Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Information Processing and Management 41 (2005) 347–359

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

A modeling approach to uncover hyperlink patterns: the case of Canadian universities

Liwen Vaughan [a,*], Mike Thelwall [b,1]

[a] *Faculty of Information and Media Studies, University of Western Ontario, London, Ont., Canada N6A 5B7*

[b] *School of Computing and Information Technology, University of Wolverhampton, 35/49 Lichfield Street, Wolverhampton WV1 1EQ, UK*

Received 18 July 2003; accepted 13 October 2003
Available online 27 November 2003

Abstract

Hyperlink patterns between Canadian university Web sites were analyzed by a mathematical modeling approach. A multiple regression model was developed which shows that faculty quality and the language of the university are important predictors for links to a university Web site. Higher faculty quality means more links. French universities received lower numbers of links to their Web sites than comparable English universities. Analysis of interlinking between pairs of universities also showed that English universities are advantaged. Universities are more likely to link to each other when the geographical distance between them is less than 3000 km, possibly reflecting the east vs. west divide that exists in Canadian society.

Keywords: Hyperlink patterns; University Web sites; Mathematical modeling

1. Introduction

University Web sites in many nations are large multifaceted communication devices, and are increasingly central for a wide variety of purposes from attracting new students to providing online library catalogues. In terms of research, university Web sites can announce the existence and boast the achievements of individuals, research groups, institutes and departments. They can also disseminate findings either through hosting online articles or by publishing summaries, data sets or tools. The pages themselves can be created centrally, by administrators or Webmasters, or

[*] Corresponding author. Tel.: +1-519-661-2111x88499; fax: +1-519-661-3506.
E-mail addresses: lvaughan@uwo.ca (L. Vaughan), m.thelwall@wlv.ac.uk (M. Thelwall).
[1] Tel.: +44-1902-321470; fax: +44-1902-321478.

0306-4573/$ - see front matter © 2003 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2003.10.001

locally by individuals for themselves or their research team or projects. Potential benefits of an effective Web presence include greater research impact, attracting Ph.D. applicants, media interest and commercial contacts. In this context it is logical to investigate measures of the effectiveness of Web sites, both to study the communication activity that they represent and to build useful evaluation metrics.

The impact of a Web site can be investigated through analyzing the links that target it (inlinks). Although it is a relatively crude device, counts of links to academic Web sites do seem to be informative and appropriate as indicators of impact (Ingwersen, 1998; Thelwall, 2001b; Thelwall, 2002a). Even outside the academic Web space, inlinks have been found to be a measure of quality of the organization hosting the Web site. For example, Vaughan and Thelwall (2003) found links to a journal Web site to be an indicator of the quality of the journal while Vaughan and Wu (2003) found inlinks to a commercial Web site to correlate significantly with various business performance measures of the company. However, all these studies examined the nature of inlinks by correlating them with a particular measure of the quality or impact of the hosting organization. The study reported in this paper moves beyond simple correlation analysis and uses a mathematical modeling approach that allows the analysis of multiple explanatory (independent) variables simultaneously. Specifically, various measurements of the quality of a university, including student quality and faculty quality, will be used as the independent variables of the mathematical model to explain or predict inlink counts, the dependent variable. This allows us to compare faculty quality with student quality in order to find out which is the best predictor of inlinks. The modeling approach also allows us to examine another important variable for a bilingual country like Canada, the language (English vs. French) of the university, while holding other variables constant. An advantage of mathematical modeling is that the interactions among variables are taken into consideration when analyzing their contributions to the dependent variable. This helps to avoid the inclusion of spurious variables and thereby provides a better understanding of the nature of the Web hyperlinks.

2. Literature review

2.1. Background to Webometric link analysis

Many studies of hyperlinks in both information science and computer science have drawn their inspiration from citation analysis, which is part of the field of bibliometrics (Borgman & Furner, 2002). In information science, this resulted in attempts to directly transfer techniques from citation analysis to the Web (Almind & Ingwersen, 1997; Larson, 1996; Rousseau, 1997). In an influential article, Ingwersen (1998) proposed a series of measurements, Web Impact Factors, based upon using counts of links to a site, or other area of the Web, to estimate its online impact. Subsequent researchers attempted to validate these measurements, principally by correlating them with offline measures of known value, a line of research that eventually gave positive results (Chu, He, & Thelwall, 2002; Smith, 1999; Thelwall, 2001b; Thomas & Willett, 2000; Vaughan & Hysen, 2002).

