Exploiting PageRank at Different Block Level

Xue-Mei Jiang[1], Gui-Rong Xue[2], Wen-Guan Song[1], Hua-Jun Zeng[3], Zheng Chen[3], Wei-Ying Ma[3]

[1] Department of Information Management, Shanghai Commercial University,
2271 Zhongshan West Ave., 200235 Shanghai, P.R.China
cs_xmjiang@hotmail.com, swg@21cn.com

[2] Department of Computer Science and Engineering, Shanghai Jiao Tong University,
1954 Huashan Ave., 200030 Shanghai, P.R.China
grxue@sjtu.edu.cn

[3] Microsoft Research Asia, 5F, Sigma Center
49 Zhichun Road, Beijing 100080, P.R.China
{hjzeng, zhengc, wyma}@microsoft.com

Abstract. In recent years, information retrieval methods focusing on the link analysis have been developed; The PageRank and HITS are two typical ones According to the hierarchical organization of Web pages, we could partition the Web graph into blocks at different level, such as page level, directory level, host level and domain level. On the basis of block, we could analyze the different hyperlinks among pages. Several approaches proposed that the intra-hyperlink in a host maybe less useful in computing the PageRank. However, there are no reports on how concretely the intra- or inter-hyperlink affects the PageRank. Furthermore, based on different block level, inter-hyperlink and intra-hyperlink can be two relative concepts. Thus which level should be optimal to distinguish the intra- or inter-hyperlink? And how the ratio set between the intra-hyperlink and inter-hyperlink could ultimately improve performance of the PageRank algorithm? In this paper, we analyze the link distribution at the different block level and evaluate the importance of the intra- and inter-hyperlink to PageRank on the TREC Web Track data set. Experiment shows that, if we set the block at host level and the ratio of the weight between the intra-hyperlink and inter-hyperlink is 1:4, the retrieval could achieve the best performance.

1 Introduction

In recent years, several information retrieval methods using the information about the link structure have been developed and proved to provide significant enhancement to the performance of Web search in practice. Google's PageRank[3][14] and Keinberg's HITS[12] are two fundamental algorithms by employing the hyperlink structure among the Web page. A number of extensions to these two algorithms are also proposed, such as [1][2][4][5][6][9][10][11]. All these link analysis algorithms are based on two assumptions: (1) the links convey human endorsement. If there is a

link from page *u* to page *v*, then the page *v* is deemed to be valuable to the author of page *u*. Thus, the importance of page *u* can, in part, spread to the pages besides *v* it links to. (2) Pages that are co-cited by a certain page are likely to share the same topic as well as to help retrieval.

Considering the Web is a nested structure, the Web graph could be partitioned into blocks according to the different level of Web structure, such as page level, directory level, host level and domain level. We call such constructed Web graph as the *block-based* Web graph, which is shown in Fig.1 (left). Furthermore, the hyperlink at the block level could be divided into two types: Intra-hyperlink and Inter-hyperlink, where inter-hyperlink is the hyperlink that links two Web pages over different blocks while intra-hyperlink is the hyperlink that links two Web pages in the same block. As shown in Fig1, the dash line represents the intra-hyperlink while the bold line represents the inter-hyperlink. For example, when we partition the Web graph at the directory level, the web pages in the same directory are organized as a block. The hyperlinks which link two Web pages in the same directory are called as intra-hyperlink while the hyperlinks which link two Web pages over different directories are called as inter-hyperlink. There are several analysis on the block based Web graph. Kamvar et al. [18] propose to utilize the block structure to accelerate the computation of PageRank. Further analysis on the Website block could be seen in [13][15]. And the existed methods about PageRank could be considered as the link analysis based on page level in our approach. However, the intra-link and inter-link are not discriminated to be taken as the same weight although several approaches proposed that the intra-hyperlink in a host maybe less useful in computing the PageRank [7].

