Web Page Quality Estimation Based on Linear Discriminant Function[1]

Rongwei Cen, Yiqun Liu, Min Zhang, Liyun Ru, Shaoping Ma

State Key Lab of Intelligent Technology & Systems, Tsinghua University

Beijing, China P.R.

crw@mails.tsinghua.edu.cn

Abstract

With the growth of web data, how to estimate web page quality effectively and rapidly becomes more and more important for web information retrieval and knowledge discovery. This paper analyzes the differences between retrieval target pages and ordinary pages using query-independent features. Using these features, an algorithm called Linear Page Estimation (LPE) is proposed for web page quality estimation. Based on experiments on .GOV corpus and SOGOU corpus involving 26 million pages, about 95% pages can be reduced with more than 90% retrieval target pages retained using our algorithm. Experimental results based on TREC datasets also show that retrieval performance on collections selected by our algorithm can be close to or even better than that on the whole collection.

Keywords: Line Discriminant Function; Query-independent Feature; Web Page Quality Estimation.

1. Introduction

As the increasing use of the internet technology, the explosive growth of the web data makes it more and more difficult to retrieve the web information and discover knowledge. Search engines, as one of the most useful tools for managing and accessing the web data, have to index more and more web pages. According to the report of Sullivan [1], Google indexed over 8 billion pages in November 2004, which is about 20 times as many as what it indexed in the year of 2000. However, this is only a small part of the whole web page set, which contains 20 billion surface web pages and 130 billion deep web pages 2 years ago according to the How Much Info project [2]. The huge size of the web data becomes one of the main bottlenecks which refine the storage and search speed of search engines with limited resources, as well as other web-based information tools.

Data quality becomes more welcomed than data quantity with the explosive growth of the web data. As we all know that some of the web pages are filled with noisy, unreliable, low-quality and sometimes contradictory data, which are hardly the search result pages or meet the users' needs. So it is necessary to estimate the quality of web pages and cleanse these garbage data before indexed and retrieved. However, since the cleansed data set is supposed to meet all kinds of Web search needs, the page quality estimation process, which is related with users' request intuitively, should be finished independent of queries. That is why Web page quality estimation is called one of search engine's greatest challenges by Henzinger in [4].

As the web page estimation process should be independent of user query, the features related to user query cannot be used, such as the keywords of user query. However, there are several previous works on query-independent features that will be helpful. The works of the hyperlink structure analysis have a great

[1] 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)

success, such as PageRank[8] and HITS[9], which estimate the page by the hyperlink structures without the user query information. In addition, some of the web page-self features have been investigated, such as page length, URL type, etc. In our previous work, the web key resource page judgment [5][6] and web data cleansing [7][15], both of the hyperlink structure and page self-features were used to estimate the quality of the web page.

The traditional classification learning mechanisms need the data have three elements, attribute, attribute-value and the class tag of examples, which contain positive ones and negative ones for two classification problems. To web page quality estimation, positive examples can be picked up by several assessors using techniques such as pooling [20]. However, since it is almost impossible to take a uniform sample of web pages without bias [4] and hardly to define what is not needed by the search engine user clearly, there is a lack of the representative negative examples. So, in the research of the web page estimation, there is an algorithm research for such a special classification problem besides the page query-independent feature discovery and analysis. This paper studies the query-independent features on the about 26 million Web page data from both an English corpus (.GOV applied in TREC) and a Chinese corpus (obtained from Sogou.com). In addition, an algorithm based on the linear discriminant function is proposed to estimate the web page quality.

The remaining parts of the paper are organized as follows: Section 2 gives a brief review of related work in web data estimation and page classification methods. Section 3 compares differences between retrieval target pages and ordinary pages using the query-independent feature analysis with both common-used features and newly-proposed ones. Section 4 introduces details of the page quality estimation algorithm based on the linear discriminant function. Experimental evaluation and analysis is stated in Section 5 to assess the performance of our algorithm. Finally, the conclusion and future work come.

2. Related Work

In 2002, Henzinger [10] from Google pointed out that it would be extremely helpful for search engines to be able to identify the quality of web pages independent of a given user request. In this chapter, the relative work about the page quality estimation and the features used in the process will be briefly introduced.

2.1. The Query-independent Feature

In [11], Craswell has found that, in optimal conditions, all of the query-independent methods they studied (in-degree, URL-type, and two variants of PageRank) offered a better improvement than random on a content only baseline in site finding. In the page estimation, query-independent feature has more advantages than query-dependent feature: the later one is always relative to the keywords of the search engine user request; however, the first one, independent of user query, covers all kinds of topics and can meet different users' requests. So, using the query-independent features for the web page selection, especially for the homepage location, becomes the research focus of the information retrieval [12][13].

The existing link analysis algorithms, success in many common commercial search engines and hardly considering the page-self features, rely on the assumption that if two pages are connected by a hyperlink, the page being linked is recommended by the page which links to it and the two pages share a similar topic [16]. However, there is little research on the page-self feature which is necessary in query-independent feature study. In our previous work, web key resource page judgment [5][6] and web data cleansing [7][15], some query-independent features have been proposed, including some link structure features, such as

in-link number, in-site out-link number, in-site out-link anchor ratio, etc, and some page-self features, such as page length, URL-type, etc. In this paper, with these existing features, some new features are investigated and analyzed.

