An analysis of backlink counts and Web impact factors for Chinese university Websites*

Qiu Junping Chen Jingquan
Research Center for Chinese Science Evaluation, Wuhan University, Wuhan, China 430072
Email: jpqiu@whu.edu.cn, ccjjq@163.com
Wang Zhi
Department of Statistics, North Carolina State University ,U.S.A 27606
Email: wzhi2000@yahoo.com

Keywords university Websites, backlink count, Web Impact Factor

Abstract This article aims to study the total backlink counts, external backlink counts and the Web Impact Factors (WIFs) for Chinese University Websites. By studying whether the backlink counts and WIFs of Websites associate with the comprehensive ratings and the research ratings for Chinese universities, the article demonstrates that the external backlink count can be a better evaluation measure for University Websites than WIF. The study also investigated issues about data collection by using different search engines. It shows that data collected by Alta Vista are more stable than AllTheWeb.

1 Introduction

The Internet is mainly comprised of a large number of web pages with links to one another. In some sense, the web links between sites or pages are analogous to the citations in published papers. Inspired by this analogy, McKiernan (1996) creatively proposed the term “sitation” to describe the citation phenomenon within Websites on the Internet. Ingwersen (1998) proposed the concept of Web Impact Factor (WIF) to measure the overall online impact of a domain or a Website. Whereafter, more and more researchers come to this field. A sitation analysis was carried out by Rousseau (1997) and the Lotka function was found to provide an adequate description for the distribution of domain names and links among web sites. Although no significant correlation was found between the equivalent of the AltaVista original general WIF and research output for Australian universities (Smith,1999) and neither between Web link counts and UK Research Assessment Exercise ratings for library and information science departments(Thomas and Willet, 2000), results from recent advances (Thelwall, 2001) in sitation study have demonstrated that the hyperlink structure of national university systems is strongly related to the research productivity of the individual institutions. Furthermore, a study by Thelwall (2002) shows that there is high correlation for British universities between research ratings and four different WIFs calculated from several different source domains.

In some previous studies, the links to a Website was considered equivalent to the web pages containing at least one link to the given Website, which were named by

Junping Qiu and Jingquan Chen, *An analysis of backlink counts and Web impact factors for Chinese university websites*, In: Guohua Jiang, Ronald Rousseau, Yishan Wu (Eds), Proceedings of the 9th International Conference on Scientometrics and Informetrics – ISSI 2003, Dalian University of Technology, Dalian, China, 2003, 221-229.

Ingwersen (1998) as link-pages. Consequently, the number of links for a given Website is equal to the number of link-pages. However, a link-page does not necessarily contain only one link to a given site. So the number of link-pages is not exactly the same as the number of links. Actually, the word “link” in previous studies and this paper means “backlink” or “in-link”, which is the link to a given web site or page, other than “out-link”, which is the link contained in a given web site or page while pointing to other web sites or pages.

Up to the present, few researches concern about Chinese Websites. This research aims to study whether the backlink counts and WIFs of Websites associate with the comprehensive ratings and the research ratings for universities in mainland China. Two different search engines are used respectively for data collection in the study. The authors expected to find the best measure among the external backlink counts, the total backlink counts and WIFs calculated with different methods. The stability of results from different search engines is also discussed.

2 Methodology

2.1 URLs of universities

The universities chosen for the study are the top 100 universities in Chinese University Rating 2002 (in mainland China) published by Guangdong Research Institute of Management, which is considered as the most professional and authoritative university rating in China.

The URLs or domain names of the universities were identified from lists on the China Ministry of Education Website (http://www.moe.edu.cn/) and China Education and Research Network (http://www.edu.cn/). Considering some university Websites have multiple URLs, several Websites about Chinese universities navigation had been also visited to obtain the different URLs for each university as completely as possible. All the URLs are tested by visiting the university Websites with them.

2.2 Selection of search engine

In parallel to the ISI (the Institute for Scientific Information) citation databases used in bibliometric studies, the Internet search engine databases, such as Alta Vista, AllTheWeb, Northernlight, Google etc, provide data for webometric studies. AltaVista and AllTheWeb were selected to collect data for the study. Compared to the Northernlight databases, AltaVista and AllTheWeb have more extensive coverage. The AllTheWeb index databases have extended greatly in recent years. Up to the June 2002，AlltheWeb has collected approximately 2.1 billion web pages as the largest web index database beating Google with 2.07 billion web pages. Although Google has enough coverage on Web, it has no function to distinguish internal backlink from external backlink. Also the “link:host” command provided by Google only retrieves the links to the homepage of the host Website, other than the links to all the WebPages of the site.

