Journal of Computational Information Systems
Available at http://www.JofCI.org

Learning-based Web Data Cleansing for Information Retrieval *

Yiqun LIU†, Min ZHANG, Canhui WANG, Shaoping MA

State Key Lab of Intelligent Technology and Systems, Tsinghua University
Beijing 100084, China

Abstract

With rapid growth of web information, how to select high quality web pages that cover valuable information query-independently becomes more and more important in web IR research. Based on query-independent feature analysis, we propose a data cleansing algorithm by selecting an important type of high quality pages (key resources) on the web. Study into the cleansed page set shows that the set contains only 44.3% pages of the whole collection, while involves more than 98% of hyperlinks and covers about 90% of key information. Experiments based on TREC 2003 data show that the cleansed collection outperforms the whole collection by less than a half size and 8% improvement of retrieval performance.

Keywords: Web Information Retrieval; Data Cleansing; Query-independent Features

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).

The number of pages indexed by web search engines is increasing at a high speed; however, not all pages can be collected by web information management tools. Google increased its index to 8 billion pages in November 2004, to counter the increase to 5 billion by MSN [4]. However, this amount only covers a small part of the whole page set on the web, which contains over 10 billion surface web pages and 130 billion deep web pages in February 2003 according to [11]. Furthermore, not all pages collected are useful, since the web is filled with noisy, unreliable, low-quality and sometimes contradictory data. Therefore the idea of web data cleansing via selecting high quality pages is worth paying attention to.

This kind of web data cleansing requires identifying quality of web pages independent of a given user request. Cleansing algorithms should use query-independent features so that qualified ones can be selected in advance of retrieval process. Hyperlink analysis algorithms such as PageRank [16] and HITS 错误！未找到引用源。 have gained much success in making use of link structure information (a query-independent feature) to estimate web page quality. However, query-independent page quality estimation is still called one of web search engine's biggest challenges in [12] because methods that just make use of hyperlink information meet the problem of link spam. Therefore a better page quality estimation algorithm should utilize query independent sources of information both within a page and across different pages.

In this paper we proposed a web data cleansing method based on selecting an important kind of high quality pages: key resource pages. Key resources are entries for useful Web information and valued by Web search users. In our previous research into more than 1M web pages in [18], some query-independent features (hyperlink-based or not) are found to be effective in picking up key resource pages from the whole

* Supported by the Chinese National Key Foundation Research & Development Plan (2004CB318108), Natural Science Foundation (60223004, 60321002, 60303005, 60503064) and the Key Project of Chinese Ministry of Education (No. 104236)
† Corresponding author.
Email address: liuyiqun03@mails.tsinghua.edu.cn (Yiqun Liu)

June, 2006

collection. In that work, we also developed a decision tree based algorithm to select key resources for effective topic distillation search. However, our previous method depends on prior knowledge of key resource proportion within the whole collection. This requirement limits its application and makes it difficult to be used for general web search. The main contributions of our work are:

1. A clustering-based method is developed to locate key resources using query-independent features. It is more effective than the previous method and doesn't need key resource proportion knowledge. This method is designed for general web data cleansing instead of topic distillation task only. **2.** The selected key resource set retains almost all (over 98%) hyperlink structure information although it contains less than 50% pages. **3.** Experiments with TREC 2003 web track data show that it is possible for web search tools to improve retrieval performance and reduce index size at the same time by our Web data cleansing algorithm.

The remaining part is constructed as follows: Section 2 clarifies the definition of key resource and gives a brief review of related works in web page classification. Section 3 proposed the data cleansing algorithm based on query-independent features. Hyperlink structure analysis of the cleansed page set is made in Section 4. Retrieval experiment results are shown in Section 5. Finally come discussion and conclusions.

2. Related Works

2.1. Definition of Key Resource Page

Key resource page is a kind of high quality web page which covers key information on the Web. They are resources which a human editor might list under a subject category in Web directories such as DMOZ. The conception of key resource is firstly proposed by TREC web track in 2002 [14]. It is described as target pages of topic distillation task, which aims at finding quality web pages that are most representative of a certain topic. Further experimental study in TREC [14][15] shows key resource pages either offer credible information itself or provide entries to clusters of high quality pages. In this way, key resource is like a combination of what HITS algorithm wants to locate: good hubs and authorities. However, key resource is likely to be, but not necessary to have hub/authority features in hyperlink structure.

Key resource is different from ordinary pages even if they are relevant to the same topic. The major difference is: although key resource itself may not provide detail information, it offers plenty of useful information that can be obtained via no more than one click.