2.2. The Learning Methods for Web Page Estimation

Due to the data particularity, the traditional machine learning methods cannot be well adopted and applied in this web data mode, lots of unlabelled data and some positive examples picked out. Currently, there are several learning mechanisms that can finish the task of Web page classification based on this data mode. Techniques such as semi-supervised learning [17], single-class learning [18] or one class SVM (OSVM)[19] have been applied to solve the problem. However, these methods may not be applicable for retrieval target page classification for some reasons, such as not designing for low dimensionality and high density instance space, requiring knowledge of the proportion of positive instances within the universal set, time-consuming not suitable for the real-time task of cleansing billions of pages required by search engines, etc.

Our previous researches have proposed the Improved Decision Tree [5][6], a approach based on K-means clustering [7], and a learning method based on naive Bayes classifier [15], and gained some success in the page quality estimation. However, the Improved Decision Tree requires the knowledge of positive example proportion in the corpus, which is difficult to obtain in the real environment. The cleansing algorithm based on K-means meets the problem of time efficiency and is not suitable for practical application, and the naive Bayes learning algorithm needs the features be discretized into Boolean ones by comparing the differences between the distributions of ordinary and high quality pages in advance, which may lose some real information of the features.

Unlike these algorithms, our page quality estimation is based on the linear discriminant function. According to the complexity analysis in Section 4 and the comparison experiments of the different methods in Section 5, our algorithms can be considered to be effective and efficient for the web page estimation problem.

3. Query-independent Features Analysis of Web Page

According to the above sections, we konw that only the query-independent features can be used in the estimation process of web page quality. If these query-independent features can tell the differences between the high quality web pages and ordinary ones, we can use them to cleanse the web page set or estimate the pages. Actually, our previous work showed that there were some query-independent features that can be used for the work of data cleansing or web key resource judgment [5][6][7][15].

Here we first analyze the common-used and newly-proposed features in two different web page corpuses simply, .GOV corpus [14] and SOGOU corpus (obtained from Sogou.com), and find out the effective features which can distinguish the high quality pages from the ordinary ones. The details of these two corpuses are shown in Table 1.

Table 1 Statistics of the .GOV and SOGOU corpuses

	#Page	Language	Total size	Average docsize	Domain limit	Crawling time
.GOV	1,247,753	English	18.1G	15.2k	.gov	Early 2002
SOGOU	24,833,521	Chinese	372.5G	15.0k	No limit	Nov. 2005

.GOV corpus, which is made up of 1.2M English web pages, is a TREC (Text Retrieval Conference, http://trec.nist.gov/) test collection and it is applied in TREC 2002-2004 Web tracks. This corpus is

collected from .gov domain only, so the quality of the overall pages is higher than the SOGOU corpus, which is made up of 24.8M Chinese Web pages and crawled from all domains. However the .GOV corpus from only one domain cannot represent the whole WWW environment and the conclusions on the .GOV may not be so applicable.

We build retrieval target page sample sets for each corpus as the high quality page, and can compare the differences between the ordinary page corpus and the retrieval target page set. The sample set for .GOV, is selected from TREC2002-2004 Web track answers and the set for the SOGOU corpus is labeled by 3 assessors using pooling technology [20]. There are total 4375 retrieval target pages for .GOV and 6404 pages for SOGOU.

With our previous work [5][6][7][15], about the key resource study and query-independent feature analysis, this paper adopts the following query-independent features for page quality estimation, some of which are common-used, others are newly-proposed. These features can be divided into two different kinds, the link structure related features and the page-self features.

1) *In-link number: namely in-degree, the number of in-links of a web page;*
2) *Out-link number: namely out-degree, the number of out-links of a web page;*
3) *Pagerank value: the Pagerank value calculated according to algorithm in [8];*
4) *In-site out-link number: the number of links from a web page to other pages in the same site;*
5) *In-site out-link anchor text ratio: the ratio of the in-site out-link anchor text taking the page whole text;*
6) *Page length: the number of words in a web page;*
7) *Anchor text length: the number of a web page anchor text words;*
8) *URL classification: the type of the URL, ROOT, SUBROOT, PATH and FILE [21];*
9) *URL format: whether a URL contains a question mark [15];*
10) *Title length: the number of words in a web page title field;*
11) *Image number: the number of images in the page;*
12) *GBK code: whether the encode of a web page is GBK [15];*
13) *Cluster: the number of mirror copies of a web page [15].*

By investigating these features into the corpuses and retrieval target page sample sets, we find out that these two kinds of pages have different distributions. See Table 2 and Table 3 below.