The Chinese search engine Baidu (www.baidu.com) also has the "link:host" command. However, after December 2002 the "link:host" command no longer works on Baidu. So the Baidu search engine was not included in the study.

2.3 Search method

For some reason, AltaVista cannot be visited in China for several months. So it was in the United States that we carried out searches with AltaVista. Restricted by time, we only performed searches for the Websites of the top 50 universities in China and didn't perform the "site: host" search to retrieve pages in the host sites.

We used the advanced text searching mode provided by AltaVista, and selected the "count only" option. Two search queries were used for every university Website. For instance, the following two retrieval commands have been applied to Nanjing University Website:

```
1 link:www.nju.edu.cn
2 link:www.nju.edu.cn AND NOT host:www.nju.edu.cn
```

By means of the first command the sum of links to Nanjing University Website can be obtained, which is the total backlink counts. And the external backlink counts of Nanjing University Website can be identified by the second one.

Because of university merger or other reasons, some university Websites have two URLs or more. For these Websites with multiple URLs, the elaborate Boolean search strings were used. Take Fudan University Website as an example, which has two URLs: "www. fudan .edu.cn" and "www. shmu.edu.cn", the following elaborate Boolean searches were conducted:

```
1 link:www.fudan.edu.cn OR link:www.shmu.edu.cn
2 (link:www.fudan.edu.cn OR link:www.shmu.edu.cn) AND NOT
   host:www.fudan.edu.cn AND NOT host:www.shmu.edu.cn
```

The first Boolean string retrieves the number of links to Fudan University Website. And the second one retrieves the number of external links to Fudan University Website.

Differed from Alta Vista , the advanced search in AllTheWeb is provided with search filters. This search mode allows for the use of word filters, domain filters, IP-address filters and result restrictions etc in a specific search. The word filters could meet our needs in this study. For example, choose word filters and enter

"must include www. nju.edu.cn in the link to url"

The total backlink count for Nanjing University website will be obtained.

And enter

"must include www. nju.edu.cn in the link to url and
must not include www. nju.edu.cn in the url"

The external backlink count for Nanjing University website will be retrieved.

For those university websites with more than one URL, the multiple searches were needed. For example, Donghua University has two URLs: "www.dhu.edu.cn" and "www.ctu.edu.cn". To gain the total backlink count for Donghua University website, the search filters were conducted twice. First, enter “must include www.dhu.edu.cn in the link to url”, and the result is 464. Then enter “must include www. ctu.edu.cn in the link to url”, and the result is 2243. The sum (2707) of 464 and 2243 is the total backlink count.

The search filters also need to be conducted twice to obtain the external backlink count.

Enter

“must include link:www.dhu.edu.cn in the link to url and
must not include: www.dhu.edu.cn in the url and
must not include: www.ctu.edu.cn in the url”

The number of hits is 493.

Enter

“must include link:www.ctu.edu.cn in the link to url and
must not include: www.dhu.edu.cn in the url and
must not include: www.ctu.edu.cn in the url”

The number of hits is 2231.

The sum (2724) of 493 and 2231 is the external backlink count of Donghua University website.

It is particularly worth noting that the results of AllTheWeb are inconsistent in the cases of the counts are very large (usually larger than 5,000). In these cases, AllTheWeb gave two different result counts for each search: a large number and a small one. For example, in the search for the total backlink count of Wuhan University Website, we entered “must include www.whu.edu.cn in the link to url”. Then the first page of the results showed “Displaying results 1-10 of 17,964 web pages found”, while the second page of the results showed “Displaying results 11-20 of 4,213 web pages found”. Following pages were clicked, the results showed 4,213 constantly. Which is the result? 17,964 or 4,213? The same question appears when try to use the homologous queries for the top 25 universities. Sometimes the small number of results does not appear until the sixth or seventh page of the results is examined, while the first page always shows the large number.

