MFCRank: A Web Ranking Algorithm Based on Correlation of Multiple Features

Yunming Ye[1], Yan Li[1], Xiaofei Xu[1], Joshua Huang[2], and Xiaojun Chen[1]

[1] Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055, China
yym_sjtu@yahoo.com.cn
[2] E-Business Technology Institute, The University of Hong Kong, Hong Kong
jhuang@eti.hku.hk

Abstract. This paper presents a new ranking algorithm ***MFCRank*** for topic-specific Web search systems. The basic idea is to correlate two types of similarity information into a unified link analysis model so that the rich content and link features in Web collections can be exploited efficiently to improve the ranking performance. First, a new surfer model ***JBC*** is proposed, under which the topic similarity information among neighborhood pages is used to weigh the jumping probability of the surfer and to direct the surfing activities. Secondly, as *JBC* surfer model is still query-independent, a correlation between the query and *JBC* is essential. This is implemented by the definition of *MFCRank* score, which is the linear combination of *JBC* score and the similarity value between the query and the matched pages. Through the two correlation steps, the features contained in the plain text, link structure, anchor text and user query can be smoothly correlated in one single ranking model. Ranking experiments have been carried out on a set of topic-specific Web page collections. Experimental results showed that our algorithm gained great improvement with regard to the ranking precision.

Keywords: Ranking, Search Engine, Link Analysis, *PageRank*, Web.

1 Introduction

The enormous volume of the Web presents a big challenge to Web search, as there are always too many results returned for specific queries, and going through the entire results to find the desired information is very time-consuming for the user. To improve the information retrieval efficiency, Web search engines need to employ a suitable page ranking strategy to correctly rank the search results so that the most relevant (or important) pages will be included in the top list of the search results.

In traditional information retrieval, ranking measures, such as *TF*IDF* [1], usually rely on the text features alone to rate plain text documents. This strategy can give poor results on the Web, due to the fact that the indexed Web document collection is so enormous and diverse that the text alone is not selective enough to limit the number of search results to a manageable size. An

A. Gelbukh (Ed.): CICLing 2006, LNCS 3878, pp. 378–388, 2006.

important characteristic that differentiates Web ranking from traditional ranking is that the former offers more features to be exploited. Besides plain text features, HTML tags, anchor text, hyperlinks among pages and meta data, all provide rich information for Web ranking. Effectively exploiting these features is critical for the success of any ranking strategy. In recent years, various link-based ranking methods have been developed to exploit hyperlink information for improving the search results. Among them, *PageRank* [2, 3] and *HITS* [4] are the two best-known algorithms. It has been testified that proper utilization of link information is very helpful for Web search, where the success of *PageRank* in *Google*'s search engine is one well-known example.

However, the initial *PageRank*-like algorithms purely depend on the link structure information to rank the search results, and can't effectively integrate the multiple features of the Web pages. Thus, they are not robust enough and suffer from various topic drift problems [5]. Recently, integrating the text features with link structure features for Web ranking has been a very active research topic. Several algorithms have been proposed, including Richardson's query-dependent *PageRank* [6], Haveliwala's topic-sensitive *PageRank* [7], the personalized *PageRank* [8], and similarity ranking method for queries of 'related pages' [9, 10]. When combining the content features with link information, these previous approaches mainly focused on utilizing the similarity relationship between the user query and the retrieved pages (text features), while the topic similarity information among ***neighborhood pages***[1] has not been used in computing the rank scores of indexed pages. We argue that to improve the accuracy of ranking algorithms, the topical similarity information among neighborhood pages should be consolidated into the link analysis model, because this similarity information can be a good measurement in computing the rank score for a given page. This is similar to a real-world scenario: when evaluating a man, the opinions from the people with more similar background will be more valuable than those from irrelevant communities. Based on this intuition, we develop a new Web ranking algorithm, *MFCRank*[2], which can effectively combine both the similarity information among neighborhood pages and the query similarity information into one ranking model.