Table 2 Differences in average values of features between ordinary pages and high quality pages in .GOV corpus

Page features	Ordinary pages	High quality pages
Page length	6182.04	11164.91
Anchor text length	21.91	358.69
URL classification	3.85	3.46
In-Site out-link number	15.44	24.41
In-Site out-link anchor text ratio	0.054	0.071
URL format	0.185	0.055
In-link number	8.03	152.05
Out-link number	20.96	35.78
Pagerank value	0.94	13.79
Title length	4. 023	5. 34

Image number	8. 21	15. 04

Table 3 Differences in average values of features between ordinary pages and high quality pages in SOGOU corpus

Page attributes	Ordinary pages	High quality pages
In-link number	32.615	6008.913
URL classification(root, path ,file)	1.905	1.831
Anchor length per link	0.187	2.194
Page length	514.2	500.42
Cluster	1260.1	15559.2
Pagerank value	3572.7	15518.1
Non-GBK code	0.142	0.007
URL format	0.133	0.023

Table 2 and Table 3 show the differences in average values of features between ordinary pages and high quality pages in .GOV and SOGOU corpus. Taking the GBK code feature for example, the average value of the ordinary pages is 0.142, which means about 14.2% pages are non-GBK encoded, and the average value of the high quality pages is 0.007, which means that about 0.7% pages are non-GBK encoded. This can be illustrated by the reason that the SOGOU corpus comes from a Chinese search engine and the users mainly pay more attention to web pages written in simplified Chinese.

Fig 1 shows that the different distributions of retrieval target and ordinary pages in .GOV corpus using the title length feature. Titles are the 'names' of the pages and contain significant information of the pages. From this figure, we can see that there is about 28 percent of the ordinary corpus which has nothing in the title fields (title length values 0). However the value is only 11% in the retrieval target page set. One of the main reasons may be that the page makers pay more attention to the high quality page than ordinary one, and design the title text by himself instead of generating it automatically which have nothing or "untitled" or "untitled document" in the title fields.

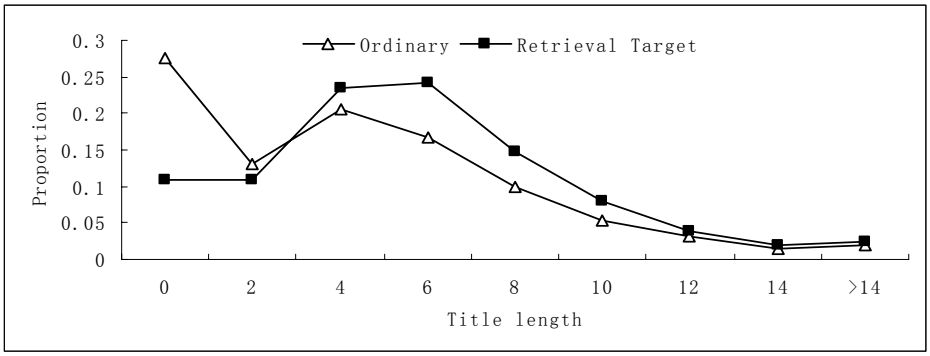


Fig 1 The different title length distributions of retrieval target and ordinary pages in .GOV corpus

Based on the statistical analysis mentioned above, we use these features in our new algorithm, which will be introduced in the next section, for estimating the web page quality.

4. The page quality estimation algorithm based on linear classification

Linear classifiers are the simplest ones as far as implementation is concerned, and directly related to many known techniques such as correlations and Euclidean distances. However, in the Bayes sense, linear classifiers are optimum only for normal distributions with equal covariance matrices. So, in many

applications of pattern recognition or classification, the assumption of equal covariance is not appropriate. Many attempts have been made to design the best linear classifiers for normal distributions with unequal covariance matrices and non-normal distributions. Of course, these are not optimum, but in many cases the simplicity and robustness of the linear classifier more than compensate for the loss in performance [22].

Here, we can clearly get neither the equal covariance, nor the normal distributions of the high quality or the low quality pages, so the linear classifier is not optimum if we use it to process the problem of web page quality classification. However, the method we propose based on the linear classification can make the page estimation process simply and easily. Since there are not the negative samples to learn in the page estimation process, we cannot design the general linear classifier to learn and process directly. Using the statistics methods, a new linear discriminant function is designed for web page quality estimation, according to which the page quality is calculated and compared. In the next part of this section, we will discuss how new method based on the linear classification is designed using the Euclidean distances from general linear classifier. At last, we will get the discriminant function for the web page quality estimation which we want.

