Data Collection System for Link Analysis

Bo Yang* and Jian Qin**

* National Science Library, Chinese Academy of Sciences, Postal address, 100080 Beijing, China
Phone: (8610) 82626611, E-Mail: yangb@mail.las.ac.cn

** School of Information Studies, Syracuse University, Syracuse, Postal address, 13244 NY, United States
Phone: (315)443-5642, E-Mail: jqin@syr.edu

Abstract ***–The study presented in this paper exploited several possible ways to meet the needs of link analysis in Webometrics and developed a prototype (LinkDiscoverer) that collects data from both real-time links and search engines. The prototype consists of two parts: a crawling part for collecting real-time link data from a given domain or site and a search engine part for harvesting link data from search engines by using specific search commands. An experiment was conducted to evaluate the performance of LinkDiscoverer on link analysis. The results show that the LinkDiscoverer's functions can well satisfy the needs for link analysis. This study contributes to data collection methods and selection strategy in Webometrics.***

Keywords: ***link analysis, data collection, Web crawler, Webometrics.***

I. INTRODUCTION

The last decade has seen a rapid development in the theories, technologies, and methodologies of Web link analysis. It has great potentials for applications in crime investigation, financial swindling prevention, Web mining (Web search service, commercial competitive information analysis, etc.) and communication (Adibi et al., 2004).

Web link analysis prompted the advent of "Webometrics" (Almind & Ingwersen, 1997). Link analysis has been used as the base for Webometrics research for Web document link patterns (Jepsen et al., 2004), Web impact factor for journals and link patterns between entities on the Internet (Goodrum et al., 2001; Thelwall, 2001a, 2002, & 2003). Web links offer a valuable source not only for developing informetric theory but also for studying link patterns between network entities. However, variability and wide application of link technologies as well as the complexity of link motivation (Kim, 2000) can hamper the effectiveness and reliability of link data. Imbalance of link quality and convenience in creating links are also covert reliability issues in link data. The use of different data harvesting tools and strategies may lead to very different, even opposite, results. These factors make link data harvesting extremely challenging in link analysis. Since the credibility of link analysis heavily relies on the reliability of original data (Bar-Ilan, 2000), it is critical to ensure quality and consistency of the data harvested from the Web. The diversity of data harvesting strategies on different sample sets depends on the flexibility of data harvesting tools. Optimizing the harvesting tools, therefore, would be the first consideration in link analysis study. The performance of a link-harvesting tool can be evaluated based on the following criteria:

- Effectiveness in harvesting links pointing to key resources in the sample set,
- Compatibility of data organization and classification with data analysis tools, and
- Ability to customize different data harvesting strategies, such as choosing data harvesting depth and scope according to different purposes.

Few link-harvesting tools, either commercial or shared software, currently available can satisfy all the above requirements alone. This paper introduces a link-data collection system for link analysis, LinkDiscoverer, as a solution to the problems and errors caused by data harvesting tools. The system is capable of reflecting not only the major webometric indicators mentioned in previous studies, but also some new relationships among link entities.

II. RELATED WORK

Strategies and tools used in link data collection can be divided into the following groups:

- Using popular Internet search engines, such as AltaVista、Google, etc. as data collection tools,
- Combining third-party software and self-developed tools, such as Offline Explorer plus webStat (Duan, 2005), and
- Using self-developed tools alone, such as CheckWeb (Magnusson, 2004), Mike Thelwall's crawler (Thelwall, 2001b), and Lawrence, Bollacker and Gilles's CiteSeer (Giles, Bollacker & Lawrence, 1998).

