Web Data Cleansing for Information Retrieval using Key Resource Page Selection

Yiqun Liu, Canhui Wang
State Key Lab of Intelligent technology & systems
Tsinghua University
Beijing, China P.R.
Liuyiqun03@mails.tsinghua.edu.cn

Min Zhang, Shaoping Ma
State Key Lab of Intelligent technology & systems
Tsinghua University
Beijing, China P.R.
{z-m, msp}@tsinghua.edu.cn

ABSTRACT
With the page explosion of WWW, how to cover more useful information with limited storage and computation resources becomes more and more important in web IR research. Using web page non-content feature analysis, we proposed a clustering-based method to select high quality pages from the whole page set. Although the result page set contains only 44.3% of the whole collection, it is related with more than 98% of links and covers about 90% of key information. Link property and retrieval affects are also observed and experiment results show that key resource selection method is more suitable for the job of data cleansing and the result page set outperforms the whole collection by smaller size and better retrieval performance.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms
Experimentation

Keywords
Web data cleansing, Non-content feature, Web IR.

1. INTRODUCTION
The explosive growth of data on the Web makes information management and knowledge discovery increasingly difficult. In one hand, not all pages can be collected by web information management tools. In the other hand, not all pages collected are useful and not all information on these pages is high-qualified, since the web is filled with noisy, unreliable, low-quality and sometimes contradictory data. It would be extremely helpful for web search engines to identify the quality of web pages independent of a given user request, so that they can index more high-quality pages with limited resources. This is called one of web search engine's challenges. Link-based approaches such as PageRank [3] and HITS [3] can partly solve the problem. However, they only use link structures of the web and a better estimate should require additional non-content sources of information both within a page and across different pages.

In this paper we proposed a web data cleansing method. Non-content features including but not limited to link analysis features are involved in this method. Clustering-based algorithms are adopted to select a certain kind of high quality pages called key resources. Main contributions of our work are:

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-051-5/05/0005.

[1] A non-content feature study is conducted to draw a clear picture of the differences between high quality pages and ordinary web pages. [2] A cluster-based method is proposed to automatically select important web pages according to whether they have chance to be key resources. This method makes use of both prior knowledge and page's non-content features. [3] The possibility of achieving better retrieval performance with a pre-selected page set is discussed.

2. NON-CONTENT FEATURES OF KEY RESOURCE PAGES
Key resource page is a kind of high quality web page which provides as much credible information as possible on a certain topic [2]. It is different from ordinary pages even if they are relevant to the same topic. From relevant but not key resource pages one only gets information covering a small part of knowledge and can't read more. However, key resource page either provides credible information, or offers plenty of useful information that can be obtained via only one click.

Feature selection is the key point in the procedure of separating key resource pages. Data preparation should be a query-independent process for web search engines so only non-content features can be adopted, which are shown in Table 1.

Table 1 Differences in non-content feature average value between ordinary pages and key resource pages[2]

Average Value	.GOV	Key Resource
In-degree	9.94	153.12
URL length (in 4 categories)	3.85	3.07
In-site out-link anchor text rate	0.06	0.12
In-site out-link number	17.58	37.70
Document Length (in words)	7037.43	9008.02

In-site out-link is defined as an out-link navigating to another page located in the same site. This kind of link is specially treated because key resource pages should have enough in-site out-link to connect to other pages in the same site and enough in-site out-link anchor text to give a brief view of these pages.

There are differences in non-content feature distribution between ordinary and key resource pages. It means that these features can tell the two kinds of pages from each other.

[1] Our work is supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108) and Natural Science Foundation (60223004, 60321002, 60303005)

[2] .GOV (ordinary page set) is a crawl of 1.25M Web pages from .gov domain composed of over 18G data. Training set is composed of relevant answers of TREC 2002's web track.

3. A CLUSTER-BASED KEY RESOURCE SELECTION ALGORITHM

Although we can get several positive examples for key resource pages, there are too many reasons for one page to be a non key resource page. Negative instance collecting will be very difficult or even impossible. That is why supervised classification analysis can't be performed for key resource selection. A clustering-based algorithm is therefore used to combine features to do this task. Figure 1 show how the algorithm performs on .GOV corpus.