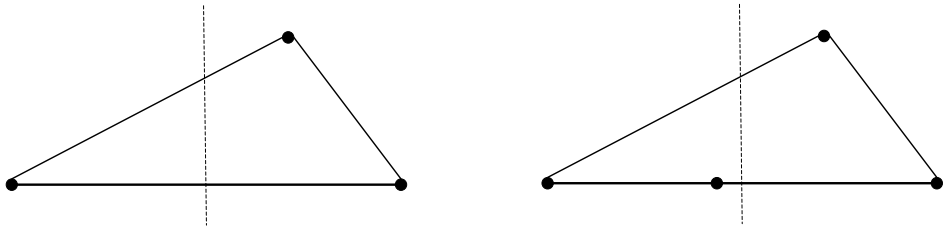


Fig 2 A linear classifier based on the Euclidean distance

Here, X is a vector, which is used to present a page and grouped by the values of features mentioned in Section 3. From the left part of Fig 2, if we have known the expected vector of the positive samples (M_+) and negative samples (M_-), we can use the distances between sample X and $M_+ | M_-$ to classify X. In the Euclidean space, we can get the middle dashed line as the decision boundary which we can use to classify the samples. If we have $d_{XM+} < d_{XM-}$, namely X on the right of the dashed, X will be judged into the positive one. Otherwise X will be judged into the negative one. In addition, the discriminant function $g(X) = d^2_{XM-} - d^2_{XM+}$ can be used to judge the class of X: if $g(X) > 0$, then X is judged into the positive one, otherwise X is judged into the negative one. This function gives the matrices: the bigger value of $g(X)$ is, the more possibility of that X is judged into the positive one. We can use the definition of the distance to rewrite the function $g(X)$ as:

$$\begin{aligned} g(X) &= d^2_{XM-} - d^2_{XM+} = (X - M_-)^T g(X - M_-) - (X - M_+)^T g(X - M_+) \\ &= 2(M_+ - M_-)^T gX + \| M_- \|^2 - \| M_+ \|^2 \end{aligned} \tag{1}$$

In equation (1), $\| M_- \|^2 - \| M_+ \|^2$ is the difference between the distance from the negative center to the origin and the one from the positive center to the origin. If the web page corpus is given, then $\| M_- \|^2 - \| M_+ \|^2$ is a constant. We rewrite (1) as:

$$g(X) = 2(M_+ - M_-)^T gX + C_1 \tag{2}$$

In equation (2), $(M_+ - M_-)$ is a vector which point from the negative center to the positive one. M_+ can be estimated using the retrieval target pages which we have. However, we don't have the low quality pages to estimate M_-. Luckily, we have the corpus of all pages, and the center of all pages can be calculated. We also know that the three centers, M_{all}, M_+ and M_-, are on the same beeline(see the right part of Fig 2). So, the direction of the vector $(M_+ - M_-)$ can be obtained only by the M_{all} and M_+, namely $W = \frac{(M_+ - M_{all})}{\| (M_+ - M_{all}) \|}$ and $(M_+ - M_-) = \| M_+ - M_- \| W$, which $\| M_+ - M_- \|$ is a positive constant. We rewrite (2) as:

$$g(X) = C_2 W^T gX + C_1 \quad which \;\; C_1 \;\; and \;\; C_2 \;\; are\ constants,\ and \;\; C_1 > 0 \tag{3}$$

From equation (3), we know that we cannot get the value of $g(X)$ as the value of M_- is not known. However, if we just compare the values of $g(X)$ in a given web page corpus, the constants, C_1 and C_2, can be ignored, which wouldn't affect the comparative results, so we only need to pay attention to the value of $W^T gX$. We use a new symbol $f(X)$ to rewrite the discriminant function:

$$f(X) = W^T gX \quad where \;\; W = \frac{(M_+ - M_{all})}{\| (M_+ - M_{all}) \|} \tag{4}$$

We can use anchor equation which is equivalent to (4):

$$f(X) = (M_+ - M_{all})^T gX \tag{5}$$

Equation (5) is a discriminant function which can map a sample *X* (or a web page) to a numerical value which can be used to compare the page quality, namely gives each page a score. In a fixed corpus such as .GOV/SOGOU, we can use $f(X)$ to compare any two pages of the corpus or rank the pages according to the page score. $f(X)$ is called as Linear Page Estimation function (LPE).

Equation (5) has a directly mean, a sample *X* is projected onto the line generated by the center of all samples M_{all} and the positive center M_+. From equation (5), we know that the training step is to calculate the values of M_{all} and M_+, and the testing step is to scan the data set (*X*) once. So the complexity of training is N+M, and the complexity of testing is N, if the dataset has N pages and M positive samples.

5. Experiments and Discussions

5.1. Page Quality Estimation

To our knowledge, there has been little research towards the evaluation of web page quality estimation. In our previous works, Liu proposed several methods to evaluate the web page cleansing [7][15] and web key resource page judgment [5][6]. The retrieval target page recall ratio was used in [5][6], which should be given a fixed portion of the high quality pages or the web key resource pages in the corpus. Though this simple method can prove the effectiveness of a certain method, it has difficulties in telling which method has a better performance. Actually, there is a tradeoff between size and high-quality page recall. In [15], Liu proposed a new metric called High-quality Page Average Recall (AR) which is mean of the recall scores after each page counted.

$$AR = \sum_{i=1}^{\#(TestCollection)} Recall(i) \Big/ \#(TestCollection) \text{ [15]} \quad (6)$$

We can rewrite the AR expression based on the definition of AR more strictly (see equation 7). In equation (7), p is the proportion of the high quality pages in all data corpus.

$$AR = \int_0^1 Recall(p)dp \quad (7)$$

In Section 3, we have introduced the .GOV corpus, SOGOU corpus and some high quality page set. Here, we use about half of the high quality pages in SOGOU corpus for training the page quality estimation algorithm and the others for testing. For .GOV, we use the TREC 2002-2003 Web Track answers for training, which contained the TD and NP tasks[12][13], and TREC 2004 Web Track for testing, which is a mixed task with TD, NP, HP queries [3].