Popular search engines are sophisticated and capable of Web page grabbing, document indexing, and parallel retrieval. They generally cover massive numbers of Web pages, which makes them suitable as a tool for harvesting link data. Harvesting link data from popular Internet search engines has become a dominant method and approach (Ingwersen, 1998; Vaughan & Thelwall, 2002; Musgrove et al., 2003; Tang & Thelwall, 2003). However, studies have confirmed problems such as low reliability, low stability, and low update frequency existing in link data collected from popular Internet search engines (Rousseau,

2001; Bar-Ilan, 1999; Thelwall, 2005). Although popular Internet search engines have no obvious language bias, their regional coverage trends are skewed. For instance, Web information originated from the U.S. gets much higher coverage than that from Chinese Mainland, Taiwan, and Singapore. This is partially due to early adoption of information technology, various linking trends of sites in different countries and regions, and deep-rooted social and political reasons (Vaughan & Thelwall, 2004). The unexpectedly small number of links returned by Google is an indication that its system has functional limitations for link analysis study. This conclusion has been confirmed by previous study (Qu, Yang & Yan, 2005). Finally, there is a low level overlap ratio in link data from major Internet searching engines. An integrated analysis of the first 1000 links returned by Yahoo、MSN、AltaVista, and AlltheWeb (MSN returns the first 250) revealed that the overlap ratio of the four search engines was less than 40 percent (according to the testing data in this study). Jepsen et al.'s study (2004) also has similar findings.

As a frequently used Web information retrieval tool, popular Internet search engines typically have sophisticated technologies, wide coverage of the Internet, quick response time, and low information retrieval cost. These characteristics make search engines attractive and competitive over smaller Web crawling systems for link analysis, especially in calculating inlink values of given websites. Internet search engines remain the main tool used to collect link data.

While using a single search engine can cause data distortion due to functional limitations for link data collection, such limitations may be minimized by properly combining the data collected from more than one search engine. In theory, there is comparability, even some measurable unity, between the data from search engines and the data from self-developed crawling tools. Although the two are at different levels, the authors argue that the relatively wider coverage of search engines could compensate for the shortcomings of crawling software. In addition, self-developed crawling software could be applied to check the stability, reliability, and Web coverage of search engines.

Having taken into account the pros and cons of search engine and self-developed data harvesting tools, LinkDiscoverer developed in this study combines the merits of the two link data collection approaches. It allows the researchers to gather real-time link data from sampled websites and modify strategies on data harvesting depth and scope in order to collect the most significant link data with the least cost. The combination of the results returned by more than one search engines makes the link analysis research less reliant on the reliability of a single search engine. The prototype LinkDiscoverer achieved these goals by implementing two functional modules—Crawling Part and Search Engine (SE) Part.

III ARCHITECTURE

LinkDiscoverer has two parts (Fig.1): Crawling Part and SE Part. The Crawling Part is used to collect real-time link data from a given domain or site, and the SE Part is used to harvest link data from search engines by specific search commands. The SE Part allows user to specify one or more search engines to automatically submit search command and conduct return results from the selected search engines. All records are stored after duplicate links are removed. The Crawling Part can obtain data about linkage relationships between each case in the sample set and outlink data from each sampled website without any functional limitation on search engines, while the SE Part is used to harvest inlink data from all the other websites on Internet. The two modules can be applied separately or synchronously to reflect the reliability of data collected by different methods.

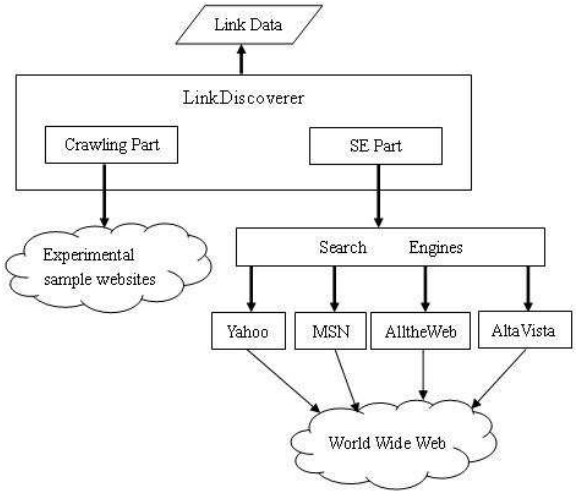

Fig.1 Architecture of LinkDiscoverer

3.1 Crawling Design

3.1.1 System Frame

Crawling Part is similar to the crawler system that search engines usually use. In the crawling process, every thread executes its own crawling scheme independently, provided that one or more start URLs (seeds) are given and crawling rules and page choosing policy are properly set.

