Information Research, Vol. 8 No. 3, April 2003

What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation

Mike Thelwall
School of Computing and Information Technology
University of Wolverhampton
Wolverhampton WV1 1EQ, UK

Abstract

Analogies between Web links and citations have been used in information retrieval to improve search engine query matching and in information science to develop link metrics for academic and other Web spaces. The purpose of this paper is to begin a fine-grained process of differentiating between creation motivations for links in academic Web sites and citations in journals on the basis that they are very different phenomena. A sample of 100 random inter-site links to UK university home pages was used as a starting point for a qualitative exploration and four new types of motivation are postulated. The term'ownership' is coined for links acknowledging authorship or co-authorship of a resource,'social' for links with a primarily social reinforcement role, 'general navigational' for those with a generalinformation navigation function and 'gratuitous' for those that serve no communication function at all. It is argued that all of these form a role unique to the Web, albeit in varying degrees. Compared to citer motivations they are relatively trivial and instead of being primarily socio-cognitive, none are cognitive and the gratuitous are not even social.

Introduction

The promise of Web links as a new information source in the area of information science has been cogently expressed by many authors (Ingwersen, 1998; Davenport & Cronin, 2000; Cronin, 2001; Borgman & Furner, 2002) yet data validity is a major issue when conducting any kind of simple link counting (Bar-Ilan, 2001a; Björneborn & Ingwersen, 2001; Thelwall & Harries, 2003). If hyperlinks are to fulfil their potential, then considerable groundwork must be invested in deepening understanding of the variety of motivations for their creation. This is a necessary precursor to the overarching theory that would be needed to provide some validity to link analysis studies. The logical starting point for a theory of scholarly hyperlinking is indeed the more mature field of citation analysis.

Research in this area has included many studies of author motivations for citation creation. The principal aim of this paper is to sharply differentiate academic Web linking from citation practice by identifying common types of link motivations that are clearly new and unique to the Web. This has already been achieved by Kim (2000) for e-journal articles but these form only a tiny percentage of inter-university links.

Web links represent both anarchy and order. The unofficial leader of Web standards, the World Wide Web Consortium, imposes no rules on hyperlink creation. Its role is merely 'to lead the World Wide Web to its full potential by developing common protocols that promote its evolution and ensure its interoperability' (W3C, 2003). Whilst in practice there are some limitations on Web page authoring, varying from strictly enforced organisational policies for standardised link structures to legal requirements concerning trademark infringement (Oppenheim, 2001) few would argue with the contention that Web linking is essentially an unregulated phenomenon. Yet there is order in the chaos: search engines such as Google and AltaVista successfully use the link structure of the Web to optimise search results (Brin & Page, 1998; AltaVista, 2002); counts of links between university Web sites in several countries correlate significantly with research ratings (Thelwall, 2001a; Smith & Thelwall, 2002; Thelwall & Tang, 2003); the topology of the Web obeys striking mathematical laws (Broder *et al*, 2000; Thelwall &

Wilkinson, 2003). The challenge for researchers in many fields is now to harness whatever order there is so as to be able to extract meaning from the chaos. In information retrieval this appears to have been achieved by the commercial search engines (which do not allow academic validation of their methods) but not yet replicated in academic studies (Hawking *et al.*, 2000; Gao *et al.*, 2001

From one perspective, Information Retrieval (IR) researchers have a less problematic task than those from two other related fields: Webometrics and communication networks (Garrido & Halavais, 2003; Park *et al.*, 2002). In IR, algorithms are required that are 'only' statistically more effective at retrieving useful information whereas the other two must also be concerned with data validity, which is a key issue that a growing body of research is throwing into relief as both complex and necessary. To illustrate this point, it is known that counts of links to UK university Web sites correlate strongly with their research productivity. This has been ratified from different perspectives (Thelwall, 2002b) and with increasingly complex metrics (Thelwall, 2001a; Thelwall, 2002a). Moreover, there is a stronger correlation with research productivity from links more closely related to research (Thelwall, 2001a; Thelwall & Harries, 2003). Despite this, no causal connection between research and link creation is claimed. One study, illuminating along the way the difficulty in classifying link motivations, has suggested that over 90% of inter-university links are related in some way to informal scholarly communication, but less than 1% are equivalent to scholarly citations (Wilkinson *et al.*, 2003).

Background and literature review

Citation analysis

Citation analysis is a logical starting point for an investigation of hyperlink motivation issues because of the similarities between the two as inter-document connections. Borgman and Furner (2002) see both as specific instances of general linking phenomena. Citer motivations have been extensively studied, driven by the need to explore the validity of using citations in various bibliometric measurements. The traditional approach is to investigate the connected pair of documents to find out what they have in common or that which makes one worthy of being cited by the other. In an ideal model, a citation might represent a finding in the earlier paper that was subsequently used or built upon in the later one. Investigations have revealed other trends at work, however, for example with some types of articles being more likely to be cited than others, such as review articles. Such findings undermine somewhat the ideal model.