First, we use the retrieval target page recall to evaluate the algorithm. Table 4 shows the high quality corpus size and its corresponding retrieval target recall for .GOV and SOGOU corpuses. From the statistics in Table 4, we can see that we can retain most retrieval target pages and reduce corpus size. About 95% pages are reduced by the high quality page estimation process while over 92% retrieval target pages are retained for SOGOU corpus. However, for .GOV corpus, it needs about 50% pages be retained while about 96.8% retrieval target pages are left. It can be explained by the fact that the page quality of .GOV corpus is much higher than SOGOU corpus. The recall in training set is only about 0.83 which is lower than the testing set's for the reason that the page quality in training set, which only contains the TD and NP pages, may be lower than the testing set containing the TD, NP and HP pages using our features and algorithm.

Table 4 High Quality Corpus Size and Target Page Recall

Corpus	High Quality Corpus Size(Percentage of original set)	Retrieval Target Page Recall(Training set)	Retrieval Target Page Recall (Testing set)
.GOV	50%	83.18%	96.84 %
SOGOU	5%	92.19%	92.26%

Fig 3 shows the high quality page size / recall curves. According to the definition of High Quality Page Average Recall (AR), we can know that the AR value is just the area that the curve surrounds. The dashed line show the recall when the page is picked randomly with out using our algorithm. From the figure, we can see that the area surrounded by the real curve is much bigger than the area surrounded by the dashed line, which just means the effectiveness of our quality page estimation algorithm.

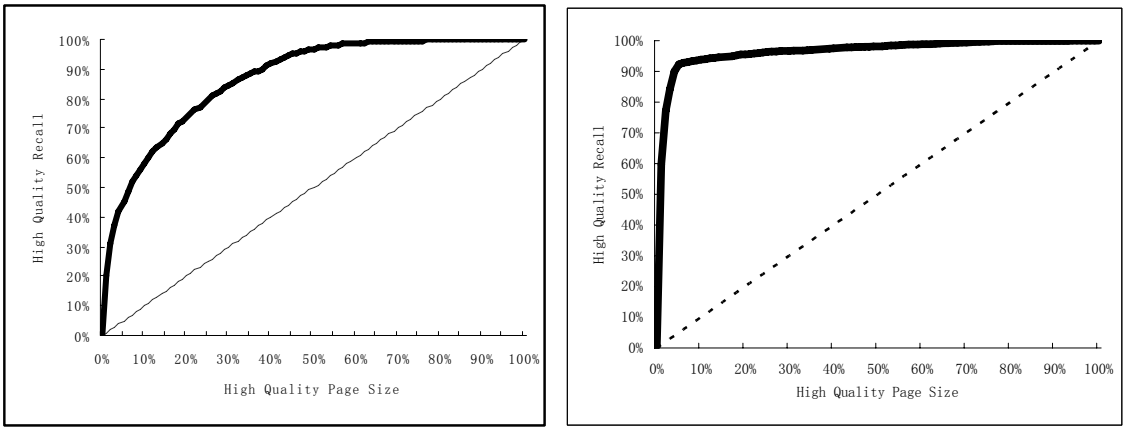


Fig 3 Size of high quality page and its corresponding retrieval target page recall curves on the .GOV (left figure) and SOGOU (right figure) corpus

When we know the recall value of each high quality page size, we can calculate the AR values which is shown in the Table 5. From the table, we can see the AR value of the .GOV is lower than the one of SOGOU, which means the quality of .GOV is higher than SOGOU's, stated many times.

Table 5 The AR value on .GOV and SOGOU corpus

Corpus	Average Recall
.GOV	0.8707
SOGOU	0.9667

Fig 4 shows the AR value of the different page type based on the mixed query task in the TREC 2004 Web Track, which contains three query types, HP, NP and TD [3]. The HP has the highest AR value (0.959), but the NP has the much lower AR value (0.783), which means that the homepage can be distinguished easily using our algorithm and all of the current features, however, it is harder for the named page.

