Linked Title Mentions: A New Automated Link Search Candidate[1]

Pardeep Sud, Mike Thelwall
Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK.

Many webometric studies have used hyperlinks to investigate links to or between specific collections of websites to estimate their impact or identify connectivity patterns. Whilst major commercial search engines have previously been used to identify hyperlinks for these purposes, their hyperlink search facilities have now been shut down. In response, a range of alternative sources of link data have been suggested, but all have limitations. This article introduces a new type of link that can be identified from commercial search engines, linked title mentions. These can be found by querying title mentions in a search engine and then removing those not associated with a relevant hyperlink. Results of a proof of concept test on 51 U.S. library and information science schools and four other sets of schools suggest that linked title mentions may tend to give better results than title mentions in some cases when used for site inlinks but may not always be an improvement on URL citations. For links between or co-inlinks to specified pairs of academic websites, linked title mentions do not generally provide an improvement over title mentions, but they do over URL citations in some cases. Linked title mentions may also be useful for sets of non-academic websites when the alternatives give too few or misleading results.

1. Introduction

A major task within the field of webometrics is to identify links between or to specific collections of organisational websites, such as universities (Aguillo, Granadino, Ortega, & Prieto, 2006; Lepori, Barberio, Seeber, Aguillo, 2013; Ortega & Aguillo, 2009; Seeber, Lepori, Lomi, Aguillo, & Barberio, 2012; Thelwall & Harries, 2004), academic departments (Chu, He, & Thelwall, 2002; Li, Thelwall, Musgrove, & Wilkinson, 2003), organisations contributing to a research area (Park, 2010; Thelwall, Klitkou, Verbeek, Stuart, & Vincent, 2010), academic journals (Smith, 1999; Vaughan & Hysen, 2002), political parties (Romero-Frias & Vaughan, 2010), or businesses (Vaughan & Yang, 2012). The links found can be used to identify the most important websites within the group examined, at least in terms of online impact. Alternatively, the objective of a link analysis study could be to identify clusters or groupings within a network of websites either to infer similar offline groupings or to understand the overall web presence of the collection (e.g., Heimeriks, Hoerlesberger, & Van den Besselaar, 2003). The earliest webometric studies used the search engine AltaVista's advanced hyperlink search features, which were later available in Bing and Yahoo!. For a while, hyperlinks could also be automatically searched for using advanced queries submitted to the Applications Programming Interface (API) of Yahoo!, which made it possible to conduct large scale link analyses quickly and efficiently. Since then, however, all major search engines have ceased to provide useful and general hyperlink searches and so webometricians have designed and assessed alternatives.

As discussed in the literature review section, although there are some commercial sources of link data, there are currently two main free ways in which to gather hyperlink data: personal web crawlers and the Alexa.com website. The former is time-consuming and is impractical for identifying the links between collections of large websites or for identifying links to websites from the rest of the web. The latter cannot be automated and cannot be directly used to identify the links to websites without their own canonical domain name (e.g., links to nd.edu can be found from Alexa but not links to math.nd.edu because it is a derivative domain name) and cannot be used to identify the links between websites (e.g. the number of links from nd.edu to ndu.edu.lb). In response to the difficulties in obtaining hyperlink counts for many webometric purposes, two new methods have been developed to identify alternative types of implicit or explicit inter-document connections, *URL citations* and *title mentions*, as described in the literature review. The former is a somewhat unnatural type of connection and can be used for irrelevant purposes and the latter can produce too many false matches in some cases.

In response to the need for link data and the limitations of current sources of hyperlink data and alternatives to them, this article introduces a new way to use web search engines to identify hyperlinks to or between websites. The method, *linked title mentions*, has two stages. First, commercial search engine queries identify web pages that mention the target website name. Second, these web pages are checked for hyperlinks to the target website, removing all web pages without any such links. Although this method does not identify all hyperlinks to the target website it has the advantage that it can be fully automated (and is now available free in the program Webometric Analyst, lexiurl.wlv.ac.uk) and can, in theory, be used to identify all types of hyperlinks needed for webometric purposes (i.e., site inlinks, co-inlinks, direct links), except when there are more results than the search engine will return (1000 is normally the maximum per query).

2. Literature Review: Methods of counting links

Interest in counting hyperlinks to websites began when the now defunct commercial search engine AltaVista introduced a range of hyperlink search commands, including `linkdomain:` (Ingwersen, 1998; Rodríguez i Gairín, 1997). This allowed any web user to request a list of web pages linking to a specific website (called *inlinks* in webometric terminology). Combined with the `host:` command (or the `site:` or `domain:` commands) it allowed users to request a list of pages in a specified website that linked to a second specified website. Using repeated queries of these types it was possible to compare the number of links pointing to each one of a set of websites, for example to find the most visible or highest impact site. AltaVista's Hit Count Estimates (HCEs) on the top of the first page of results gave an estimate of the number of linking pages so that the user did not need to count the number of results themselves. Using an analogy with bibliometric citations and the journal Impact Factor, counts of links to websites could be used to estimate their web impact and hence identify the most successful or visible sites, in some sense (Ingwersen, 1998). Furthermore, counts of links between each pair of websites within a pre-defined set could be used to create a web network of the extent of (hyperlink) connectivity between the sites (Larson, 1996).

Hyperlink searches became faster when Yahoo! introduced AltaVista's search operators and made them available for automated access by computer programs via the Yahoo! Search API. This made large scale webometric studies possible by constructing a list of websites of interest and then writing a program to submit the necessary searches to

Yahoo!. This automation also made it possible and quick to process the results, for example by counting the number of different websites matching a search rather than the number of URLs matching it. Bing later also introduced hyperlink searches and its own API. In contrast, Google's contribution to link search was (and still is) the link: command, which returns a list of pages linking to a given URL. It is not useful for webometric purposes, however, because its results are unreliable, are based upon a small sample of the linking pages known to Google (Fishkin, 2009), and are not supported by an API. For a short period both Bing and Yahoo! provided automated access to hyperlink queries, for example with the version 2.0 of the Bing API (Thelwall & Sud, 2012). Since then, however, both search engines have withdrawn their support for hyperlink searches and research has turned to alternative methods of obtaining similar data. Moreover, currently there is only one (partly) free API for a major search engine, that of Bing (the Bing Search API, replacing the version 2.0 API), and it does not return Hit Count Estimates in its API. Therefore, all methods using automatic searches must now be based upon the complete list of URLs matching a query (limited to 1000 URLs per query).