We sent a Bug Report at December 04, 2002 9:50 PM to Task Team of Fast Search & Transfer, Inc. who gave the answer quickly. The Task Team explained that different operational circumstances could cause the different number sets that users see. That tends to happen while they are merging in a new index. As it merges in, there are certain areas that are not available for a period of time while they are copying in the information. While one page of results is delivered to user from a completely merged in area, the other page may not be able to access all the available information, temporarily, when the new information is being merged in. And they suggested that if the problem appears again, the large number should be considered . For the case of the Wuhan University Website, its total backlink count is 17,964, not 4,213. However, the number of 17,964 only appears on the first page of the result, while the other pages all give the number of 4,213 as the total backlink count steadily. Further, the low number is far closer to the results in Google and Altavista than the large one. So it seems that the small number is more down-to-earth and we decided to use it. We sent our viewpoint to the Task Team but have not received their answer any more.

One month later, we used AllTheWeb to conduct experimental searches again and found the problem still exists. The only difference is that the small number is more difficult to find than last time. It usually doesn't appear until the fortieth or fiftieth page of the results is examined.

2.4 Calculation for WIF

The impact factor is a measure of the frequency with which the "average article" in a journal had been cited in a particular year or period. The annual JCR impact factor is a ratio between citations and recent citable items published (Garfield, E. 1994). Thus, the impact factor of a journal is the number of current year citations divided by the source items published during the previous two years in that journal. For example:

Let
A= total cites in 2002
B= 1992 cites to articles published in 2000-2001 (this is a subset of A)
C= number of articles published in 2000-2001, then
D= B/C = 2002 impact factor

Considering the dynamic real-time nature of link, it is difficult and unnecessary to identify the time when the links to a given Website were created. Therefore the calculation for Web Impact Factor (WIF) does not consider the time axis, other than the journal impact factor. The WIF introduced by Ingwersen is the ratio of the number of backlinks to a site, divided by the number of pages at the site. In this article we name it as WIF_p.

Let
E= the number of external backlinks to a given Website
P= the number of pages in the given Website, then
$WIF_p = E/P$

The WIF has some modification by Mike Thelwall (2001) in his WIF calculation for British university Website, and the denominator of the calculations was changed to the number of faculty members working for the university associated with the Website. In this article we name it as WIF_f.

Let
E= the number of external backlinks to a university Website
S= the number of faculty members in the university, then
$WIF_f = E/S$

Probably it is the difference between university Websites and academic journal Websites that motivates Thelwall to modify the calculation for WIF. Differed from academic journal Websites, not every page in university Websites has academic content. And maybe the university reputation rather than the content of the university Website constitutes the incentive for the links to university Website. So a measure of average backlinks per page in a Website might not be a useful tool for Web site evaluation. Based on the same motivation, we proposed a new expression for WIF as following, which we name it as WIF_c:

Let

E= the number of external backlinks to a university Website

C= the number of colleges or departments in the university, then

$WIF_c = E/C$

The number of faculty members and the number of colleges or departments in universities can be obtained from university Websites. The three versions of calculations for the WIF will be used in this study. The WIFs, the total backlink count and the external backlink count will be compared with the comprehensive ratings and the research ratings respectively for universities in mainland China. Kendall correlation tests will be conducted. The correlation analysis will be undertaken with SAS (Statistical Analysis System).

2.5 Results

Table 1 shows Kendall correlation coefficients measured among total backlink counts, external backlink counts, three WIFs of the 98 university Websites, general ratings and research ratings of the universities. This is the result after excluding the data of Zhongshan University and Beijing Normal University, of which the backlink counts are far higher than other universities. The backlink counts and pages of the university Websites are all retrieves by AllTheWeb. “TLINK” in the table represents the total backlink counts, and “ELINK” represents the external backlink counts.

Table 1

	AllTheWeb TLINK	AllTheWeb ELINK	WIF_p	WIF_f	WIF_c
General score	0.58060 (p<.0001)	0.58382 (p<.0001)	-0.02982 （p=0.6669）	0.03261 （p=0.6343）	0.35283 (p<.0001)
Research score	0.57146 (p<.0001)	0.57889 (p<.0001)	-0.01294 （p=0.6040）	0.09701 （p=0.8519）	0.36679 (p<.0001)

Table 2 shows Kendall correlation coefficients measured among several variables for the top 50 universities in China. WIFs in the table are all calculated with the data from Alta Vista. Limited by the search time in America, only data for 50 university Websites have been retrieved and the page counts for each Website are absent. So the WIF_p has not been calculated.