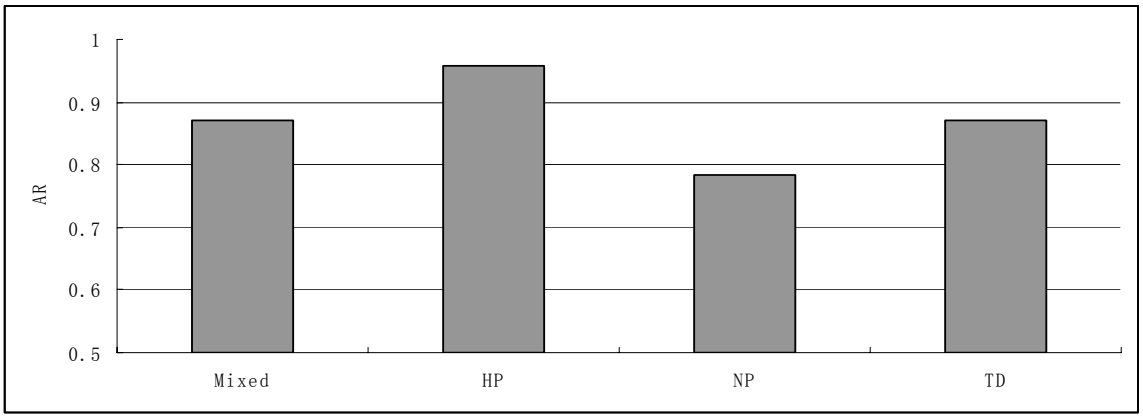


Fig 4 AR value on .GOV (Mixed, HP, NP, TD) and SOGOU corpus

5.2. Effectiveness of Query-independent Features

In this part, we analysis the query-independent features and find out which one is the most important or effective for our web page estimation algorithm. We use AR to evaluate our algorithm with different features. Fig 4 shows the AR values of each feature.

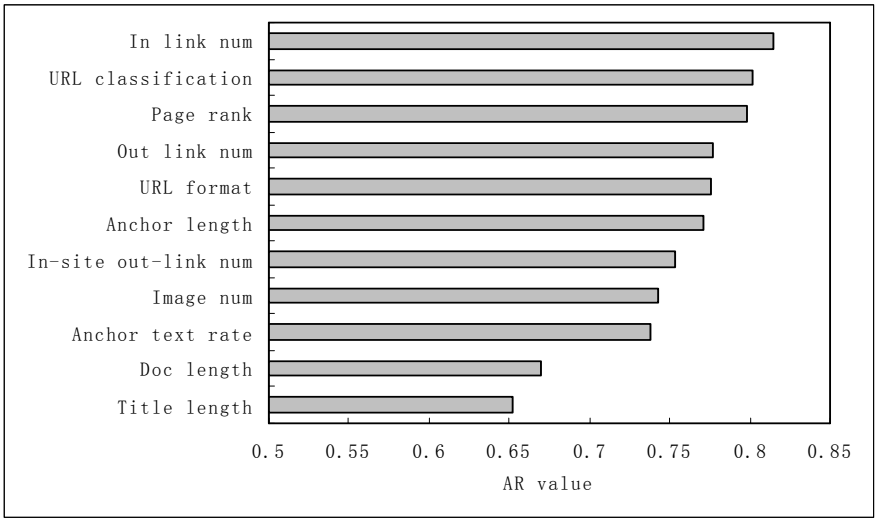


Fig 4 The effectiveness of every query-independent feature in .GOV corpus

We can see from Fig 4 that the features, In-link number, URL classification and Page rank, play the most important roles in the page quality estimation on .GOV corpus. But if we use one feature, even the most important feature, the AR value is only 0.815, which is far from the AR value with all features (0.871). It suggests that the other features should not be discarded in the process of the page quality estimation using our algorithm.

5.3. Comparison with Improved Decision Tree Algorithm

Here, we compare our Linear Page Estimation algorithm (LPE) with the Improved Decision Tree algorithm (IDT) proposed by Liu in [5][6] and Bayes learning algorithm in [15]. The IDT, which requires the knowledge of positive example proportion in the corpus before learning, is difficult to train automatically and change the positive example proportion flexibly. For both the IDT and Bayes learning, the attribute value should be dispersed into Boolean, which may lose the real value information. The distribution of each attribute value should be observed to find out the appropriate threshold for the discrete process in advance. Here we use the Improved Decision Tree trained with five features (page length, in-link number, URL classification, in-site out-link anchor text ratio, in-site out-link number) in [5] due to the same corpus. The size of high quality pages is come from the result of the Improved Decision Tree, because it is difficult to fix the proportion for the Improved Decision Tree exactly before learning and testing, which is one of the main disadvantages of the IDT.

Fig 5 gives the results with different methods and different features. From the figure, we can see that the result of the Improved Decision Tree is higher than the LPE's with five features. This may be explained by the reason that the features used in IPE have the same weight, which can be chosen by the information gain in IDT, and the factor of the projection line is not the optimization for the distribution is unknown and the positive center is estimated by the samples. However LPE can add more features easily, which have some difficult for IDT, and the result is higher (LPE with all features). Comparing the LPE and Bayes learning algorithm with all features, we can see that LPE has higher recall than the one of Bayes learning, which may be explained by the reason that in the discrete process into Boolean, the information of feature value may be lost.