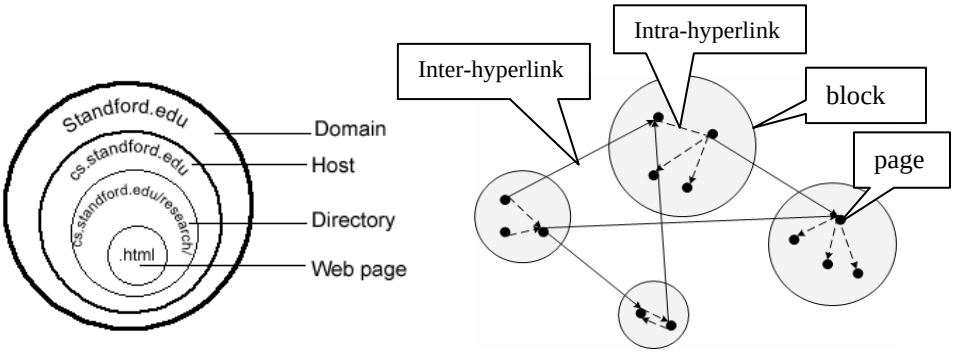

Fig. 1. Block-based Web graph

Since the intra-hyperlink and inter-hyperlink are two relative concepts based on different block level, our motivation is to analyze the importance of the intra- and inter-hyperlink to PageRank as well as find the optimal block level. Intuitively, we consider a hyperlink from a page *u* to page *v*, if *u* and *v* belong to different host then *v* will be more valuable for *u* taken an objective view, but if *u* and *v* belong to the same host then the link is considered to be probably made for convenience of the Web browsing. Therefore it will be useful for the link analysis to analyze the hyperlink, dividing them into links inside block (Intra-hyperlink) and links between the blocks (Inter-hyperlink).

In this paper, we first analyze the distribution of hyperlink at the four different block levels. Then, we construct the corresponding Web graph by leveraging the weight of the inter-hyperlink and intra-hyperlink of the block. By assigning different weight, we want to disclose which type of hyperlink are more useful to PageRank algorithm and how the ratio set between the intra-hyperlink and inter-hyperlink could achieve best performance on searching. Furthermore, we want to know which level the PageRank algorithm could be an adaptive block to leverage the intra-hyperlink and inter-hyperlink.

The contribution of this work can be summarized as follows.

- We first propose to construct the block based Web graph by partitioning Web graph into the block according to page level, directory level, host level and domain level, respectively.
- Based on the block based Web graph, the intra-hyperlink and inter-hyperlink are discriminated to be set different ratio in calculating PageRank, from which ratio 1:4 is found to be the best to ultimately improve the performance of the Web search.
- We also evaluate four different segmentation of the Web graph, and show that the host block is best level to distinguish the intra-hyperlink and the inter-hyperlink.

The rest of this paper is organized as follows. In Section 2, we review the PageRank algorithm. In Section 3, we present the characteristics about the Web graph structure. In Section 4, we show our ranking algorithm. Our experimental results are presented in Section 5. Finally, conclusions and future works are discussed in Section 6.

2 PageRank

The basic idea of PageRank is that if page *u* has a link to page *v*, then the author of *u* is implicitly conferring some importance to page *v*. Intuitively, Yahoo! is an important page, reflected by the fact that many pages point to it. Likewise, pages prominently pointed to from Yahoo! are themselves probably important. Then how much importance does a page *u* confer to its outlinks?

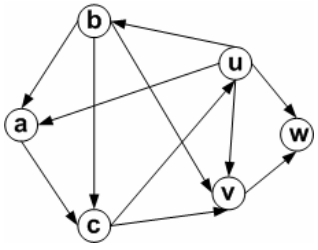

Fig. 2. Web Graph

Let N_u be the outdegree of page *u*, and let *PR* (*u*) represent the importance (i.e., PageRank) of page *u*. Then the link (*u*, *v*) confers $PR(u)/N_u$ units of rank to *v*. This simple idea leads to the following iterative fixpoint computation that yields the rank

vector over all of the pages on the Web. If n is the total number of pages, assign all pages the initial value 1/ n. Let B_v represent the set of pages pointing to v. In each iteration, propagate the ranks as follows:

$$\forall_v PR^{(i+1)}(v) = \sum_{u \in B_v} PR^{(i)}(u) / N_u \quad \textbf{(1)}$$

We continue the iterations until *PR* stabilizes to some threshold. The final vector *PR* contains the PageRank vector over the Web. This vector is computed only once after each crawl of the Web; the values can then be used to influence the ranking of search results.

