Text Characteristics of English Language University Web Sites

Mike Thelwall[1]
School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK. E-mail: m.thelwall@wlv.ac.uk
Tel: +44 1902 321470 Fax: +44 1902 321478

The nature of the contents of academic Web sites is of direct relevance to the new field of scientific Web intelligence, and for search engine and topic-specific crawler designers. We analyze word frequencies in national academic Webs using the Web sites of three English-speaking nations: Australia, New Zealand and the U.K. Strong regularities were found in page size and word frequency distributions, but with significant anomalies. At least 26% of pages contain no words. High frequency words include university names and acronyms, Internet terminology, and computing product names: not always words in common usage away from the Web. A minority of low frequency words are spelling mistakes, with other common types including non-words, proper names, foreign language terms or computer science variable names. Based upon these findings, recommendations for data cleansing and filtering are made, particularly for clustering applications.

Introduction

The world's university Web sites contain an enormous quantity of information, ranging from preprints to administrative and recreational pages, created by faculty, students and support staff (Middleton, McConnell & Davidson, 1999). This goldmine can easily be exploited for many academic-related purposes such as exploring topics with online articles and course notes, identifying active scholars and their publications from their personal home pages, and finding courses from online prospectuses. A commercial search engine such as Google is likely to intermediate between users and university Web sites, particularly for new or infrequent information needs. Yet there is much information in the Web that is not apparent from its individual pages, and can only be identified through large-scale investigations of factors such as the relationships between documents. This concept will be familiar to information scientists, particularly those versed in relational bibliometrics (Borgman & Furner, 2002) but the extraction of information from the Web is primarily the task of the multidisciplinary specialisms Web mining and Web intelligence, closely affiliated to computer science, artificial intelligence and statistics (Menasalvas, Segovia & Szczepaniak, 2003). Web mining is typically applied to the whole Web or a very restricted subset of it, such as a single site. The importance of academic information on the Web is a powerful argument for the development of specific Web mining techniques to deal with the world's university Web sites and other Web sources of scholarly information. We call the study of Web mining and Web intelligence techniques for academic-related sites *scientific Web intelligence* (Thelwall, 2004). This new field can combine insights from Web mining, Web intelligence and relational bibliometrics. Much of the field of webometrics, a field of quantitative studies of Web phenomena deriving from bibliometrics, can already claim to be scientific Web intelligence. This is particularly true of webometric link analysis studies of academic Web sites (e.g., Aguillo, 1998; Almind & Ingwersen, 1997; Heimeriks, Hörlesberger & van den Besselaar, 2003; Smith, 1999). One of the goals of scientific Web intelligence is to produce visualizations of academic Web spaces, using text and/or link data to construct maps to illustrate relationships between and within disciplines. This is a standard problem for knowledge domain-specific literatures (e.g., Börner, Chen, & Boyack, 2003) but is more complex for Web data because of its heterogeneous nature.

In this (methods-oriented) paper we seek to underpin the formation of the new field of scientific Web intelligence by a basic study of the text of university Web sites in three different English-speaking nations. The aim is exploratory rather than hypothesis testing: to find out how word use in academic Web spaces differs from that in standard written English. Although word frequency analyses are capable of giving directly useful results (e.g., Boynton, Glanville, McDaid & Lefebvre, 1998; Leydesdorff & Curran, 2000; Lindsay & Gordon, 1999; Oakes, 1998; Rousseau, 1999; Sanderson & van Rijsbergen, 1999; Vilensky, 1998), the primary aim of our research is to aid text-based data cleansing. A second use is in building intuition for future

[1] *This is a preprint of an article to be published in the Journal of the American Society for Information Science and Technology © copyright 2004 John Wiley & Sons, Inc.* http://www.interscience.wiley.com/

scientific Web intelligence applications: knowledge of the words used in academic Webs will suggest the kinds of information that can be extracted and perhaps also the problems that are likely to be encountered. The scope of this study is national rather than international for purely practical reasons: the difficulty in data collection. However, a good understanding of text characteristics on the national level lays the foundations for future research, such as knowledge domain visualisation, that will exploit Web data on an international level.