A large proportion of search engine user requirement can be regarded as topic distillation according to user log mining (75% in AltaVista log mining by Broder [1], 85% in Yahoo! log mining by Rose and Levinson [3]). If key resource pages can be selected query-independently, it will meet a large number of web search user need. That supports our idea of applying key resource selection in data cleansing.

2.2. Web Page Classification without Negative Examples

Key resource selection is a kind of web page classification problem. It is different with ordinary classification process in the lack of negative examples because uniform sampling of negative examples is difficult and sometimes not possible. This classification should be performed with only a few positive examples (some manually collected key resources) and unlabeled data (the whole page collection we use). Because traditional classification approaches use both fully-labeled positive and negative examples in classification, approaches should be developed to deal with this kind of web page classification problem.

In 1998, Denis defined the PAC learning model for positive and unlabeled examples in [7]. Later work [6][8] showed that k-DNF and C4.5 can gain experiment success in solving this problem. Our work in [18] proved that key resource page selection is possible with some positive examples and unlabelled data. However, these methods are not applicable for all kinds of Web page classification problems because positive instance proportion within the universal set is required, which is not available in many problem settings. One-class SVM (OSVM) [13] and Positive Example Based Learning framework (PEBL) [9] are

based on the strong mathematical foundation of SVM. They both distinguish one class of data from the rest in the feature space without negative examples. PEBL framework also makes use of unlabelled data and therefore gains better performance. However, the PEBL framework is designed for topic-dependent problems and it is not suitable for key resource selection task, which uses query-independent features only.

3. Key Resource Selection Method

3.1. Comparison of Non-content Features between Key Resources and Ordinary Pages

We use query-independent features proposed in our previous work [18] to separate key resource pages from ordinary pages. The features are in-link count (in-degree), document length, URL-type, in-site out-link number and in-site out-link anchor rate. Key resource training examples are collected from relevant qrels (answers labeled by assessors) of TREC 2002's topic distillation task. We also use relevant qrels of TREC 2003's topic distillation task for testing. .GOV corpus which is a crawl of 1.25M Web pages from .gov domain composed of more than 18G data is adopted as the unlabelled page set. The two types of web pages have completely different average values in these non-content features as shown in Table 1.

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

	Ordinary Page Set	Key Resource Page Set
In-degree	9.94	153.12
URL type*	3.8516	3.0734
In-site out-link anchor text rate	0.0618	0.1240
In-site out-link number	17.58	37.70
Document Length (in words)	7037.43	9008.02

In Table 1, key resource pages have more in-links and it meets with the conclusion by Kraaij et al. [17] that entry pages tend to have a higher number of in-links than other pages. Key resources also tend to have PATH type URL for a similar reason: they are important pages and should be given a non-FILE URL type. Although the difference in document length between to page sets are not so obvious, further analysis show that only 1.12% key resource pages have fewer than 1000 words. The corresponding percentage in ordinary page set is 16.08%. It means that document length can also help tell the two kinds of pages from each other. In-site out-link is defined as the out-link navigating to another page located in the same site. This kind of link is specially treated in our method because it is believed that key resource pages are mainly entry pages or index pages for certain web sites. Web site entry page 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. Key resource pages have much larger in-site out-link number and anchor text rate than ordinary pages.

3.2. A Clustering-based Algorithm to Select Key Resources

Learning algorithms should be performed to combine the query-independent features described in the previous section to separate key resources from other pages. However, as we proposed in Section 2.2, both positive and negative data are needed in a traditional learning algorithm. Key resource page samples, which are easily obtained from TREC qrels or Web directories, can be used as positive examples. But it is difficult to sample non-key resource pages uniformly because there are various reasons for one page not being a key resource page. How to deal with the lack of negative examples becomes a dominant problem.

Our method is based on K-means clustering algorithm. According to [5], K-means is an algorithm for partitioning N data points into K disjoint subsets containing data points so as to minimize the

* In order to calculate the average value for URL type, different values are given to each type of URL as follows: ROOT type = 1, SUBROOT type = 2, PATH type = 3 and FILE type = 4.

sum-of-squares criterion. It is applied in key resource selection because it performs well with a fixed cluster number (2 for this task) and low-dimensional feature space (5 dimensions according to Section 3.1). In our algorithm, information of negative examples is estimated using both key resource training set and unlabelled data. Supposing *Key* stands for key resource training set and *R* is the proportion of key resource pages in the whole page set, the data cleansing algorithm can be described as follows:

[1] Choose 2 initial cluster centroids:

$$M_1 = \frac{1}{|S_{Key}|} \sum_{X_i \in S_{Key}} X_i$$

$$M_2 = \frac{M(whole\ \ page\ set) - R \cdot M_1}{1 - R}$$

[2] In the k^{th} iterative, instance X will be assigned to the j^{th} cluster $S_j^{(k)}$ if :

$$\left\| X - M_j^{(k)} \right\| = \min \left\| X - M_i^{(k)} \right\| \quad (j = 1,2)$$

[3] For $S_j^{(k)}$, caculate $M_j^{(k+1)}$, which is defined as:

$$M_j^{(k+1)} = \frac{1}{N_j} \sum_{X \in S_j^{(k)}} X \quad (j = 1,2)$$

[4] If $M_j^{(k+1)} = M_j^{(k)} \quad (j = 1,2)$, exit.