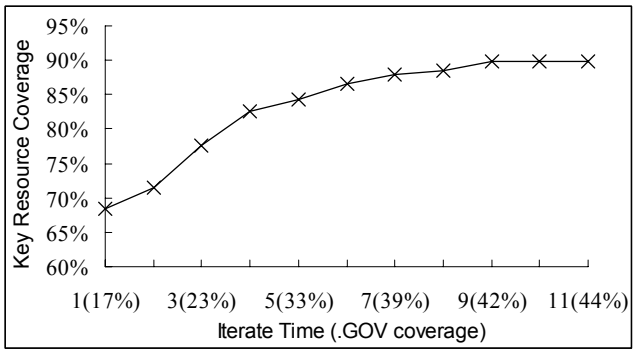

Figure 1. Key Resource Coverage varies with iterate times.

At first, .GOV coverage is assigned an initial value (1/6) to calculate the negative instance set centroid. Key resource test set is composed of relevant answers of TREC 2003's topic distillation task (516 pages related with 50 topics). From Figure 1 we can see that it is possible to cover 90% test set pages with about 44% of all pages.

4. EXPERIMENTS AND DISCUSSIONS

4.1 Link Analysis Experiment Results

Page set developed with our key resource selection method covers a majority of links in the whole collection. There are altogether 1247753 pages and 10185630 hyperlinks in .GOV corpus and this page set are related with over 98% links. It means although more pages are outside the page set in .GOV, there would be almost no hyperlinks if pages inside the set were taken. This hyperlink distribution shows that key resource pages would be the top scored ones in link structure analysis.

4.2 Retrieval Experiment Results

According to the query log analysis of Alta Visa by Broder in [1], web search engine queries are divided into 3 categories, which are Navigational, Informational and Transactional. In order to simulate web search user activity as close as possible, we build a query set in which each type of query has the same coverage as in query log analysis. Details are shown in Table 2.

Table 2 Query type distribution in our test query set

Type of query	Query log analysis in [1]	Test query set
Navagational	20%	10 (20%)
Informational	50%	40 (80%)
Transactional	30%	

We randomly selected 10 topics from TREC 2003's known item search task and 40 topics from its topic distillation task to build this query set. Corresponding relevant answers are used to judge the effectiveness of retrieval experiment results.

Corresponding retrieval Experiment results are shown in Figure 2 and Table 3. BM25 weighting in full text index and default parameter tuning are applied in both page sets.

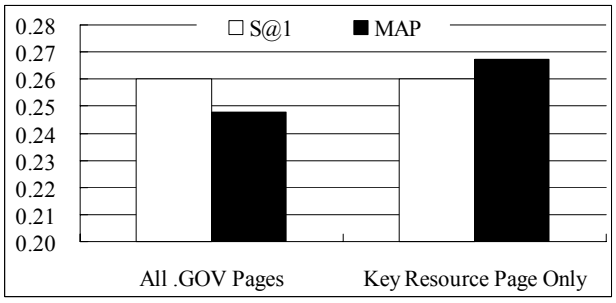

Figure 2 Retrieval performances in the measure of success rate at 1st document and mean average precision.

Table 3 Retrieval performance for different type of queries

Page Set	Informational queries (P@10)	Navigational queries (MRR)
.GOV	0.1025	0.7443
Key resource	0.1275	0.7278

From these experiment results, we can get the following conclusions: First, web page cleansing via selecting key resources is useful because the page set after cleansing outperforms entire page set by smaller size and better retrieval performance. It is composed of less than half of .GOV pages but get better overall performance in the measure of MAP. Second, the entire page set gets higher MRR for navigational type queries according to Table 3 because it is unavoidable to reduce part of useful information in the process of data cleansing. However, navigational queries only cover a small part (20%) of web search request and overall performance improves with the data cleansing method.

5. CONCLUSIONS AND FUTURE WORK

In this paper, a key resource selection approach based on K-means clustering for web data cleansing is proposed. The whole collection, which is 19G .GOV web data set, is divided into two sets. By doing so, only 44.3% pages are selected as key resources, which will be kept for indexing and retrieval by IR systems. This set contains 98% links of the whole collection. Retrieval experiments get better overall performance although this page set has much fewer pages than the whole collection.

Future study will focus on following aspects: Is it possible to identify navigational search destiny pages query-independently so that we can improve performance for this kind of queries? Can we rank pages instead of filter them with a similar non-content feature analysis process for web data cleansing? Is there a best trade off between page set size and retrieval performance?

6. REFERENCES

[1] A. Broder, A taxonomy of web search. SIGIR Forum Volume 36 Number 2, 2002.

[2] D. Hawking and N. Craswell. Overview of the TREC-2003 Web track. NIST Special Publication: SP 500-255, The Twelfth Text Retrieval Conference (TREC 2003), 2003.

[3] J. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of the ACM, Volume 46 (5), 1999.

[4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Proceedings of the 7th World-Wide Web Conference (WWW7), 1998.