Web mining tends to be dominated by computer science, with an emphasis on constructing software to implement new algorithms and evaluating the finished problem as a whole. Data cleansing, a necessary part of data mining (Hand, Mannila & Smyth, 2001), is rarely mentioned, although it is particularly relevant to Web data (Thelwall, 2005). The treatment of raw data in Web mining is often cursory or unexplained. For example, in an influential Web graph analysis article (Broder, Kumar, Maghoul *et al.*, 2000), data cleansing is described in very general terms and in only one sentence, "The crawl proceeds in roughly a BFS manner, but is subject to various rules designed to avoid overloading Web servers, avoid robot traps (artificial infinite paths), avoid and/or detect spam (page flooding), deal with connection time outs, etc." This lacks detail, even though for this kind of Web mining the details are potentially critical (Cothey, 2005). Our exploratory data analysis represents an information science contribution to scientific Web intelligence and also in the long term gives the potential for 'scientific' to have a double meaning: mining part of the scientific Web but also using techniques that will be based upon scientific principles. This is possible because studies restricted to the academic Web have more potential for a systematic approach than those aimed at the general Web.

Background

Two research fields can provide useful information to set the context of this study: Webometrics and corpus linguistics. Webometrics provides knowledge about Web crawling and data cleansing. University Web sites contain aspects that are undesirable for Web mining, such as duplicated pages and copies of Web sites from other sources (mirror sites). A crawling technique has been developed to avoid these. It is partly automatic, for identifying duplicate pages within a site, and partly manual, for identifying sites mirrored from elsewhere in the world. The crawl-time data cleansing also involves the exclusion of statistical reports, such as server logs, and content not created by the host institution, such as e-journals (Thelwall, 2001; Thelwall, 2003). Webometrics also provides important context information through the general finding that Web phenomena are often counterintuitive and exhibit huge anomalies. Exploratory Web research needs to be careful in formulating experiments and drawing conclusions.

Corpus linguistics (e.g., McEnery & Wilson, 2001; Mitkov, 2003) is concerned with analyzing defined language corpora, both written and verbal. Objectives include comparative analyses of linguistic constructs across languages and contexts. Corpus linguistics typically involves analysis of complex linguistic phenomena for theory development and word semantics and usage for the purposes of dictionary building or foreign language teaching. Simple word frequency counts are also sometimes investigated (Oates, 1998), the approach adopted here.

For scientific Web intelligence, the purpose of text analysis is to extract meaning from Web documents and patterns of interrelated meanings between documents. Some natural features of language serve to make this more difficult. The first is polysemy; multiple meanings for a word. E.g. 'can' can be a verb or an unrelated noun. The process of determining which meaning a given word has in a particular context is known as semantic disambiguation or word sense identification. Automated processes for this have been developed and word sense annotated corpora are available (e.g., Miller, 1995). The second language problem is synonymy: identical or similar meanings for multiple words. There are two forms of synonymy. Similar words can be of different origin, e.g. liquid and fluid; or they can be related, e.g., stop, stopping, stops, stopped and stopper. Two approaches have been used to get round this problem. In information retrieval word stemming algorithms such as Porter's (1980) are sometimes used. These truncate words to stems, which are supposed to convey the root meaning. In the above example, 'stop' would be a logical stemming of stop, stopping, stops, stopped and stopper. Lemmatisation (McEnery, & Wilson, 2001) is a similar approach, more favored by corpus linguistics, in which words are replaced with a 'headword' or 'lexeme' – the related word that could be looked up in a dictionary. This is a more complex process than simple stemming: for example the lexeme is 'go' for each of: go; going; and went, but no stemming algorithm would be able to derive 'go' as the stem of 'went'. All automatic stemming and word sense annotation algorithms make mistakes and are likely to be less reliable on academic words than on standard English, simply because academic words are less frequent. For exploratory research with large corpora they make unnecessary changes in the raw data and so they will not be used for our study. This argument should not preclude their use in future research, however, but such research should explicitly compare the analysis results with and without automated corpus annotation or stemming.

