The Ninth International Conference on Web-Age Information Management

Weighting Links Using Lexical and Positional Analysis in Web Ranking [*]

Yi Zhang,Yexin Wang, Lidong Bing, Yan Zhang [†]
Key Laboratory of Machine Perception (Ministry of Education)
Peking University
100871 Beijing, China
{zhangyi,wangyx,bingld,zhy}@cis.pku.edu.cn

Abstract

Link analysis has been widely used to evaluate the importance of web pages. Popular link analysis algorithms are mainly based on the link structure between pages. However, a web page usually contains various links such as for navigation, decoration or nepotism, which are irrelevant to the topic of the web page and can not reflect the actual voting relations between pages. In order to improve the performance of web ranking, we bring out one filtering algorithm to recognize and eliminate these unrelated links using Content Lexical and Positional analysis. Experimental results on different web domains show that our filtering model can efficiently detect the irrelevant links and effectively help to build a good link graph for the ranking calculation.

1. Introduction

Recently, global search engines such as Google and AltaVista have been widely used to retrieve web information. According to [2], nowadays web users spend totally 13 million hours per month interacting with Google alone, which shows their great importance in our daily life.

These search engines mainly adopt the ranking techniques on the analysis of link structure between pages. Link analysis algorithms, which convey the relative importance of web pages by exploiting the web link structure, effectively help web search engines to evaluate the quality of the pages and rank the correlative results. There are many well known link analysis algorithms, such as PageRank [6, 19] and HITS [18, 4]. Many extensions to these two algorithms are also proposed, such as [5, 22, 8, 16, 14, 24].

The traditional PageRank [6, 19] corresponds to the standing probability distribution of a random walk on the web graph which the web viewer simply keeps clicking on successive links at random or gets tired and jumps to another page. Then each page is assigned a score of importance. In HITS algorithm [18, 4], each web page has both a hub score and an authority score. The hub score of one page is the sum of all the authority scores of pages that are pointed by it and its authority score is the sum of all the hub scores of pages that point to it. Finally, each web page could obtain the authority and hub scores separately, which can reflect its importance.

All these link analysis algorithms are based on two assumptions. Firstly, a link conveys human endorsement. If there exists a link from page A to page B and these two pages are authored by different people, then probably the first author think the second page is valuable. Thus the importance of a page can be propagated through the link. Secondly, pages that are co-cited by a certain page are likely related to the same topic. However, these two assumptions do not hold in many cases nowadays because the change of the link structure. A web page usually contains various types of links, which are irrelevant to the topic of the web page and can not indicate the actual voting relations between pages.

Some people have noticed the fact that a single web page often contains multiple links in different blocks of the page layout, which can not deliver the importance between pages and put forwards several algorithms to solve the problems[7, 10, 15, 20, 23, 21, 12]. These methods mainly filter the irrelevant links after analyzing the page content or the page layout. However, without considering the lexical information, these algorithms seem low efficient and do not effectively remove the irrelevant links when dealing with a large amount of web pages.

In the paper, we propose a link analysis algorithm based on lexical and positional analysis. Firstly,we extract the keyword candidates from the words sets using the lexical knowledge. Then we treat the content of the web page hierarchically into several levels referring to their positions. We use the non-linear weighting function to compute the similarity between the pages and determine whether the page

[*] Supported by National Natural Science Foundation of China (with Grant No.60673129 and 60573166) and Beijing Natural Science Foundation under Grant No.4073034

[†] Corresponding author

978-0-7695-3185-4/08 $25.00 © 2008 IEEE
DOI 10.1109/WAIM.2008.75

IEEE computer society

is relevant to its parent's to remove the irrelevant links. Finally, we implement the PageRank algorithm on the new web graph. Based on the filtering model, the new link analysis algorithm is capable of removing most of the unrelated links and greatly decreasing the calculation cost.

The rest of this paper is organized as follows. Before describing the Content Lexical and Positional filtering algorithm (in the following discussion, we call CLP for short), we first present some definitions and statistical results on the changes of the link structure nowadays in Section 2. Section 3 represents the related work. And in Section 4, we elaborate each step of the ranking framework and describe how to build the cleaned web graph for ranking. Section 5 evaluates the method based on experimental results. Finally, we give concluding remarks and future work in Section 6.

