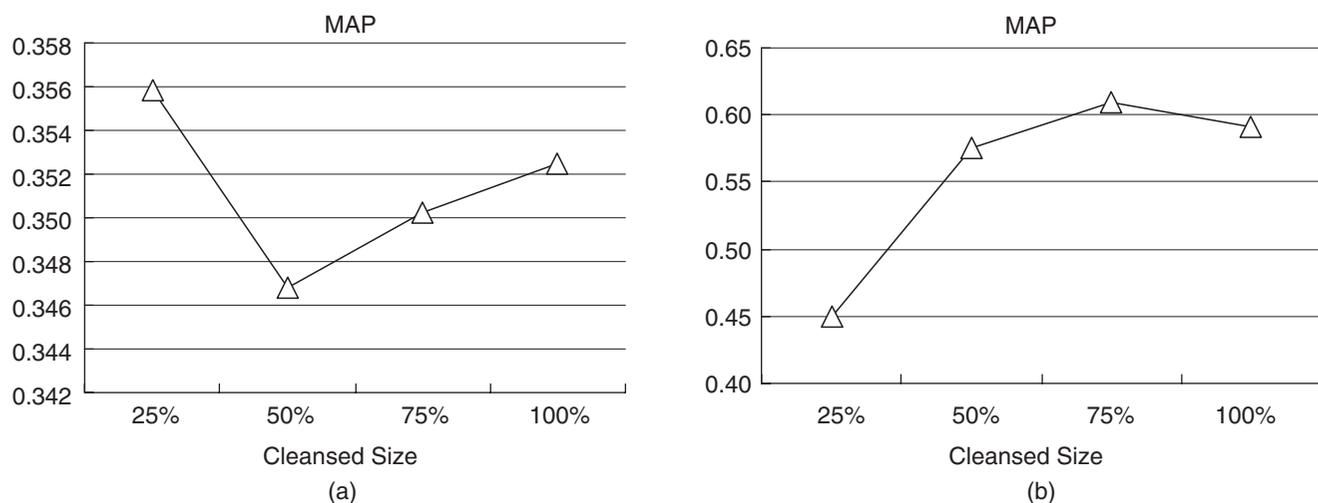

FIG. 8.   Detail retrieval experiment results for TREC 2004 tasks. (a): informational/transactional type searches; (b): navigational type searches.
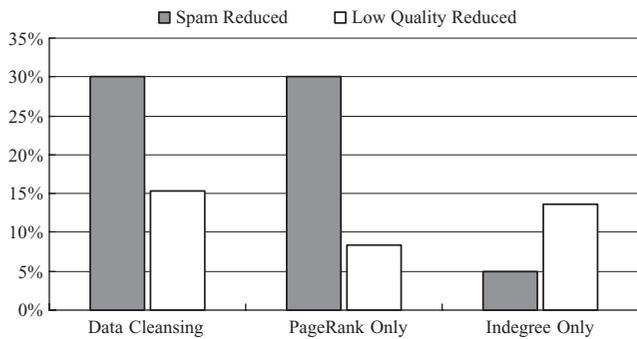
FIG. 9. Effectiveness of reducing low-quality and spam pages using data cleansing.

the whole Web collection. Second, the features adopted in experimental environments may not be so easily obtained for practical applications. For example, it is much more difficult to obtain hyperlink-related features in a practical Web environment because the link graph is much more complex than that of a Web page corpus. Last, the methods that work well for corpora may not be applicable for practical Web search applications if they are too time-consuming or involve too many parameters to be tuned.

Our data cleansing method is effective in the corpus-based experiments, but its effectiveness in practical Web search applications relies on the following aspects.

Reliability of the data cleansing performance. Uniform sampling of Web pages is a challenge for many Web application studies, so the reliability of our data cleansing method relies on the degree to which our experiment settings are similar to the practical Web environment. To our knowledge, our method is built on the largest Web page corpus and corresponding retrieval target page training set ever used by Web IR researchers. The SOGOU corpus contains 37 million Web pages and covers almost 5% of pages of the whole Chinese Web collection. Experiments on .GOV also prove that data cleansing can be effective with various language environments or Web page sizes.

Although the experiment settings are different from a practical Web search application environment, we believe that they are currently the closest to that environment. Parameter tuning and feature selection may vary with different application requirements, but the basic idea is reliable and generally applicable for real-world Web search engines.

Availability of the query-independent features applied in the data cleansing method. The differences between retrieval target pages and ordinary pages exist because of the wide spread of unimportant, outdated, contradictory, and spam data in the WWW. Once these kinds of low-quality data exist on the Web, our learning-based query-independent data cleansing method can be a reliable and effective way to solve this problem.

Query-independent features selected in our cleansing algorithm are collected from a practical search engine's operation process. They are all adopted in the preprocessing, indexing, or ranking stage of a search engine. PageRank feature is important in result ranking, number of copies is a key factor for result clustering, and these features should be collected by search engines even without the cleansing process. So the feature collection process does not require additional workload for the search engine system.