MFCRank is based on the correlation of multiple features in a Web document collection. The ranking algorithm consists of two correlation steps: (1) First, similar to the random surfer model in *PageRank*, we propose a new surfer model, i.e. *JBC* surfer model, which uses the similarity information of neighborhood pages to weigh the jumping probability of the surfer. The surfer in the new model is not 'random' any more, but directed by the neighborhood similarity information, and therefore the first step is the correlation between the text features of pages with the link structure features. (2)However, *JBC* surfer model is still query-independent as *PageRank*, therefore a correlation between the query and the *JBC* surfer model is essential. This is implemented by the definition of *MFCRank* score, which is the linear combination of *JBC* score and the similarity

[1] Two pages are neighbors to each other if they are connected by at least one hyperlink.
[2] *MFCRank* stands for '**M**ultiple **F**eatures **C**orrelation **Rank**ing'.

value between the query and the matched pages. Through the two correlation steps, the features contained in plain text, link structure, anchor text and the query are combined in a single ranking model.

We have implemented *MFCRank* in a topic-specific Web search platform. Ranking experiments were carried out on a set of topic-specific Web page collections. Experimental results show that the *MFCRank* algorithm gains great improvement w.r.t. the ranking precision.

The rest of this paper is organized as follows. In Section 2, we discuss the random surfer model in *PageRank*. Section 3 describes the *MFCRank* algorithm, including the *JBC* surfer model and the definition of *MFCRank* score. Experimental results and analysis are presented in Section 4. In Section 5 we draw the conclusions and points out some avenues for future work.

2 Random Surfer Model in *PageRank*

The Web is logically a directed graph $\boldsymbol{G}=(\boldsymbol{V}, \boldsymbol{E})$, where $\boldsymbol{V}$ is a set of nodes representing pages and $\boldsymbol{E}$ is a set of directed edges representing hyperlinks. Assume that the Web graph is strongly connected, that is, from any node u there is a directed path to another node v. Imagine a Web surfer starting from a random page, clicking the hyperlinks on pages forever, and picking a link on a page at random to move on to the next page. Occasionally, the surfer will not follow the hyperlinks on the page (or when a page has no out-links), but jump to a random page with some small probability ε. In this random surfer model, the probability that the surfer visit some page (node) d_i at one point of time can be defined as:

$$P(d_i) = \frac{\varepsilon}{|V|} + (1-\varepsilon) * \sum_{d_j \in B(d_i)} \frac{P(d_j)}{|F(d_j)|} \tag{1}$$

where $|\boldsymbol{V}|$ is the total number of the nodes in $\boldsymbol{V}$, $\mathrm{B}(d_i)$ is the set of nodes linking to node d_i, that is, $\mathrm{B}(d_i)$ and d_i are neighborhood pages. $|F(d_j)|$ denotes the total number of the nodes d_j links to. The probability $\mathrm{P}(d_i)$ is the *PageRank* score for page d_i, and formula (1) defines the page ranking strategy in *PageRank*. Pages with greater *PageRank* score will get higher ranks in search results.

In fact, *PageRank* is query-independent. The *PageRank* score is assigned to each page independent of a specific user query. At query time, this score is used with or without some query-dependent ranking criteria to rank all pages matching the query. The *PageRank* score is a measure for distinguishing important (high-quality) pages from unimportant (low-quality) pages, and its computation is completely based on the link structure information without considering the content of pages. However, important pages may not be relevant. An elegant ranking algorithm should give high rank scores to pages with both high relevance and great importance. This presents two requirements for more effective ranking strategies:

(1) First, the content information in pages should be combined with link information in the definition of rank score in order to improve the scoring accuracy and robustness;
(2) Secondly, the rank score should also be smoothly correlated with user query, so that it is query-dependent.

The development of *MFCRank* algorithm follows the two requirements.

3 Web Ranking Based on Multi-feature Correlation

There are two correlation steps in *MFCRank* algorithm. In the first step, through a new surfer model, *JBC*, the content similarity information among neighborhood pages is correlated with link information to define the query-independent rank score, i.e. *JBC* score. The second step is the definition of query-dependent *MFCRank* score, which combines the *JBC* score with the similarity value between user query and the matched pages.