It is technically possible to bypass APIs and submit automated queries to search engines and then automatically scrape the pages for results, bypassing API rate limits and giving access to HCEs. This does not seem to be ethical for Bing, however, since it bans crawling search results in its robots.txt (see: www.robotstxt.org) file (http://www.bing.com/robots.txt, accessed August 8, 2013) and provides an API alternative. Similarly, Google bans automated queries in its terms of service (https://support.google.com/webmasters/answer/66357, accessed August 8, 2013).

2.1 Web crawling

An alternative way to access hyperlink counts for a collection of websites is to create a personal web crawler and use it to crawl the websites and then identify and count the hyperlinks between them (Cothey, 2004; Thelwall, 2001; Thelwall, 2008). This method is fully under the control of the researcher, unlike commercial search engines, and can give complete lists of relevant hyperlinks in the pages crawled, whereas commercial search engines hide some of their results for various reasons (Bar-Ilan & Peritz, 2009; Mettrop & Nieuwenhuysen, 2001; Rousseau, 1999). The complete list of hyperlinks also allows more complex link counting to be conducted, which can improve the quality of the results (Thelwall, 2002). For example, instead of counting the number of interlinking web pages, counting the number of interlinking web domains can remove anomalies caused by pairs of web domains that extensively interlink. Alternative counting options are built into the free web crawler SocSciBot (socscibot.wlv.ac.uk; Thelwall, 2009), for example. Web crawling has disadvantages of scope and a technical crawling limitation, however:

- It is impractical to crawl websites that are too large, such as those of major universities or large corporations. For example, SocSciBot has a limit of 1,000,000 pages per website. Hence it cannot be used to identify the hyperlinks between collections of websites if any large sites are included, unless they are only partially crawled.
- It is impractical to crawl the whole web and so personal web crawlers cannot calculate the number of hyperlinks pointing to a set of websites from the rest of the web, which is commonly needed for web impact studies.
- Personal web crawlers may not crawl some websites extensively because they have sections that are hidden but that commercial search engines may find due to links from other parts of the web (Thelwall, 2000).

2.2 Alexa.com

An important current free source of hyperlink data is the Alexa.com website. This reports lists of hyperlinks to websites and gives an estimate of the total number of hyperlinks as well, similar to the HCE of search engines. The data is not from a web crawl but is based upon evidence gathered from users that surf the web with the Alexa toolbar installed. This data source has the advantage that irrelevant links in obscure or spam parts of the web are likely to be ignored but has the disadvantage that there may be biases caused by the demographics of Alexa.com users, such as poor coverage of pages in languages not supported by the Alexa toolbar. Nevertheless, Alexa inlink data seems to be better than URL citations (Vaughan & Yang, 2012) although it has disadvantages in some contexts:

- Alexa's data is only available for websites hosted on their own domain name rather than on a subdomain or a directory within a larger domain. For example a search in Alexa.com for cybermetrics.wlv.ac.uk (a research group website) will yield results only for its host website wlv.ac.uk for the host university and a search for www.med.cam.ac.uk (Department of Medicine) only reveals results for the whole of the University of Cambridge. This excludes studies with academic websites smaller than entire universities.
- Alexa only lists the top 100 inlinking websites, whereas commercial search engines return up to 1000 matches for queries (e.g., for URL citation or title mention searches). This is not a problem for web impact studies because Alexa's HCE is for inlinking sites, which should give better results than a count of inlinking pages, but it does not allow effective random sampling of inlinking sites for content analysis purposes, as is a standard part of many link analyses (Thelwall, 2006).
- Alexa does not reveal the number of links between a pair of specified websites. For example it is impossible to ask it for a list of pages in www.wlv.ac.uk that link to pages in www.ox.ac.uk. As a result, it can only be used for web impact studies and not for web network studies unless the websites involved are so small that they all have less than 100 hyperlinks to them and these could be manually identified in the Alexa site inlink lists.

2.3 Analytics sites and commercial data sources

Some web analytics sites, such as Google Analytics, Bing Webmaster Tools and Blekko.com, report lists or counts of hyperlinks to a website that were clicked on to arrive at the site by its visitors. With the exception of Alexa.com and commercial data sources, however, these sites seem to only provide this data to site owners and hence it cannot be used for webometric studies of multiple sites without the permission of the site owners, which seems to be possible in limited cases (Eccles, Thelwall, & Meyer, 2012; Jonkers, De Moya Anegon, Aguillo, 2012).

Commercial hyperlink data may also be an appropriate alternative but has the disadvantages that it costs money to access, unless an agreement is reached, and the owning company may not be stable in the long term – although this was also a problem for Yahoo! and AltaVista. Commercial hyperlink data sources are thus impractical for small scale research projects and educational uses, which are an important part of webometrics.

At least three commercial link sources seem to have value for webometric research. Majestic SEO (www.majesticseo.com) is a commercial link database that is designed for search engine optimisation users and its huge database, having apparently crawled 600 billion URLs, makes it a valuable resource for link analysis. Ahrefs (ahrefs.com) and Open

Site Explorer (www.opensiteexplorer.org) are similar large commercial link database sites. The Ranking Web of World Universities, a webometric initiative, uses a combination of Ahrefs and Majestic SEO for its link data (webometrics.info/en/Methodology, as of 10 January 2014), for example.

2.4 URL citations

One response of webometricians to the removal of hyperlink searches from the major commercial search engines has been to devise alternative types of query to identify inter-document connections. The first such query was the URL citation (Kousha & Thelwall, 2006; Stuart & Thelwall, 2006). A URL citation is a mention of a URL in the text of a web page, with or without a hyperlink. URL citations can be identified through commercial search engine searches by simply entering the URL in quotes as a phrase search, normally without the initial http://. For example the following query in any major search engine will match URL citations to the BBC news website: `"news.bbc.co.uk" -site:bbc.co.uk` from pages outside of the bbc.co.uk website. All web pages matching this query will display the text *news.bbc.co.uk* somewhere but may not necessarily hyperlink to a page within the news.bbc.co.uk site.