Other research has taken a different approach by looking at the relationship between the authors themselves, discovering factors unrelated to content such as tendencies to cite compatriots (Herman, 1991) or colleagues (see Cronin & Shaw, 2002). Many authors have attempted to develop schemes for classifying the context or motivation for creation of a citation but the difficulty of this task is highlighted by the wide variety of approaches reported in Liu's (1993) review, and the fact that motivations and connections between documents are typically multiple and overlapping (Leydesdorff, 1998). Cronin (1984, chapter 5) illustrates how far back this recognition goes and different classification approaches to dealing with it. He has also argued that citations must be considered in relation to four interested groups (Quality Controllers, Educators, Consumers, Producers) to fully understand their use, adding a new dimension of complexity (Cronin, 1984, chapter 7).

Two dimensions of citation practice that are not related to the need to persuade the referees and readers that an article is of high quality are convenience and self-publicity. A citation may partly reflect the author's ability to access the document (Lawrence, 2001) and to read the language that it is written in (Yitzhaki, 1998). These are both convenience issues but neither would be *primary* motivations except in peculiar circumstances. Hyland's (2003) recent in-depth study of self-citation shows that this practice can be part of an individual's wider strategy to promote themselves. This suggests an exclusively social motivation, although purely promotional citations seem unlikely to pass peer review. Nevertheless Hyland finds self-citation to be sanctioned to differing extents by discipline, which hints at a degree of acceptance in some fields of the promotional motivation.

The human-centred approach, then, has identified trends independent of the documents themselves, but this is not the same as demonstrating that cited document content is irrelevant to citer motivation. The assumption must be that the cited work must be useful in some way to be cited at all, but perhaps if there is a choice about which document to cite or whether to cite one at all then other factors can come into play. Citing, therefore, is a socio-cognitive act, with the interpretation of a cognitive connection being wide enough to include knowledge not directly built upon in the argument such as methods and background information

The Web as a global hypertext

Although it is true to say that the Web is at root just a global hypertext system, from a social perspective it is radically different from classical hypertext systems in its functions and use. Hypertext systems are 'an approach to information management in which data is stored in a network of nodes connected by links' (Smith & Weiss, 1988). From this early explanation it can be seen that a classical hypertext is typically a self-contained set of information that uses links for internal navigation. Berners-Lee's (1993) conceptualisation of the Web was a system that allowed external links to resources elsewhere on the Internet with contents out of the control of the link author. This could be used for instant citation retrieval when both documents were online, for example

There are many differences between standard and Web hypertext use in practice. Firstly, links *between* Web sites are of a fundamentally different character to those inside. Whilst the latter can be solely for internal navigation, reasons for the former are more difficult to classify. In fact, all forms of Web communication behaviour are recognised as being intrinsically difficult to study (Riva, 2001). The anchor text of a hyperlink may be as uninformative as 'click here', requiring the user to interpret its context to divine its intention. Moreover, the author may not even have created the link; it could be part of a standard navigation bar that was required to be embedded in each page. Google claims a great deal of success with using the text immediately adjacent to a link in order to estimate the semantic content of the target page (Brin & Page, 1998), which does give some hope that link contexts will offer some help in motivation identification.

Web links as a new social phenomenon

Compared to citations, links between Web sites are radically different. Whereas a journal article may be rejected for publication due to failure to cite a relevant article, Web link creation is a much less formal affair. It is difficult to imagine a situation where a Web page is removed from an academic server for its failure to link to a page on an external Web site. The strongest incentive I can conceive of as being common would be fear of social sanctions if a link was missed from a list of similar types. For example, if an academic owned a page linking to the home pages of other researchers in their field, then they might be careful to avoid offending someone by leaving them out. An exception would be funded portal sites, where an agency has paid for the creation of link sites and could reasonably expect to have an appropriate number of high quality sites linked to. But these are relatively rare in academic Webs, almost by definition.

This leads to a second question: why would a]scholar *need* to create hyperlinks? Scholars need to create citations because they (typically) need to have references in their articles to justify their scholarly status (Hyland, 2000). They (typically) need to create journal articles to justify their existence. Yet many scholars create no Web pages at all and even entire disciplines seem to create virtually no links (Tang & Thelwall, 2003). It seems to me highly likely that, other than links in e-journal articles and online copies of preprints, very few hyperlinks between academic sites are created as a result of a necessity on a par with that for citations.

One answer is that most scholars are also educators. In many disciplines, education includes pointing students to a range of information sources for assimilation or evaluation. Given the importance of the Web as an information source, there *is* a necessity in general terms to identify relevant online information and point students towards it, whether this is achieved by URLs in printed handouts or links online in course Web pages.