The logic structure of Crawling Part can be divided into thread control module and Web crawling module. The thread control module is responsible for crawling rule setting and a series of operations on thread, such as appending, deleting, pause of the thread and state display.

3.1.2 Crawling Rule Setting and Link Classification

Settings for coverage, crawling depth, DNS, and timeout need to be defined before starting a new thread. LinkDiscoverer supports two crawling modes, single site crawling and domain crawling. For instance, if the initial URL is "http://www.domain.edu.cn", under single site crawling mode, LinkDiscoverer only crawls on the www

host and all the other URLs pointing to other hosts will be classified as outlink. By contrast, LinkDiscoverer will visit all the hosts in the domain called "domain" under domain crawling mode, and all URLs out of the "domain" will be classified as outlink.

3.1.3 Rules of Choosing Links and Pages

It is easy to answer the question "what is a link?" from the perspective of hypertext technology. Dale & Lewis (2005) define a link as "a connection between a Web page and another." However, it is difficult to apply this definition in the context of Webometric study. Even though a link may be technically a link, not all links are meaningful enough to warrant a link analysis.

Some Web pages are dynamically generated from a database and have no file extension but only an ID or key number in a HTTP style query. Examples include news release pages and event lists. LinkDiscoverer has the capability of capturing this type of pages.

3.2 Meta-Search Engine Design

There are several similar sites, such as uptimebot (Uptimebot, 2005) and LinkPopularity (LinkPopularity, 2005), which evaluate link popularity by directly calling several search engines. Since the final data they provide are sums of records returned from each search engine without link records, it is difficult to perform more detailed data analysis and synthesize the data from search engines. Thus, it is impossible to minimize the limitation of link data collection performed by a single search engine.

3.2.1 Search Engine Selection

The Search Engine (SE) Part of LinkDiscoverer is designed to minimize the limitation caused by a single search engine. It can parse, analyze and integrate the data returned from four mainstream search engines (Yahoo, MSN, AlltheWeb, and AltaVista) that support advanced link retrieval. Take the search query "domain.edu.cn" as an example, LinkDiscoverer automatically obtains the data returned from each search engine (optional, one or more), parses all links that point to the hosts in "domain", removes the duplicate links and stores them into the database. With a search query "www.domain.edu.cn", only the links on the "www" host in the "domain" will be returned. Since each search engine returns limited numbers of records (Yahoo、AlltheWeb、AltaVista return the first 1000, MSN returns the first 250.), LinkDiscoverer identifies and stores both links and the sum of records retrieved from search engines.

3.2.2 Search Term Selection

Previous studies on link analysis have used terms "link", "site" and "host" to submit retrieval requests to search engines, which could retrieve only the links pointing to the given host (Tang & Thelwall, 2003; Musgrove et al., 2003), but not to the given domain. This problem was solved by combining three search terms "linkdomain", "site" and "host" in LinkDiscoverer. The search term "linkdomain" is a "hidden term" and not even mentioned by three of the four search engines in their help files (Yahoo, 2007; AlltheWeb, 2007; AltaVista, 2007).

Under default condition, LinkDiscoverer submits requests to four search engines synchronously with an initial URL, reads the retrieved records page by page and outputs the sum of search results to log files. To avoid refusals by search engines as an malicious attacks, every thread in LinkDiscoverer allows a randomly limited time delay. Several possible problems may occur during the distilling process and should be taken into account, for example, there may be network problems or some search engines could be too sensitive to respond to the search request. It is necessary to monitor these problems by analyzing the log files after the distilling process is completed. In case of no error, data can be written into the database.

IV EXPERIMENT

The design principles and data harvesting rules are described in detail in the sections above. A large-scale experimental study was conducted to validate the applicability and reliability of LinkDiscoverer. The LinkDicoverer project develops strategies to address the flaws found in popular Internet search engines, investigates potential research values on LinkDiscoverer, and evaluates the feasibility for further development. Websites of universities that generally served as testing samples in Webometric studies are selected to evaluate the performance of LinkDiscoverer through comparative analysis with previous studies. What is discussed furthermore is the possibility of core resources discovery on university websites and institution evaluation by link analysis based on the data harvested from sampled websites.