Although URL citations are not hyperlinks, they are sometimes described with the more generic term *link* to reflect that they are also inter-document connections and can be used, in theory, for all webometric purposes for which hyperlinks are used. Nevertheless, URL citations are not always an adequate replacement for hyperlinks for two reasons:

- The main practical disadvantage of URL citations is that URL citations seem to be rarely used in some genres of website (Thelwall, 2011), causing data sparseness and reducing the power of any statistical tests.
- URLs can also be mentioned frequently in automatically generated pages, such as web server log reports and some types of spam pages (Thelwall, 2011).

2.5 Title mentions

Title mentions, sometimes also called web mentions, in the context of links refer to mentions in one web page of an organisation owning a different website. For example, a title mention of the BBC News website would be any occurrence of the phrase *BBC News* in a web page from a different site. Title mentions can be identified through queries in commercial search engines. For example, the Bing query `"BBC News" -site:bbc.co.uk` matches web pages outside the main BBC website that contain the phrase *BBC News*. Something similar to title mentions has been used for a long time: searches for the titles of academic articles as phrase searches (Vaughan & Shaw, 2003; Kretschmer, Aguillo, 2004).

Like URL citations, title mentions are not hyperlinks but are also a kind of inter-document connection and are hence links in the general sense of the term. Title mentions have an advantage over URL citations in that they are more natural and seem to be more frequent on the web (Thelwall & Sud, 2011; Thelwall, 2011). Nevertheless, both can be used for many types of network (Thelwall, Sud, & Wilkinson, 2012; Ortega, Orduña-Malea, & Aguillo, 2014) and there are some important disadvantages of title mentions:

- Title mentions can generate false matches if the phrases used do not uniquely identify the target organisation. In some cases it is impossible to construct a unique phrase for an organisation because two organisations have the same name or because one organisation has a longer version of the name of another (e.g., Cambridge University and

Cambridge University Press; Imperial College, Imperial College of Professional Studies, and Imperial College for Business).

- An organisation may be mentioned in many different ways and with different versions of its name and so many non-standard mentions may be missed by any specific query. To partially combat this, multiple versions of an organisation's name may be combined to capture common variations. This may particularly affect non-English nations that have English equivalent names that are commonly used (Ortega, Orduña-Malea, & Aguillo, 2014).
- The above problems may affect the websites in a collection differently and so comparing the results for a web impact study can be unfair.

3. Linked Title mentions

Introduced in this article, linked title mentions are title mentions with the extra stage that the pages matching the title mention search are subjected to a second check and rejected if they do not contain a hyperlink to the website of the organisation whose name they mention. This extra check ensures that the pages mention the correct organisation and means that all the matching pages contain a hyperlink to, in addition to a title mention of, the target organisation. It is possible to check for a link directly by downloading web pages matching the title mention search, then extracting and checking their hyperlinks. This gives an indirect way to identify appropriate hyperlinking pages from commercial search engines and ensures that there are no incorrect matches, hence improving on the accuracy (precision) that could be obtained using title mention searches alone. Linked title mentions may also improve on URL citations by being more natural and hence computer-generated pages should be less prevalent. There are some disadvantages, however:

- There will be fewer linked title mentions than title mentions due to the hyperlink filtering stage and this reduction in the amount of data reduces statistical power, making the results more susceptible to anomalies or random variations. This will be particularly the case if it is common to mention an organisation without hyperlinking to it.
- Linked title mentions may be much less common than URL citations in some contexts, again causing a loss of power due to data sparseness.

Finally, linked title mentions, like other methods based upon automatic search engine queries, can use two methods to get extended results lists: query splitting (Thelwall, 2008) and search market variation (Wilkinson & Thelwall, 2013).

4. Research Questions

The effectiveness of the linked title mention method is likely to vary by context. Intuitively, it seems likely to be most useful when title mentions return too many false matches and when URL citations do not return enough matches or return too many irrelevant matches (e.g., from server log file statistics or spam), and for the contexts discussed above when Alexa cannot be used. Since webometric studies have been applied for many different topics and types of site, a comprehensive evaluation would need to compare linked title mentions with the alternatives for a range of different data sets (Thelwall, 2011). It is no longer possible to conduct comprehensive evaluations without paying for access to search engine data, however, due to Bing's API restriction of 5000 queries per month; since each search may consume up to 20 queries to obtain up to 1000 matches in 20 pages of 50 each, the effective limit is 250 searches per month if most searches have a large number of results.

Because of the above issues, this study does not attempt a comprehensive evaluation but instead presents a proof of concept by evaluating linked title mentions in the context of the most commonly used type of data in webometrics: links between academic websites. The following research questions are therefore framed for evaluating whether linked title mentions *can* be a reasonable alternative to title mentions and URL citations rather than evaluating *the contexts in which* linked title mentions are likely to be a reasonable alternative.

1. Can linked title mentions provide an improved alternative to URL citations and title mentions for links to academic websites? Here, "improved" means giving higher correlations with established rankings.
2. Can linked title mentions provide an improved alternative to URL citations and title mentions for links between, or co-inlinks to, pairs of academic websites? Again, "improved" means giving higher correlations with established rankings.

In the above research questions there are many ways in which one link metric could be an improvement over another link metric. For example, it could give more links or a higher proportion of relevant links. Nevertheless, the most common way to assess the value of a link metric is to check its correlation with an established academic ranking scheme or measure of research productivity (Thelwall, 2006; Thelwall, & Sud, 2011). If a link metric is valid and if links relate to research to some extent then a positive correlation could be expected between link counts and any ranking or research productivity measure. Such correlations have been found many times before, confirming that links do associate with both research productivity and academic rankings, including the US & World News rankings, even though these are not primarily rankings of research.

The strength of the correlation with an existing research productivity metric or ranking is a reasonable way to compare two alternative link metrics because a poor quality metric with few links, many spam links or many irrelevant links would presumably be essentially random and hence have a correlation of close to zero with any metric or ranking scheme. Hence, the larger the correlation, the lower the proportion of irrelevant content a link metric is likely to reflect.