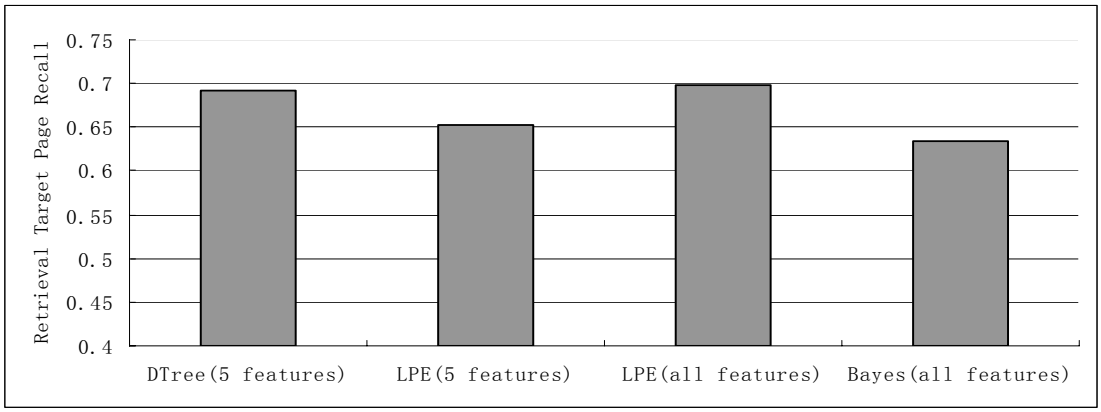
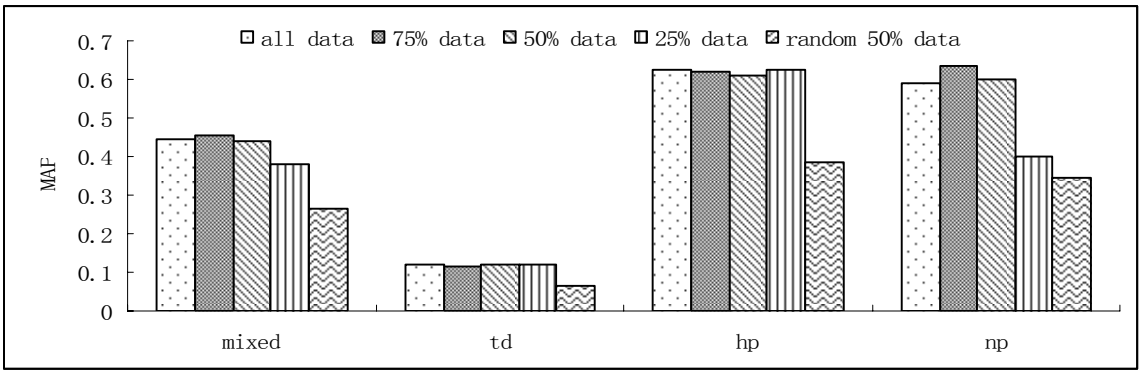


Fig 5 Comparison between different algorithms in .GOV corpus

5.4. Retrieval on Different Size of High Quality Pages

One of the main goals of the high quality page estimation is to use part of the web page corpus to build the index and retrieval which can meet more users' needs with limited storage and computation resources. So, we do the retrieval experiments with different sizes of .GOV corpus to see whether our algorithm can improve the retrieval effectiveness or get the close results with much less documents. The queries and results of Trec 2004 Web Track are used to take the retrieval experiments, which are divided into three types, topic distillation, named page and home page, and contain 225 queries, 75 queries for each type. We experiment and analysis the performance of retrieval on the three different types and the mixed one which contain all of the 225 queries. The data sets are selected by our algorithm according to the value of LPE, which contain 75% data, 50% data and 25% data. We also use the whole set and 50% data selected randomly for comparison. The MAP and P@10 metrics are used to evaluate the performance, which are always used in retrieval evaluation. See the results in Fig 6.



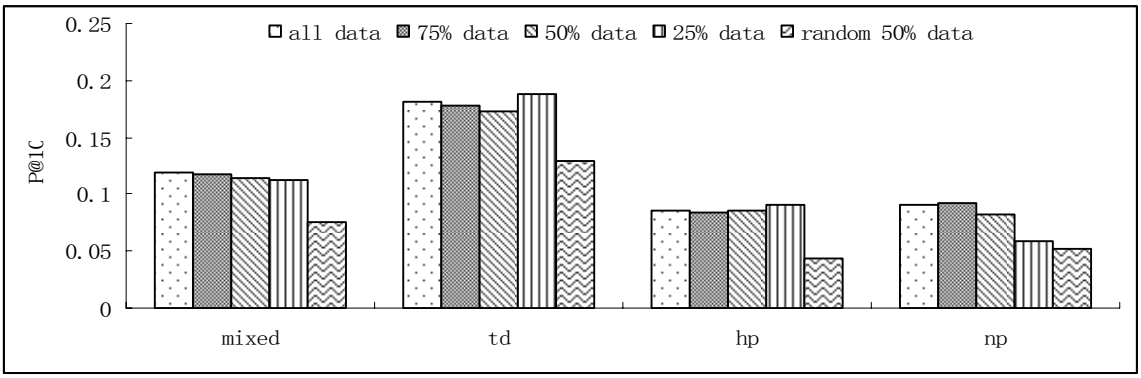


Fig 6 The performance of the retrieval with different size of .GOV corpus (all, 75%, 50%, 25%, random 50%), evaluated by MAP and P@10

According to Fig 6, we can see the performances of different search task with different data size. First, the performances on the part data set (75%, 50%, 25%) selected by our algorithm are close to or even better than the one on the whole collection, and the one on the half data set selected randomly is much lower than the ones on the other data sets, even the 25% data selected by our algorithm, which shows the effectiveness of page quality estimation process. Second, our algorithm shows perfectness for TD and HP search task. However, the performance drops quickly while the data set becomes smaller for NP search task (See the NP search result on data 25%). It is can be illustrated by Fig 4 that our algorithm can not easily distinguish the named page using current features as the home page or the page for topic distillation.