4.1 Sampling

An ideal way to evaluate the influence of an institutional website is using the the entire Internet as the data source, because the results would be considered objective and complete. This would be very difficult, however, if not impossible, due to the issues mentioned above and current technical limitations. It becomes vital, therefore, to select proper study objects and scale and sound technical strategy so that the conclusion may be reliable and valid. In this study, the top 50 universities in mainland China were selected according to the generally accepted university ranking list (2005) published by China Academy of Management Science (Xinhuanet, 2005). Some cases were excluded for the reason of varying domain names caused by university mergers. The final usable number of sample websites is 47.

4.2 Data Collection

Two data collection processes were applied to 47 university websites respectively to harvest real-time link data and

those from search engines. The real-time data collection process was divided into three periods (the first two were for testing purpose) from October 20th to November 15th 2006. After careful supervision during the first two data collection periods and data analysis on every sampled website, the Crawling Part was improved and data collection strategies such as crawling depth, filtering rules, and other options were altered according to the characteristics on each sample. The last data collection period lasted from November 5th to November 15th 2006. It is better to start the SE part data collection after the same job on the Crawling Part was finished about one month to make the two data sets comparable.

In the end, the number of cases in the SE part was reduced to 47 for some exceptions and that in the Crawling Part was reduced to 44. After data cleaning, the number of links collected from the websites of 47 universities by Crawling Part was 1,696,890 and that from the websites of 44 universities returned by Yahoo, MSN, AlltheWeb and AltaVista in SE Part was 84,890.

4.3 Data analysis

Studies on website link characteristics may uncover hidden correlations among links and such correlations can be measured by indicators such as link distribution, link layering, link depth, and regional link tendency, in addition to those mentioned in the former studies such as inlink, outlink, selflink, WIF, and link density (Ingwersen, 1998; Duan, 2004). Several chief indicators that have not been explored before will be analyzed in this section in order to demonstrate the availability and reliability of LinkDiscoverer and the advantages of the methodology of improving data harvesting tool.

4.3.1 Distribution of Outlink

The scientific power of linked countries/regions is reflected in distribution analysis of outlink by the statistical data, as well as the collaborative closeness between the linking one and the linked. Taking .edu link for example, the top 10 countries/regions that attracted the most inlinks from 47 Mainland China universities were U.S, Taiwan China, UK, Germany, Japan, Canada, France, Australia, Hong Kong China, and Italy.

4.3.2 Depth of Link Coverage

The outlinks deeper than 5 layers on the 47 universities websites that pointed to universities and academic institutions consisted of only 0.8% of the links pointing to those organizations, accounting for a small fraction (0.27%) of all outlinks, and the outlinks deeper than 5 layers to universities and academic institutions on 32 in 47 sample websites were less than ten. It seems that most valuable Web resources tend to keep the layers up to 4 or fewer.

4.3.3 Inlink Coverage

Three domestic regions—Beijing, Shanghai, and Jiangsu—were the most concentrated locations of universities among 47 and were selected to perform inlink analysis. The results showed that the universities in Beijing, Shanghai, and Jiangsu received 269.3, 164.5, and 110.1 inlinks on average respectively. This was coherent with the generally accepted rank in research and education strength, and inlink coverage tends to be regional-centered based on the attraction by research and education strength.

4.3.4 Correlation between Inlink and Other Indicators

The statistical result for the link data collected by Crawling Part showed that the Spearman correlation coefficient between university rankings on comprehensive strength (URCS) and inlink was 0.769(Table 1), which was significant at the 0.01 level. The Spearman correlation coefficient between the integrated analysis of the link data harvested from four search engines combined by SE Part and URCS was 0.616(Table 2), which is by far the best result based on the data harvested by search engines. The correlation coefficient between Web Impact Factors (measured by inlink number) and URCS was as high as 0.769, and that between the host numbers and URCS was 0.715(Table 3).