Applicability of the data cleansing method. The applicability of the cleansing method depends on two factors: efficiency and effectiveness. Data cleansing should be efficient so that billions of Web pages can be evaluated before they are placed in data indexes. The algorithm should also be effective in reducing both the index size and the reaction time while retaining high retrieval performance.

The time complexity of the cleansing algorithm depends on the learning algorithm we selected. Because naive Bayes learning is adopted in our algorithm, the cleansing process (judging whether the page should be cleansed or not) is with liner complexity. So its efficiency fits well for Web search applications.

Effectiveness of the cleansing algorithm is decided by several aspects. According to Retrieval Performance of the Cleansed Page Set, retrieval performance depends on both the cleansed set size and the retrieval target page recall of the cleansed set. A possible way to reduce the reaction time and obtain high performance for search engines is to use a hierarchical indexing structure such as the one shown in Figure 10.

As shown in Figure 10, a large number of user requests can be satisfied with a frequently visited page set. This page set is from the data cleansing results of the original set and only contains a small proportion of pages in the whole indexable collection. Experiments with the SOGOU corpus show that the cleansed set may contain less than 5% of pages but fulfill a large fraction of user requests. When the cleansed set
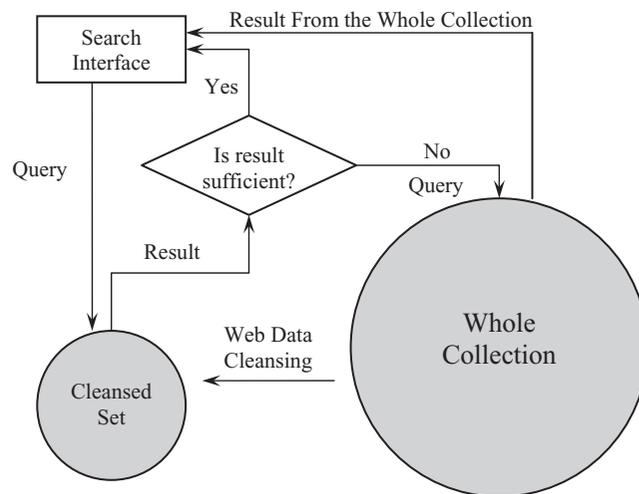

FIG. 10. A hierarchical indexing architecture for search engines based on data cleansing.

fails to give sufficient answers for a certain query (such as a named page type query whose retrieval target is reduced in data cleansing), the system turns to the whole collection and returns retrieval results to the search users.

Conclusions and Future Work

We have shown that by using a Web data cleansing algorithm, it is possible to reduce the Web data size significantly while retaining most high-quality pages. Our algorithm, based on analysis into large-scale Web corpora, exploits the differences between high-quality pages and ordinary pages on the Web. We combine machine learning techniques and descriptive analysis on the query-independent features of retrieval target pages to provide a better understanding of the relationship between user requests and the index structure of Web IR tools.

In the near future, we hope to extend this work's framework to include other applications such as low-quality page reduction and personalized Web search. We also plan to work on a hierarchical storage model for Web IR tools according to our findings in this paper.

Acknowledgments

References

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In Proceedings of the Seventh World Wide Web Conference (pp. 107–117). Brisbane. Australia: Elsevier Science Publishers B.V.

Broder, A. (2002). A taxonomy of Web search. SIGIR Forum, 36(2), 3–10.

Buckley, C., & Voorhees, E.M. (2004). Retrieval evaluation with incomplete information. In Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 25–32). Sheffield, England: ACM Press.

Buttler, D., Liu, L., & Pu, C. (2001). A fully automated object extraction system for the World Wide Web. In Proceedings of the International Conference on Distributed Computing Systems (pp. 361). Washington, DC: IEEE Computer Society.

Buyukkokten, O., Garcia-Molina, H., & Paepcke, A. (2001). Accordion summarization for end-game browsing on PDAs and cellular phones. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 213–220). New York: ACM Press.

Cai, D., He, X., Wen, J., & Ma, W. (2004). Block-level link analysis. Microsoft Technical Report MSR-TR-2004-50. Retrieved June 17, 2005, from ftp://ftp.research.microsoft.com/pub/tr/TR-2004–50.pdf

Cai, D., Yu, S., Wen, J., & Ma, W. (2003). VIPS: A vision-based page segmentation algorithm. Microsoft Technical Report MSR-TR-2003-79. Retrieved June 17, 2005, from ftp://ftp.research.microsoft.com/pub/tr/tr-2003–79.pdf

Craswell, N., & Hawking, D. (2004). Overview of the TREC 2003 Web track. In E.M. Voorhees and Lori P. Buckland (Eds.), NIST Special Publication 500-261: The 13th Text REtrieval Conference (TREC 2004). Washington, DC: Department of Commerce and National Institute of Standards and Technology.

Craswell, N., Hawking, D., & Robertson, S. (2001). Effective site finding using link anchor information. In Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 250–257). New York: ACM Press.