5. Data and Methods

To address the research questions, it would be sufficient to test any coherent set of academic websites if the results are positive (i.e., if the new link metrics gave higher correlations with established metrics than did the old link metrics). For improved credibility, this should be a previously-used set of academic websites. In this study, US Library and Information Science (LIS) school websites were chosen as the primary data set following their use in similar previous papers (Thelwall & Sud, 2011; Thelwall, Sud, & Wilkinson, 2012). Entire universities would be inappropriate because these would have too many URL citations and title mentions to be collected from the Bing API, the only source of automated queries from a major search engine. This is also an appropriate case study because not all LIS schools have their own domain name and so Alexa cannot give link data for this set. The ranking used for the correlation comparison was that of the U.S. News & World Report, which is an established ranking scheme based upon peer review (see below) and so is appropriate for this purpose.

A single test is insufficient to give convincing evidence of the value of a link metric because its utility is likely to vary by context. Hence, a secondary study was conducted of US

schools in an additional four subject areas (mathematics, economics, English, and psychology) using the same approach.

5.1 Constructing lists of names and URLs for LIS schools

There is a nationally recognised ranking of US LIS schools that can be used for comparison purposes, the U.S. News & World Report Library and Information Studies master's degree program rankings 2013[2]. This ranks LIS schools (and similar entities) on the basis of their master's degrees using a survey of program deans. This is not a perfect ranking for the purpose here because it is not based upon research and some schools received too few ratings to be ranked but it seems to be reasonable nevertheless, as argued in the Analysis subsection below. It has been previously used for the same purpose, with unrated schools ranked last (Thelwall & Sud, 2011), and unrated schools are again ranked last here.

An alternative way to obtain research-related rankings would be to identify all articles (e.g., in the Web of Science or Scopus) published by each school over a specific time period and then rank the schools based upon total publications or citations to their publications. This would have the advantage of being more directly research-related than the U.S. News & World Report rankings but the disadvantage that book-based research, which is important in some areas of library and information science, is not well recognised in citation databases (e.g., see White, Boell, Yu, et al., 2009). Moreover, links to a school may reflect its reputation overall, which may include the reputation of senior faculty whose main research was published many years ago and perhaps in a different school. Hence, on balance, citations do not seem to be clearly superior to the U.S. News & World Report rankings and so the latter was used as a more convenient (and more easily reproducible) source.

The official website(s) of each US LIS school were identified, starting from the list made for previous papers (Thelwall & Sud, 2011; Thelwall, Sud, & Wilkinson, 2012) and checking for changes with Google searches and links in the U.S. News & World Report LIS rankings pages. Current and former (when found) school URLs were included. Former URLs were included because old pages linking to a previous version or alternative version of a website still represent valid attempts to link to a school, even if broken. For example, two URLs were used for University of Maryland College of Information Studies, www.clis.umd.edu and ischool.umd.edu. URLs were truncated to the shortest possible form uniquely identifying the appropriate website to make the URL citation searches as powerful as possible. In most cases this was a domain name (e.g., www.slis.indiana.edu) but in some cases this included a path (e.g., www.ou.edu/cas/slis/).

Appropriate names for each US LIS school were identified, again starting from the list made for a previous paper (Thelwall & Sud, 2011) and checking for changes with Google searches and the name reported in the U.S. News & World Report LIS rankings pages. Queries typically took the form of a phrase search for a school in conjunction with a phrase search for its university (e.g., `"Texas Woman's University" "School of Library and Information Studies"`) rather than an individual single query containing both because testing suggested that the two were often mentioned together in different orders (e.g., *Texas Woman's University School of Library and Information Studies* or *School of Library and Information Studies, Texas Woman's University*) and sometimes

[2] http://grad-schools.usnews.rankingsandreviews.com/best-graduate-schools/top-library-information-science-programs/library-information-science-rankings

separately (e.g., *The School of Library and Information Studies at Texas Woman's University*). Care was taken to avoid false matches as far as possible (replicating the previous approach used for the same data (Thelwall & Sud, 2011)). Multiple queries were constructed when different ways of describing a school or its university were commonly found (e.g., `"Wayne State University" "School of Library and Information Science"`, and `"Wayne State University" "Library and Information Science Program"`). All the searches give incomplete sets of results, however, since schools are likely to be sometimes mentioned without the name of their university.

5.2 Site inlink queries

Webometric Analyst (Thelwall, 2009) was used to download and process the data. Webometric Analyst is free software that is primarily designed for automatically submitting sets of queries to search engines (currently only Bing, but previously also Google and Yahoo! when they had APIs) and processing their results to generate either networks of interlinking between sets of websites or summaries of the links to sets of websites. It also uses a range of other APIs as part of its suite of Webometric services. For automated Bing searches, users are required to enter their Microsoft Windows Azure key, entitling them to 5000 free queries per month (additional monthly queries can be purchased from Microsoft). Webometric Analyst requires a list of URLs and/or names of the websites to be analysed, as described below, and then automatically constructs and submits appropriate queries of the type requested by the user. Webometric Analyst takes care of paging through all the results for a query and combining the results in different ways (e.g., counting unique URLs or domains in query results), as well as drawing networks from the results or creating a set of web pages with tables and lists that summarise the link results.

The results from the above section were formatted for reading by Webometric Analyst[3]. This was achieved by listing all the title queries first, each separated by a pipe character |, followed by a tab and then a list of the URLs, each again separated by a pipe character, as in the example below with one query and two URLs.

- `"University of Washington" "Information School"<tab>ischool.uw.edu|ischool.washington.edu`

Webometric Analyst was then used to construct and submit searches to the Bing API for inlink counts and for links between the websites with each of the three methods tested (URL citations, title mentions, and linked title mentions) using the Wizard, and selecting the advanced wizard option to select alternative types of link search for each of the three types. If there were multiple URLs then these were excluded from the results of all site inlink searches. For example, in the case of the university of Washington, the following two URL citation searches were submitted.

- `"ischool.uw.edu" -site:uw.edu -site:washington.edu`
- `"ischool.washington.edu" -site:uw.edu -site:washington.edu`

In cases where there were multiple queries for the same university Webometric Analyst automatically combined the results of the different queries and eliminated duplicate URLs.