Clustering is a key data mining technique that will be the foundation for many scientific Web intelligence approaches. The basics of text clustering are described here to show the importance of the word frequency characteristics of document sets. Although there are many different clustering techniques, most rely upon some measure of the inter-document similarity (Jain, Murty, & Flynn, 1999). A standard approach to inter-document similarity measurement is to use the Vector Space Model (VSM) (Baeza-Yates, & Ribeiro-Neto, 1999). With the VSM, documents are represented as word frequency vectors. These are lists of the number of occurrences of the words in the document. For example, suppose there are three documents to be clustered. Document A contains the single word "hello", document B contains "Hello world" and C contains "welcome, welcome". The vocabulary for this document collection is the list all words used: hello, world, welcome. Relative to this vocabulary, A can be represented as the word frequency vector (1,0,0), meaning 1 occurrence of "hello" and 0 occurrences of "world" and "welcome". Similarly, B is (1,1,0) and C is (0,0,2). Now A, B, and C can compared using mathematical similarity measures, and we would intuitively expect that any reasonable measure would show that A and B are closer to each other than to C, since they share no words in common with C. One often-used measure is the cosine similarity measure (Baeza-Yates, & Ribeiro-Neto, 1999). For two vectors their cosine similarity is obtained by multiplying together word frequencies for each word and then dividing by the square root of the sum of the squares of all the word frequencies for each of the two documents.

For A and B this calculation is:

$$\frac{1\times1+0\times1+0\times0}{\sqrt{(1^2+0^2+0^2)(1^2+1^2+0^2)}}=0.707$$

For A and C, and for B and C the equivalent calculations give a cosine similarity measure of 0. In fact any pair of documents with no words in common will have a cosine similarity measure of zero. A pair of documents will have a cosine similarity measure close to the maximum of 1 if they use mainly the same words and with similar frequencies (relative to document size). If they use mainly different words, and with different frequencies for the words that they have in common, they will have a cosine similarity measure of close to the minimum of 0. The highest frequency words within any document will have the biggest influence on its similarity with other documents. Documents with many occurrences of an unusual word or many different unusual words will have low cosine similarity measures with most other documents. Weighting schemes are frequently used to modify the standard cosine measure. These typically lower the importance of common words (i.e., those that occur in many different documents).

Domain clustering is an alternative clustering approach, based upon collating all the words of all documents belonging to the same domain name (host) into a single conceptual multi-page document, with one word frequency vector (i.e., the sum of the frequency vectors of all pages in the domain). This is an attractive approach for clustering national academic Webs because clustering is computationally expensive and impractical for the millions of pages in a typical country but practical for the thousands of domains.

Research Questions

The aim is to identify word frequency information about English language university Webs that will be useful in the design of scientific Web intelligence applications. The following four questions drive the investigation.

1. In English language university Webs, what are (a) the distributions of Web page sizes (in terms of the number of words per page) and (b) the distributions of word frequencies? For part (b), how do the distributions of word frequency counts compare with those from standard non-Web English?
2. What proportion of words in English language university Webs are not standard English words, and how does this vary with word frequency?
3. What types of words in English language university Webs are not standard English words? How does type vary with word frequency (c.f. Weeber, Vos, & Baayen, 2000)?
4. Are there words in English language university Webs that are anomalous in the sense of having significantly higher frequencies in the Web corpora than in standard English?