Table 2

	Alta Vista TLINK	Alta Vista ELINK	WIF_f	WIF_c
General score	0.51776 (p<.0001)	0.55429 (p<.0001)	0.10041 (p=0.3035)	0.39429 (p<.0001)
Research score	0.52103 (p<.0001)	0.56735 (p<.0001)	0.19837 (p= 0.0421)	0.43020 (p<.0001)

Compared with data in table 2, Table 3 shows Kendall correlation coefficients based upon the data retrieved by AllTheWeb for the top 48 Chinese universities. This is the result after excluding the exceptional data of Zhongshan University and Beijing Normal University.

Table 3

	AllTheWeb TLINK	AllTheWeb ELINK	WIF_p	WIF_f	WIF_c
General score	0.48936 (p<.0001)	0.50288 (p<.0001)	0.06915 (p=0.4881)	-0.00887 (p=0.9292)	0.35461 (p=0.0004)
Research score	0.51596 (p<.0001)	0.52594 (p<.0001)	0.04255 (p=0.6697)	0.09220 (p=0.3553)	0.39539 (p<.0001)

Tables 1 to 3 show that total backlink counts and external backlink counts, especially the latter, have significant correlation with university ratings. WIF_c is also correlated with university ratings, but is weaker than backlink counts do. However, differed from Thelwall’s study (2001) for British universities, WIF_p and WIF_f have not been found significantly correlated with Chinese university ratings.

Comparing Table 2 with 3, the results based on the two different search engines are consistent, while stronger correlation of the results can be seen when Alta Vista is used.

Table 4 shows Kendall correlation coefficients measured among several variables for 80 Chinese universities after excluding the data of 18 merged universities as well as the exceptional data of Zhongshan University and Beijing Normal University. We hope to determine the influence of the university merger on the results. The results are based on the data retrieved by AllTheWeb.

Table 4

	AllTheWeb TLINK	AllTheWeb ELINK	WIF_p	WIF_f	WIF_c
General score	0.59449 (p<.0001)	0.59725 (p<.0001)	-0.11076 （p= 0.1459）	0.09114 （p=0.2315）	0.38734 (p<.0001)
Research score	0.59785 (p<.0001)	0.61010 (p<.0001)	-0.07407 （p=0.3309）	0.16461 （p=0.0307）	0.41152 (p<.0001)

Table 5 shows Kendall correlation coefficients measured among several variables for 40 Chinese universities after excluding the data of the 10 merged universities in recent years. The results are based on the data retrieved by Alta Vista.

Table 5

	Alta Vista TLINK	Alta Vista ELINK	WIF_f	WIF_c
General score	0.54359 (p<.0001)	0.59744 (p<.0001)	0.26667 （p= 0.0154）	0.51026 (p<.0001)
Research score	0.56154 (p<.0001)	0.63077 (p<.0001)	0.34615 （p= 0.0017）	0.54359 (p<.0001)

Comparing Table 4 with Table1 as well as Table 5 with Table 2, more significant correlation was found after excluding the data of merged universities in recent years, especially for the cases of WIF_f and WIF_c.

Except for Table1, the other four tables show that total backlink counts, external backlink counts and three WIFs have slightly higher correlation with university research ratings than with university general ratings.

3 Conclusion and discussion

3.1 External backlink counts are highly correlated with Chinese university ratings

This confirms the assertions (Ingwersen, 1998; Harter & Ford, 2000; Thelwall, 2001) that external backlinks may represent a useful measure. The total backlink count and WIF_c are significantly correlated with Chinese university ratings too. Unlike the relationship above, there seems to be little correlation between WIF_p and university rating, which demonstrates the way of calculating WIF for academic journal Websites is not suitable for university Websites in China. This might be due to the lack of academic information in university Websites. If there are more introductory content than academic content in a Website, it might be improper to evaluate the Website by the measure of average external impact per page. Furthermore, the method of WIF_p calculation doesn't fit the university Websites with too many pages or few pages, because WIF would be far lower or higher than others.

In most cases presented in Table 1--5, WIF_f doesn't have significant correlation with general ratings and research ratings, which indicates the measure of average external impact per faculty member is not useful to evaluate Chinese university Websites. Generally web pages of university are organized by the colleges and departments. Accordingly, it's more appropriate to divide the external backlink counts of university Website by the number of colleges or departments. And the highest correlation between university ratings and WIF_c among three versions of WIFs supports this viewpoint.