The process can also be expressed as the following eigenvector calculation, providing useful insight into PageRank. Let M be the square, stochastic matrix corresponding to the directed Web graph G. If there is a link from page j to page i, then let the matrix entry m_{ij} have the value $1/N_j$. Let all other entries have the value 0. One iteration of the previous fixpoint computation corresponds to the matrix-vector multiplication $M \times \overrightarrow{PR}$. Repeatedly multiplying $\overrightarrow{PR}$ by M yields the dominant eigenvector $\overrightarrow{PR}$ of the matrix M . In other words, *PR* is the solution to

$$\overrightarrow{PR} = M \times \overrightarrow{PR} \quad \textbf{(2)}$$

Because M^T is the stochastic transition matrix over the graph G, PageRank can be viewed as the stationary probability distribution for the Markov chain induced by a random walk on the Web graph.

However, in practice, many pages have no in-links (or the weight of them is 0), and the eigenvector of the above equation is mostly zero. Therefore, the basic model is modified to obtain an "actual model" using *random walk*. Upon browsing a web-page, with the probability 1-ε, a user randomly chooses one of the links on the current page and jumps to the page it links to; and with the probability ε, the user "reset" by jumping to a web-page picked uniformly and at random from the collection. Therefore the ranking formula is modified to the following form:

$$\forall_v PR^{(i+1)}(v) = \frac{\varepsilon}{n} + (1-\varepsilon) \sum_{u \in B_v} PR^{(i)}(u) / N_u \quad \textbf{(3)}$$

Or, in the matrix form:

$$\overrightarrow{PR} = \frac{\varepsilon}{n} \vec{e} + (1-\varepsilon) M \overrightarrow{PR} \quad \textbf{(4)}$$

where $\vec{e}$ is the vector of all 1's, and ε ($0<\varepsilon<1$) is a parameter. In our experiment, we set ε to 0.15. Instead of computing an eigenvector, the simplest iterative method—Jacobi iteration is used to resolve the equation.

3 Block Structure of the Web

The key terminology we use in the remaining discussion is given in Table 1.

Term	Example: cs.standford.edu/research/index.html
Domain	Standford.edu
Host	cs.standford.edu
Directory	cs.standford.edu/research/
Page	cs.standford.edu/research/index.html

Table 1. Terminology using the sample URL http://cs.standford.edu/research/index.html

As we known, the whole Web is a graph structure where the Web pages represent the nodes and hyperlink between the Web pages represents the edge between nodes. Such Web graph looks like a flat. All the nodes are considered as the same level and all the hyperlinks are set equal weight in existed method about PageRank.

The fact is that the Web is also a hierarchical structure based on the organization of the Web page, for example the website structure. Hence the Web graph is also a hierarchical structure ranged from the Web page to the domain. We could partition the Web graph into the blocks according to the four different views: page level, directory level, host level and domain level. On the basis of different blocks, the hyperlink could be divided into two types: Intra-hyperlink and Inter-hyperlink.

According to the analysis on Section 1, the weight of the hyperlink should be set different weight when we compute the PageRank. The basic intuition is that the hyperlink within the same block is mainly for navigation while the link over the blocks is for recommendation. So we should give the different weight to the two types of link. In order to testify the intuition, we perform the PageRank on the modified Web graph over the four different block levels Web graph.

To investigate the structure of the block based Web, we run the following simple experiment. We take all the hyperlinks in the .GOV collection [20] , and count how many of the links are “Intra-hyperlink” and “Inter-hyperlink” at different block level.

Level	Intra-hyperlink	Inter-hyperlink
Domain	7342031 (97%)	227322 (3%)
Host	6506578 (86%)	1062775 (14%)
Directory	2956566 (39.1%)	4612787 (60.9%)
Page	0 (0%)	7569353 (100%)

Table 2. The Number of intra-hyperlink and inter-hyperlink at different level

As shown in Table 2, the number of the intra-hyperlink and the inter-hyperlink at four different level is different, so we also should evaluate which block level to distinguish the intra-hyperlink and the inter-hyperlink could achieve the higher performance than other three levels.