Data Sources

Standard non-Web English Word Frequencies

In order to compare university Web English with standard English, a reference source of standard English is needed. English use varies internationally, with national additions to the language, different national writing styles (and spellings) and different local place names. This has to be taken into account when comparing with non-British English. The construction of a coherent corpus is time-consuming and so corpora are often created as specialized projects and shared with interested parties. We selected the British National Corpus (Aston, & Burnard, 1998; Burnard, 1995), a spoken (10%) and written (90%) collection, with the written collection being a diverse set of documents chosen to represent a wide range of British English written sources. The sources represent modern English, so old sources were excluded even if still in use (e.g. nineteenth century novels). The most recent sources are from 1993 and hence predates the impact of the Web on popular culture. In the nature of language, however, full coverage of written sources is clearly impossible and the sample will necessarily be idiosyncratic in some respects, particularly in its coverage of minority special interest vocabularies – some will inevitably be absent, and others overrepresented.

In order to compare the university Web word frequencies with those from the British National Corpus, an online (unlemmatised) BNC word frequency list was used, covering just the written sources (Kilgarriff, 2003). This was in a semantically disambiguated form (i.e., different senses of the same word listed separately), and so the list was first converted to simple word frequency counts. Words which would not have been extracted from the Web corpora (e.g. containing numbers or hyphenated) were also removed. The frequency list excluded ’s and ’ endings, so these were removed (by word truncation) from the Web data sets to ensure compatibility of the frequency figures.

National University Webs

The national university Webs of Australia, New Zealand and the U.K. were chosen, three English speaking nations using essentially the same spelling system, comparable with each other and the BNC. There are no known sources of academic Web corpora, although corpus linguistics approaches have been applied to the Web, often via commercial search engines (e.g., Blair, Urland & Ma, 2002; Ghani, Jones, & Mladenić, 2001; Heyer, Quasthoff & Wolff, 2002; Keller & Lapata, 2003; Resnik & Smith, 2003), and Web corpora have been constructed for Information Retrieval (IR) purposes (e.g., Bailey, Craswell & Hawking, 2003). The Web sites of all universities in each country were crawled in 2003 to create the necessary corpora (see http://cybermetrics.wlv.ac.uk/ for a list of universities and crawl dates). The crawler used was an adaptation of a Webometric link crawler previously used for link analysis (Thelwall, 2003). A banned list was maintained so that the crawls would exclude all identified mirror sites and collections of automatically created pages, such as a server statistical files (see http://cybermetrics.wlv.ac.uk/ for full banned lists, and Thelwall [2001, 2003] for the crawl parameters). The banned list procedure is necessary data cleansing: without it the final results would consist of just Java-related terms because there are copies of the Java documentation on most university Web sites.

A program was written to extract words from each downloaded HTML page. Pages associated with server error or redirection notices were ignored, as were non-HTML pages, including those in Portable Document Format. Words were extracted from the page title and body only. Words inside tags were ignored. A word was defined as a continuous list of non-white space characters delimited by white space characters, punctuation or HTML tags. The maximum word length allowed was 25 characters. Words containing any characters other than an apostrophe and the 26 (unaccented) letters of the English alphabet were excluded after being extracted. In particular, no hyphenated words, words containing numbers, or words containing accented letters were retained.

Known Words

A baseline word set was needed to compare the four data sets against. An online British English word list was used (Atkinson, 2003). It is designed to be the basis for software spell-checkers and contains common dictionary words as well as some proper nouns. This was chosen to represent common words that any Web mining application could be expected to incorporate. Of course, no such list could be perfect or free from national biases. In an attempt to compensate for any peculiarities with the list, three additional popular online sources were used: the Cambridge Advanced Learner's Dictionary (dictionary.cambridge.org), the OneLook

Dictionary (www.onelook.com) and Encyclopedia.com. The encyclopedia is an additional source of common proper nouns.

Methods

For the first research question a program was built to calculate frequency statistics from the words extracted from the university Web sites to produce the necessary statistics.