Scholars are not simply publishing and teaching machines, however, they are human too and also engage in both purely social actions and a range of informal scholarly activities designed to publicise their research (Hyland, 2003) and promote their field. Web links can clearly help these activities by pointing others (friends, colleagues) to useful resources. Alternatively, link creation may be partly the result of a perhaps subconscious association with citations: if the use of citations helps to demonstrate a researcher's credibility, then perhaps the use of appropriate links may perform the same function.

Creating a well-used site is perhaps a special case of motivation type. This may not bring formal recognition, but may accrue symbolic capital (Cronin, 2002), including simple name recognition. This may have a symbiotic relationship with pure altruism - a site created for personal pleasure may be maintained and expanded because the owner gets positive feedback from users. Of course, all well-used sites start off without any visitors and so there is no reason why unused sites cannot have been created with the expectation of attracting many visitors.

The research question

The objective of this study is to present an argument for the existence of one or more common link motivations that are unique to the Web. The investigation is based around a study of a random collection of links from UK university pages to the home page of a different UK university. Links to institutional home pages were selected under the hypothesis that targets so general in content were likely to give rise to novel motivations. This choice is legitimised by the fact that university home pages are popular link targets, accounting for 45 of the top 100 linked pages in a recent study (Thelwall, 2002c) and so this is a numerically significant type of link. Although this is a qualitative study there is a quantitative element, which is to use a link typology to justify the discussion being of 'common' link types. It would be easy but pointless to find obscure links and discuss their creation motivations and so part of this study is a justification that the types discussed are not infrequent. In terms of the Tashakkori & Teddlie (1998) taxonomy this is a mixed model study dominated by the (qualitative) constructivist paradigm. The bias is justified by the issue addressed forming an early stage in the much larger project of link motivation typology generation, which will be as a whole a more balanced exercise in a pragmatic paradigm, with the eventual introduction of more positivist elements such as rigorous content analysis (Krippendorff, 1980).

Methods

A publicly accessible database of the UK link structure of 111 UK university Web sites was used for the basic data set (http://cybermetrics.wlv.ac.uk/database/). This was created by a particularly accurate Web crawler (Thelwall, 2001b, 2001c) but it only covers the portion of Web sites reachable by following links from the home page, the 'publicly indexable set' (Lawrence & Giles, 1999), an unavoidable type of problem (Thelwall, 2002d). A program was written to extract all links where the target was the home page of a different UK university from the source, giving 19,438 in all. Note that these are individual links rather than link pages and so a page with 110 links to other university home pages would occur 110 times. The links were then placed in a random order and the first 100 selected for investigation. Each source page was loaded into a Web browser and the context of the identified link investigated.

The investigation methodology was an inductive content analysis, based upon Krippendorff (1980) but carried out by the author and without cross-checking by additional classifiers. This makes the results subjective to the author and not the hard empirical evidence as a full content analysis approach would give. The reason for this was that as an exploratory research I wanted make the heart of the paper the qualitative analysis of the results and, therefore, the aim of the classification was merely to identify types of apparent linking motivation that were sufficiently numerous to justify a serious discussion. A major problem for a full-scale content analysis exercise is that Web pages are known to not conform to existing genres particularly well (Crowston & Williams, 2000) and even relatively identifiable new genres such as academic home pages can have the confusing factor of being spread across multiple pages (Rehm, 2002). As a result, a highly prescriptive and non-intuitive classification scheme would have had to be drawn up in order to get the necessary degree of inter-classifier consistency (see, for example, Weare & Lin, 2000). This would have defeated the purpose of the paper.

For the classification, an initial scheme of categories was drawn up based upon observations from previous Web page analysis experiments that used a similar approach (Bar-Ilan, 2001b; Thelwall, 2001a; Thelwall & Harries, 2003). The pages containing the links were then visited in order and the link motivation classified according to the scheme. When a motivation was found that did not fit the scheme, a new description was added. The list was also revised during the classification process when a category became large and it was apparent that the links could actually be reclassified into two separate sections. When this happened, the links were revisited to decide into which of the new categories they fell. Additionally, some categories had their descriptions slightly changed to accommodate new members. This is clearly a highly subjective process, essentially a clustering approach, but fit for the purpose of grouping together similar link types in a way that would facilitate a discussion of creation motivations.

Results and discussion

The results of the investigation are summarised in Table 1. One of the strangest links lists was of the Web sites of organizations that were reachable by a number 73 London bus, part of an experimental research project INCITE - Incubator for Critical Inquiry into Technology and Ethnography (Smith, 2001). The pages and links types are hopefully self-explanatory. Collaborative student support is perhaps the least clear category. This included all multi-institution initiatives to provide resources of any kind for students and included, for example, a regional careers initiative and a library-sharing scheme.