2. Preliminaries

2.1. Some Definitions

We group the links into 4 main types, which are defined according to their functional characteristics in presenting one web page. They are human-edit link, advertisement link, navigational link and nepotistic link.

Definition 1: Human-Edit Link

A link that is generated artificially by humans according to the content relevancy between two pages. Most of this kind of link could represent the recommendation from one page to another page.

Definition 2: Advertisement Link

A link that is created dynamically in order to bring a profit to a certain commercial organization. This kind of link takes up a large proportion in web pages nowadays, especially in .com web sites and they do not endorse much to the page's value. Therefore, we should not consider them in the ranking computation.

Definition 3: Navigational Link

A link that is designed to facilitate the user randomly browsing in one site. It can not deliver strong recommendations between pages.

Definition 4: Nepotistic Link

A link between pages that is present for reasons other than merit. In this paper, we classify the links such as spam links to this type.

Definition 5: Relevant Link and Irrelevant Link

Because only the Human-Edit Link could reflect the voting relations between pages and actually do contributions in the ranking calculation, we group it to be the relevant link and the other three types of links are considered to be the irrelevant links. Fig.1 shows the link distribution in a Chinese Yahoo page. Part 4 is the human-edit link area, where the links should be preserved as the relevant links. Part 1, 2, 3 are respectively areas of navigational links, nepotistic links and advertisement links. These three types of links are typically undesirable in content analysis, so it is necessary to eliminate them before the calculation.

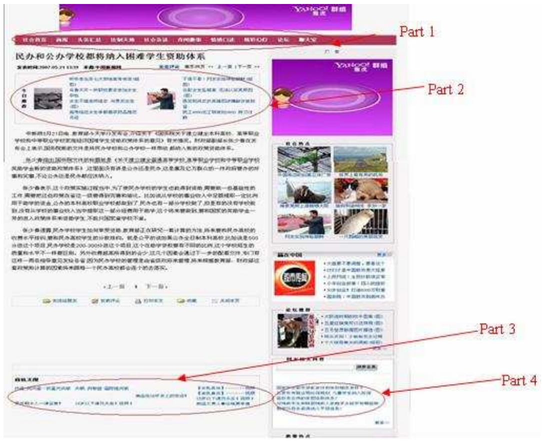

Figure 1. Link Distribution on a Page

2.2. Changes of Link Structure

The link number of web pages increases greatly these years. There are three main reasons. Firstly, there is a higher demand for web information nowadays and the number of web pages expands correspondingly. Until 2005, there are more than 2,601,901,000 Chinese web pages [3], which is more than twice of that in 2004. Secondly, the tools that can generate links automatically are popularly used by web masters. Finally, due to the broad use of search engines in guiding today's web traffic, some people who are blinded by gain aim at making their pages rank highly by playing with web page features that the search engines' ranking algorithms base on. Link spamming [11], one of the famous spamming techniques, is performed by adding a large number of in-links to one page or making the pages point to each other mutually to form a spam farm. These links could terribly deteriorate link-based ranking algorithms and lead to bad ranking results.

Seen from the distribution of different domains in Chinese 2,601,901,000 web pages [3], commercial site number takes up the largest part, accounting for more than 60% of all the sites. We randomly select 321 pages from today's web set and count the outlink number of each page. We find that the outlink number of these pages averagely reaches 59. And it is extremely more in the commercial pages, which is up to 62. The outlink number in different domains can be seen in Fig. 2.

Moreover,we humanly judge whether each of their outlinks is a relevant link and find that only averagely 11.75%

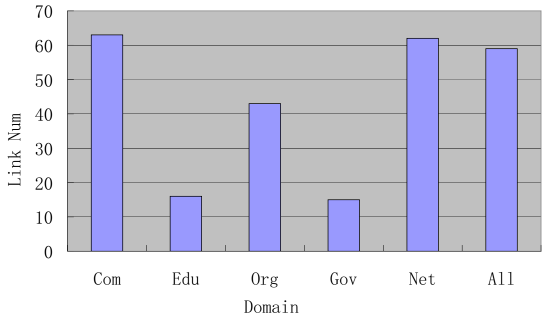

Figure 2. Average Outlink Number of Different Domains

of links are relevant and should be kept in the ranking calculation (See Fig. 3). The fact explains the requirement to eliminate the noisy links first and purify the link graph to rank the pages on the true recommendation linkage.