3.1 *JBC* Surfer Model

Similar to the PageRank algorithm, *MFCRank* defines a query-independent rank score function based on a surfer model, i.e. *JBC* (***J***umping-***B***ased on ***C***ontent) model. The basic idea of *JBC* can be described as follow. Similar to the random surfer, the *JBC* surfer starts from a random page and clicks the hyperlinks on the visited pages constantly, however, unlike the random surfer, when picking a link on a page to follow, the *JBC* surfer is not at random, but tend to choose preferentially the links of which the corresponding pages (the child pages) have higher similarity to the page being visited (the parent pages). That is, the jump probabilities from one page to a linked page are weighted based on the similarities between the parent page and the neighborhood child pages. The intuition captured by this idea is the following: when surfing the Web, it is more likely that the surfer will focus on some topic and tend to follow similar pages over a period of time, but after some time he may jump to another topic with some probability. This idea is encoded in the definition of *JBC* rank score as follows:

$$F_{JBC}(d_i) = \frac{\varepsilon}{|V|} + (1-\varepsilon) * \sum_{d_j \in B(d_i)} \lambda_{ji} \cdot F_{JBC}(d_j) \tag{2}$$

the definition is similar to formula (1), where $F_{JBC}(d_i)$ is the *JBC* rank score for page d_i, λ_{ji} represents the jumping probability from the page d_j to the page d_i (that is d_i and d_j are neighbors to each other), which is computed according to the similarity scores between neighborhood pages as:

$$\lambda_{ji} = \frac{Sim(d_j, d_i) + \sigma}{\sum_{d_j \in B(d_k)} (Sim(d_j, d_k) + \sigma)} \tag{3}$$

where $Sim(d_j, d_i)$ is the similarity of the page d_j to the page d_i. σ is a small positive value acting as a normalization factor. The value $Sim(d_j, d_i)$ can be

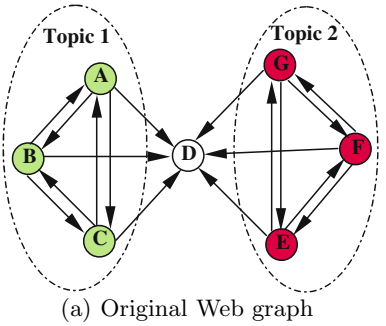
computed as the similarity of the original text features in the two pages, or by concatenating their anchor texts as the virtual pages to calculate the degree of similarity. In our experiments, we use the traditional $tf * idf$ scheme [1] as the term weighting measure to compute the similarity value.

The *JBC* rank score makes a smooth tradeoff between the relevance measure and importance measure through the correlation of content features with link features. It will be more robust to tackle the topic drift problem. The following gives an illustration.

3.1.1 An Illustration of *JBC* Ranking

Fig.1 shows a sample Web graph for computing rank scores, and the corresponding similarity matrix for connected nodes. In this graph, nodes A, B, C are in the same topic, i.e., topic 1, while topic 2 includes the nodes E, F, G. The node D is a popular page with many in-links, such as the Yahoo homepage. Although page D doesn't focus on specific topic, it may have some keywords which appear in some topics (such as the topic 1). Therefore pages like D will often be included in the result lists for many topic-specific user queries.