Else go to [2].

In this algorithm, M_2 (the negative centroid) is estimated using the whole collection centroid, the positive centroid and *R*. All centroids are adjusted after each partition. The whole collection is divided into two non-interacted sets when algorithm finishes. Figure 1 show how the algorithm performs on .GOV.

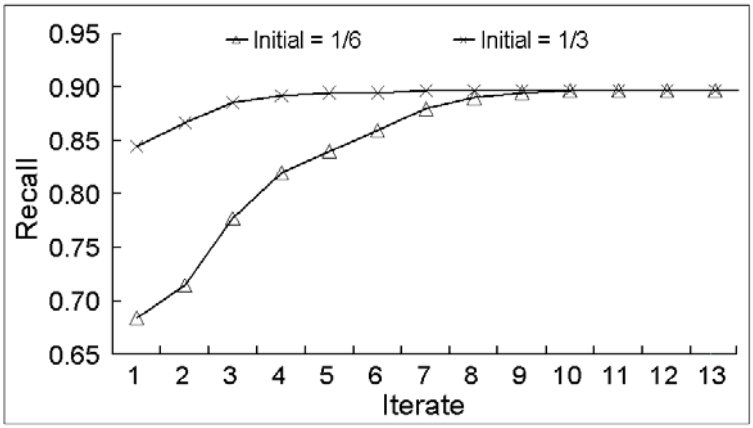

Fig. 1 Key resource selection result set recall varies with iterates

Figure 1 show that with different initial *R* values (1/6 or 1/3), Recall value increases with iterates and converges to the same position at last. Further experiments show that the convergence recall value doesn't change with many different initial values (but it requires different number of iterates before convergence). It means this algorithm doesn't require the knowledge of positive (key resource) proportion from the whole collection and it is one of its major advantages.

The result set obtained from this algorithm covers 89.70% key resource pages in .GOV according to Figure 1 (the converge point). We also found that this set contains 44.30% pages in the corpus, which means more than half web pages can be reduced with 10.30% key information loss.

4. Link Structure Analysis of Cleansed Page Sets

We analyzed the hyperlink structure of the selected page set to find out whether the pages selected by our data cleansing method are highly qualified. Page sets developed with different methods cover a majority of links of the whole page set according to our experiment results.

There are altogether 10185630 hyperlinks in .GOV corpus and 98% of them is linked to or located in a page selected by our data cleansing algorithm. It means although the amount of pages outside the key resource set (55.70%) is larger than the amount of pages inside (44.30%), there would be almost no links between the outside pages if inside pages were taken. This selected key resource set is similar with Strongly connected components (SCC) proposed by Broder et al in [2]. This phenomenon means that there is little hyperlink structure information loss during the data cleansing process.

5. Experiments and Discussions

5.1 Retrieval Tasks and Evaluation Measures

According to the query log analysis of Alta Visa by Broder in [1], web search engine queries are grouped 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 proportion as in query log analysis. TREC2003 web track topics and qrels are selected to form this testing set.

Mean Average Precision (MAP) is used as an overall performance metric because it can be applied to measure the performance of all types of queries. Topic distillation and navigational search performance are also evaluated separately with the measure of P@10 (precision at 10 documents) and MRR (Mean Reciprocal Rank) to see the performance of data cleansing method for a particular task. BM2500 weighting and default parameter tuning are used. In-link anchor text is treated as part of the page it links to improve retrieval performance according to [15].

5.2 Retrieval Experiments with Different Data Cleansing Methods

The test query set described in the previous section are retrieved on all of four page sets selected from .GOV using different data cleansing methods. They are:

1. .GOV after data cleansing using K-means clustering / decision tree learning algorithms in [18].
2. .GOV after reducing low PageRank/authority/hub value pages.