In contrast to computer science applications of link analysis, which have tended to focus on Web information retrieval applications (Arasu, Cho, Garcia-Molina, Paepcke, & Raghavan, 2001; Chakrabarti, Joshi, Punera, & Pennock, 2002; Henzinger, 2001), Webometric approaches have been more theoretical and exploratory in nature. They have included relational investigations (Polanco, Boudourides, Besagni, & Roche, 2001; Thelwall, Tang, & Price, 2003; Thelwall, 2001c, 2002b) and the assessment and development of the metrics themselves (Ingwersen, 1998; Smith, 1999; Thelwall, 2001d; Thelwall, 2002a). In contrast to citation analysis, however, their use is not yet strongly recommended for evaluative purposes (Thelwall, 2002c) because of both technical issues and problems in the interpretation of results based on an underdeveloped theoretical framework.

2.2. University Web site interlinking

A number of studies have analyzed the interlinking between the universities within a single country, and within and between two or three countries. The countries/regions covered include Australia (Smith & Thelwall, 2002; Smith, 1999), Mainland China (Thelwall & Tang, 2003), New Zealand (Smith & Thelwall, 2002), Spain (Thelwall & Aguillo, in press), Taiwan (Thelwall & Tang, 2003), and the UK (Thelwall, 2001a).

A number of patterns have emerged from these studies. First, counts of links to a university from its peers correlate significantly with measures of university research productivity (Smith & Thelwall, 2002; Thelwall, 2001a; Thelwall & Tang, 2003): universities that are more productive in research attract more inlinks. This does not imply that research causes link creation. In fact studies of reasons for link creation would refute such a claim (Thelwall, 2001b; Thelwall & Harries, 2003; Wilkinson, Harries, Thelwall, & Price, 2003). Links between universities are created for a wide variety of reasons and although about 90% seem to be related to scholarly activities, less than 1% directly target online formal academic publications (Wilkinson et al., 2003). The reason for the correlation found appears to be that more productive researchers simply publish more on the Web (Thelwall & Harries, in press).

A mathematical modeling approach has been used in one paper to explore different potential combinations of source and target university research productivity in order to find the best predictor of link counts between them (Thelwall, 2002e). The conclusions pointed to the product of the source and target university research productivity as being a much better predictor of the number of links between them than many other models, including the research productivity of either university individually or added together.

It is known that geographic factors can influence link creation within a single country; closer universities tend to interlink more (Thelwall, 2002d). This is a less strong trend than that for research productivity, however. It has only been found for the UK and a smaller scale study of the US found no evidence of a geographic trend (Tang & Thelwall, in press).

2.3. Linguistic factors on the Web

English is an international language, in some senses a world language that many non-English speakers use to communicate in a wide variety of settings (Pennycook, 1994). From the perspective of formal scholarly communication, and journal articles in particular, it should be noted that English is often regarded as the carrier for the highest quality publications, and capable of giving the most impact to a researcher's work (Moed, 2002; van Leeuwen, Moed, Tijssen, Visser, & van

Raan, 2000). In a scholarly context, therefore, it is natural to ask whether this would transfer to the Web. No studies have analyzed linguistic influences on linking between the universities of a single country, but one has looked at languages of link pages in European universities (Thelwall et al., 2003). It was found that English was a major Web language for international linking between universities throughout the European Union, accounting for about half of pages in most cases. Another trend identified was for countries that share a language to interlink in that language. French was not a particularly well-used language on the Web. Another study compared English with Chinese Web pages that linked between Mainland China and Taiwan, finding no evidence for a preference for English as a language of scholarly communication in this context (Thelwall & Tang, 2003).

