Data Cleansing for Web Information Retrieval using Query Independent Features

Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma
State Key Lab of Intelligent technology & systems, Tsinghua University
liuyiqun03@mails.tsinghua.edu.cn

Abstract

We report on a study that was undertaken to better understand what kinds of Web pages are the most useful for web search engine users by exploiting query-independent features of retrieval target pages. To our knowledge, there has been little research towards query-independent web page cleansing for web information retrieval. Based on more than 30 million web pages obtained both from TREC and from a widely-used Chinese search engine SOGOU (www.sogou.com), we provide analysis on the differences between retrieval target pages and ordinary ones. We also propose a learning-based data cleansing algorithm for reducing Web pages which are not likely to be useful for user request. The results obtained show that retrieval target pages can be separated from low quality pages using query-independent features and cleansing algorithms. Our algorithm succeeds in reducing 95% web pages with less than 8% loss in retrieval target pages. It makes it possible for web IR tools to meet over 92% users' needs with only 5% pages on the Web.

1. Introduction

The explosive growth of data on the Web makes information management and knowledge discovery increasingly difficult. The size of the web document collection becomes one of the main obstacles which stumbles 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) is increasing at a high speed. Google indexed over 8 billion pages in December 2004, which is about 20 times as many as what it indexed in the year of 2000 [5]. However, this amount still can't cover the whole page set on the web, which already contains over 20 billion surface web pages and 130 billion deep web pages almost 2 years before (in February, 2003) according to How Much Info project [4]

It is well known that web is filled with noisy, unreliable, low-quality and sometimes contradictory data so a data cleansing process is necessary before retrieval. In order to cleanse web data according to whether it is useful for a search engine user, we proposed a novel data cleansing method: First, we try to find differences between retrieval target pages and ordinary pages based on analysis in over 30 million web page data from both an English corpus (.GOV applied in TREC) and a Chinese corpus (obtained from Sogou.com). According to statistical comparison, several query-independent features are found to be able to tell the differences between the two kinds of pages. Then a learning-based algorithm is designed based on these features to cleansing web data using retrieval target page classification.

The main contributions of our work are:

1. A query-independent feature study is conducted to draw a clear picture of the differences between retrieval target pages and ordinary web pages.

2. A learning based method is proposed to automatically select high quality web pages according to whether they have chance to be retrieval target pages instead of the possibility of being visited.

The remaining part of the paper is organized as follows: Section 2 compares differences between retrieval target pages and ordinary pages with query-independent feature analysis. Section 3 introduced details of the data cleansing algorithm. Experiment evaluation is presented in Section 4 to assess the performance of our algorithm. Finally come discussion and conclusion.

2. Analysis in features of retrieval target pages

We analysis into two different web page corpuses: .GOV corpus which is made up of 1.2 M

English web pages and SOGOU corpus containing 37 M Chinese web pages.

.GOV is collected from .gov domain only so the overall page quality is generally higher than SOGOU, which is constitutive of pages crawled from all domains. We build a retrieval target page sample set for either corpus so the differences can be compared with ordinary pages. The sample set for .GOV is selected from TREC2003 - 2004 web track answers and the set for SOGOU corpus is labeled by 3 assessors using pooling technology [2]. TREC web track offers several hundreds queries collected from search engine logs or designed by assessors. Besides that, we collected 650 queries which are from SOGOU search engine logs representing users' search requests in several famous fields, such as Film/TV stars, Songs, Software, Movies, Novels, PC/TV Games, People names, News topics, Positions, Sports. With these queries, we build a retrieval target page sample set which contains 2631 pages for .GOV and 48930 pages for SOGOU. We use about half of these pages for testing the effectiveness of the data cleansing algorithm (1732 pages for .GOV, 24927 pages for SOGOU) and the others for algorithm training.

With analysis into the corpuses and corresponding retrieval target page sample set, we found out that retrieval target pages have totally different distributions with ordinary pages. The correlation values of several query independent features are shown in Figure 1 for .GOV corpus.

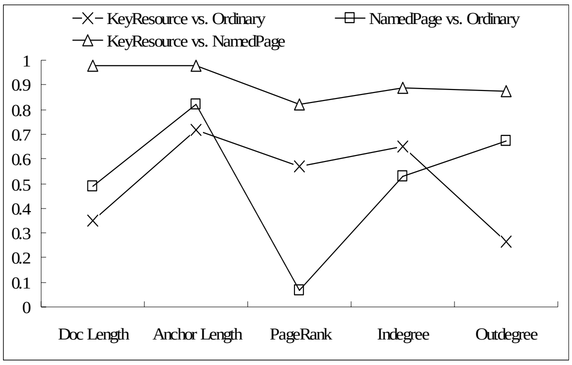

Figure 1 Differences in query-independent feature distributions represented by correlation values of different pairs of pages.