5.3 Direct link and co-inlink network queries

To create a site interlinking network, direct link searches were used for each pair of schools in the data set (i.e., 51x50=2550, excluding self-links). If either the source university or the

[3] http://lexiurl.wlv.ac.uk/images/USLISDepts2013.txt

target university or both had multiple titles or URLs then multiple searches were needed and these are also constructed and submitted automatically by Webometric Analyst and the results combined. This gave an additional 1,472 queries for the URL citation searches and 1,665 extra queries for the title mention searches. For example the following four URL citation searches were combined to give the total direct URL citations from Illinois to the University of North Texas because both schools had two URLs (one old URL in each case).

- `"www.unt.edu/slis" site:lis.uiuc.edu`
- `"www.lis.unt.edu" site:lis.uiuc.edu`
- `"www.unt.edu/slis" site:lis.illinois.edu`
- `"www.lis.unt.edu" site:lis.illinois.edu`

A second type of network was also created for each type of search, a co-inlinking network, because co-inlinks are commonly used in webometrics in addition to direct links (e.g., Holmberg, 2010; Thelwall, Sud, & Wilkinson, 2012; Vaughan, & You, 2006). A co-inlink search for a pair of schools A and B counts the number of pages that link to both A and B, excluding all pages in the websites of A and B (all versions). Again, multiple searches were needed in cases where there were multiple URLs or titles for one or both of the schools in question. For example, the following four searches were combined for the URL citation co-inlink count for the Illinois and SUNY Albany websites. Note that the initial www. at the start of most URLs was retained to avoid giving too big an advantage to organisations with their own websites and which may have multiple domains. Such multiple domains would not be captured by queries for URL that included any path information in the department URL, as in the first example below. As an example of the "unfair" advantage that deleting initial www. sections could give some schools, the Graduate School of Library and Information Science, University of Illinois has the website www.lis.illinois.edu and `"www.lis.illinois.edu"` captures URL citations of it but the query `"lis.illinois.edu"` also gets URL citations to the derivative site cirss.lis.illinois.edu for the *Center for Informatics Research in Science and Scholarship* within the school.

- `"www.lis.uiuc.edu" "www.albany.edu/cci/informationstudies" -site:uiuc.edu -site:illinois.edu -site:albany.edu`
- `"www.lis.uiuc.edu" "www.albany.edu/informationstudies" -site:uiuc.edu -site:illinois.edu -site:albany.edu`
- `"www.lis.illinois.edu" "www.albany.edu/cci/informationstudies" -site:uiuc.edu -site:illinois.edu -site:albany.edu`
- `"www.lis.illinois.edu" "www.albany.edu/informationstudies" -site:uiuc.edu -site:illinois.edu -site:albany.edu`

For all the searches, level 1 query splitting (Thelwall, 2008) was used. This extracts additional results beyond the first 1000 returned by the search engine and was necessary because the site inlinking searches sometimes returned just over 1,000 URLs.

Webometric Analyst was not used to gather title mentions directly because it gathers title mention results as the first stage of linked title mention searches and saves a copy of the title mention result in addition to the (filtered) linked title mention results. This unfiltered version was used for the title mention results, both to minimise the total number of Bing API searches used and to ensure the maximum degree of compatibility between the title mention and linked title mention results.

To check whether a title mention was also a linked title mention, Webometric Analyst downloaded the page containing the title mention, extracted all of its hyperlinks and checked each hyperlink URL to see whether it contained (as a substring, using a case-

intensive check) any version of the URL of the department associated with the mention. A title mention page was rejected as a linked title mention if no appropriate hyperlink was found in the source page.

5.4 Analysis

The site inlink counts are reported in terms of URLs and domains, as calculated by Webometric Analyst. For site inlinks and co-inlinks, the *domain inlink count* was obtained by truncating each source URL to its domain (e.g., www.albany.edu/cci/informationstudies to www.albany.edu) and then eliminating duplicate domains. Counting by domain seems to be superior to counting URLs because of the potential for URLs to be automatically replicated throughout a site. Nevertheless, URL counting is still commonly used in Webometrics and so both approaches are reported here. Also, URL counting is the only method normally used for networks because domain counting usually gives binary data, and can give wrong results if some sites in the set examined have multiple domains, so domain counting was not applied to the data set relating to links between academic websites.

The site inlink searches were compared against the U.S. News & World Report rankings using Spearman correlations (due to skewed data) with the assumption that accurate inlink counts would have a high correlation with this ranking, based upon previous research showing that inlink counts tend to correlate with research productivity (Thelwall, 2002; Thelwall, & Harries, 2004). The assumption here is that schools with high research productivity are likely to be more highly ranked in the U.S. News & World Report site. Although there is no direct evidence for this claim, it seems that in academic contexts widely varying ranking schemes tend to produce reasonably similar outcomes (Aguillo, Bar-Ilan, Levene, & Ortega, 2010; Thelwall & Kousha, in press). For the network searches, there is no way to assess the entire network for accuracy. Instead, we calculated inlink counts for each node in the network from the network data (by loading the network into the Webometric Analyst Network drawing component and requesting centrality statistics from the Stats menu) and correlated these with the U.S. News & World Report rankings. This is a similar test to that for the site inlink searches except that the links included in the network results are all from the other LIS schools and the figure is made for each university by combining the results of many (at least 50) different direct link searches. For reporting simplicity, the rankings list was inverted to convert all correlations with it from negative to positive.

5.5 Secondary tests

To check whether linked title mentions might work better for other types of academic web sites, the direct link and co-inlink searches were repeated (in March-October 2013) for the 50 highest ranked mathematics, economics, English, and psychology schools in the US & World News site. These were chosen to represent major subject areas that tended to be sited within single schools rather than split into multiple schools or combined with other subjects into single schools. The above procedures were repeated for these new datasets.

6. Results

The results are reported separately for the site inlinks and for the networks, with the primary data set of US LIS schools discussed first and in the most detail.

6.1 Site inlinks for US LIS schools

From Table 1, it is clear that linked title mentions are about an order of magnitude less frequent than title mentions and about a quarter as frequent as URL citations for site inlink searches on this data set. Nevertheless, Table 2 shows that linked title mentions have the highest correlation with U.S. News & World Report ranks, despite the smaller amount of data used for them, suggesting that the filtering is effective at removing irrelevant or low quality matches.

Table 1: Descriptive statistics for site inlink searches using six different link counting methods (N=51).