3. Data sources

3.1. Web pages

The initial candidates for the study were all Canadian university Web sites. First, URLs of all Canadian universities were obtained from an online list (Association of Universities & Colleges of Canada, 2002) and the completeness of the set verified and supplemented using an unrelated print media source (Johnston, 2002). Data collection took place in the summer of 2002 so the 2002 version of these sources was used. A list of 74 universities and colleges was compiled. Canada is a bilingual country and Canadian universities have different language models. The language breakdown of universities in the study is 56 English, 16 French, and 2 bilingual. Each university's Web site was crawled by a specialist information science Web crawler (Thelwall, 2001a) to record link information. It was designed to cover sites accurately and to check for duplicate pages exhaustively. The crawler found pages by following links iteratively from the home page. When a university's home page did not contain any HTML links, a page that contained links to all departmental home pages was sought and used as an alternative starting point. In line with previous studies, some Web pages were excluded on the basis of being mirror sites or huge online databases with only internal links. The total number of pages crawled was 3,930,113.

3.2. University profile data

The main data source for university profiles is the 2003 version of Maclean's Guide to Canadian universities (Johnston, 2003). Maclean's is the leading newsmagazine of Canada, comparable to Time in the US. Maclean's publishes an annual guide to Canadian universities to help students choose which ones to apply for. The guide provides data that measures various aspects of a university such as faculty quality (e.g. research grants), student body caliber (e.g. average entering grade), and university library quality. Maclean's then ranks universities based on these data. Although the ranking results are not without controversy, the accuracy of the Maclean's data is not usually disputed and the data are generally accepted as the most comprehensive and reliable source for Canadian universities. For this reason, Maclean's objective (raw) data rather than the ranking scores were used. The Maclean's 2003 version was used for the study because these data were collected during the spring and summer of 2002 (Johnston, 2003), around the time university Web site data were collected. The following Maclean's data were used.

1. Measures of faculty research
   - Average number of research grants from the Social Science and Humanities Research Council of Canada per 100 eligible faculty members (referred to as **social_science_grant** below).
   - Average number of research grants from both the Natural Sciences and Engineering Research Council and the Canadian Institutes of Health Research per 100 eligible faculty members (referred to **science_grant** below).
   - Number of full-time professors, per 1000, who have won national awards (referred to as **faculty_award** below).
2. Measures of student quality
   - Number of students, per 1000, who have won national awards (referred to as **student_award** below).

Maclean's provided these data for 47 out of 74 universities in the study (all major universities are included in the 47). As a result, only 47 universities are included in the data analysis below. We also needed data on faculty size (number of academic staff) to reach a normalized link count (details below). Maclean's does not provide such data. The 16th edition of the International Handbook of Universities (Taylor, 2001) was used instead for this data. To investigate the geographic distance factor in the hyperlink patterns, distances among universities were needed. The official Web site of Natural Resources Canada (2003) was used to obtain the longitude and latitude data for each university and these were then converted into the geographic distances between universities.

4. The inlink model and its underlying hypothesis

The dependent variable of the model is inlink counts (an inlink is a link coming into a site from other sites). The independent variables are social_science_grant, science_grant, and student_award (see Section 3.2). Multiple regression analysis was used to develop a model that best describes the relationship among all these variables. No variables are significantly non-normally distributed so the normality requirement of multiple regression is satisfied. The approximately normal distribution of these variables means that more universities are in the middle or average range while fewer universities are either on the very high or low range of the variables (e.g. having either very large or very small research grants). In addition, the "language" variable was introduced as an independent variable to investigate the language factor in Web linking patterns. To simplify the model building process, the two bilingual universities were omitted from the study. The language variable thus has two values (English vs. French) and is treated as a dummy variable in the regression model. See Chapter 14 of Kleinbaum, Kupper, Muller, and Nizam (1998) for a detailed discussion of dummy variables. Essentially, a dummy, or indicator, variable is any variable in a regression equation that takes on a finite number of values so that different categories of a nominal variable can be identified. The term dummy reflects the fact that the values taken on by such variables (in this case, 0 for English and 1 for French) do not indicate meaningful measurements but rather the categories of interest (Kleinbaum et al., 1998, p. 317). To summarize, the regression model to be tested is

$$\mathbf{inlink} = \boldsymbol{b_1 X_1} + \boldsymbol{b_2}\mathbf{X_2} + \boldsymbol{b_3 X_3} + \boldsymbol{b_4 X_4}$$

where X_1 is social_science_grant; X_2 is science_grant; X_3 is student_award; X_4 is language; b_1, b_2, b_3, b_4 are coefficients for the corresponding independent variables.