For the second research question, in order to compare statistics for words of differing frequencies, random samples of size 100 were taken from each corpus for words of frequency 1, 2, 3, 4, 5, 10, 100 and 1000. In cases when there were not enough words of the correct frequency, words were successively added using a list of those with just higher or lower frequencies. For example when there were not enough words of frequency 1000 then extras were added with frequency 999, 1001, 998, 1002, etc. Each word in the list was checked for being recognized in the known words data sources described above.

For the third research question, a word type classification scheme was devised and applied. The initial two categories were errors and specialist academic words. Other categories were added to the scheme in order to fit words that did not appear in the two main categories. The author devised the classifications manually. A specially written interface for the university data sets was built to help the process by allowing the rapid identification of words in the context of the texts from which they originated (a concordance program in the terminology of corpus linguistics, but for computer science it is a search engine interface with keyword sensitive context summaries). For the BNC text the concordance program SARA was used (Aston & Burnard, 1998).

For the fourth research question, a program was written to merge the word frequency lists from the four different primary sources.

Results

Question 1: Page Size and Word Frequencies

Table 1 reports the total number of words and total unique words for the four corpora used. This shows the enormous size of two of the Web corpora compared to the BNC, itself considered to be a large corpus. For comparison, the TREC WT10g Web IR test corpus contains 1.69 million pages (Bailey, Craswell & Hawking, 2003).

Table 1. Corpus statistics

Corpus	Total words	Total unique words	Pages
BNC	85,744,013	326,424	N/A
Australian academic Web (au)	656,653,108	1,542,589	205,513
New Zealand academic Web (nz)	60,998,403	385,093	2,152,386
United Kingdom academic Web (uk)	1,349,418,614	3,522,664	4,864,271

All the corpora exhibit word frequencies that follow a power law (Figure 1 and Figure 2), as would be expected by previous results (Zipf, 1949; c.f. Li, 1992). Frequencies are found on both the x and y axes of these graphs. For example, the point in Figure 1 with x coordinate 1 gives the information that there are over 100,000 words that occur only once in the entire BNC corpus. Word frequency studies for the general Web do not seem to exist. Search engines implicitly use word frequency information for page ranking with the vector space model (Baeza-Yates, R., & Ribeiro-Neto, 1999) and to suggest spelling corrections, but choose not to report them. In a small experiment Baldi, Frasconi and Smyth (2003) found a power law for word frequencies in a single university Web site. Power law phenomena have been previously found on the Web for Web linking (Barabási & Albert, 1999; Rousseau, 1997).

The main difference between the BNC and the Web corpora was that the former was a 'purer' power law with less anomalous behavior. This can be seen from a comparison of Figure 1 and Figure 2. The other two Web corpora have graphs more similar to Figure 2 than Figure 1 (not shown).

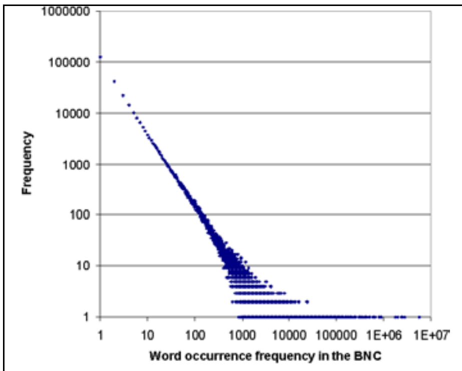


Fig. 1. The power law-like distribution of word frequencies in the British National Corpus.

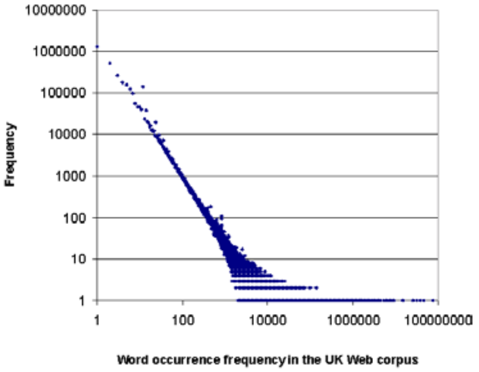


Fig. 2. The power law-like distribution of word frequencies in the UK academic Web corpus.