In Figure 1, we use five query-independent features to compare the differences between retrieval target page and ordinary page. 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 Page in [1]), Indegree (number of in-links), Outdegree (number of out-links). We can get the following conclusions from the stats 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.

2. The two kinds of retrieval target page (named page and key resource page) have a similar distribution in these query-independent features. The correlation values of named page and key resource page are not lower than 0.8, which means the two kinds of pages are positively correlated. Although these two kinds of retrieval target pages come from different retrieval requests, they share a lot in common. So it is reasonable to treat retrieval target pages as a whole instead of separately.

Based on the statistical analysis mentioned above, we found a number of query-independent features in which retrieval target page behaves differently from ordinary page. These features are listed in Table 4 and our learning-based data cleansing algorithm depends on them to cleanse web data for information retrieval tools.

3. Learning based Web data cleansing algorithm

In this paper, we adopt naïve Bayesian learning algorithm 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 calculate explicit probabilities for whether a web page can be a retrieval target page so that we can estimate web page quality according to the probabilities.

If we adopt query-independent feature *A*, the probability of one web page *p* being a retrieval target page can be calculated by $P(p \in Target\ \ page \mid p\ has\ \ feature\ \ A)$. We can use Bayes theorem to rewrite this expression as

$$P(p \in Target\ page \mid p\ has\ \ feature\ A) = \frac{P(p\ has\ \ feature\ A \mid p \in Target\ page)}{P(p\ has\ \ feature\ A)} \times P(p \in Target\ \ page) \qquad (1)$$

In Equation (1), $P(p \in Target\ \ page)$ is the proportion of retrieval target pages in the whole page set. As mentioned in the preceding part of this paper, this proportion is difficult to be estimated in many cases, including our problem of retrieval target page classification. However, if we just compare the values of $P(p \in Target\ \ page \mid p\ has\ \ feature\ \ A)$ in a given web page corpus, $P(p \in Target\ \ page)$ can be regarded as a constant value and it wouldn't affect the comparative results. So in a fixed corpus such as .GOV/SOGOU, we can rewrite equation (1) as:

$$
\begin{aligned}
&P(p \in Target\ \ page \mid p\ has\ \ feature\ A) \\
&\propto \frac{P(p\ has\ \ feature\ A \mid p \in Target\ \ page)}{P(p\ has\ \ feature\ A)}
\end{aligned}
\qquad (2)
$$

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

$$
\begin{aligned}
&\frac{P(p\ has\ \ feature\ A \mid p \in Target\ \ page)}{P(p\ has\ \ feature\ A)} \\
&= \frac{\#(p\ has\ \ feature\ A \cap p \in Target\ \ page)}{\#(Target\ \ page)} \bigg/ \frac{\#(p\ has\ \ feature\ A)}{\#(CORPUS)}
\end{aligned}
\qquad (3)
$$

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:

$$
\begin{aligned}
&\frac{\#(p\ has\ \ feature\ A \cap p \in Target\ \ page)}{\#(Target\ \ page)} \\
&= \frac{\#(p\ has\ \ feature\ A \cap p \in Target\ \ page\ sample\ set)}{\#(Target\ \ page\ sample\ set)}
\end{aligned}
\qquad (4)
$$

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

$$
\begin{aligned}
&P(p \in Target\ \ page \mid p\ has\ \ feature\ A) \\
&\propto \frac{\#(p\ has\ \ feature\ A \cap p \in Target\ \ page\ sample\ set)}{\#(Target\ \ page\ sample\ set)} \bigg/ \frac{\#(p\ has\ \ feature\ A)}{\#(CORPUS)}
\end{aligned}
$$

All terms in this equation can be obtained by statistical analysis into a web page corpus, so we can calculate the probability of being a retrieval target for each page according to this equation.

4. Experiment results

4.1. Evaluation methods

To our knowledge, there has been little research towards evaluation of web page cleansing for IR. In order to solve this problem of data cleansing evaluation, we proposed a new metric called High Quality Page Average Recall (*AR*). Average recall of high quality page is mean of the recall scores after each high quality page counted.

$$
AR = \sum_{i=1}^{\#(HighQuality)} Recall(i) \bigg/ \#(HighQuality) \qquad (8)
$$

Similar with the famous IR evaluation metric Average Precision, it is a summary measure of a ranked page list. When web pages in a certain corpus are ranked using a certain data cleansing method, *AR* is calculated according to the given ranking. It contains both cleansed-size-oriented and recall-oriented aspects.

If one data cleansing method doesn't work at all, the ranking given by this method will be a random sequence of the corpus and the corresponding *AR* for this method will be 1/2. If another data cleansing method gives all high quality pages top scores, *AR* for this method will be 1. *AR* will be close to 0 if all high quality pages are placed at the end of the ranking.

4.2. Data cleansing experiment results

As mentioned in Section 3, we keep about half of the retrieval target samples for testing our data cleansing algorithm. The effectiveness will be examined by the following means: First, we will see whether this algorithm can pick up high quality pages. Then we will compare the effectiveness of the query-independent features applied in algorithm to find out which one plays the most important role in data cleansing. At last, we will check out whether this algorithm can separate low quality pages as well.