Type of page/type of link	Count
General list of links to all university home pages	16
Regional university home page link list	2
Personal bookmarks	2
Subject-based link list	5
Other link lists	6
Personal home page of lecturer	
/ link to degree awarding institution	8
/ link to previous employer	6
/ link to collaborator's institution	3
/ other	3
Collaborative research project page/ link to partner site	17
Other research page	
/ link to collaborator's institution	3
/ link to institution of conference speaker	2
/ link to institution hosting conference	2
/ other	3
Link to home institution of document author e.g. in mirror site	7
Collaborative student support	
/ link to partner institution	6
/ link to institution for access to information	4
Other type of page	5

Table 1. A categories for source pages for 100 random links to external UK university home pages

The text around each link was also investigated to see whether an explicit description of the content of the target site was given. This occurred in only four cases, for example one where the target site was said to contain an online prospectus. This is far from evidence for general disinterestedness about the target site contents, however. In the context of the large number of university link lists, it is reasonable to suppose that page authors would be able to assume that their visitors would know the kind of content to expect from any UK university Web site.

Four motivation categorisations will be defined and analysed as a result of the analysis of the contexts of the links. These will not cover all of the links in the table, only some groups that appear to have motivations that are different from those normally associated with citations.

General navigational links

The purpose for links in university home page lists appears to be as a starting point for browsing to find a range of information. Their utility is derived from the range of information that can be accessed by starting from them. Essentially, the information given by such pages is the domain names of UK universities. Links will be described as *general navigational* if their primary creation motivation is to allow the visitor to start with the link and then to browse to find a wide variety of non-subject specific information. The emphasis here is on the generality of the link target, so that a link to the home page of a department or research group would not count as general navigational even though some navigation would probably be needed in most cases to get to content.

Are general navigational links unique to the Web? Perhaps the most closely related identified common citer motivation is 'Setting the background to the present study' (Peritz, 1983) or variations of this such as 'Part of relevant literature, serves no explicit role in the analysis' (Cole, 1975). A reference to a literature review might be a common item to fit in this category. This is perhaps part citation - the reader may be expected to read the cited review - and part navigational - the reader may be expected to use the review as a starting point to retrieve more specific articles on the topic of her choice. An unalloyed navigational citation would be one where the target contained little or no information other than pointers to other documents, with these documents not having content cognitively related to the original document. Examples would include contents pages or indexes of books or journals. These are clearly far from being mainstream citation targets. Occasionally there are also navigation-based 'articles', such as Schubert's (2001) bibliography of Scientometrics. It seems clear, however, that navigational citations will be general in only exceptional cases: their target is expected to have a typically subject-based focus, or perhaps an interdisciplinary topic based theme. In summary the difference between general navigational links and navigational citations is that the latter have a cognitive connection to the source document whereas the former do not.

Ownership links

Seventeen partner institution links were in the pages of collaborative research projects, with more being on collaborative student support pages. These were often in the form of a row of university crests placed as a navigation bar either at the top or bottom of every page of the site created by the collaborative project or consortium initiative. An example of this was a page in a project site, part of a collection all containing a link to four university's home pages at the bottom right hand corner (home page at: http://www.ucl.ac.uk/epd/herdu/vdml/index.htm). The purpose of these links appears to be an implicit acknowledgement of project co-ownership or site content co-authorship. This is particularly important in the context of a consortium project where the site is hosted by the server of one of the partners. Having a clickable link to the home page of all partners on all site pages conveys the clear message of acknowledging the importance of all and the reassurance of not attempting to claim undue credit for the work. Such links, along with other types, have previously been termed 'credit links' (Thelwall, 2002c). The closest analogy for these links in bibliometrics is with paper co-authorship. They are acknowledging co-authorship of the project contents or co-ownership of the project. The hyperlink is not a necessary component of such acknowledgement, however, especially because in the cases here it is targeted at the university home page rather than those of individual researchers or even a collaborating department. Project co-membership information can also be placed in the acknowledgement section of a paper, but I can find no evidence of any similar phenomenon in the citation literature. A relatively modern possible near match is the citing in an article of the Web site of a joint project. Even this is of a different type, however, since the project site presumably contains at least some information related to the citing publication.

Ownership links will be defined to be those that acknowledge authorship, co-authorship, ownership or co-ownership of the host Web page(s) or associated project. This definition encompasses all the links to the home institution of a document author in the case of mirror sites and remotely hosted talks and papers. In all cases found in the sample set the link was not essential to the attribution of ownership as this could be inferred from the text or university crest image, it served only to emphasise the attribution.

Social links

One link type identified in Table 1 may have social origins: that of links in general research or personal home pages to the institution of a collaborator or collaborating group. These appear to have no particular function, but may be performing a social reinforcement role. For example, they can be seen to be conferring the implicit compliment: 'We have recognised your site and think it important enough to link to'. Those created with the apparent primary purpose of reinforcing social ties will be termed *social links*. Definitively identifying social links is difficult from any page analysis since the question addressed is whether the primary intention is a social one, and social interactions are not really the purpose of the Web. This category of link is in many ways the most interesting but also the one for which the attribution of motivation seems most tenuous. Given the low numbers involved a larger scale study with author interviews is really needed to verify that this motivation really exists in non-trivial numbers.