TABLE 1. Correlation between WIF (Crawling Part) and URCS

Correlations

			rank	inlink
Spearman's rho	rank	Correlation Coefficient	1.000	.769**
		Sig. (2-tailed)	.	.
		N	47	47
	inlink	Correlation Coefficient	.769**	1.000
		Sig. (2-tailed)	.	.
		N	47	47

**. Correlation is significant at the 0.01 level (2-tailed).

TABLE 2. Correlation between WIF (SE Part) and URCS

Correlations

			Rank	Yahoo	MSN	AlltheWeb	AltaVista	MAX	AVG
Spearman's rho	Rank	Correlation Coefficient	1.000	.130	.616**	.366*	.367*	.583**	.598**
		Sig. (2-tailed)	.	.402	.000	.014	.014	.000	.000
		N	44	44	44	44	44	44	44

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

TABLE 3. Correlation between host numbers *(Crawling Part) and URCS*

Correlations

			rank	host
Spearman's rho	rank	Correlation Coefficient	1.000	.715**
		Sig. (2-tailed)	.	.
		N	47	47
	host	Correlation Coefficient	.715**	1.000
		Sig. (2-tailed)	.	.
		N	47	47

**. Correlation is significant at the 0.01 level (2-tailed).

The sampled institutions were evaluated by host numbers analysis (selflink) and inlink analysis (inlink attracted from the other samples) respectively. The results received from the two different ways were quite coherent and the correlation is significant. On the one hand, it is not coincident that two statistical methods and indicators

validate the theoretic presumption that host numbers analysis and inlink analysis are valuable for institution evaluation; on the other hand, the reliability of LinkDiscoverer on institution evaluation is approved by better regularity of the link data.

4.3.5 Use of Link Data in Resource Ranking and Type Identifying

In the process of selecting key websites, some websites providing open access resources were identified, such as www.ncbi.nlm.nih.gov (free molecular biology information) and arXiv.org (more than 4 million e-print resources in the fields of physics, mathematics, non-linear science, computer science, and quantitative biology).

There were a wide variety of university types among the 47 universities, including comprehensive universities, normal universities, medical universities, and technical/engineering universities. In theory, universities in the same type would be similar in linking interest and hence have similar link behavior. The similarity of link behavior can be helpful in type identification and sample clustering based on link behavior. The data harvested by LinkDiscoverer showed that the similarity of link behaviors among sample cases did present similar linking interests based on a high correlation coefficient. It is therefore possible to identify the discipline areas of an institution by examining to which institution types the website belongs.

4.4 Summary

In the experiment presented above, 47 universities websites were selected and several facets and indicators were measured to evaluate the feasibility and reliability of LinkDiscoverer in link analysis. LinkDiscoverer was able to effectively identify and harvesting core links pointing to core resources on universities websites through analyzing outlink distribution (section 4.3.1) and using link data in resource ranking (section 4.3.5). LinkDicoverer was also able to overcome some limitations of offline browsing tools and search engines in terms of data organization and classification, and the Crawling Part improved the efficiency of data collection through identifying link types online and storing link data in the format of link analysis in the end. In the SE Part, record numbers returned from every search engine were recorded and duplicate inlinks were removed. The modes of data organization and classification in LinkDiscoverer paved the way for further data analysis by data mining tools. LinkDiscoverer allows for configuration of data harvesting breadth, depth, and scope so that users can customize flexible data harvesting strategies.

The analysis results mentioned above suggest that in both the traditional field of Web impact factor studies (Ingwersen, 1998; Duan, 2004) and the new ones such as link analysis, the data collection system designed in the theoretic frame of this study—LinkDiscoverer was able to perform quantitative link analysis. Furthermore, LinkDiscoverer offers innovative methods for link analysis and several link patterns that were first discovered by this study, including link layering, regional link tendency, and WIF. Link clustering and other potential themes worthy further discussion in link analysis were also detected.