Now assume that the user query is on topic 1, and the matched pages include page A, B, C and D. The matched list is ranked according to their rank scores. Table 1 shows resulting rank scores for all nodes in the given Web graph, which are computed according to formula (1) and (2) (During computation, the damping factor ε is set to 0.5). In *PageRank* scoring, the rank score of page D is the highest, as *PageRank* score is defined to bias for the strongly connected pages (namely important pages). Therefore, page D will be ranked as the No.1 in the top list, although it is not relevant to topic 1, which is a typical topic drift problem. The problem is solved in the *JBC* scheme, as shown in Table



(a) Original Web graph

	A	B	C	D	E	F	G
A		0.80	0.90	0.20	-	-	-
B	0.80		0.86	0.16	-	-	-
C	0.90	0.86		0.26	-	-	-
D	0.20	0.16	0.26		0.16	0.16	0.16
E	-	-	-	0.16		0.90	0.80
F	-	-	-	0.16	0.90		0.86
G	-	-	-	0.16	0.80	0.86	

(b) Similarity matrix

Fig. 1. A Web graph and its similarity matrix for computing rank scores

Table 1. Ranking scores of the nodes in the sample Web graph

Algorithm	A	B	C	D	E	F	G
PageRank	0.12901	0.12901	0.12901	**0.22597**	0.12901	0.12901	0.12901
JBC Model	0.14438	0.14196	0.14814	**0.12119**	0.14776	**0.15076**	0.14580

1, the rank score of page D becomes lower than those of pages A, B, C which are relevant to the user query. This is due to the correlation of topic (content) locality [11] information with link information in the *JBC* surfer model.

This example implies that to determine the rank score for a given page, more authoritative pages (experts), that is, those with greater similarity to the given page (such as on the same topic), should play more important roles (i.e. contributing more weights according to formula (2)). It is clear that *JBC* surfer model considers both the link structure information and the topical relevance information, which makes it more robust in dealing with topic drift problems.

3.2 The Definition of *MFCRank* Score

JBC ranking scheme is still query-independent as PageRank, which may bring up another topic drift problem. For Fig.1, assume the user query is still about topic 1, and unfortunately the pages in topic 2 are also included in the matched result list. According to *JBC* ranking, the irrelevant page F will get higher rank score than the relevant pages A, B, C. This problem is because that the definition of *JBC* rank score is query-independent and biased for the pages of the topics with stronger topical locality [11], which we call 'topical winner-take-all effect'. Since the simple keyword matching is not selective enough to filter out the irrelevant pages w.r.t the user query, a more effective query analysis technique should be developed to aid the *JBC* ranking scheme. The definition of *MFCRank* score is under this motivation.

The *MFCRank* score is the rank score defined in *MFCRank* algorithm. It is the linear combination of *JBC* score and the similarity value between the query and the matched pages, defined as follows:

$$F_{MFC}(d_i^Q) = (1-\mu)F_{JBC}(d_i) + \mu \cdot Sim(d_i, Q) \cdot F_{JBC}(d_i) \tag{4}$$

where $F_{MFC}(d_i^Q)$ is the *MFCRank* score for page d_i w.r.t the user query Q, $Sim(d_i, Q)$ is the similarity of the page d_i to the user query calculated through $tf*idf$ scheme, μ is a bias factor between the query-independent *JBC* rank score and the query similarity score. Previous works have shown that the anchor texts of Web page are very informative and descriptive for the original page [12, 13]. To accelerate the ranking process, when computing the value $Sim(d_i, Q)$, we didn't use the original content of the pages, but concatenated the anchor texts of pages to construct the virtual pages, and $Sim(d_i, Q)$ was computed as the similarity between the virtual page of d_i and the user query Q.

Through the definition of *MFCRank* score, the *MFCRank* ranking strategy becomes query-dependent. The online query analysis results is closely correlated with the offline link analysis results, which will make the ranking strategy more accurate and robust.

3.3 Computation Scalability

The computation cost in *MFCRank* includes two parts. First is the offline pre-computation of the *JBC* score vector for the pages in the indexed document

collection. This is similar to the iterative computation of *PageRank* score vector in *PageRank* algorithm, which has been proved to be scalable in practical use [2]. An overhead in *JBC* is that the similarity matrix of the indexed pages should be pre-computed before computing the *JBC* score vector. Assume the number of the indexed pages is n, if there is a hyperlink between any two pages, i.e. any two pages are neighbors, it will need $n^2/2$ times to calculate the similarity between two pages. This will take enormous computation when n is a large value because computing similarity of pages is very costly. Fortunately, in practice each page always has a very limited neighbor pages, and the average number of neighbors for each page is a small constant k (such as 11). Therefore, computing the similarity matrix will cost $k \cdot n$ times of the similarity computation of two pages, which is scalable to very large page collections.