It is recognised that social factors can play an important role in citer motivation for journals (Case & Higgins, 2000). In fact a major function of academic writing is to conform to disciplinary discourses (Hyland, 2000) and so the overall direction of scholarly writing and citing is fundamentally a social one. A socially motivated reference must still have at least an ostensible relationship to the content of the cited article in order to satisfy the quality control of the referees (and subsequent readers) however, whereas social links do not imply any such target content connection. There is a related concept concerning communication without content from linguistics: phatic communion. It was popularised by Jakobson (1960) but originates from the anthropologist Malinowski (1923) who used the term to describe, 'a type of speech in which the ties of union are created by a mere exchange of words'.

Typically no real information is exchanged but social relationships are reinforced. Alternative descriptions of phatic communion are 'small talk', 'just chatting', and 'passing the time of day' (Stubbs, 1983: 101), but these do not explicitly convey the social reinforcement aspect. A revealing perspective is to see it as 'a concern with the act of communication itself' (Eagleton, 1996). Studies of aspects of the phatic have found that in some contexts it is highly formulaic, e.g., in conversation openings, but that social norms in this are culture-specific (Jaworski, 1990). This mirrors both social and ownership links, which are often in standard links bars, emphasised with a clickable logo or in a sentence with a series of links to institutions or parts of one institution. Phatic communion is normally different to hyperlink creation, however, since it is interactive. Link creation would be a similarly interactive phenomenon in contexts where it is expected to be reciprocated, perhaps in academics' personal home pages, but the logical targets of such exchanged links would be the hosting pages themselves rather than university home pages.

The term phatic has also been used in a related but different context that can also transfer to Web contexts where reciprocal links are not expected. Individual comments can be used to bestow implicit compliments (Boyle, 2000) when something is said which cannot be interpreted as a compliment from its linguistic structure but only from an analysis of the social context (e.g., an ethnomethodological perspective). Boyle explicitly applies a wider interpretation of the phatic for this than in its original meaning, but as alluded to above, some links can be conceived as conferring implicit compliments. Note that Manovich (1996) has previously imported the term phatic to the Internet from the Web for the status and screen construction messages from the browser during the time-delayed downloading of Web pages, but this is clearly a different context.

Gratuitous links

Fourteen of the pages were personal home pages of academics that contained clickable links to the home page of an institution where they had either previously worked or obtained an academic qualification. Here is an example with the clickable link underlined.

> I was an undergraduate at the University of East Anglia (1991-4) graduating with a …

Although it is possible to imagine a context where almost any link would be found useful, these links seem to be there primarily because the page author knows that the organisation has a Web site and thinks it appropriate to include a link to it, rather than because it would be of value to the viewer of the page. These links typically occur in short paragraphs where every online organisation mentioned has a link. Here is a longer example.

> I was a Teaching Assistant in the Department of Computer Science and Technology and a Research Assistant in the State Key Laboratory of Intelligent Technologies and Systems, Tsinghua University, Beijing, China (1997 ~ 1998).

These links are probably also not phatic in the wider sense of just carrying an implicit message in the link that is additional to that contained in the text. The term *gratuitous link* will be used for those without any discernable communication motivation behind their creation. In other words, the link is not expected to be used, nor does it play any other identifiable communication role.

Issues in link motivation attribution

There are other potential explanations for link creation motivations than those modelled above, some of types that would be difficult to identify even through direct methods such as author interviews.

- An expression of technical competence:'We know where your site is and know how to create a link to it.'
- A formulaic/genre following activity:'Web sites of this kind always link to partner institutions' Web site home pages.'
- Part of a learning exercise: the creation of personal pages is a common first exercise in HTML authoring courses, and in this context it is natural to use HTML features such as linking, merely to learn how to use them. It is worth mentioning that one of the links in the sample set was actually in an online HTML authoring exercise with the instruction to create a link to the home page given, another gratuitous link but probably not of a common type.

As with citer motivations, the probable overlap in causes is an additional complication. The first two are also available in a combination: a demonstration of competence in knowledge of, and conforming to, genre. Extensive publishing of practice pages could also produce a recognisable genre in this relatively new publishing arena, an overlap between the last two categories.