6. Conclusions and Future Work

In this paper, an algorithm of the web page quality estimation (LPE) is proposed based on the Linear Discriminant Function, which can give a score to each page. Using LPE, it is possible to reduce web data size significantly, and retains the most high quality pages for other usage, such as information retrieval. At the same time, the features, common-used and newly-proposed, are studies to exploits the differences between the high quality pages and ordinary pages on the large scale web page corpus.

In the near future, we hope to extend this work to include other machine learning algorithms and new effective features. The pages can be estimated in some special area, not only general page quality, but also personalized web search or vertical search. We also plan to work on a hierarchy storage model for Web IR tools according to page score given by the page quality estimation process.

Reference

[1] Sullivan, D. (2005). Search Engine Sizes. Search engine watch Web site articles. Retrieved December 10, 2005, from http://searchenginewatch.com/reports/article.php/2156481

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

[3] N. Craswell, D. Hawking. Overview of the TREC 2004 Web track. In E. M. Voorhees and Lori P. Buckland, eds. NIST Special Publication 500-261: The Thirteenth Text REtrieval Conference (TREC 2004). Washington: Department of Commerce and National Institute of Standards and Technology, 2004.

[4] Henzinger, M.R., Motwani, R. & Silverstein, C. (2003). Challenges in Web Search Engines (pp. 1573-1579). Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Press.

[5] Liu Y., Zhang M., Ma, S. (2005). Web Key Resource Page Judgment Based on Improved Decision Tree Algorithm, Journal of Software, Vol.16(11)

[6] Liu Y., Zhang M., Ma, S. (2004). Effective Topic Distillation with Key Resource Pre-selection (pp. 129 - 140), Lecture Notes in Computer Science, Volume 3411, London: Springer-Verlag, 2004.

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

[8] Brin S. & Page L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. Proceedings of the Seventh World Wide Web Conference (WWW7), Brisbane, also in a special issue of the journal Computer Networks and ISDN Systems, Volume 30, issues 1-7.

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

[10] Henzinger, M.R., Motwani, R. & Silverstein, C. (2003). Challenges in Web Search Engines (pp. 1573-1579). Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence. San Francisco: Morgan Kaufmann Press.

[11] Craswell N, Hawking D. Query-Independent evidence in home page finding. ACM Trans. on Information Systems (TOIS), 2003,21(3):286-313.

[12] Hawking D, Craswell N. Overview of the TREC-2002 Web track. In: Voorhees EM, Buckland LP, eds. NIST Special Publication 500-251: The 11th Text REtrieval Conf. (TREC 2002). Washington: Department of Commerce, National Institute of Standards and Technology, 2002.

[13] Hawking D, Craswell N. Overview of the TREC 2003 Web track. In: Voorhees EM, Buckland LP, eds. NIST Special Publication 500-255: The 12th Text REtrieval Conf. (TREC 2003). Washington: Department of Commerce, National Institute of Standards and Technology, 2003. 78-92.

[14] TREC, http://es.csiro.au/TRECWeb/govinfo.html, 2002.

[15] Y. Liu, M.Zhang, L. Ru, S. Ma. The 1st China-Kyoto Student Workshop on Digital Content and Web Computing. Data Cleansing for Web Information Retrieval using Query Independent. Being, China. 2006.

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

[17] 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.

[18] Denis, F. (1998). PAC Learning from Positive Statistical Queries (pp.112-126). Proceedings of the 9th international Conference on Algorithmic Learning theory . M. M. Richter, C. H. Smith, R. Wiehagen, & T. Zeugmann, Eds. Lecture Notes In Computer Science, vol. 1501. London: Springer-Verlag, 1998.

[19] Manevitz, L. M. & Yousef, M. (2002). One-class svms for document classification. J. Machine Learning. Res. 2: 139-1.

[20] Hawking, D. & Craswell, N., (2005). Very Large Scale Retrieval and Web Search, in Ellen Voorhees & Donna Harman (Ed.), TREC: Experiment and Evaluation in Information Retrieval, MIT press, 2005.

[21] Kraaij W, Westerveld T, Hiemstra D. The importance of prior probabilities for entry page search. In: Ricardo BY, ed. Proc. of the 25th Annual Int'l ACM SIGIR Conf. on Research and Development in Information Retrieval. New York: ACM Press, 2002. 27-34.

[22] K. Fukunaga, Introduction to Statistical Pattern Recognition, Second Edition, Academic Press, New York, 131-153.