The hypothesis underlying this model is that the number of inlinks a university Web site receives (normalized by the total number of faculty members) can be predicted or explained by faculty research profile (as measured by research grants), student quality (as measured by national awards students received), and the language of the university. It should be noted that the model used throughout the paper is the standardized regression model, which means that units of different independent variables are standardized to make the regression coefficients more comparable. This explains why the model does not have the coefficient b_0 that is usually present in an unstandardized model. Regression analysis of this model will test the hypothesis by indicating whether the coefficients b_i associated with these variables are statistically significant or not. The analysis will also indicate the relative importance of the variables. To summarize, the purpose of testing the model is to find out (1) whether the suspected independent variables are genuinely influential; (2) whether universities attract links related to their faculty quality, their student quality, and their languages; and (3) whether a sensible model can be built by combining the suspected independent variables.

5. Data analysis and results

5.1. Data preparation

Web site data downloaded by the crawler can be analyzed in various ways to extract link data. There are two types of links: inlinks (links coming in to a site) or outlinks (links going out from a site). The two types are strongly correlated (the Pearson correlation coefficient was over 0.8) so only one type needs to be analyzed. Inlinks were analyzed in this study because inlink data are available through commercial search engines so they have wider applications. There are various ways of counting links depending on which document model is used, i.e. what is the counting unit (page, directory or domain). Thelwall (2002a) examined different alternative document models (ADMs) and found domain level counting to be the best. To determine the best document model for this study, links were counted using all these three units and then each correlated with the "best overall" ranking of the university by Maclean's (Johnston, 2003). Domain level counting was found to have the highest correlation with the "best overall" ranking so this counting method was used.

Larger universities are likely to have more Web pages and thus attract more links to their sites. To test the hypothesis that link counts are correlated with the quality of the university, as measured by faculty and/or student quality, link counts need to be normalized by university size. This was achieved through normalizing the inlink counts by faculty size (i.e. inlink counts divided by the number of faculty members). Faculty size, rather than student body size, was used as a measure of university size because previous studies have found that links target very few student pages, even if students were allowed to post pages on their university's server (Wilkinson et al., 2003). An alternative is to normalize inlink counts by the total number of pages of a university. However, the number of Web pages are highly influenced by artificial Web design decisions on how to divide up pages and thus is not a good measurement for this purpose (Thelwall, 2001b, 2001d).

5.2. Inlink model testing results

Three commonly used methods for multiple regression analysis are stepwise regression, forward selection, and backward elimination. All three methods were tried and the resulting model is the same:

$$\textbf{inlink} = \textbf{0.52 science\_grant} - \textbf{0.27 language}$$

The multiple correlation coefficient of this model is 0.57, which is statistically significant at the 0.01 level. The model shows that inlink counts correlate positively with research. Specifically, the number of science and engineering research grants is a better indicator of inlinks than the number of social science research grants. It should be noted that social_science_grant correlated with inlinks too (Pearson's r is 0.38). However, social_science_grant does not contribute significantly to predicting inlinks once science_grant is present in the model. The negative coefficient of the language variable means that French sites attracted fewer links than comparable English sites (when the value of the language variable increases from 0, English, to 1, French, the link count decreases). Between the two variables in the model, science_grant is the more important one (contributes more in predicting inlink counts) based on the fact that this was the first variable that entered the model in the stepwise regression and that it has a higher semi-partial correlation coefficient. The variable that measures student quality (student_award) is excluded from the model, which means that faculty quality is a better predictor of inlinks than student quality. This is understandable given that Web pages that attract links are generally created by faculty. This also supports our decision above to normalize inlink counts by faculty size.