Denis, F. (1998). PAC Learning from Positive Statistical Queries. In M.M. Richter, C.H. Smith, R. Wiehagen, & T. Zeugmann (Eds.), Proceedings of the Ninth International Conference on Algorithmic Learning Theory: Lecture Notes in Computer Science (Vol. 1501, pp. 112–126). London: Springer-Verlag, 1998.

Embley, D.W., Jiang, Y., & Ng, Y.-K. (1999). Record-boundary discovery in Web documents. In Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data (pp. 467–478). Philadelphia: ACM Press.

Hawking, D., & Craswell, N. (2002). Overview of the TREC 2002 Web track. In E.M. Voorhees and Lori P. Buckland (Eds.), NIST Special Publication 500-251: The 11th Text REtrieval Conference (TREC 2002). Washington, DC: Department of Commerce and National Institute of Standards and Technology.

Hawking, D., & Craswell, N. (2003). Overview of the TREC 2003 Web track. In E.M. Voorhees and Lori P. Buckland (Eds.), NIST Special Publication 500-255: The 12th Text REtrieval Conference (TREC 2003) (pp.78–92). Washington, DC: Department of Commerce and National Institute of Standards and Technology.

Hawking, D., & Craswell, N. (2005). Very large-scale retrieval and Web search. In Ellen Voorhees & Donna Harman (Eds.), TREC: Experiment and evaluation in information retrieval. Cambridge, MA: MIT Press.

Hedger, J. (2005). Google takes backhanded bow out of size war with Yahoo. Retrieved October 17, 2005, from http://www.searchenginejournal.com/?p=2277

Henzinger, M.R., Motwani, R., & Silverstein, C. (2003). Challenges in Web search engines. In Proceedings of the 18th International Joint Conference on Artificial Intelligence (pp. 1573–1579). San Francisco: Morgan Kaufmann.

Kaasinen, E., Aaltonen, M., Kolari, J., Melakoski, S., & Laakko, T. (2000). Two approaches to bringing Internet services to WAP devices. Computer Networks: The International Journal of Computer and Telecommunications Networking, 33(6), 231–246.

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

Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (2006). Core algorithms in the CLEVER system. ACM Transactions on Internet Technology, 6(2), 131–152.

Liu, Y., Zhang, M., & Ma, S. (2004). Effective topic distillation with key resource pre-selection. Lecture Notes in Computer Science, 3411, 129–140.

Liu, Y., Wang, C., Zhang, M., & Ma, S. (2005). Web data cleansing for information retrieval using key resource page selection. Special Interest Tracks and Posters of the 14th International Conference on World Wide Web (pp. 1136–1137). New York: ACM Press.

Lyman, P., & Varian, H.R. (2003). How much information 2003? Retrieved June 18, 2005, from http://www.sims.berkeley.edu/how-much-info-2003

Manevitz, L.M., & Yousef, M. (2002). One-class SVMs for document classification. Journal of Machine Learning Research, 2, 139–154.

Mitchell, T. (1997). Bayesian learning. In Machine learning. New York: McGraw-Hill Education.

Nigam, K., McCallum, A.K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. Machine Learning, 39(2–3): 103–134.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1998). The PageRank citation ranking: Bringing order to the Web. Retrieved October 17, 2005, from http://citeseer.ist.psu.edu/page98pagerank.html.

Robertson, S.E., Walker, S., Hancock-Beaulieu, M.M., & Gatford, M. (1994). Okapi at TREC-3. NIST Special Publication 500-225: Overview of the Third Text REtrieval Conference (TREC-3) (pp. 109–126). Washington, DC: Department of Commerce and National Institute of Standards and Technology.

Rose, D.E., & Levinson, D. (2004). Understanding user goals in Web search. In Proceedings of the 13th international Conference on World Wide Web (WWW '04) (pp. 13–19). New York: ACM Press.

Song, R., Liu, H., Wen, J., & Ma, W. (2004). Learning block importance models for Web pages. In Proceedings of the 13th International Conference on World Wide Web (pp. 203–211). New York: ACM Press.

Sullivan, D. (2003). Search engine sizes. Search engine watch Web site articles. Retrieved December 10, 2005, from http://searchenginewatch.com/reports/article.php/2156461

Sullivan, D. (2005). Searches per day. Search engine watch Web site articles. Retrieved December 10, 2005, from http://searchenginewatch.com/reports/article.php/2156481

Wong, T.L., & Lam, W. (2004). A probabilistic approach for adapting information extraction wrappers and discovering new attributes. In Proceedings of the Fourth IEEE International Conference on Data Mining (pp. 257–264). Washington, DC: IEEE Computer Society.

Wong, W., & Fu, A.W. (2000). Finding structure and characteristics of Web documents for classification. In Proceedings of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (pp. 96–105). Dallas, TX: ACM Press.

Yi, L., Liu, B., & Li, X. (2003). Eliminating noisy information in Web pages for data mining. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 296–305). New York: ACM Press.

Yu, H., Han, J., & Chang, K.C. (2004). PEBL: Web page classification without negative examples. IEEE Transactions on Knowledge and Data Engineering 16(1), 70–81.