4 PageRank at Different Level

In this section, we consider the different weight of the hyperlink according to the link that belongs to the same block or across two different blocks. First, we construct a matrix to represent the link graph, where each represents the weight of the link; then, a modified link analysis algorithm is used on the matrix to calculate the importance of the pages. Finally, we re-rank the Web pages based on two kinds to re-ranking mechanics: order-based re-ranking and score-based re-ranking.

4.1 Matrix Construction

The Web can be modeled as a directed graph $G = (V, E)$ where $V = \{p_i \mid 1 \le i \le n\}$ is the set of vertices representing all the pages in the web, and E encompasses the set of links between the pages. $l_{i,j} \in E$ is used to denote that there exists a link between the page p_i and p_j.

We propose to construct a new *block-based* Web graph instead of the original *page-based* Web graph. This new graph is a weighted directed graph $G' = (V, E')$, where V is same as above and E' encompasses the links between pages. Furthermore, each link $l_{i,j} \in E'$ is associated with a new parameter w_{ij} denoting the weight of the page p_j to page p_i, where the weight is calculated according to the hyperlink that is in the block or between the block. In this paper, we tune the weight of the link by value w_{ij}:

$$w_{ij} = \begin{cases} \alpha & p_i, p_j \text{ in same block} \\ \beta & p_i, p_j \text{ in different blocks} \end{cases} \quad (5)$$

Where α and β are the parameters. In this paper, we tune the α and β as the ratio between inter-hyperlink and intra-hyperlink to evaluate how the different types of the links affect the PageRank of the pages.

4.2 Modified PageRank

After obtaining the block-based Web structure, we apply a link analysis algorithm similar to PageRank to re-rank the web-pages. We construct a matrix to describe the graph. In particular, assume the graph contains n pages. The $n \times n$ adjacency matrix is denoted by A and the entries $A[i, j]$ is defined to be the weight of the links $l_{i,j}$.

The adjacency matrix is used to compute the rank score of each page. In an "ideal" form, the rank score PR_i of page p_i is evaluated by a function on the rank scores of all the pages that point to page p_i :

$$PR_i = \sum_{j:l_{ji} \in E} PR_j \cdot A[j,i] \quad (6)$$

This recursive definition gives each page a fraction of the rank of each page pointing to it—inversely weighted by the strength of the links of that page. The above equation can be written in the form of matrix as:

$$\overrightarrow{PR} = A\overrightarrow{PR} \quad (7)$$

However, in practice, many pages have no in-links (or the weight of them is 0), and the eigenvector of the above equation is mostly zero. Therefore, the basic model is modified to obtain an "actual model" using *random walk*. Upon browsing a web-page, with the probability 1-ε, a user randomly chooses one of the links on the current page and jumps to the page it links to; and with the probability ε, the user "reset" by jumping to a web-page picked uniformly and at random from the collection. Therefore the ranking formula is modified to the following form:

$$PR_i = \frac{\varepsilon}{n} + (1-\varepsilon) \sum_{j:l_{j,i} \in E} PR_j \cdot A[j,i] \quad (8)$$

or, in the matrix form:

$$\overrightarrow{PR} = \frac{\varepsilon}{n}\vec{e} + (1-\varepsilon)A\overrightarrow{PR} \quad (9)$$

where $\vec{e}$ is the vector of all 1's, and ε ($0<\varepsilon<1$) is a parameter. In our experiment, we set ε to 0.15.

4.3 Re-ranking

The re-ranking mechanism is based on two types of linear combination: the score based re-ranking and the order based re-ranking. The score based re-ranking uses a linear combination of content-based similarity score and the PageRank value of all web-pages:

$$Score(w) = \lambda Sim + (1- \lambda)\, PR \quad (\lambda \in [0, 1]) \quad (10)$$

where *Sim* is the content-based similarity between web-pages and query words, and *PR* is the PageRank value.

The order based re-ranking is based on the rank orders of the web-pages. Here we use a linear combination of pages' positions in two lists in which one list is sorted by similarity scores the other list is sorted by PageRank values. That is,

$$Score(w) = \lambda O_{Sim} + (1- \lambda) O_{PR} \quad (\lambda \in [0, 1]) \quad (11)$$

where O_{Sim} and O_{PR} are positions of page w in similarity score list and PageRank list, respectively.