5.3. Model validation through triangulation

One of the ways to validate a model is to split the data set into two sets; developing a model using one set and then testing the model using the other set. This method is not feasible in this study due to the size of data set (not enough universities to do this). An alternative method was used in which the robustness of the underlying constructs of the model was tested. The main variable in the model above is science_grant which is a measure of faculty quality. Another variable that measures faculty quality, faculty_award (see Section 3.2 for details of this variable), was then used to replace the science_grant variable. In other words, the model with independent variables faculty_award and language was tested. The "enter" method rather than the "stepwise" method of regression analysis was used in testing this model because here we are not trying to select variables but rather confirm a model with given variables. The resulting model is

$$\textbf{inlink} = \textbf{0.47 faculty\_award} - \textbf{0.26 language}$$

The multiple correlation coefficient of this model is 0.53, which is also statistically significant at the 0.01 level. Note that the regression coefficients of this model are very similar to those of the model above. The relative importance of the two variables is the same too, i.e. the faculty quality is a better predictor of inlinks than the language variable. This shows that the model developed in the study is fairly robust.

5.4. Model adequacy checking

The following examinations were carried out to check how well the models fit the data. First, standardized residuals were inspected to identify possible outliers in dependent variables and the Cook's distances were calculated to identify influential data points. The criteria used were standardized residual > 3 (Montgomery, Peck, & Vining, 2001, p. 133) or Cook's distance > 1 (Montgomery et al., 2001, p. 212) signals a problem. No outliers or influential data points were detected when these criteria were applied to the two models presented above. Next, graphical analysis of residuals, which is a very effective way to investigate the adequacy of the fit of a regression model (Montgomery et al., 2001, p. 138), was carried out by constructing a normal probability plot of the standardized residuals. Both models showed very good fit according to the standards in Montgomery et al. (2001, p. 139). Finally, multicollinearity, a problem when independent variables are highly correlated with each other, was examined (one of the undesirable consequences of multicollinearity is that regression coefficients will be unstable from sample to sample). Using the criteria that VIF (variance inflation factor) > 10 is a reason for concern (Stevens, 1996, p. 77), none of the models have the multicollinearity problem.

5.5. Modeling of interlinks

The models tested so far examined factors that affect the inlinks to a site. Interlinks (links between pairs of universities) were then investigated using a graphical modeling approach. Since faculty size, faculty quality (as measured by science_grants), and language were shown to affect inlink counts in the models reported above, these factors were taken into consideration in interlinking analysis in the following ways. First, interlinks were analyzed in four separate groups: English to English, French to French, English to French, and French to English. Second, the interlink count between university A and university B was normalized by the formula

$$\frac{\text{Link count from A to B}}{(\text{science\_grant A}) \times (\text{faculty size A}) \times (\text{science\_grant B}) \times (\text{faculty size B})}$$

This formula for normalization was found to be effective in an early study on interlinking (Thelwall, 2002d). Geographical distance between universities was examined as a possible factor that affects the tendency for two universities to link to each other. This decision was based on a study of UK universities, which found an inverse relationship between distance and tendency to link (Thelwall, 2002d).

Fig. 1 is the graphical summary of the data. Note that each distance group of Fig. 1 represents a varying number of universities and that missing bars represent missing data (i.e. no two universities have this distance between them) rather than bars of zero height. The vertical axis is the normalized interlink count averaged over the universities within the group.

The dominance of English in interlinking is evident in Fig. 1. Interlinking between English language sites is the highest in each distance group while interlinking between French universities

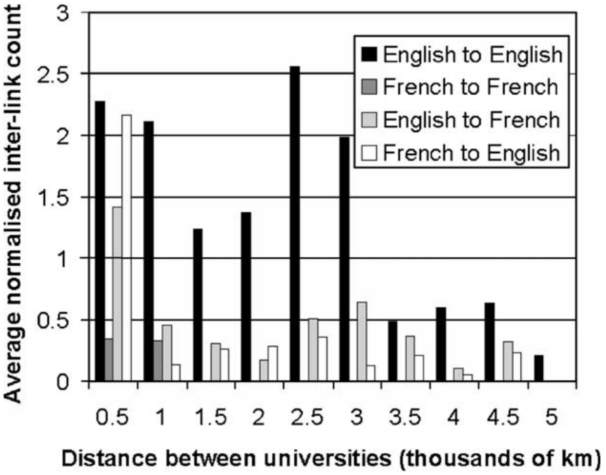

Fig. 1. Average interlink counts among Canadian universities.