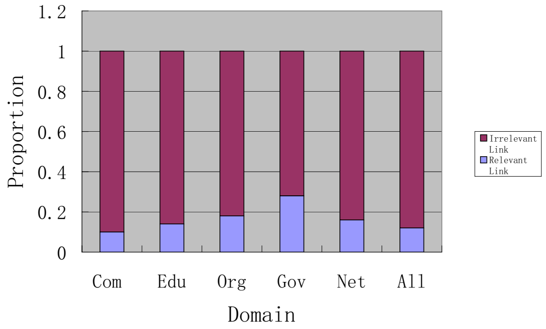

Figure 3. Relevant and Irrelevant Link Distribution on Different Domains

3. Related Work

For years, there has been a great deal of work on link analysis [6, 19, 18, 4, 5, 22, 8, 16, 14, 24]. All link analysis algorithms are based on the consideration that web is treated as a graph, with the pages represented by nodes and the links between pages functioning as the edges, connecting the nodes. The number and quality of edges between nodes usually represent relevancy or endorse some authority.

Kleigberg proposes a ranking algorithm of finding authoritative pages on a given topic [18, 4]. The algorithm is known as HITS.It assigns two scores to a page, authority score and hub score. Hubs and authorities exhibit a mutually reinforcing relationship.PageRank [6, 19] simulates a random walk on the link graph and assigns each page a score of importance. HITS and PageRank are two of the most popular algorithms in ranking web pages.

Block-level link analysis begins to utilize the web page'structure and each page is partitioned into blocks using the vision based page segmentation algorithm [7]. By extracting the page-to-block, block-to-page relationships from link structure and page layout analysis, they construct a semantic graph over the WWW such that each node exactly represents a single semantic topic. This graph can better describe the semantic structure of the web. However, lack of content analysis partially influences the filtering performance of unrelated links.

Nicholas Kushmerick have described the ADEATER system, a browsing assistant that automatically learns advertisement-detection rules, and then applies those rules to remove advertisements from Internet pages during browsing [15]. In controlled experiments, ADEATER can achieve a very high level of accuracy. But it is only effective to remove advertisement links.

Brian D. Davison explores some of the issues surrounding the question of what links to keep, and reports high accuracy in initial experiments to show the potential after using a machine learning tool to automatically recognize the nepotistical links [10]. However, the method only considers several features to remove irrelevant links. It does not deal with the links unrelated to the topic of the page.

Some Chinese researchers also begin to apply the positional analysis to the web situation. Haifeng Li has put forward two models, which show a better performance than the traditional vector space model[12] in the paper retrieving.

4. CLP

In the following, we will describe our CLP ranking framework in details (see Fig. 4). Since Chinese words are more challenging due to their homonymy and uncertainty, we use Chinese in our experiments. Five main steps constitute the framework: namely, Main Content Extraction, Chinese POS Tagging, Positional Analysis, Relevant Model Calculation and Ranking.

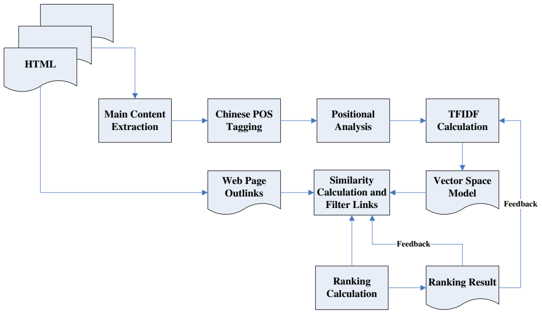

Figure 4. The CLP Ranking Framework

4.1. Main Content Extraction

Web pages usually contain many cluttered information including navigation bars, related readings, advertisement links, copyright notices, responsibility statement and timestamps. Such information items are functionally useful for human viewers and necessary for the Web site owners. However, they are not informational for identifying the topic of the page and worse can cripple the performance of content analysis. Therefore, we use ridge algorithm [25]to extract the primary content of the web page.