We have conducted the experiments that the order based re-ranking could achieve higher performance than the score based re-ranking. So in this paper, we just impose the order based re-ranking method for evaluation.

5 EVLUATION

In this section, several experiments were performed to compare four block level link analysis algorithms, i.e. domain-based PageRank, host-based PageRank, directory-based PageRank and traditional PageRank.

5.1 Experimental Environment

By 1999, link analysis and Web search in general have become a "hot" topic and a special "Web Track" in the annual TREC benchmarking exercise [20] was dedicated to Web search related tasks [19]. Topic Distillation task [6] is mainly evaluated using the measure of precision at Top 10 and the goal of which is to find a small number of key resources on a topic as opposed to the more conventional (ad-hoc) listing of relevant pages. Topic Distillation, although not that far removed from ad-hoc is perhaps more suited to Web search evaluation because it has been found that over 85% of users never look beyond the first page of the results from any Web search [17].

In order to support the experiments of participants, TREC distributes test collections that consist of three components: a set of documents, a set of queries (called topics) and a set of relevance judgments for each query.

In this paper, we do the analysis on the .GOV collection used in 2002(and 2003 also) TREC Web track, which are better reflect today's WWW. The collection consists of 1,247,753 documents from a fresh crawl of the Web pages made in early 2002. Among them, 1,053,372 are text/html, which are used in our experiment. Finding for TREC-2002 illustrate that for some participants, the application of link analysis did indeed improve retrieval performance in the new Topic Distillation task. Link analysis can provide some extra useful information for ranking. This situation is much like the real world Web search. So the corpus and queries are very suitable in evaluating different link analysis algorithm.

They are totally 50 queries. The number of relevant pages (based on human judgment) for each query ranged from 1 to 86 with average 10.32. Among them 31 queries have less than 10 relevant pages, so the average P@10 is a litter bit low.

5.2 Relevance Weighting

In our experiments, we use BM2500 [16] as the relevance weighting function. It is of the form:

$$\sum_{T \in Q} \omega \frac{(k_1+1)tf\,(k_3+1)qtf}{(K+tf)(k_3+qtf)} \qquad (12)$$

where Q is a query containing key terms T, tf is the term frequency in a specific document, qtf is the frequency of the term within the topic from which Q was derived, and w is the Robertson/Sparck Jones weight of T in Q. It is calculated by

$$\log \frac{(r+0.5)/(R-r+0.5)}{(n-r+0.5)/(N-n-R+r+0.5)} \quad (13)$$

where N is the number of documents in the collection. n is the number of documents containing the term, R is the number of documents relevant to a specific topic, and r is the number of relevant documents containing the term. In our experiment, the R and r are set to zero. In the equation 12, K is calculated by

$$k_1((1-b)+b\times dl/avdl) \quad (14)$$

where dl and $avdl$ denote the document length and the average document length. In our experiments, we tune the k_1=4.2, k_3=1000, b=0.8 to achieve the best baseline (we took the result of using relevance only as out baseline). The average precision is 0.1285. The P@10 is 0.112. Compared with the best result of TREC 2003 anticipants (with P@10 of 0.128), this baseline is reasonable. TREC 2003 did not report the average precision.

5.3 Performance of 50 Queries on .GOV Collection

In order to evaluate the importance of the intra-hyperlink and inter-hyperlink based on the four block levels, we construct the Web graph by tuning the ratio between α and β from 5:1 to 1:10 and calculate the p@10 average precision of 50 queries on the Web TREC data.

Meanwhile, we combined the relevance rank with PageRank of four levels. We chose the top 1000 results according to the relevance, and then we sorted these 1000 results according to their PageRank values. Thus, we get two ranking list. One is according to the relevance and the other is according to importance. We tune the parameter λ in equation 11 from 0.76 to 0.9. The results of P@10 precision with λ are shown in Fig 2, Fig 3, and Fig 4. When the ratio between α and β is set as 1:1, the algorithm of the other three levels are performed as traditional PageRank algorithm.