V Discussion

LinkDiscoverer is different from other link analysis tools such as CheckWeb and Offline Explorer in several ways:

- Functionality. Both CheckWeb and offline browser have limitations in information extraction and page filtering, whereas the most key functions for link analysis are provided by LinkDiscoverer.
- Flexibility and extensibility. LinkDiscoverer allows for customizing data collection strategies. In contrast, neither CheckWeb nor Offline Explorer supports extendible URL filtering mechanism.
- Efficiency. Large-scale link data analyses requires offline browsers to work together with data mining tools to parse out link data. Page download and data extraction can be completed almost simultaneously in LinkDiscoverer. Pages were are discarded after data extraction, which can save dish space and improve the efficiency of data collection.
- Data reproducibility. None of the current tools can reproduce link data. Although the link data for revalidating the result can not be reproduced, the data collection policies and methods can. Right policies and methods do not necessarily lead to a right result, while a wrong policy and method will be doomed to produce wrong results. Objective conclusion on whether a link data collection system is effective or not could be drawn on the basis of crawling policies and scope of resource objects (Cothey, 2004), which were clearly defined in LinkDiscoverer.

LinkDiscoverer performed well in functionality, flexibility, extendibility and efficiency as it was presented in section 4 and satisfied the basic needs of link analysis, and even made some breakthroughs in analytical methodology.

VI CONCLUSION

The paper describes a data collection system for link analysis, LinkDiscoverer, which is aimed at resolving some of the problems of data collection in link analysis, such as link data harvested from specific domains, non-hypertext files filtering, control of crawling depth and scope and combined use of multiple search engines. Further studies are however needed to address issues such as website boundary demarcation, discrimination of multi-domain institutions, link filtering and processing of script links. Further developments of LinkDiscoverer will investigate and seek solutions to the remaining problems.

REFERENCES

[1] Adibi, J., Chalupsky, H., Grobelnik, M., Mladenic, D. & Milic-Frayling, N. (2004). KDD-2004 Workshop Report Link Analysis and Group Detection (LinkKDD-2004). ACM SIGKDD Explorations Newsletter, 6(2), 136-139.

[2] AlltheWeb. (2007). AlltheWeb.com: Frequently Asked Questions - Query Language. Retrieved Jun. 24, 2007 from http://www.alltheweb.com/help/faqs/query_language.

[3] Almind, T.C. & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological approaches to 'Webometrics'. Journal of Documentation, 53(4), 404-426.

[4] AltaVista. (2007). AltaVista - Special search terms. Retrieved Jun. 24, 2007 from http://www.altavista.com/help/search/syntax.

[5] Andrei, Z.B., Najork, M. & Wiener, J.L. (2003). Efficient URL Caching for World Wide Web Crawling. 12th International World Wide Web Conference, (pp.679-689). New York: ACM Press.

[6] Bar-Ilan, J. (2000). Data collection methods on the Web for informetric purposes: A review and analysis. Scientometrics, 50(1), 7-32.

[7] Bar-Ilan, J. (1999). Search engine results over time - A case study on search engine stability, Cybermetrics, 2/3(1). Retrieved Jun. 24, 2007 from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html.

[8] Cothey, V. (2004). Web-Crawling Reliability. Journal of the American Society for Information Science and Technology, 55(14), 1228-1238.

[9] Dale, N. & Lewis, J. (2005). Computer Science Illuminated (2nd Ed.). Beijing: China Machine Press.

[10] Duan, Y.F. (2004). Research on Web Link Analysis and Website Evaluation. Ph.D. dissertation, Wuhan University, China.

[11] Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. Proceedings of the third ACM conference on Digital libraries, (pp. 89-98). New York: ACM Press.

[12] Goodrum, A.A., McCain, K.W., Lawrence, S., & Giles, C.L. (2001). Scholarly publishing in the Internet age: a citation analysis of computer science. Information Processing and Management, 37(5), 661-675.

[13] Ingwersen, P. (1998). The Calculation of Web Impact Factors. Journal of documentation, 54(2), 236-243.