In this algorithm, a web page is treated as a sequence of content cells, where each cell owns its score according to the Mountain Model [25]. Primary content cells are distinguished from those cluttered content cells by the features processed only by primary cells. The algorithm is site-independent and both the accuracy and time efficiency could satisfy our need.

4.2. Chinese POS Tagging

Losee finds that web users mostly query nouns or noun phrases to search engines after investigating the search engine "Spies" [17]. Besides, Chowdhury has proved that building the index only on nouns can greatly improve the search efficiency [9]. Then, it can be assumed that words with different syntactical functions vary in expressing the topic of the web page and POS Tagging shows a great potential to enhance the retrieval performance.

Therefore, we introduce the POS tagging technique after obtaining the primary content. Chinese part-of-speech (POS) tagging assigns one POS tag to each word in a Chinese sentence. However, since words are not demarcated in a Chinese sentence, it requires word segmentation as a prerequisite. We perform Chinese POS tagging strictly after word segmentation. Then we assign POS tags on a word-by-word basis, making use of word features in the surrounding context according to the Chinese dictionary.

4.3. Positional Analysis

Even the primary content can express the page's topic to some extent. However, the words in title and in the last sentence of the body are absolutely relevant to the topic in different degrees. Therefore, it is necessary to weigh the words in different positions dissimilarly.

Until now, nobody has given a scientific solution on how to assign authorities to the words according to the positions in one article, even less for a web page. Referring to the statistics by [13], a web page can be partitioned in 12 informational areas, which are web page title (Title), article title(Bt), Html tag(Html), the first sentence in first(Ds1),second (Ds2),third (Ds3)paragraph, the last sentence in first(Dw1),second (Dw2),third (Dw3)paragraph, the other sentences except for the ones mentioned before in first(Sd),last(Wd) paragraph and other paragraphs(Qt). The proportion among them probably is :

Bt:Html:Sd:Ds1:Title:Dw1:Qt:Wd:Ds2:Dw2:Ds3:Dw3 = 5:5:5:4:4:4:2:2:2:2:2:2

From the statistical expression above, it is obviously seen that contents in some positions can not convey as much information as others. Weighing all the primary content equally in a page is not suitable. Therefore, after the main content has been segmented and tagged, we group the words in several bags according to their positions, which can express the importance to the topic of the page differently.

The proportion we mentioned above can be treated as the initial value for each words bag. And the weight will be regulated according to the ranking performance dynamically.

4.4. Relevant Model Calculation

We eliminate the links according to their content relevancy to the parent's. Each web page can be conveniently represented in several high-dimensional vectors for different positions where the $TFIDF$ value of each salient tagged word in that position is considered as one feature.

Phrase Frequency/Inverted Documents Frequency which is usually called $TFIDF$ of each term is calculated in accordance with the definition of Equation 1 to decide which words should be chosen as the better candidates of salient phrases for each page in different position. Intuitively, more frequent phrases are more likely to be better candidates of salient phrases, while phrases with higher document frequency might be less informative to represent a distinct topic.

Phrase Frequency / Inverted Document Frequency

$$TFIDF = f(w) \cdot log \frac{N}{|D(w)|} \tag{1}$$

where f represents term frequency calculation, N is the number of the documents in the dataset and D expresses document frequency.

After we obtain the vectors for positions with M terms according to their $TFIDF$ values, we use the vector-space model and the metric defined by the cosine coefficient, which is the cosine of the angle formed by the vector-space representation of the two connected pages P and Q on that position. (See Equation 2)

Vector Space Model

$$sim(P_i, Q_i) = \frac{P_i \cdot Q_i}{|P_i| * |Q_i|} \tag{2}$$

where P_i and Q_i are vectors for two linked pages on position i.