The second part of computation is to calculate the *MFCRank* scores for each matched page w.r.t. the user query, which is performed online. The main time cost of this part is the computation of the similarity between the user query and the virtual documents of the matched pages (anchor text concatenation), which is similar to the computation in traditional $tf * idf$ ranking. We have developed a fast anchor text index to speed up the retrieval of anchor text for computing the query similarity. Although there is some overhead for online computation, it is affordable for practical applications.

4 Experiments

4.1 Experiment Setup

MFCRank has been implemented in a topic-specific Web search platform-*TopSearch* [14], which is a scalable and configurable platform for building topic-specific search systems. Experimental study was performed on this platform. In *TopSearch*, a focused crawling system *iSurfer* [15] was employed to collect Web pages of specific topics from the Web. The crawled pages were used to build the topic-specific page collections for the search experiments. Each collection has its own topic (such as '*Chinese history*'), therefore it can be regarded as a search engine on a certain topic (such as a topic-specific search engine on '*Chinese history*').

For each topic-specific collection, we built inverted full-text index and other auxiliary indexes, such as anchor text index, before doing search experiments. A fast link graph data structure was constructed on each page collection as well, upon which the rank score vectors for evaluated ranking algorithms (PageRank and MFCRank) were computed. As the design of *TopSearch* has carefully considered the encoding and language problems, it can process both English and Chinese Web pages efficiently. This feature greatly facilitates our multi-language experiments.

In traditional information retrieval research, the precision and recall are the main evaluation metrics. However, it is difficult to get the recall for Web search, as measuring the size of relevance set from a large Web collection is almost impossible. We employ the precision in the top K list of the search results as the

main performance evaluation metric. Typical values for K include 10, 30, and 50, which represent the search results user may pay attention to. The precision will be simply calculated as the percentage of relevant pages for all the experiments.

4.2 Experimental Results on Topic-Specific Collections

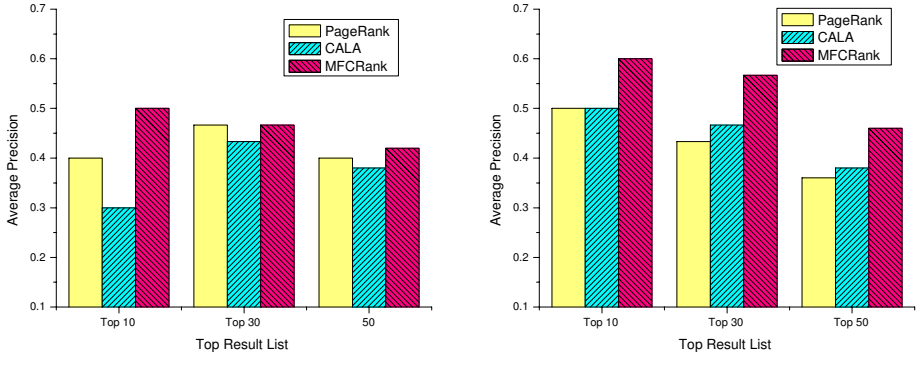
We built a set of topic-specific collections, each of which consists of tens of thousands pages crawled by *iSurfer*. For each collection, we used a list of user queries (about 60 queries) to search the corresponding collection. Table 2 shows the size of the collections and the number of corresponding user queries performed.

Three ranking algorithms were implemented for performance comparison: (1) the original *PageRank* algorithm, (2) ranking algorithm *CALA* [16], which defines the rank score based on the linear combination between *PageRank* and online anchor text analysis, and (3) the *MFCRank* algorithm. In previous work, **CALA** has been testified experimentally to have higher precisions than *PageRank* in general Web search [16].