	URL citations (URLs)	URL citations (domains)	Title mentions (URLs)	Title mentions (domains)	Linked title mentions (URLs)	Linked title mentions (domains)
Mean	164.0	130.6	297.4	242.3	34.2	31.7
Median	117	90	298	242	30	27

Table 2: Spearman correlations between U.S. News & World Report ranks and site inlink counts for US LIS schools using six different link counting methods (N=51, sig p<0.001 in all cases).

	Rank	URL citations (URLs)	URL citations (domains)	Title mentions (URLs)	Title mentions (domains)	Linked title mentions (URLs)	Linked title mentions (domains)
Rank	1.000	0.631	0.640	0.507	0.528	0.667	0.668
URL Citation (URLs)		1.000	0.997	0.613	0.636	0.878	0.880
URL Citation (domains)			1.000	0.626	0.649	0.895	0.897
Title mentions (URLs)				1.000	0.996	0.747	0.745
Title mentions (Domains)					1.000	0.764	0.762
Linked title mentions (URLs)						1.000	0.998
Linked title mentions (Domains)							1.000

6.2 Web networks for US LIS schools

From the medians in Table 3, linked title mentions are 20 times less frequent than title mentions for direct links and about 46 times less frequent than title mentions for co-inlinks. Linked title mentions are more frequent than URL citations for direct links and about 2.5

times less frequent for co-inlinks. Table 4 shows that linked title mentions have the lowest correlation with U.S. News & World Report ranks for co-inlinks, although the difference with URL citations is not large, and have the middle correlation with ranks for direct links. The zero median for direct link URL citations (Table 3) and the lowest correlation in Table 4 being between Rank and direct link URL citations both suggest that URL citations are not useful for identifying links between US LIS schools. The median in particular suggests that US LIS schools rarely refer to each other with a URL in the text of their web pages, perhaps because URLs in pages do not fit well with the official web presences that presumably form a significant part of these websites.

Table 3: Descriptive statistics for indegree centrality for web networks of US LIS schools using six different link counting methods (N=51).

	URL citations (direct links)	Title mentions (direct links)	Linked title mentions (direct links)	URL citations (co-inlinks)	Title mentions (co-inlinks)	Linked title mentions (co-inlinks)
Mean	1.9	149.2	6.6	42.0	401.9	14.0
Median	0	100	5	18	320	7

Table 4: Spearman correlations between U.S. News & World Report ranks and inlink counts for web networks of US LIS schools using six different link counting methods (N=51, sig p<0.001 in all cases).

	Rank	URL citations (direct links)	Title mentions (direct links)	Linked title mentions (direct links)		URL citations (co-inlinks)	Title mentions (co-inlinks)	Linked title mentions (co-inlinks)
Rank	1.000	0.570	0.747	0.647		0.681	0.767	0.661
Direct URL citation		1.000	0.528	0.609		0.702	0.625	0.652
Direct title mention			1.000	0.769		0.754	0.951	0.717
Direct linked title mention				1.000		0.864	0.789	0.755
URL citation co-inlinks						1.000	0.784	0.807
Title mention co-inlinks							1.000	0.779
Linked title mention co-inlinks								1.000

6.3 Secondary tests: English, mathematics, economics, and psychology

To check whether linked title mentions might work better for other types of academic web sites, the direct link and co-inlink searches were repeated (in March-October 2013) for the 50 highest ranked mathematics, economics, English, and psychology schools in the US & World News site but in these cases the results were more mixed (tables 5 and 6). For convenience, the US LIS school results above are repeated in the new tables.

Table 5: (Site inlinks) Spearman correlations between U.S. News & World Report ranks and site inlink counts for US schools using six different link counting methods (N=50 or 51, sig p<0.001 in all cases). The highest correlation for each subject is in bold.

Subject	URL citations (URLs)	URL citations (domains)	Title mentions (URLs)	Title mentions (domains)	Linked title mentions (URLs)	Linked title mentions (domains)
Economics	0.453	**0.454**	0.327	0.204	0.411	0.430
English	0.250	0.245	**0.398**	0.321	0.112	0.120
LIS	0.631	0.640	0.507	0.528	0.667	**0.668**
Maths	**0.353**	0.337	0.308	0.240	0.277	0.273
Psychology	0.043	0.054	0.267	0.256	0.343	**0.360**

Table 6: (Web networks) Spearman correlations between U.S. News & World Report ranks and inlink counts for web networks of US schools using six different link counting methods (N=50 or 51, sig p<0.001 in all cases). The highest correlation for each subject and each type of link is in bold.

Subject	URL citations (direct links)	Title mentions (direct links)	Linked title mentions (direct links)		URL citations (co-inlinks)	Title mentions (co-inlinks)	Linked title mentions (co-inlinks)
Economics	0.439	**0.733**	0.516		**0.398**	0.306	0.375
English	0.226	**0.547**	0.370		0.321	**0.458**	0.372
LIS	0.570	**0.747**	0.647		0.681	**0.767**	0.661
Maths	0.602	**0.742**	0.540		0.448	**0.614**	0.559
Psychology	0.175	**0.630**	0.304		0.106	**0.218**	0.202

7. Discussion

The primary results suggest that for site inlinks to US LIS schools, linked title mention searches are at least as good as the available alternatives and may be better because of their higher correlation with the master's program rankings. To assess the extent to which the matches removed by the filtering stage of the linked title mention process were false, two random URLs per school from the site inlink results were examined for title domains that were rejected by the title mention test: i.e., they matched a search for a LIS school but did not contain a hyperlink to the matching school. The random selection was made using a random number generator. The results were as follows for the 102 URLs.

- 12 URLs were technically incorrect. They did not match the original title mention search or the page had disappeared. Some of these were dynamic pages that may have previously contained the title mention text.
- 14 URLs were technically correct but conceptually incorrect. They matched the original title mention search but did not mention the LIS school. For example, one of the URLs returned for the query `"University of Illinois" "Graduate School of Library and Information Science" -site:uiuc.edu -site:illinois.edu` was http://www.lakeforest.edu/academics/catalog/directory/, which mentioned the Graduate School of Library and Information Science of a different university and the University of Illinois in a different context. This page contained a list of affiliations.
- 76 URLs were technically and conceptually correct. These matched the original title mention search and mentioned the appropriate LIS school.