Figure 3 illustrates the hooked power law (c.f., Pennock *et al*., 2002) that is typical for the Web page size (number of words) graphs of for each of the three Web corpora. It is perhaps less surprising that the graph is hooked than that Web pages with few words are so frequent. This can be explained by the number of Web pages containing only graphics or words buried in graphics, as well as frameset pages. Pages that do not contain any words, not shown due to the log-log scale used, account for 612,501 (28%) of pages in the Australian corpus, twice as many as any other word count frequency. There were similar results for the U.K. (26%) and New Zealand (27%). A significant proportion of pages in each corpus contain few words. In the Australian corpus, for example, 35% of pages contain 10 words or less. The hooked shape may indicate two competing tendencies (c.f. Pennock et al., 2002), perhaps a pure power law associated with document sizes reflecting varying information communication needs, and a non-power law for documents that do not have a main purpose of conveying textual information.

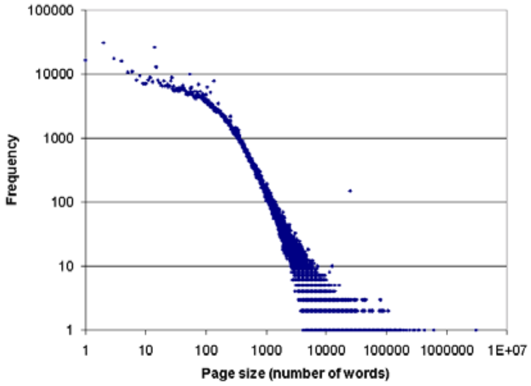


Fig. 3. The 'hooked power law' distribution of word frequencies in the Australian Web corpus

Question 2: The Proportion of Known English Words

Figure 4 shows that low frequency words are mostly unknown (i.e., not in any of the dictionaries nor the encyclopedia). In three of the corpora over a fifth of words occurring only once are known words. In all three Web corpora there are very high frequency unknown words.