Then we can define similarity between two pages P and Q as a function $sim(P, Q)$. This function allows us to decide whether a link is relevant by measuring the similarity between the web page and the page it links after referring to the similarities on all the positions. Classical methods basically use the linear term weighting method to compute the similarity between two pages by simply linearly adding the similarities on each position together. (See Equation 3)

Linear Weighting Model

$$sim(P, Q) = \sum_{i=1}^{n} (\lambda_i \cdot sim(P_i, Q_i)) \qquad (3)$$

Where n is the number of considered positions, λ_i is the weight for position i and the similarity between P and Q is the linear sum of the similarities in all positions.

However, we observe that the similarity between two web pages is not strictly a linear sum of independent components. Through the analysis between the term frequency and the similarity, we find that the similarity between two pages does not increase linearly with the term frequency, especially on two different positions. The increasing curves for different positions vary because their importance to the topic is distinct. The similarity increasing rate with term frequency in title is much more sharply than that in body. Moreover, the similarity between two pages based on term frequency inclines to reach a stable value as the term frequency increases.

Therefore, we utilize the non-linear function to satisfy the characters of the non-linear phenomenon. The functions in Fig. 5 are always used in non-linear systems and we adopt them as the weighting functions in the calculation of different positional vectors. The similarity between pages is described in Equation 4.

$$f1(n) = \frac{n}{n+1}$$

$$f2(n) = \frac{e^n}{e^n + 1} \quad (Sigmoid\ Function)$$

$$f3(n) = 1 - e^{(-n)}$$

$$f4(n) = \frac{2}{\pi} \times \arctan(n) \quad \ldots$$

Figure 5. Some Usually Used Non-Linear Functions

Non-Linear Weighting Model

$$sim(P, Q) = \sum_{i=1}^{n} (\lambda_i f(sim(P_i, Q_i)) \cdot sim(P_i, Q_i)) \qquad (4)$$

Where $f(sim(P_i, Q_i))$ is the non-linear function to balance the weight for position i. f will be chosen differently for different positions. The meanings of other variables have been discussed in Equation 3.

4.5. PageRank

When given a threshold α, after filtering the irrelevant links according to their similarity, the web can be thought as a reduced graph with the pages as nodes and the links as edges. Every page has several forward links(out-edges) and back links (in-edges). We use the random surfer version of PageRank algorithm [4] to rank all the pages. The more back links the page owns, the more importance the page obtains. The PageRank of a page A is defined as follows:

$$PR(A) = \frac{\epsilon}{n} + (1 - \epsilon) \times \sum_{i=1}^{n} PR(T_i)/C(T_i) \qquad (5)$$

where ϵ is a damping factor, which is usually set between 0.1 and 0.2. In our paper, we use 0.15 that is usually adopted; n is the number of nodes of the whole page link graph; T_i points to A and $C(T_i)$ is the number of out-edges of page T_i.

4.6. Regulate Parameters

In the ranking framework, two groups of parameters need to be decided. Firstly, the weights λ_i on different positions should be assigned in similarity calculations. Secondly, the threshold α should be considered when deciding which link will be kept. Their initial values can be acquired by the statistical records. Then we introduce the feedback mechanism to dynamically train better values so that they can be regulated in accordance with the ranking performance.

5. Experiments

5.1. The Data Set

We conduct our experiments on the subset of CWT200g Data Set, the biggest Chinese web collection with 197GB web pages [1].The documents were crawled by TianWang Search Engine on Nov.2005. Every page in the collection has a “text /html” or “text/plain”MIME type received from the HTTP server response message. After cleaning the subset, the collection consists of 5,838,073 web pages from 5,846 sites.

5.2. Evaluation Metric

In order to measure the retrieval performance of CLP, we evaluate them in two parts. Firstly, we compare the ranking performance of CLP with PageRank when only using nouns or verbs in vector calculation and figure out their influences on ranking. Then we discuss the filtering rate of unrelated links and the performance when applying different POS words and threshold α to observe the impact of setting various parameters.