Table 1
Total pairs of sites in each of the language groups in Fig. 1

English to English	French to French	English to French	French to English
1640	12	164	164

is much lower. However, the small number of French universities as shown in Table 1 means that the interpretation of the results for the French–French interlinking should be cautious. It is interesting to note that in the first distance group (under 500 km, e.g. within the province of Quebec, a French Canadian province) there are more French to English links than English to French links. Given the dominance of English, it would be reasonable to attribute the lower interlinking for pairs where one university is English and the other is French to the lower inlinks to French sites in general, rather than a tendency to link to same-language sites. All of this is consistent with the finding that French language sites receive fewer inlinks than their comparable English sites, which is revealed in the regression model.

There is no clear trend in Fig. 1 that interlinking decreases as distance increases. Given the noticeable bump in the middle of Fig. 1, it seems more likely that there is a pattern of two groups: those within 3000 km are more likely to link to each other while those more than 3000 km apart are much less likely to interlink.

6. Discussion

Hyperlink patterns of Canadian university Web sites were analyzed through a mathematical modeling approach. Various factors that could potentially affect the number of inlinks to a Web site were investigated and a parsimonious regression model involving only two independent variables was successfully developed. The two most important factors that determine the number of inlinks, in the order of importance, were faculty quality and the language of the university.

Faculty quality was first measured by faculty research grants. The number of medical, science, and engineering grants was found to be a better predictor of inlink counts than the number of social science and humanities research grants, although the two are correlated. This finding is consistent with that from other studies (Tang & Thelwall, in press), which showed that the Web pages of social science and humanities researchers attract significantly less links than those of science researchers. The caliber of the student body, as measured by the average number of students who won national awards, was not a significant predictor in inlink counts. This echoes earlier studies (Thelwall, 2001b; Wilkinson et al., 2003) that found faculty, rather than students, to be the main attractors for inlinks.

The robustness of the model was tested through model triangulation, where the construct faculty quality was measured by a different variable, the average number of professors who won national awards. The test result confirmed the model developed. Various regression diagnostics were carried to evaluate the adequacy of the model and the result shows that the model fits the data well.

Perhaps the most important finding that has real life implications is that cultural factors along linguistic lines are a significant factor in inlink attraction. French universities received a significantly lower number of inlinks to their Web sites than comparable English universities ("comparable" is determined by holding other variables constant). A visit to the French university Web sites revealed that 50% of the homepages did not have any sign that an English version of the Web site was available. The homepages that had a link to an English version did not display this option clearly. For example, one homepage had a tiny "welcome" sign on the top right corner that did not clearly look like a clickable link. Many links on English pages led to a French page. This would exacerbate the difficulties for English speakers accessing the site and could partially explain the lower English to French linking than English to English linking. It is difficult to be specific about the impact of linguistic problems because many links target home pages and do not necessarily mean that the targeted site has been visited, or that its contents are being endorsed (Thelwall, 2002f; Thelwall, 2003). There are also probably embedded cultural factors at work, with one sociologist claiming English speakers to have very little knowledge of Francophone Canada (Stebbins, 2000). There are known cultural connections between Quebec and France (e.g. Balthazar, 1995; Corbett, 1990), a known low Web publishing country (Thelwall et al., 2003), and so it would be reasonable to assume that each country would influence the other in many ways.

It will help to understand the language situation on the Web sites of French Canadian universities if one is aware of the language law in Quebec, a province of Canada where the official language is French. The Charter of the French Language (Government of Quebec, 2003) makes the use of the French language compulsory in commercial advertising and this applies to Web sites too. There are even "cyber-inspectors" to enforce the law (The Office québécois de la langue française, 2003). Although this does not apply to university Web sites, Chapter 88.1 of the Charter requires that before October 1, 2004, every Quebec institution that provides college instruction adopt a policy applicable to college-level instruction regarding the use and quality of the French language (Government of Quebec, 2003). The finding of this study has implications for the universities in setting up their policies, suggesting that at least a partially bilingual policy would be in the best interests of universities. The importance of inlinks are threefold: more inlinks means more visible on the Web and potentially more traffic to the site; sites with more inlinks are better covered by search engines (Vaughan & Thelwall, in press); sites with more inlinks will be ranked higher in search results (major search engines all use a link-based ranking algorithm). The disadvantages of fewer links are obvious.