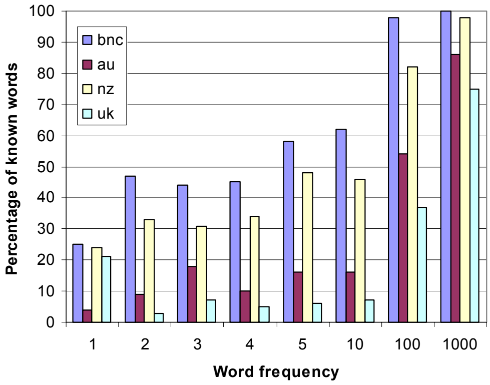


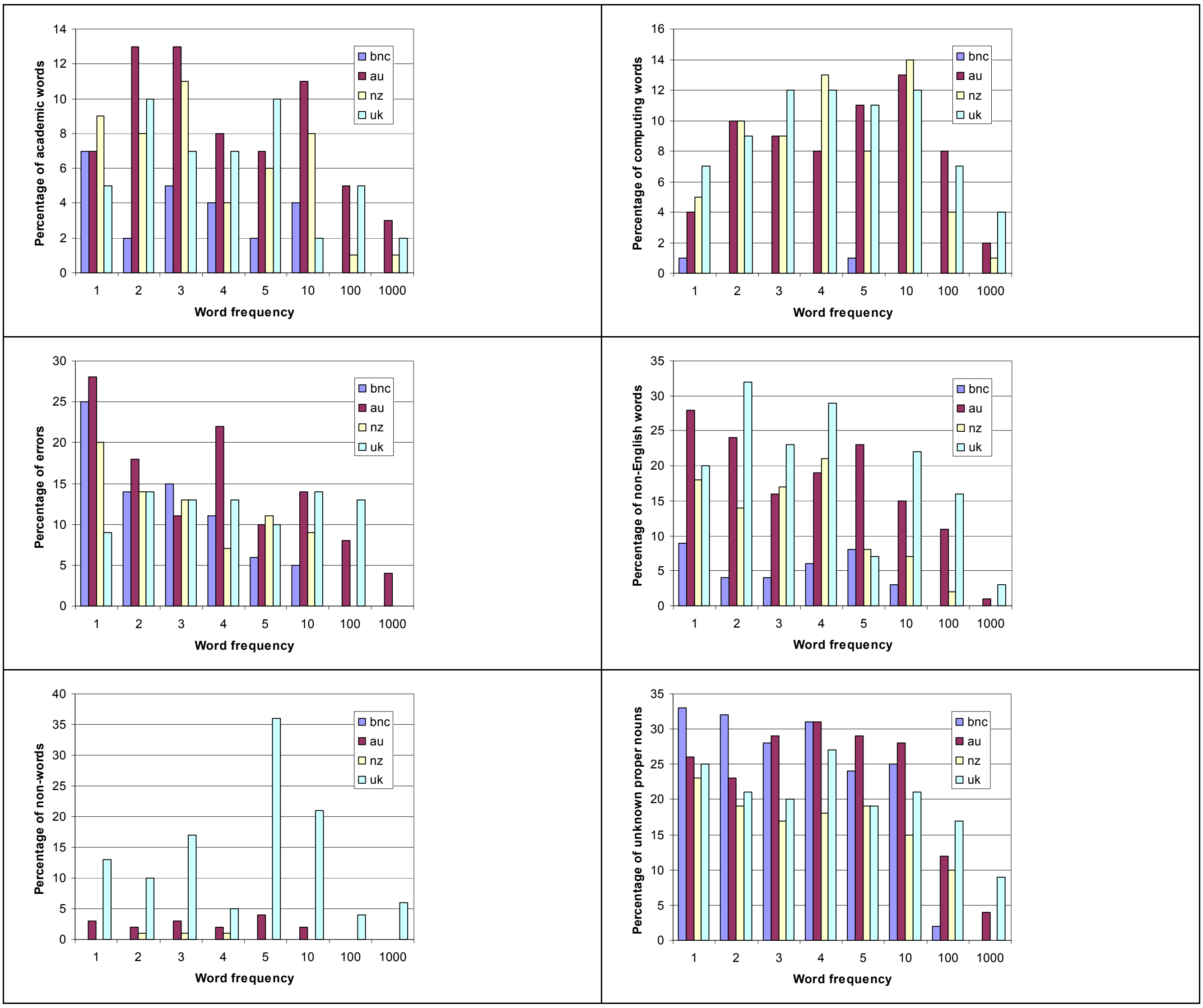
Fig. 4. The percentage of known words from a random sample of 100 at each frequency level.

Question 3: Unknown Word Types

Figures 5 to 10 show the types of words that were outside the list of known words.