5.3. Experimental Results

During the computation of CLP, we firstly use function $f(1)$ in Fig.5 to weigh the title similarity between two pages and add other contents' similarities linearly. We set M=7 and each vector is expressed in 7 words with the top $TFIDF$ values. Considering the noises brought by extracting main content from the web page, we simply treat the positions into three levels, Title, Ds1 and other parts, whose initial content weights are mentioned in Section 4 and α is set 0. In ranking calculation, we set the damping factor $\epsilon = 0.15$ and the iterative precision 10^{-4}. If the distance between two adjacent iterative values for each page or site is equal or smaller than 10^{-4}, the calculation process will terminate.

5.3.1 Top 1000 Ranking URLs

We take top 1000 ranking URLs of each result to compare their retrieval performance. Fig. 6 tells their capability to recommend sites. CLP_N stands for CLP algorithm based on nouns and CLP_V is short for CLP using verbs. PageRank only introduces 7 sites in the top 1000 URLs while CLP averagely recommends more than 160 sites and it can be protected from the effect of link spamming to some extent.

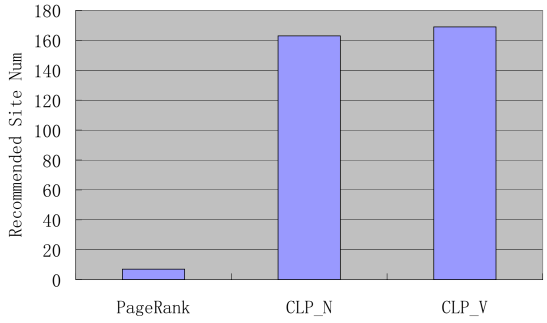

Figure 6. Variety in Sites Recommendation

Moreover, for each ranking result, we evaluate the distribution of URLs in their recommended sites. The distributions are shown in Fig. 7, Fig. 8 and Fig. 9 which are respectively in accordance with traditional PageRank, CLP_N and CLP_V. From the figures, it is obviously seen that CLP_V offers a more equal chance to each site while the URLs' distribution of PageRank fluctuates much heavier, which shows it can easily be deluded by some spamming sites.

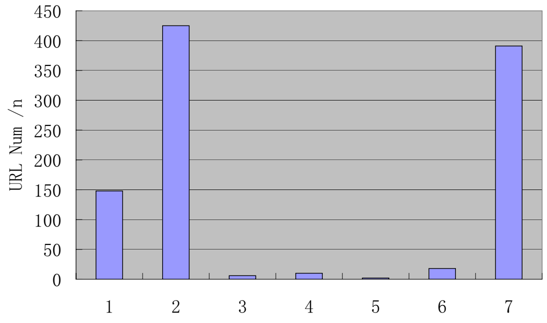

Figure 7. URL Distribution of PageRank

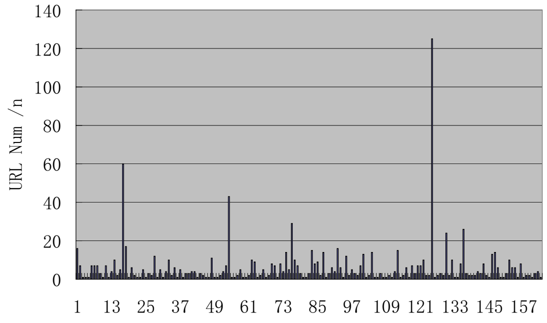

Figure 8. URL Distribution of CLP_N

Finally, we humanly label each top 1000 URLs on the site's aggregation and compare their spam rates. In the 7 sites recommended by PageRank, one site where 391 URLs come from is judged to be a spam site. Fig. 10 can suggest that the labeling results of CLP_N and CLP_V. $T1$ is the set containing sites appearing in CLP_N but not in CLP_V, while $T2$ includes the sites in CLP_V but not in CLP_N. From this figure, we learn that the average quality of sites in CLP_V looks better than that of CLP_N. We mainly ascribe its better performance to two reasons. Firstly, verbs can express the intention, the process or the result of one topic more explicitly than nouns. In addition, nouns are confusing sometimes due to their homonymy and uncertainty so that some irrelevant links are not removed when calculating the similarity. Experimental results also show that when setting α = 0, 67.15% of links are eliminated using CLP_V while in CLP_N, only 61.3% of links are removed. This illuminates that using verbs can remove more links than relying on nouns.