Based on the results of inlink modeling, the patterns of interlinks between pairs of universities were examined while taking into consideration faculty size and faculty research. Overall, English universities are better interlinked than French universities. There is no clear pattern that interlinking between universities decreases as geographical distance between them increases, which was found to be true for UK universities (Thelwall, 2002d). However, there is a fairly clear binary divide around 3000 km. There are many more interlinks when the distance is within 3000 km and fewer when distance is beyond 3000 km. This might represent the east vs. west divide that exists in the Canadian society politically, socially, and economically (Cameron, 1991; Jackson & Jackson, 2001, Chap. 3). Canada is the second largest country in the world by geographical size but its population is relatively small. Canadian universities do not spread out evenly throughout the country. They are more concentrated in east and west with fewer in the middle. If east were more likely to link to east and west more likely to link to west, then the linking pattern would be consistent quantitatively with that shown in Fig. 1. This finding is different from that of Katz (1994), who plotted a similar geographic graph using university–university co-authorship between 1984 and 1990. Katz found a smoother geographic degradation in co-authorship rather than a binary divide. A likely explanation for this is that formal collaboration, as represented by co-authorship, is less influenced by the east-west divide than the informal communication represented by hyperlinks. It is also possible, however, that a binary divide has become more pronounced in the time gap between the two studies.

Acknowledgements

We thank Zhihong Tang and Elena Saunders for their help in data collection.

References

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide Web: methodological approaches to 'Webometrics'. *Journal of Documentation, 53*(4), 404–426.

Arasu, A., Cho, J., Garcia-Molina, H., Paepcke, A., & Raghavan, S. (2001). Searching the Web. *ACM Transactions on Internet Technology, 1*(1), 2–43.

Association of Universities and Colleges of Canada (2002). The directory of Canadian Universities—University Websites. Retrieved April 24, 2002 from http://www.aucc.ca/english/dcu/universities/universitysites.html.

Balthazar, L. (1995). *Pour un renforcement de la solidarité entre Francophones au Canada; réflexions théoriques et analyse historique, juridique et sociopolitique*. Québec: Conseil de la langue française.

Borgman, C., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology, 36*, 3–72.

Cameron, N. E. (1991). *From East and West, Regional Views on Reconfederation*. No. 6 of Canada Round series. Toronto: C.D. Howe Institute.

Chakrabarti, S., Joshi, M. M., Punera, K., & Pennock, D. M. (2002). The structure of broad topics on the Web, WWW2002, Available: http://www2002.org/CDROM/refereed/338/.

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.

Corbett, N. L. (1990). *Langue et identité: le français et les Francophones d'Amérique du Nord*. Québec: Presses de l'Université Laval.

Government of Quebec (2003). The charter of the French language. Retrieved June 5, 2003 from http://www.olf.gouv.qc.ca/english/charter/index.html.

Henzinger, M. (2001). Hyperlink analysis for the Web. *IEEE Internet Computing, 5*(1), 45–50.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation, 54*(2), 236–243.

Jackson, R. J., & Jackson, D. (2001). *The fabric of Canadian society: ethno-linguistic and regional cleavages. Politics in Canada: culture, institutions, behaviour and public policy* (5th ed.). Toronto, Canada: Prentice-Hall.

Johnston, A. D. (Ed.). (2003). *The Maclean's guide to Canadian universities 2003*. Toronto, Canada: Rogers Publishing.

Johnston, A. D. (Ed.). (2002). *The Maclean's guide to Canadian universities 2002*. Toronto, Canada: Rogers Publishing.

Katz, J. S. (1994). Geographical proximity in scientific collaboration. *Scientometrics, 31*(1), 31–43.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariable methods*. Pacific Grove, CA: Duxbury Press.

Larson, R. R. (1996). Bibliometrics of the World Wide Web: an exploratory analysis of the intellectual structure of Cyberspace. In *Proceedings of the 59th ASIS annual meeting*, Baltimore, Maryland (pp. 71–78). Medford, NJ: Learned Information Inc./ASIS.

Moed, H. F. (2002). Measuring China's research performance using the Science Citation Index. *Scientometrics, 53*(3), 281–296.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2001). *Introduction to linear regression analysis* (3rd ed.). New York, USA: John Wiley & Sons Inc.