Conclusions

The methodology used here is based upon relatively thin evidence – one person's interpretation of the pages themselves – and so is far from being rigorous. In particular, other (citer) motivation studies have also interviewed authors (Kim, 2000; Hyland, 2003) to get a more complete picture. I shied away from sending unsolicited emails to the page authors to ask them a bizarre question about a link that they had created possibly years ago for a reason that I suspected was relatively trivial. I would suspect that it would be difficult to get a high enough response rate and perhaps also that authors would be reluctant to admit that there was no strong reason for creating the link. As a result the analysis was based on just the pages themselves, and this is an important limitation. The same criticism has been made of early citation typology studies (Borgman & Furner, 2002). Nevertheless, arguments have been presented for the existence of genuinely new types of linking motivation. General navigational links are probably the most clear-cut case, but ownership links map more comfortably to authorship attributions than citations. Other links appear to perform a primarily social reinforcement function, whereas some perform none at all. Other possible explanations include the expression of technical competence, genre following or learning Web authoring. Motivations probably overlap in the same way as for citations, and perhaps in many cases there will not be one clear primary motivation. The situation is further confounded by the informality of the medium: genre following is an example of a motivation classification where the author might not even have consciously thought about why the link should have been created. The clearest case for novelty is essentially this: citations are *primarily* a socio-cognitive act, whereas all of the link types found do not have a cognitive dimension and one of them (the gratuitous) does not seem to have a primarily social motivation. In addition, navigation links appear unique in their very generality and ownership links have taken the role normally played by co-authorships and acknowledgements in academic articles. Perhaps the social links are the most difficult to justify as novel. The question here is one of degree: I would argue that these are created primarily for social reasons, and without any cognitive aspect, whereas citations must have at least an ostensibly cognitive role.

The relatively trivial motivations deduced for the links investigated cause problems for attempts to apply bibliometric techniques to the Web or to use links or link counts to infer relationships between individuals or organisations.

From the perspective of university Web link metrics, an important question is how to interpret counts of links. In Cronin's (2001) semiotic terminology, the question is of the significance to the signified (the link target) of the sign (the link), and how this can be interpreted when aggregated. The exact meaning of a hyperlink, even in the restricted case considered here, varies with context. In the case of general or regional university links lists, no direct relationship between source and target is implied, but for social links an association of some kind is precisely what is present in the sign.

Traditional bibliometric relationships were also found to some degree: that of the link target containing information referenced by the source, (e.g. navigational links) or acknowledging co-authorship. As a result, this tends to confirm that sums of links will give a quantity that does not directly measure any one entity, despite their association with research productivity (Thelwall, 2002a, 2002b). It may still be the case, however, that general navigational and gratuitous links form the background noise from which large-scale aggregation can still identify trends from social and other links. This would give a combination of direct evidence of the production of useful general information and evidence of social ties associated with research and other collaborative initiatives.

As a final point, investigating motivations for link creation for coherent sets of Web documents would form a useful investigative project for students in bibliometrics courses, but on ethical grounds I would not advocate allowing students to email page authors from other institutions to question their motivations. A list of all 19,438 links to the home pages of other UK universities from the same data set has been placed online at http://cybermetrics.wlv.ac.uk/database/stats/data_only/uk_2002_homePageLinks.txt in a randomised order, and the list is also available on the journal site.

Acknowledgements

The referees are warmly thanked for their helpful comments and encouragement.

This work was supported in part by a grant from the Common Basis for Science, Technology and Innovation Indicators part of the Improving Human Research Potential specific programme of the Fifth Framework for Research and Technological Development of the European Commission. It is part of the WISER project (Web indicators for scientific, technological and innovation research) (Contract HPV2-CT-2002-00015).