Fig.2 shows the query results on the English collections 'American History' and 'American African History'. Fig.3 presents the results on the Chinese collections 'Travel in China' and 'Travel in Beijing' (the user queries are in Chinese). For each collection, we compute the average precision of the top 10, top 30, and top 50 result lists for all user queries.

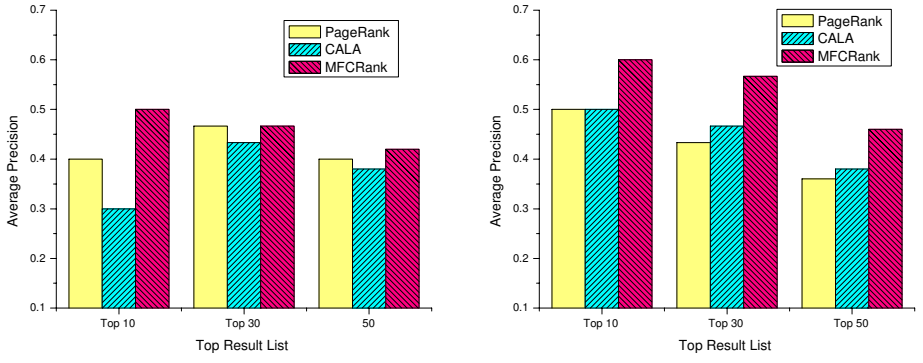
Table 2. A statistics of the user queries performed

Topic of The Collection	Number of Pages	Number of Queries
American History	251,820	62
American African History	82,218	60
Travel in China	321,336	65
Travel in Beijing	182,215	60



(a) Collection: 'American History' (b) Collection: 'American African History'

Fig. 2. The search results on two English collections



(a) Collection: 'American History' (b) Collection: 'American African History'

Fig. 3. The search results on two Chinese collections

In most of the queries, *CALA* and *MFCRank* outperformed *PageRank*. This testified that combining online query analysis information into link analysis model is very useful for improving the ranking performance. The more important observation is that *MFCRank* got precisions higher than *CALA* persistently, which demonstrate the effectiveness of the *JBC* surfer model w.r.t the random surfer model also used in CALA. As relevance is a more important evaluation metric than the importance metric in topic-specific Web search, we believe that integrating content analysis techniques into the ranking strategy is very essential.

4.3 Tradeoff Between Online Content Analysis and Offline Link Analysis

When implementing the *MFCRank* in the search system, an interesting issue is to determine the value of the bias factor in formula (4). The factor μ means the tradeoff between online content analysis and offline link analysis in deciding the final rank scores for matched pages. Our assumption was that the optimal value depended on the 'topical broadness' of the topic-specific page collection searched, which should be set higher for narrow topics and lower for broad topics. To test this assumption, we have carried out some elementary experiments.

Fig.4 shows the results on two page collections with different topical broadness, where the collection 'Travel in China' has greater topical broadness than that of the collection 'Travel in Beijing'. We set μ to different values and prepared about 100 queries to search the collections. The precision of the search results (the top 50 in our experiments) was calculated w.r.t the value of μ. As shown in Fig.4, to get optimal search performance, the factor should be set near 0.4 for the collection 'Travel in China', and 0.6 for the collection 'Travel in Beijing'. The elementary results demonstrated our initial assumption. This observation means that: for the collections with narrower topics, the online content analysis will