Natural Resources Canada (2003). *The atlas of Canada*. Retrieved May 30, 2003 from http://atlas.gc.ca/site/english/find_a_place/index_html.

Pennycook, A. (1994). *The cultural politics of English as an international language*. London: Longman.

Polanco, X., Boudourides, M. A., Besagni, D., & Roche, I. (2001). *Clustering and mapping Web sites for displaying implicit associations and visualising networks*. University of Patras. Available: http://www.math.upatras.gr/~mboudour/articles/Web_clustering&mapping.pdf.

Rousseau, R. (1997). Situations: an exploratory study. *Cybermetrics*, 1(1). Available: http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html.

Smith, A., & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics, 54*(3), 363–380.

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

Stebbins, R. A. (2000). *The French enigma: survival and development in Canada's Francophone societies*. Calgary: Detselig Enterprises Ltd.

Stevens, J. (1996). *Applied multivariate statistics for the social sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Tang, R., & Thelwall, M. (in press). Disciplinary differences in US academic departmental Web site interlinking. *Library and Information Science Research*.

Taylor, A. C. M. (Ed.). (2001). *International handbook of universities* (16th ed.). New York, NY: Palgrave.

The Office québécois de la langue française (2003). *Frequently asked questions on the Charter of the French language and Web sites*. Retrieved June 5, 2003 from http://www.oqlf.gouv.qc.ca/english/infoguides/faqs/faqs_anglais.html#frequently.

Thelwall, M., & Aguillo, I. (in press). La salud de las Web universitarias españolas. *Revista Española de Documentación Científica*, *26*(3).

Thelwall, M., & Harries, G. (2003). 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*(7), 594–602.

Thelwall, M., & Harries, G. (in press). Do better scholars' Web publications have significantly higher online impact? *Journal of the American Society for Information Science and Technology*.

Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic Web use in Western Europe. *Scientometrics, 56*(3), 417–432.

Thelwall, M., & Tang, R. (2003). Disciplinary and linguistic considerations for academic Web linking: an exploratory hyperlink mediated study with Mainland China and Taiwan. *Scientometrics, 58*(1), 153–179.

Thelwall, M. (2001a). A Web crawler design for data mining. *Journal of Information Science, 27*(5), 319–325.

Thelwall, M. (2001b). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology, 52*(13), 1157–1168.

Thelwall, M. (2001c). Exploring the link structure of the Web with network diagrams. *Journal of Information Science, 27*(6), 393–402.

Thelwall, M. (2001d). Results from a Web Impact Factor crawler. *Journal of Documentation, 57*(2), 177–191.

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). An initial exploration of the link relationship between UK university Web sites. *ASLIB Proceedings, 54*(2), 118–126.

Thelwall, M. (2002c). Research dissemination and invocation on the Web. *Online Information Review, 26*(6), 413–420.

Thelwall, M. (2002d). Evidence for the existence of geographic trends in university Web site interlinking. *Journal of Documentation, 58*(5), 563–574.

Thelwall, M. (2002e). A Research and institutional size based model for national university Web site interlinking. *Journal of Documentation, 58*(6), 683–694.

Thelwall, M. (2002f). 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. (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: http://informationr.net/ir/8-3/paper151.html.

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

van Leeuwen, T., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & van Raan, A. F. J. (2000). First evidence of serious language-bias in the use of citation analysis for the evaluation of national science systems. *Research Evaluation, 8*(2), 155–156.

Vaughan, L., & Hysen, K. (2002). Relationship between links to journal Web sites and Impact Factors. *ASLIB Proceedings, 54*(6), 356–361.

Vaughan, L., & Thelwall, M. (2003). Scholarly use of the Web: What are the key inducers of links to journal Web sites? *Journal of the American Society for Information Science and Technology, 54*(1), 29–38.

Vaughan, L., & Thelwall, M. (in press). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management*.

Vaughan, L., & Wu, G. (2003). Link counts to commercial Web sites as a source of company information. In G. Jiang, R. Rousseau, Y. Wu (Eds.), *Proceedings of the 9th international conference of scientometrics and informetrics* (pp. 321–329), Beijing, China, August 25–29, 2003.

Wilkinson, D., Harries, G., Thelwall, M., & Price, E. (2003). Motivations for 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.