References

- AltaVista (2002). "AltaVista advanced search tutorial - link popularity." http://help.altavista.com/adv_search/ast_haw_popularity (19 Jul. 2002).
- Bar-Ilan, J. (2001a). "Data collection methods on the Web for informetric purposes - a review and analysis." *Scientometrics*, **50**(1), 7-32.
- Bar-Ilan, J. (2001b). "How much information the search engines disclose on the links to a Web page? A case study of the "Cybermetrics" home page." In: *Proceedings of the 8th International Conference on Scientometrics & Informetrics*, vol 1. pp. 63-73, Sydney: Bibliometric & Informetric Research Group.
- Björneborn, L. & Ingwersen, P. (2001). "Perspectives of Webometrics." *Scientometrics*, **50**(1), 65-82.
- Berners-Lee, T. (1993). "World Wide Web seminar." http://www.w3.org/Talks/General/Concepts.html (19 Jul. 2002).
- Borgman, C & Furner, J. (2002). "Scholarly communication and bibliometrics." In: Cronin, B. (ed.), *Annual Review of Information Science and Technology* Vol. 36, pp. 3-72. Medford, NJ: Information Today Inc.
- Boyle, R. (2000). "'You've worked with Elizabeth Taylor!': phatic functions and implicit compliments." *Applied Linguistics*, **21**(1), 26-46.
- Brin, S. & Page, L. (1998). "The anatomy of a large scale hypertextual Web search engine." *Computer Networks and ISDN Systems*, **30**(1-7), 107-117.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). "Graph structure in the Web." *Journal of Computer Networks*, **33**(1-6), 309-320.
- Case, D. O. & Higgins, G. M. (2000). "How can we investigate citation behavior?: a study of reasons for citing literature in communication. *Journal of the American Society for Information Science*, **51**, 635-645.
- Cole, S. (1975). "The growth of scientific knowledge: theories of deviance as a case study." In: Coser, L. A. (ed.), *The idea of social structure: papers in honor of Robert K. Merton*. New York, NY: Harcourt, Brace, Jovanovich, pp. 175-220.
- Cronin, B. (1984). *The citation process*. London: Taylor Graham.
- Cronin, B. (2001). "Semiotics and evaluative bibliometrics." *Journal of Documentation* **56**(4), 440-453.
- Cronin, B. & Shaw, D. (2002). "Identity-creators and image-makers: using citation analysis and thick descriptions to put authors in their place," *Scientometrics* **54**(1), 31-49.
- Cronin, B. & Shaw, D. (2003). "Banking (on) different forms of symbolic capital.*" Journal of the American Society for Information Science*, **53**(4), 1267-1270.
- Crowston, K. & Williams, M. (2000). "Reproduced and emergent genres of communication in the world wide Web." *Information Society*, **16**(3), 201-15.
- Davenport, E. & Cronin, B. (2000). "The citation network as a prototype for representing trust in virtual environments." In: Cronin, B. & Atkins, H. B. (eds.). *The Web of knowledge: a festschrift in honor of Eugene Garfield*, pp. 517-534. Metford, NJ: Information Today Inc. (ASIS Monograph Series)
- Eagleton, T. (1996). *Literary theory: an introduction* (2nd Ed). Oxford: Blackwell.
- Gao, J., Walker, S., Robertson, S., Cao, G., He, H., Zhang, M. & Nie, J-Y (2001). "TREC-10 Web Track Experiments at MSRA". In: *NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)*. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec10/papers/msra.trec10.pdf. (27 February 2003)
- Garrido, M. & Halavais, A. (2003, forthcoming). "Mapping networks of support for the Zapatista Movement: applying social network analysis to study contemporary social movements." *In*: M. McCaughey & M. Ayers (eds). *Cyberactivism: online activism in theory and practice.*. London: Routledge.
- Hawking, D., Bailey, P. and Craswell, N. (2000). "ACSys TREC-8 experiments." In: *NIST special publication 500-246: the eighth Text REtrieval Conference (TREC 8)*. Gaithersburg, MD: National Institute of Standards and Technology. http://trec.nist.gov/pubs/trec8/papers/acsys.pdf (27 February 2003)
- Herman, I. L. (1991). "Receptivity to foreign literature: a comparison of UK and US citation behavior in librarianship and information science." *Library & Information Science Research*, **13**(2), 37-47.
- Hyland, K. (2000). *Disciplinary discourses: social interactions in academic writing*, Harlow: Longman.
- Hyland, K. (2003). "Self-citation and self-reference: credibility and promotion in academic publication". *Journal of the American Society for Information Science*, **54**(3), 251-259.