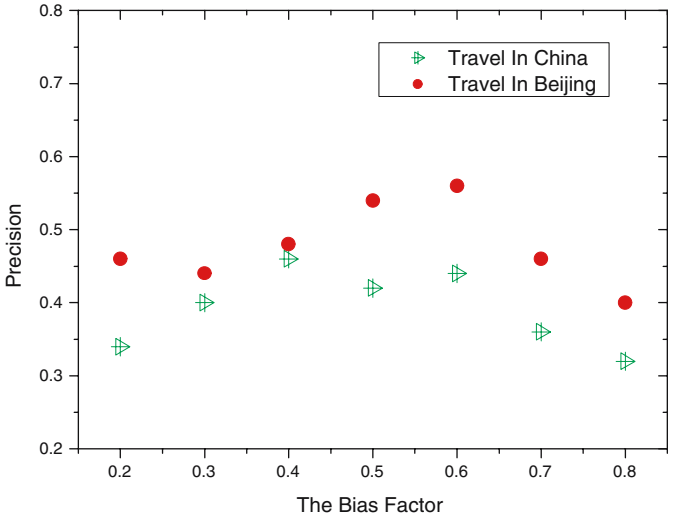


Fig. 4. The results of the experiments on the bias factor μ

play a more important role in page ranking as the link graphs of the collections are usually too dense and provide less discriminative information for ranking algorithm. On the contrary, link analysis is more important for the collections with greater topical broadness.

5 Conclusions

In this paper we have proposed a new ranking algorithm *MFCRank* for topic-specific Web search engines. Its intuition is to correlate two types of similarity information in a unified link analysis model so that the rich content and link features in Web collections can be exploited efficiently to improve the ranking performance. First, a new surfer model *JBC* is proposed, under which the topic similarity information between neighborhood pages is used to weigh the transition probability of the surfer. Secondly, a linear correlation between the query analysis and *JBC* model is designed to endow the ranking algorithm with query-dependent capability. We implemented *MFCRank* in a topic-specific Web search platform. Ranking experiments have been carried out on some topic-specific collections. Experimental results showed that the *MFCRank* algorithm gained better ranking precisions.

In the future, we will use more refined link context analysis methods to improve the stability of the online query analysis, such as by employing the word disambiguation technique. Moreover, the topical characteristics in topic-specific search engine should be studied further to clarify the relationships between the topical regularities and the ranking performance.

References

1. Baeza-Yates, R., Ribeiro-Neto: Modern Information Retrieval. ACM Press Series/Addison Wesley, New York (1999)
2. Page, L.: Pagerank: Bring order to the web. In: Stanford Ditital Libraries Working Paper. (1997)
3. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems **30** (1998) 107–117
4. Kleinberg, J.: Authoritative sources in a hyperlinked environment. In: Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms. (1998) 668–677
5. Chakrabarti, S., Dom, B., Raghavan, P.: Automatic resource compilation by analyzing hyperlink structure and associated text. In: Proceedings of the 7th International WWW Conference. (1998)
6. Richardson, M., Domingos, P.: The intelligent surfer: Probabilistic combination of link and content information in pagerank. In: Advances in Neural Information Processing Systems 14. (2002)
7. H., H.T.: Topic-sensitive pagerank. In: Proceedings of the 11th International WWW Conference. (2002)
8. Jeh, G., Widom, J.: Scaling personalized web search. In: Proceedings of the 12th International WWW Conference. (2003)
9. Jeh, G., Widom, J.: Simrank: A measure of structural-context similarity. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2002)
10. Fogaras, D., Racz, B.: Scaling link-based similarity search. In: Proceedings of the 14th International WWW Conference. (2005)
11. Brian, D.D.: Topical locality in the web. In: Proceedings of the 23rd Annual International Conference on Research and Development in Information Retrieval (SIGIR 2000). (2002)
12. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: Proceedings of the 13th International WWW Conference. (2004)
13. Eiron, N., McCurley, K.S.: Analysis of anchor text for web search. In: Proceedings of of the 26th Annual International ACM SIGIR'03 Conference on Research and Development in Information Retrieval. (2003)
14. Ye, Y., C.L., X., L.L., Z.: The anatomy of a scalable topic-specific web search platform. In: Technical report in ICE research center. (2005)
15. Ye, Y., Ma, F., Lu, Y., Chiu, M., Huang, J.: isurfer: A focused web crawler based on incremental learning from positive samples. In: Proceedings of the 6th Asia-Pacific Web Conference. (2004)
16. Zhang, L., F.Y., M., Ye, Y.: Cala: A web analysis algorithm combined with content correlation analysis method. Journal of Computer Science and Technology **18** (2003) 21–25