[14] Jepsen, E.T., Seiden, P., Ingwersen, P., Björneborn, L., & Borlund, P. (2004). Characteristics of Scientific Web Publications: Preliminary Data Gathering and Analysis. Journal of the American Society for Information Science and Technology, 55(14), 1239-1249.

[15] Kim, H.J. (2000). Motivations for hyperlinking in scholarly electronic articles: A qualitative study. Journal of the American Society of Information Science and Technology, 51(10), 887-899.

[16] LinkPopularity. (2007). LinkPopularity.com: The Free Link Popularity Service. Retrieved Jun. 25, 2007 from http://www.linkpopularity.com/.

[17] Magnusson, C. (2004). CheckWeb. Retrieved Sep. 20, 2004 from http://www.algonet.se/~hubbabub/how-to/checkweben.htm.

[18] MSN. (2007). Web Search Help: Search Builder and advanced search options. Retrieved Jun. 25, 2007 from http://search.msn.com/docs/help.aspx?t=SEARCH_REF_AdvSrch Operators.htm#2A.

[19] Musgrove, P.B., Binns, R., Page-Kennedy, T. & Thelwall, M. (2003). A method for identifying clusters in sets of interlinking Web spaces. Scientometrics, 58 (3), 657-672.

[20] Najork, M. & Wiener, J.L. (2001). Breadth-first search crawling yields high-quality pages. Proceedings of the 10th international conference on World Wide Web, (pp. 114-118). New York: ACM Press.

[21] Patterson, A. (2004). Why writing your own search engine is hard. ACM: Queue, 2(2), 48-53.

[22] Qu, W.Q., Yang, B. & Yan, S.L. (2005). Study on the Relationship between JIF &WIF of Agricultural Journals. Chinese Journal of Scientific and Technical Periodical, 16(5), 658-661.

[23] Rousseau, R. (2001). Evolution in time of the number of hits in keyword searches on the Internet during one year, with special attention to the use of the word euro. Proceedings of the 8th International Conference on Scientometrics & Informetrics, (pp. 619-627). Sydney: ISSI.

[24] Schildt, H. & Holmes, J. (2003).The art of java. California: McGraw-Hill/Osborne.

[25] Tang, R. & Thelwall, M. (2003). U.S. academic departmental Web-site interlinking in the United States Disciplinary differences. Library & Information Science Research, 25(4), 437-458.

[26] Thelwall, M. (2001b). A Web Crawler Design for Data Mining. Journal of Information Science, 27(5), 319-325.

[27] Thelwall, M. (2003). Can Google's PageRank be used to find the most important academic Web pages?. Journal of Documentation, 59(2), 205-217.

[28] Thelwall, M. (2002). A comparison of sources of links for academic Web impact. Journal of Documentation, 58(1), 66-78.

[29] Thelwall, M. (2001a). Extracting Macroscopic Information from Web Links. Journal of the American Society for Information Science and Technology, 52(13), 1157-1168.

[30] Thelwall, M. (2004). Methods for reporting on the targets of links from national systems of university Web sites. Information Processing and Management, 40(1), 125-144.

[31] Uptimebot. (2005). Link popularity check at Uptimebot. Retrieved Dec. 10, 2005 from http://www.uptimebot.com/.

[32] Vaughan , L. & Thelwall, M. (2004). Search engine coverage bias: evidence and possible causes. Information Processing & Management, 40(4), 693-707.

[33] Vaughan, L. & Thelwall, M. (2002). Web Link Counts Correlate with ISI Impact Factors: Evidence from Two Disciplines. Proceedings of the 65th Annual Meeting of the American Society for Information Science and Technology, (pp. 436-443). Maryland: Asist.

[34] Xinhuanet. (2005). Top 100 in Mainland China University Rankings 2005. Retrieved Dec. 20, 2005 from http://news.xinhuanet.com/edu/2005-06/20/content_4801454.htm.

[35] Yahoo. (2007). Yahoo! Help–Search. Retrieved Jun. 27, 2007 from http://help.yahoo.com/help/us/ysearch/basics/basics-04.html.