This suggests that about three quarters of the matches rejected by the title mention hyperlink check were valid in the sense of mentioning the correct LIS school. About half of the remainder appeared to be incorrect matches to the original searches (i.e., technically incorrect matches) and about half were technically correct but conceptually incorrect because they did not mention the correct school. The improved correlation for the linked title mention searches might therefore be due, at least in part, to the removal of some of the incorrect matches, leaving a higher proportion of correct matches in the remainder.

Taking into account the four secondary cases, the linked title mention correlations are highest in two out of five cases in Table 5 (LIS and psychology). This suggests that the

linked title mention strategy can be useful, although not universally so. The differences in correlations between disciplines, including differences in which type of data has the highest correlation with the U.S. News & World Report ranks are probably due to disciplinary differences in the quantity and types of links that are created, although small differences could also be due to statistical variations in the data.

In contrast to the case for site inlinks, from the perspective of data from links between websites in networks linked title mentions do not seem to be an improvement on the alternatives for links between specific pairs of websites or for co-inlinks to specific pairs of websites, for US LIS schools and for the four secondary subject areas (Table 6). A possible contributing reason is that false matches might not be as common as for site inlink searches because all of the searches for pairs of websites contain some information about both websites and this may help to give extra implicit context to the search that may reduce false matches. Another possible reason is that the large percentage reduction in the numbers of matches from title mention searches to linked title mention searches reduces the total number of matches too far so that the statistical power is reduced and random variations dominate the results.

More generally, it seems likely that linked title mentions will work well for some types of non-academic website, although it is not straightforward to evaluate them because of the lack of a natural ranking to follow. As an example, a web network study of UK newspapers would need to contend with the simple and common names of some of them (e.g., The Guardian, The Independent, The Star, The Sun, The Times), probably leading to many false matches. An extreme example for title co-inlink searches is `"the star" "the sun" -site:thesun.co.uk -site:thestar.co.uk`, which gives mainly astronomy-related results.

Finally, whilst the substantial positive correlations between the U.S. News & World Report ranks and inlink counts indicate that the different types of link counts are able to a large extent to correctly rank LIS schools and the top 50 maths, economics, English and psychology schools, it seems likely that the accuracy will be greater near the top of the ranking than nearer the bottom. This is because the differences in link counts are greater near the top of the ranking than near the bottom. The U.S. News & World Report ranks are probably also more reliable near the top than near the bottom, however, making it difficult to be sure about this. Overall, then, link counts are likely to be most reliable for ranking more important websites than for ranking less important websites.

8. Conclusions

Linked title mentions, which are title mentions that also have a relevant hyperlink, have been assessed for a set of academic websites as a proof of concept. The results of the two positive case studies (US LIS and psychology schools) suggest that linked title mentions can be *valuable for academic site inlink searches* but the results of all five case studies suggest that linked title mentions may *not be useful or necessary for direct links between or co-links to pairs of academic websites*. Linked title mentions may not be useful for direct links and co-inlinks because, within the narrow context of a set of sites with a specific academic focus, title mentions were already very accurate.

Site inlinks (e.g., for web impact studies) Since Alexa is already a good source of inlink data for websites, the main recommendation is that linked title mentions or URL citations should be used for site inlink searches for academic websites when at least one of the websites does not have its own domain name. The positive results are based on linked

title mentions giving the best results in two out of the five case studies, with URL citations giving the best results in the other two. The choice of these two methods depends on the extent to which the organisations involved have unambiguous and easily searchable names and so this finding should be treated with caution in future studies. Nevertheless, if title mention searches are used in future then it is a simple step to also use linked title mention searches (e.g., a single option button click in Webometric Analyst) and so it makes sense to always conduct both and use a strategy to decide which is best. For example if there are too few linked title mentions then it can't be used but if there are too many false matches with title mention searches then linked title mentions may be preferable. The same strategy could also be used for sets of non-academic websites.

Direct links and co-inlinks (e.g., for network diagrams or network analyses) Linked title mentions are not recommended for these types of links, at least in the context of subject-based academic websites since all five cases gave negative results. If there is reason to believe that many matches will be false for a collection of a different type of web site, however, then linked title mentions might still prove useful. In such a case, the same strategy could be adopted for direct links and co-inlinks as above: calculate linked title mentions simultaneously with title mentions and decide which to use based upon the number of linked title mentions and the amount of false links in the title mentions.

Intuitively, it seems likely that the linked title mention search will work in some contexts and not others. It would be useful to test it on a wide range of different types of website in order to determine general rules about when it is successful. Although this is currently difficult due to the Bing API restrictions, if many future studies use linked title mentions then a general pattern may emerge about the situations in which it is useful, extending the current results for sets of academic schools.

Finally, it would be useful to test linked title mention searches against Alexa site inlinks for a set of websites with their own domain names. It seems likely that Alexa would be the better source for this purpose since it would presumably return much higher inlink counts because the links do not have to associate with organisation names, in contrast to linked title searches.

9. Acknowledgements

This paper is supported by ACUMEN (Academic Careers Understood through Measurement and Norms) project, grant agreement number 266632, under the Seventh Framework Program of the European Union.

10. References

Aguillo, I. F., Bar-Ilan, J., Levene, M., & Ortega, J. L. (2010). Comparing university rankings. *Scientometrics*, 85(1), 243-256.

Aguillo, I. F., Granadino, B., Ortega, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetrics indicators. *Journal of the American Society for Information Science and Technology, 57*(10), 1296-1302.

Bar-Ilan, J., & Peritz, B. C. (2009). A method for measuring the evolution of a topic on the web: The case of "Informetrics". *Journal of the American Society for Information Science and Technology, 60*(9), 1730-1740.

Chu, H., He, S., & Thelwall, M. (2002). Library and information science schools in Canada and USA: A webometric perspective. *Journal of Education for Library and Information Science, 43*(2), 110-125.

Cothey, V. (2004). Web-crawling reliability. *Journal of the American Society for Information Science and Technology, 55*(14), 1228-1238.