Experimental results are shown in Table 2 and we can get the following conclusions:

Table 2 Retrieval performance of web page sets using different cleansing methods

Cleansing Method	MAP	P@10 for topic distillation queries	MRR for navigational queries
No data cleansing	0.2476	0.1025	0.7443
K-means based	0.2674	0.1275	0.7278
D-Tree based	0.2306	0.1200	0.6000
Reducing low PageRank	0.2214	0.1134	0.6533
Reducing low authority	0.2398	0.1100	0.6700
Reducing low hub	0.2481	0.1250	0.6357

1. With the evaluation metric of MAP, K-means based method obtains the highest score. It means its corresponding cleansed page set gets better retrieval performance than other data cleansing methods.

2. Web data cleansing by selecting key resources outperforms entire page set by smaller size and better retrieval performance. The cleansed set is composed of less than half of .GOV pages but gets better overall performance with the measure of MAP. However, the whole collection gets higher MRR for navigational type queries according to Table 2. It can be explained by the fact that key resource result set abandons a large number of unimportant pages and some of them aren't useless for all queries (especially for named-page finding queries described in [14]). It is unavoidable to reduce part of useful information in the process of data cleansing but results are still encouraging because overall performance improves and navigational rankings don't depress too much with a significantly smaller page amount.

3. Topic distillation obtains better results with all data cleansing methods while navigational queries got worse rankings if data cleansing is involved. It gives us a clear idea that two types of queries emphasize a different aspect in web information retrieval. Precision is more important for topic distillation because relevant pages are too many for such a query and only high-quality pages are useful for users. Meanwhile navigational task has a higher requirement in recall because only a few pages can meet users' need.

6. Conclusion

Given the vast amount of information on the World Wide Web, a typical short query of 1-3 words submitted to a search engine usually get a result list of tens of thousands web pages, while only a tiny part of these pages is useful for users. Web data cleansing with key resource selection based on K-means clustering makes it possible to get better retrieval performance with fewer pages indexed.

Future study will focus on following aspects: How well does this method work for a page set with billions of pages? Is it possible to identify known-item search destiny pages query-independently so that we can improve performance for this kind of queries?

References

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

[2] A. Z. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. WWW9/Computer Networks, 33(1--6):309--320, 2000.

[3] D. E. Rose and Danny Levinson, Understanding User Goals in Web Search. In the 13th World Wide Web conference (WWW 2004), 2004.

[4] Danny Sullivan, Search Engine Sizes. In search engine watch website; September 2, 2003; Online at: http://searchenginewatch.com/reports/article.php/2156481

[5] Eric W. Weisstein. K-Means Clustering Algorithm. From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/K-MeansClusteringAlgo-rithm.html

[6] F. DeComite, F. Denis, and R. Gilleron, Positive and Unlabeled Examples Help Learning, Proc. 11th Int'l Conf. Algorithmic Learning Theory, 219-230, 1999.

[7] F. Denis, PAC Learning from Positive Statistical Queries, Proc. 10th Int'l Conf. Algorithmic Learning Theory pp. 112-126, 1998.

[8] F. Letouzey, F. Denis, and R. Gilleron, Learning from Positive and Unlabeled Examples, Proc. 11th Int'l Conf. Algorithmic Learning Theory, 2000.

[9] Hwanjo Yu, Jiawei Han, and Kevin Chen-Chuan Chang. PEBL: Web Page Classification without Negative Examples. IEEE Transaction On Knowledge and Data Engineering, Vol. 16, NO. 1, January, 2004.

[10] Jon M. Kleinberg, Authoritative sources in a hyperlinked environment. Journal of the ACM, 1999,46(5):604-632.

[11] Lyman. Peter and Hal R. Varian, How Much Information, 2003. Retrieved from http://www.sims.berkeley.edu/how-much-info-2003 on April 2th, 2004.

[12] Monika R. Henzinger, Rajeev Motwani and Craig Silverstein, Challenges in Web Search Engines. In proceedings of the International Joint Conference on Artificial Intelligence, 2003.

[13] Manevitz, L. M., Yousef, M., One-class svms for document classification. J. Machine Learning. Res. 2, 2002.

[14] D. Hawking, N. Craswell: Overview of the TREC-2002 web track. In NIST Special Publication 500-251: The Eleventh Text REtrieval Conference (TREC 2002).

[15] N. Craswell, D. Hawking: Overview of the TREC-2003 web track. In NIST Special Publication 500-255: The twelfth Text Retrieval Conference (TREC 2003).

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

[17] W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In 25th ACM-SIGIR conference on research and development in information retrieval, pages 27--34.

[18] Y. Liu, M. Zhang, S. Ma: Effective Topic Distillation with Key Resource Pre-selection. In Asia Information Retrieval Symposium AIRS 2004: 129-140.