However, the significance of WIF_c is depressed by the fact that among all variables external backlink counts have the highest correlation with Chinese university ratings. Moreover, the high correlation between WIF_c and university ratings might be attributed to the similar college and department count for all universities. According to the data retrieved by AllTheWeb, the variation coefficient of college and department count is 34.53, lower than those of Web page count (126.47) and faculty members count (49.83).

On the other hand, the difference and disproportion of colleges in different universities might account for the phenomenon that WIF_c haven't shown any advantage vs. simple backlink counts as a measure for the evaluation of university Websites. Another reason might be that some colleges or departments haven't established their own Web pages. Also it's notable that many links to university Websites just link to the homepage of the websites other than to pages of colleges or departments. We hope to verify this point in the future research.

3.2 The backlink counts of university Websites can be related to the reputation (especially the academic reputation) of the university

The total backlink counts and external backlink counts are highly correlated with general ratings and research ratings of Chinese universities, especially with the latter. The higher the university status or reputation (especially the academic reputation) is, the more links point to its Website. Differed from academic journal Websites, the incentive of creating links to the university Websites is based on the university status rather than the Website content. It is also supported by the fact that WIF_p is not a good measure for university website evaluation.

3.3 The influence of university merger on the results

In recent years, the URLs of some Chinese university Websites have changed because of university mergers in China. This brought some trouble to our study. In some cases, the URLs of the original universities have been reserved and all point to the new university Website. Thus, the backlink count of the Website of the new university will be larger than other universities at the same level in university ratings, such as Wuhan University and Fudan University. In some cases, some URLs of the original universities were abandoned and only one URL is available for the new university, then the backlink count of the university Website will be less than other universities at the same level, such as Zhejiang University, Shandong University etc. This demonstrates that links to university Websites haven't been updated in time. All of this could affect the accuracy of the results. Excluding the data of universities merged in recent years, the correlations are much stronger evidently, as showed in Table 4 and Table 5.

3.4 The stability of search engines

Results from different search engines (Alta Vista and AllTheWeb) are consistent and stable in the study. With the largest web index on the Internet today, AlltheWeb returned more comprehensive set of search results than Alta Vista. However, the stability and reliability of Alta Vista seems better than AllTheWeb. No exceptional data have been found in the results of Alta Vista while some have been found in that of AllTheWeb. Furthermore, users of AllTheWeb must keep cautious when the number of hits is larger than 5,000 in the backlink search, which was described in the above paragraphs.

* This article is supported by National Natural Science Foundation of China (70273032).

References

Garfield, E. (1994). The Impact Factor, http://www.isinet.com/isi/hot/essays/journalcitationreports/7.html

Harter, S. P., & Ford, C. E. (2000). Web-Based Analyses of E-journal Impact: Approach, Problems, and Issues. Journal of the American Society for Information Science and Technology, 52(13):1159-1176

Ingwersen, Peter. (1998). 'The calculation of Web Impact Factors', Journal of Documentation, 54(2), 236-243.

McKiernan, G. (1996). CitedSites(sm): Citation Indexing of Web resources. http://www.public.iastate.edu/~CYBERSTACKS/Cited.htm

Oppenheim, C. (1995), "The correlation between citation counts and the 1992 Research Assessment Exercise Ratings for British Library and Information Science University Departments", Journal of Documentation, 51(1), 18-27.

Rousseau, Ronald.(1997). Sitations: an exploratory study, Cybermetrics, 1(1) http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Smith, A.G. (1999), "A tale of two web spaces: comparing sites using Web Impact Factors", *Journal of Documentation*, 55(5), 577-592.

Thomas, O. and Willett, P. (2000), "Webometric analysis of departments of librarianship and information science", *Journal of Information Science*, 26(6), 421-428.

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

Thelwall, M. (2002) A Comparison of Sources of Links for Academic Web Impact Factor Calculations, Journal of Documentation, 58(1), 60-72.

A revised version of this contribution has been published as:

Junping Qiu, Jingquan Chen and Zhi Wang (2004). An analysis of backlink counts and Web impact factors for Chinese university websites. Scientometrics, 60(3), 463-473.