Eccles, K. E., Thelwall, M., & Meyer, E. T. (2012). Measuring the web impact of digitised scholarly resources. *Journal of Documentation, 68*(4), 512-526.
Fishkin, R. (2009). Google link: command - Busting the myths, *The Moz Blog*. Retrieved August 8, 2013 from: http://moz.com/blog/google-link-command-busting-the-myths
Heimeriks, G., Hoerlesberger, M., & Van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. *Scientometrics*, 58(2), 391-413.
Holmberg, K. (2010). Co-inlinking to a municipal Web space: a webometric and content analysis. *Scientometrics*, *83*(3), 851-862.
Ingwersen, P. (1998). The calculation of web impact factors. *Journal of Documentation, 54*(2), 236-243.
Jonkers, K., De Moya Anegon, F., Aguillo, I.F. (2012). Measuring the usage of e-research infrastructure as an indicator of research activity. *Journal of the American Society for Information Science and Technology*, 63(7), 1374-1382.
Kousha, K., & Thelwall, M. (2006). Motivations for URL citations to open access library and information science articles. *Scientometrics, 68*(3), 501-517.
Kretschmer, H., Aguillo, I.F. (2004). Visibility of collaboration on the Web. Scientometrics, 61(3),405-426.
Larson, R. R. (1996). *Bibliometrics of the world wide web: An exploratory analysis of the intellectual structure of cyberspace*. *Proceedings of the ASIST Annual Meeting, 33*, 71-78.
Lepori B., Barberio V., Seeber M., Aguillo I. (2013). Core-periphery structures in national higher education systems. A cross-country analysis using interlinking data. *Journal of Informetrics*, 7(3), 622–634.
Li, X., Thelwall, M., Musgrove, P. B., & Wilkinson, D. (2003). The relationship between the WIFs or inlinks of computer science departments in UK and their RAE ratings or research productivities in 2001. *Scientometrics, 57*(2), 239-255.
Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines - fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623-651.
Ortega, J. L., & Aguillo, I. F. (2009). Mapping world-class universities on the web. *Information Processing & Management, 45*(2), 272-279.
Ortega, J. L., Orduña-Malea, E., & Aguillo, I. F. (2014). Are web mentions accurate substitutes for inlinks for Spanish universities? *Online Information Review*, 38(1), 59-77.
Park, H. W. (2010). Mapping the e-science landscape in South Korea using the webometrics method. *Journal of Computer-Mediated Communication, 15*(2), 211-229.
Rodríguez i Gairín, J. M. (1997). Valorando el impacto de la información en internet: AltaVista, el "citation index" de la red (evaluating the impact of internet information: AltaVista, the "citation index" of the web). *Revista Española De Documentación Científica, 20*(2), 175-181.
Romero-Frias, E., & Vaughan, L. (2010). European political trends viewed through patterns of web linking. *Journal of the American Society for Information Science and Technology, 61*(10), 2109-2121.
Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*, Retrieved January 21, 2014 from: http://cybermetrics.cindoc.csic.es/articles/v2i1p2.pdf.
Seeber, M., Lepori, B., Lomi, A., Aguillo, I., & Barberio, V. (2012). Factors affecting web links between European higher education institutions. *Journal of Informetrics, 6*(3), 435-447.
Smith, A. G. (1999). A tale of two web spaces; comparing sites using web impact factors. *Journal of Documentation, 55*(5), 577-592.
Stuart, D., & Thelwall, M. (2006). Investigating triple helix relationships using URL citations: A case study of the UK west midlands automobile industry. *Research Evaluation, 15*(2), 97-106.
Thelwall, M. (2000). Web impact factors and search engine coverage. *Journal of Documentation, 56*(2), 185-189.
Thelwall, M. (2001). Results from a web impact factor crawler. *Journal of Documentation, 57*(2), 177-191.
Thelwall, M. (2002). Conceptualizing documentation on the web: An evaluation of different heuristic-based models for counting links between university web sites. *Journal of American Society for Information Science and Technology, 53*(12), 995-1005.
Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*(1), 60-68.
Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology, 59*(1), 38-50.
Thelwall, M. (2009). Introduction to webometrics: Quantitative web research for the social sciences. San Rafael, CA: Morgan & Claypool.
Thelwall, M. (2011). A comparison of link and URL citation counting. *ASLIB Proceedings, 63*(4), 419-425.

Thelwall, M., & Harries, G. (2004). Do the web sites of higher rated scholars have significantly more online impact? *Journal of the American Society for Information Science and Technology, 55*(2), 149-159.
Thelwall, M., Klitkou, A., Verbeek, A., Stuart, D., & Vincent, C. (2010). Policy-relevant webometrics for individual scientific fields. *Journal of the American Society for Information Science and Technology, 61*(7), 1464-1475.
Thelwall, M. & Kousha, K. (in press). ResearchGate: Disseminating, communicating and measuring scholarship? Journal of the Association for Information Science and Technology.
Thelwall, M., & Sud, P. (2011). A comparison of methods for collecting web citation data for academic organisations. *Journal of the American Society for Information Science and Technology, 62*(8), 1488-1497.
Thelwall, M. & Sud, P. (2012). Webometric research with the Bing Search API 2.0. *Journal of Informetrics*, 6(1), 44-52.
Thelwall, M., Sud, P., & Wilkinson, D. (2012). Link and co-inlink network diagrams with URL citations or title mentions. *Journal of the American Society for Information Science and Technology*, 63(4),805-816.
Vaughan, L. & Hysen, K. (2002). Relationship between links to journal Web sites and impact factors. *ASLIB Proceedings*, 54(6), 356-361.
Vaughan, L. & Shaw, D. (2003). Bibliographic and Web citations: What is the difference? Journal of the American Society for Information Science and Technology, 54(14),1313-1322.
Vaughan, L., & Yang, R. (2012). Web data as academic and business quality estimates: A comparison of three data sources. *Journal of the American Society for Information Science and Technology, 63*(10), 1960-1972. doi:10.1002/asi.22659
Vaughan, L., & You, J. (2006). Comparing business competition positions based on Web co-link data: The global market vs. the Chinese market. *Scientometrics*, 68(3), 611-628.
White, H.D., Boell, S.K., Yu, H., Davis, M., Wilson, C.S., & Cole, F.T. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60(6), 1083-1096.
Wilkinson, D., & Thelwall, M. (2013). Search markets and search results: The case of Bing. *Library and Information Science Research*, 35(4), 318-325.