5.3.2 Filtering Rate of Links

This section tells us the relation between the filtering rate of links when using CLP_V after setting different thresholds α. Seeing from Fig. 11, we can notice that with the increase of α, the filtering rate of links increases slower. However, it sharply goes up when α=0.8. It can be deduced that the link

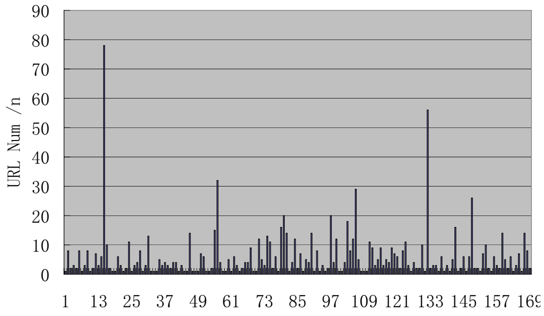

Figure 9. URL Distribution of CLP_V

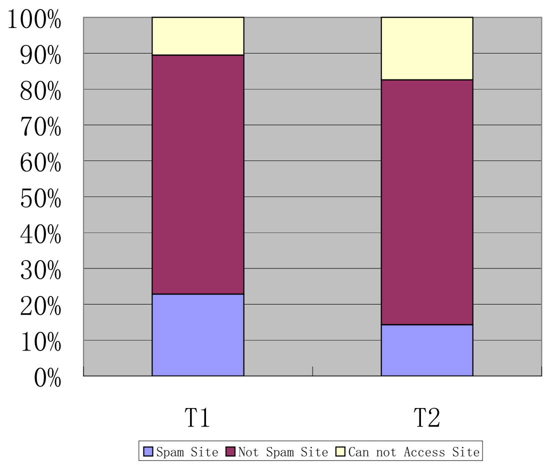

Figure 10. Comparison of CLP_N and CLP_V

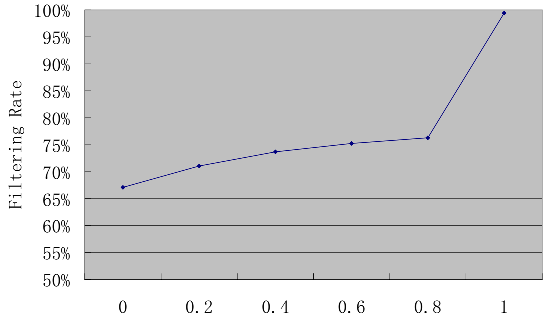

Figure 11. Filtering Rate on Different Thresholds

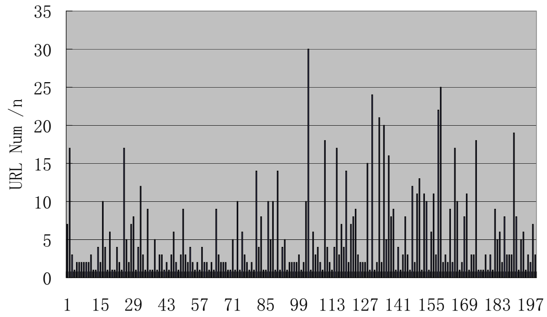

Figure 12. URL Distribution of CLP_V when α=0.8

pairs whose similarity is 0 take up the biggest ratio, 67.1% of all the link pairs and the link pairs that their similarity above 0.8 is secondly up to 21.7%. This fact shows that probably fewer than 20% of outlinks of one page are highly relevant, which provides another evidence to the statistical result in Section 2.

Furthermore, we compare each top 1000 URLs and find that the result seems better when we set α=0.8. It can recommend more good sites, show a better URLs' distribution (See Fig. 12) as well as decrease the spam rate. However, we can not promise that the ranking quality will be better when we set α=0.85. However, we can affirmatively say when comparing with the other 4 values, choosing α around 0.8 is more advisable.