- Ingwersen, P. (1998). "The calculation of Web impact factors." *Journal of Documentation*, **54**(2), 236-243.
- Jakobson, (1960). "Linguistics and poetics." In Sebok, T. (ed.) *Style in language*. pp. 350-377. Cambridge, MA: MIT Press
- Jaworski, A. (1990). "The acquisition and perception of formulaic language and foreign language teaching." *Multilingua* **9**(4), 397-411.
- Kim, H. J. (2000). "Motivations for hyperlinking in scholarly electronic articles: a qualitative study." *Journal of the American Society for Information Science*, **51**(10), 887-899.
- Krippendorff, K. (1980). *Content analysis: an introduction to its methodology*. Beverly Hills, CA: Sage.
- Lawrence, S. & Giles, C. L. (1999). "Accessibility of information on the Web". *Nature*, **400**, 107-109.
- Lawrence, S. L. (2001). "Online or invisible?", *Nature*, **411**(6837) 521.
- Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.
- Leydesdorff, L. (1998). "Theories of citation?" *Scientometrics*, **43**(1), 5-25.
- Liu, M. (1993). "The complexities of citation practice: a review of citation studies." *Journal of Documentation*, **49**(4), 370-408.
- Malinowski, B. (1923). "The problem of meaning in primitive languages", In: C.K. Ogden & I.A. Richards (eds.), *The meaning of meaning*, pp. 296-346. London: Routledge & Kegan Paul.
- Manovich L. (1996). "Global algorithm 1.3: the aesthetics of virtual worlds: report from Los Angeles." *CTheory.net* http://www.ctheory.net/text_file.asp?pick=34 (10 July 2002)
- Oppenheim, C. (2001). "LISLEX: legal issues of concern to the library and information science sector." *Journal of Information Science*, **27**(4), 277-286.
- Park, H. W., Barnett, G.A. & Nam, I. (2002). "Hyperlink-affiliation network structure of top Web sites: examining affiliates with hyperlink in Korea." *Journal of the American Society for Information Science*, **53**(7), 592-601.
- Peritz, B. C. (1983). "A classification of citation roles for the social sciences and related fields." *Scientometrics*, **5**(5), 303-312.
- Rehm, G. (2002). "Towards automatic Web genre identification." In: *35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4. January 07 - 10, 2002, Big Island, Hawaii*. New York, NY: Institute of Electrical and Electronics Engineers, Inc. [Also available at the author's site: www.uni-giessen.de/~g91063/pdf/HICSS35-rehm.pdf (27th February 2003)]
- Riva G. (2001). "The mind over the Web: the quest for the definition of a method for internet research." *CyberPsychology & Behavior*, **4**(1), 7-16.
- Schubert, A. (2001). "Scientometrics: a citation based bibliography 1997-2000." *Scientometrics*, **50**(1), 99-198.
- Smith, A. & Thelwall, M. (2002). "Web impact factors for Australalasian universities" *Scientometrics*, **54**(3), 363-380.
- Smith, J. & Weiss, S. (1988), "Hypertext." *Communications of the ACM*, **31**(7), 816-819.
- Smith, S. (2001). *Introduction*. http://www.soc.surrey.ac.uk/incite/mapping/, (17 July 2002).
- Stubbs, M. (1983). *Discourse analysis: the sociolinguistic analysis of natural language*. Oxford: Basil Blackwell.
- Tang, R. & Thelwall, M. (2003, forthcoming). Disciplinary differences in US academic departmental Web site interlinking. *Library and Information Science Research*.
- Tashakkori, A. & Teddlie, C. (1998). *Mixed methodology*. London: Sage.
- Thelwall, M. (2001a). "Extracting macroscopic information from Web links." *Journal of the American Society for Information Science and Technology*. **52**(13), 1157-1168.
- Thelwall, M. (2001b). "A Web crawler design for data mining." *Journal of Information Science*, **27**(5), 319-325.
- Thelwall, M. (2001c). "A publicly accessible database of UK university Website links and a discussion of the need for human intervention in Web crawling." Wolverhampton: University of Wolverhampton, School of Computing and Information Technology. http://www.scit.wlv.ac.uk/~cm1993/papers/a_publicly_accessible_database.pdf (27 February 2003)
- Thelwall, M. (2002a). "Conceptualizing documentation on the Web: an evaluation of different heuristic-based models for counting links between university Web sites." *Journal of the American Society for Information Science and Technology*, **53**(12), 995-1005.
- Thelwall, M. (2002b). "A comparison of sources of Links for academic Web impact factor calculations." *Journal of Documentation*, **58**, 60-72.
- Thelwall, M. (2002c). "The top 100 linked pages on UK university Web sites: high inlink counts are not usually directly associated with quality scholarly content." *Journal of Information Science*, **28**(6), 485-493.

- Thelwall, M. (2002d). "Methodologies for crawler based Web surveys." *Internet Research: Electronic Networking and Applications*, **12**(2), 124-138.
- Thelwall, M. & Harries, G. (2003, forthcoming). "The connection between the research of a university and counts of links to its Web pages: an investigation based upon a classification of the relationships of pages to the research of the host university." *Journal of the American Society for Information Science and Technology*, **54**(4).
- Thelwall, M. & Tang, R. (2003, forthcoming)." Disciplinary and linguistic considerations for academic Web linking: An exploratory hyperlink mediated study with Mainland China and Taiwan." *Scientometrics*
- Thelwall, M. & Wilkinson, D. (2003, forthcoming). "Graph structure in some national academic Webs: power laws with anomalies." *Journal of the American Society for Information Science and Technology*.
- W3C (2003). "About the World Wide Web Consortium (W3C)". Cambridge, MA: World Wide Web Consortium. http://www.w3.org/Consortium/ (27 February 2003).
- Weare, C., & Lin, W. Y. (2000). "Content analysis of the World Wide Web - opportunities and challenges". *Social Science Computer Review*, **18**(3), 272-292.
- Wilkinson, D., Harries, G., Thelwall, M. & Price, E. (2003). "Causes of academic Web site interlinking: Evidence for the Web as a novel source of information on informal scholarly communication." *Journal of Information Science*, **29**(1), 59-66.
- Yitzhaki, M. (1998). "The language preference in sociology: measurements of 'language self-citation,' 'relative own language preference indicator, ' and 'mutual use of languages'". *Scientometrics*, **41**, 243-254.

Find other papers on this subject.

hyperlinks "academic Web sites"

Scholar Search

hyperlinks "academic Web sites"

Google Search

How to cite this paper:

Thelwall, Mike (2003) "What is this link doing here? Beginning a fine-grained process of identifying reasons for academic hyperlink creation."*Information Research*, **8**(3), paper no. 151 [Available at: http://informationr.net/ir/8-3/paper151.html]

Check for citations, using Google Scholar