4.2.1. High Quality Page Classification. Table 1 shows the cleansed corpus size and its corresponding retrieval target page recall for .GOV and SOGOU corpuses.

Table 1 Cleansed Corpus Size and Corresponding Target Recall using Data Cleansing Algorithm

	GOV	SOGOU
Cleansed Corpus Size (Percentage of original set)	52.00%	4.96%
. Retrieval Target Page Recall (Training set)	95.53%	92.73%
Retrieval Target Page Recall (Test set)	93.57%	92.37%

From the statistics in Table 1, we can see that our data cleansing algorithm can retain most retrieval target pages as well as significantly reduce corpus size. More than 92% retrieval target pages remains in the cleansed corpus. However, there is a major difference in cleansed sizes when the algorithm is applied to different corpuses. Over 95% pages are reduced by the algorithm for SOGOU corpus while only 48% pages are regarded as unimportant for .GOV. It may be explained by the fact that data quality of .GOV corpus is much higher than SOGOU. .GOV is crawled in 2002 and its pages are limited to .gov domain, whose content is more reliable than the whole Web. SOGOU corpus is collected in 2005; currently a lot more spam and low quality pages appear on the Web and the crawled pages are not limited to a certain domain. It is reasonable to find more high quality pages in .GOV than in SOGOU.

Compared with .GOV corpus, SOGOU corpus is more likely practical application environment for a Web information retrieval system. According to our experimental results, it is possible to satisfy over 90% user request with only 5% pages in the corpus, and the 5% pages can be selected query-independently using our data cleansing algorithm. It means that Web IR tools can apply a hierarchy structure to their data index. The cleansed page set can be placed into a high-level, frequently-used, fast-accessible index, which can meet most users' request. The other pages which are reduced by our algorithm can be placed into low-level indexes, because they are not so important for users.

4.2.2. Effectiveness of query-independent features. According to the definition of High Quality Page Average Recall (*AR*) in Section 5.1, we can calculate the *AR* value for our algorithm is 0.9064, which means this algorithm is effective because the value is close to 1.0000.

Further, we want to find out which query-independent feature is the most important in our algorithm. We also want to answer the question: Does this cleansing function come 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 the learning algorithm.

In order to answer this question, we tests *AR* values for our data cleansing algorithm, each time with one single feature dropped out. The experimental results are shown in Table 2.

Table 2 The effectiveness of query-independent features in data cleansing

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

We can see from Table 2 that when PageRank or Indegree is dropped out, the AR value drops the most badly. It means these two features, especially PageRank, plays the most important role in data cleansing. But our further experiment results in Figure 2 suggest that the other features should not be discarded in our algorithm.

According to Figure 2, we can see that the performance gets worse when only PageRank is applied to rank pages in the data cleansing process. Data cleansing algorithm which combined other features can gain better performance. It accords with the conclusion of Henzinger [3] that a better page quality estimation algorithm (than PageRank) should involve other sources of information than hyperlink structure analysis.

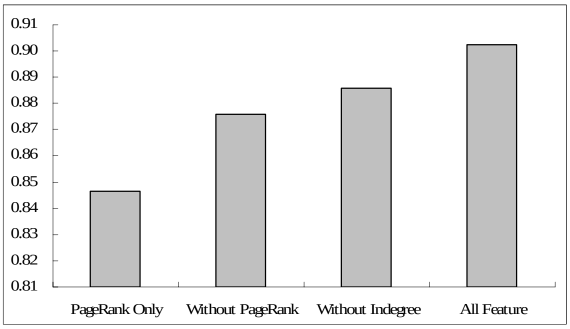

Figure 2 Effectiveness of PageRank and other features in data cleansing

5. Conclusions and future work

We have shown that by using a web data cleansing algorithm, it is possible to significantly reduce web data size and retains most high quality pages. Our algorithm, based on analysis into large scale web corpuses, exploits the differences between high quality pages and ordinary pages on the Web. We combine machine learning techniques and descriptive analysis to 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 to include other algorithms such as low quality page reduction and personalized Web search. We also plan to work on a hierarchy storage model for Web IR tools according to our findings in this paper.

6. References

[1] Brin S. & Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7).

[2] Hawking, D. & Craswell, N., (2005). Very Large Scale Retrieval and Web Search, in TREC: Experiment and Evaluation in Information Retrieval, MIT press, 2005.

[3] Henzinger, M.R., Motwani, R. & Silverstein, C. (2003). Challenges in Web Search Engines. Proceedings of the 18th International Joint Conference on Artificial Intelligence.

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

[5] Sullivan, D. (2003). Search Engine Sizes. Search engine watch web site articles. Retrieved December 10, 2005, from http://searchenginewatch.com/reports/article.php/2156461.