As we can see from the Fig 3, Fig 4, and Fig 5, the different weight of the intra-hyperlink and inter-hyperlink is sure to affect the performance of the P@10 precision. Generally, the performance is better when the weight of the intra-hyperlink is lower than that of the inter-hyperlink.

From the Fig 3, the directory-based PageRank could achieve the highest performance when we set the ratio between α and β as 1:7 while λ as 0.81.

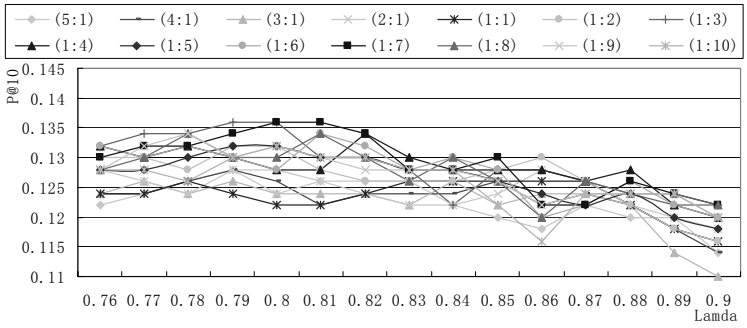

Fig. 3. Performance of directory-based PageRank

From the Fig 4, host-based PageRank could achieve the highest performance when we set the ratio between α and β as 1:4 while λ as 0.83.

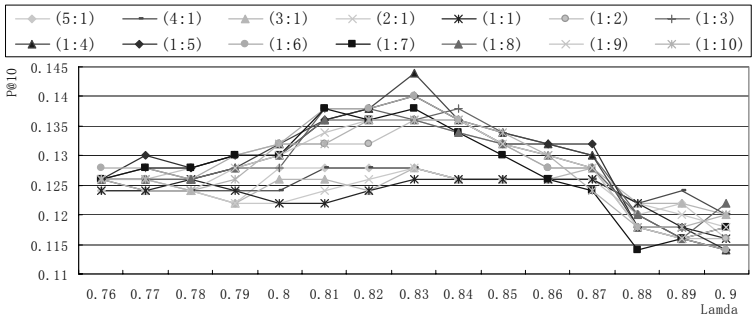

Fig. 4. Performance of host-based PageRank

From the Fig 5, domain-based PageRank could achieve the highest performance when we set the ratio between α and β as 1:5 while λ as 0.81.

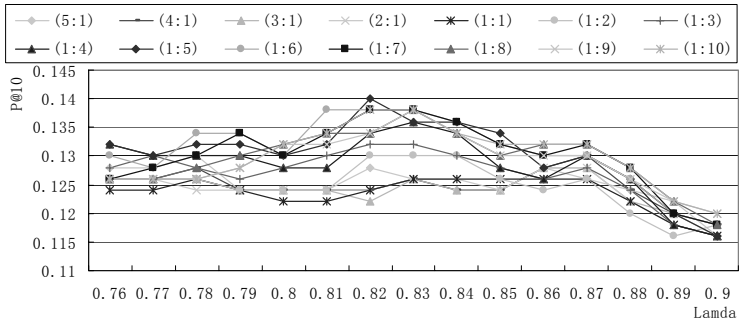

Fig. 5. Performance of domain-based PageRank

Furthermore, we conduct the experiments to get which block level to leverage the intra-hyperlink and inter-hyperlink could achieve the better performance than the other three methods. As shown in Fig 6, the host-based PageRank could achieve highest performance than other three block levels. To understand whether these improvements are statistically significant, we performed various t-tests. For the p@10

improvement, compared with PageRank, both other three block levels based PageRank are significant (p-value is 0.028, 0.000916 and 0.000674, respectively).

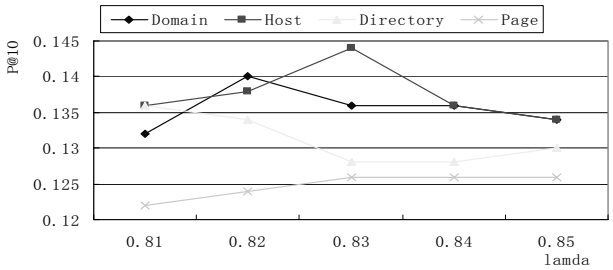

Fig. 6. Comparison of four block levels' PageRank

From the above experiments, we could infer that the inter-hyperlink should have more importance than the intra-hyperlink when we calculate the importance of the Web pages. Generally, we should set the ratio between the intra-hyperlink and the inter-hyperlink should be great than 1:3. Furthermore, if we distinguish the link from the host block level, the link analysis could be getting the highest performance when applying to the Web search.