6. Conclusion

Nowadays, most of the search engines take advantage of link analysis techniques to rank the web pages. All link analysis algorithms are based on one assumption: the links convey human endorsement. Nevertheless, the changes of the web structure have brought a negative effect on the ranking calculation. This paper, based on the solution of such a problem, puts forward an irrelevant link filtrating algorithm called CLP, which utilizes lexical and positional analysis with the combination of a non-linear weighting model. Experimental results show that this algorithm can effectively eliminate most of unrelated links and improve the ranking performance. Moreover, we also discuss the ranking influence when applying different POS words and thresholds. Search engines can benefit from the analysis.

In future work, we will emphasize on deeply investigating the relationship between ranking performance and the content's lexical, positional information so as to find a good combination to effectively deal with the vertiginous web.

References

[1] Chinese web test collection with 200 gb web pages. http://www.cwirf.org/.

[2] Nielsen netratings search engine ratings. http://searchenginewatch.com/reports/article.php/2156451.

[3] Web information resource investigation in china. www.cnnic.net.cn/download/2005/20050301.pdf.

[4] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 104–111, Melbourne, AU, 1998.

[5] a. W. H. Brian Amento, Loren Terveen. Does ”authority” mean quality? predicting expert quality ratings of web documents. In *The 23th Annual International ACM SIGIR Conference (SIGIR’2000)*, July 2000.

[6] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc.of 13th International World Wide Conference*, May 1998.

[7] D. Cai, X. He, J.-R. Wen, and W.-Y. Ma. Block-level link analysis. In *The 27th Annual International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR2004)*, July 2004.

[8] S. Chakrabarti. Integrating the document object model with hyperlinks for enhanced topic distillation and information extraction. In *WWW2001*, May 2001.

[9] M. M. Chowdhury A. Improving information retrieval systems using part of speech tagging. Technical report, Institute for Systems Research, University of Maryland, 1998.

[10] B. D.Davison. Recognizing nepotistic links on the web. In *AAAI-2000 Workshop on Artificial Intelligence for Web Search*, July 2000.

[11] Z. Gyongyi, P. Berkhin, H. G. Molina, and J. Pedersen. Link spam detection based on mass estimation. Technical report, Stanford University, 2005.

[12] Y. W. Haifeng Liu, Qian Wang. Research on the weighting of posiiton in text retrieval based on web situation. *Journal of The China Society For Scientific and Technical Information*, 25(3), 2007.

[13] H. Z. Hanqing Hou, Chengzhi Zhang. Research on the weighting of indexing sources for web concept mining. *Journal of The China Society For Scientific and Technical Information*, 1, 2001.

[14] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW2002*, May 2002.

[15] N. Kushmerick. Learning to remove internet advertisements. In *Proc. of 3rd International Conference On Autonomous Agents*, July 1999.

[16] R. Lempel and S. Moran. The stochastic approach for link-structure analysis (salsa) and the tkc effect. In *WWW2000*, July 2000.

[17] L. R. M. Natural language processing in support of decision-making:phrases and part-of-speech tagging. *Information Processing and Management*, 37(6), 2001.

[18] J. M.Kleinberg. Authoritative source in a hyperlinked environment. *ACM*, 46(5):604–622, 1999.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[20] W. Y. H. qianLI Jian-fu. Entropy-based link analysis algorithm for web structure mining. *Computer Engineering and Design*, 9, 2006.

[21] M. J. Soumen Chakrabarti and V. Tawde. Enhanced topic distillation using text, markup tags, and hyperlinks. In *Proc of the 24th annual international ACM SIGIR conference*, July 2001.

[22] P. R. Soumen Chakrabarti, Byron Dom and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *WWW1998*, Sept 1998.

[23] D. N. G. Suhit Gupta, Gail Kaiser. Dom-based content extraction of html documents. In *Proc of the 12th World Wide Web Conference New York*, May 2003.

[24] G. Xue, Q. Yang, H. Zeng, Y. Yu, and Z. Chen. Exploiting the hierarchical structure for link analysis. In *The 28th Annual International ACM SIGIR Conference (SIGIR’2005)*, August 2005.

[25] L. B. W. Zhang and H. Wang. Primary content extraction with mountain model. In *Proc of the IEEE CIT2008*, July 2008.