6 Conclusion and Future Work

In this paper, we argued that the hyperlink should have different weight while traditional PageRank algorithms ignored this fact. Based on the hierarchical organization of Web pages, we could divide the Web graph into the blocks according to four levels: domain level block, host level block, directory level block and page level block. We tune the ratio of the intra-hyperlink that inside a block and the inter-hyperlink cross blocks to evaluate the performance of searching. The experimental results show that when the ratio of weight is set to 1:4, the system could achieve the best performance. Meanwhile, the host level block could achieve the higher performance than other three levels segmentation.

7 Reference

[1] B. Amento, L. Terveen, and W. Hill. Does "authority" mean quality? predicting expert quality ratings of web documents. In Proc. of ACM SIGIR 2000, pages 296--303.

[2] K. Bharat and M. R. Henzinger. Improved algorithms for topic distillation in a hyperlinked environment. In Proc. of the ACM-SIGIR, 1998.

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual Web search engine", In The Seventh International World Wide Web Conference, 1998.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P.Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink struc-

ture and associated text. In Proc. of the 7th Int. World Wide Web Conference, May 1998.
[5] S. Chakrabarti, Integrating the Document Object Model with hyperlinks for enhanced topic distillation and information extraction, In the 10th International World Wide Web Conference, 2001.
[6] S. Chakrabarti, M. Joshi, and V. Tawde, Enhanced topic distillation using text, markup tags, and hyperlinks, In Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval , ACM Press, 2001, pp. 208-216.
[7] Christof Monz, Jaap Kamps, and Marrten de Rijke, The University of Amsterdam at TREC 2002.
[8] Brian D. Davison. Recognizing nepotistic links on the Web. In Artificial Intelligence for Web Search, pages 23--28. AAAI Press, July 2000.
[9] G. Flake, S. Lawrence, L. Giles, and F. Coetzee, Self-organization and identification of web communities, IEEE Computer, pp. 66-71, 2002.
[10] D. Gibson, J. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In Proceedings of the 9th ACM Conference on Hypertext and Hypermedia (HYPER-98), pages 225--234, New York, June 20--24 1998. ACM Press.
[11] T.H. Haveliwala. Topic-sensitive PageRank. In Proc. of the 11th Int. World Wide Web Conference, May 2002.
[12] J. Kleinberg, Authoritative sources in a hyperlinked environment, Journal of the ACM, Vol. 46, No. 5, pp. 604-622, 1999.
[13] Krishna Bharat, Bay-Wei Chang, Monika Rauch Henzinger, Matthias Ruhl. Who Links to Whom: Mining Linkage between Web Sites. 1st International Conference on Data Mining (ICDM) 51-58 (2001).
[14] L. Page, S. Brin, R. Motwani, and T. Winograd, The PageRank citation ranking: Bringing order to the web, Technical report, Stanford University, Stanford, CA, 1998.
[15] Nadav Eiron and Kevin S. McCurley, Locality, Hierarchy, and Bidirectionality on the Web, Workshop on Web Algorithms and Models, 2003.
[16] S. E. Robertson, Overview of the okapi projects, Journal of Documentation, Vol. 53, No. 1, 1997, pp. 3-7.
[17] Silverstein C, Henzingger M, Marais J and Moricz M. Analysis of a Very Large AltaVista Query Log. Digital SRC Technical Note 1998-014.
[18] S. Kamvar, T. Haveliwala, C.Manning and G. Golub, Exploiting the Block Structure of the Web for Computing PageRank. In Proc. of the 12th Int. World Wide Web Conference, May 2003.
[19] Hawking D, Overview of the TREC-9 Web Track. In Proc. of the 9th Annual TREC Conference, pp.87-102.
[20] TREC http://trec.nist.gov/.