- Academic words are those found in an academic context with a specialist meaning. For example, the Latin name of a plant would be counted as an academic word if used in an academic context (e.g., in a discussion of the classics), but the same word would be counted as Latin if used outside of an academic context (e.g., in one case in a list of plants in the grounds of a university).
- Computing words are the names of variables and functions in computer programs (e.g., SetDefaultAuthenticator), the names of variables outside of programs (mostly database table names and lists of program parameters in online documentation), and usernames.
- Words are classed as errors if they are spelling mistakes, slang words, or two words joined together (e.g., videoproject).
- Non-English words are foreign words not used as proper nouns. Common languages include French, Gaelic, German, Italian, Latin, Maori, Spanish, and Welsh. Other languages represented by at least one word include Chinese, Cornish, Greek, Hebrew, Japanese, Medieval English, Sanskrit, Sumerian, Turkish and Eastern European languages. Words in any language not using the standard 26 letters of the alphabet for English were not extracted by the word counting program and so could not be included. A few of these words did appear, however, in versions encoded with the standard English 26 letters of the alphabet (e.g., Chinese Pinyin).
- Non-words are collections of characters that are not intended to form a dictionary word or a proper noun in any language and are not variable names. These are alphabetic data encodings but not words in the normally accepted sense. Non-words are typically descriptions of chemical or biological compounds (e.g., the protein sequence allgsqfgqg). There are several collections of pages that describe biological compounds through long lists of non-words.
- Proper nouns include personal names, place names and product names. In many cases a personal name is also a place name or a word in English or a foreign language. Words were classified based upon a random choice of text location, and so context was used to determine category.

A second classifier classified 10% of the words based upon the same categories, but with a different random selection of context with which to classify the word. The agreement rate was 88%, which is not high (Neuendorf, 2002), but reflects partly the difficulty in assigning classifications by context and partly the fact that the same word in different pages can match different categories.



Figs. 5 to 10. The classification of different word types in the four corpora.

Question 4: High Frequency Anomalies

Tables 2 to 4 show the highest frequency words in each academic Web corpus that are not found in the BNC corpus. There are three patterns. First, standard Web terminology is highly represented. This reflects the BNC predating the popularization of the Web as well as the large amount of Web documents that discuss or refer to the Web. Second, the names of computing companies and software products are evident, particularly, but not exclusively (e.g., Linux), those relating specifically to the Web. Third, university name abbreviations are present. The other words represented are a common Maori word, two searchable databases in Waikato University library: ProQuest and Epnet, and CRICOS, which is frequently used in the context of giving a code to describe Australian university courses – “the CRICOS code”.

Table 2. The most common words in the Australian corpus that are not in the BNC corpus

Word	Origin	Frequency
linux*	Computer operating system	217,585
url*	Web term	170,142
website*	Web term	145,312
homepage*	Web term	144,137
html*	Web term	137,627
cricos	Commonwealth Register of Institutions and Courses for Overseas Students	107,856
unsw	University of New South Wales	84,718
uwa	University of Western Australia	72,732
rmit	Royal Melbourne Institute of Technology	54,487
netscape*	Web browser company	50,009

* Word appears in more than one of Tables 2 to 4.

Table 3. The most common words in the New Zealand corpus that are not in the BNC corpus

Word	Origin	Frequency
proquest	ProQuest - Commercial digital library system	24,986
website*	Web term	12,857
cellml	CellML – an XML based modeling language to store and exchange computer-based biological models	8,881
html*	Web term	8,127
homepage*	Web term	7,246
url*	Web term	6,968
linux*	Computer operating system	6,847
vuw	Victoria University of Wellington	6,375
atu	Maori word	5,722
epnet	Commercial digital library system	5,530

* Word appears in more than one of Tables 2 to 4.

Table 4. The most common words in the U.K. corpus that are not in the BNC corpus

Word	Origin	Frequency
linux*	Computer operating system	306,532
website*	Web term	300,847
html*	Web term	253,913
url*	Web term	246,873
homepage*	Web term	198,289
faq	Web term – Frequently Asked Questions	79,622
xml	Web computing term –eXtensible Markup Language	78,032
netscape*	Web browser company	74,462
webmaster	Web term	68,686
twiki	TWiki - a Web-based collaboration platform	65,704

* Word appears in more than one of Tables 2 to 4..

Tables 5 to 7 show the words in each academic Web that have unexpectedly high frequencies – defined to be a frequency of at least 1 in both corpora and a high Web frequency divided by BNC frequency. The reverse exercise was also conducted but is not shown, i.e., high frequency BNC words that rarely appear in each of the academic Webs. The highest frequency term found in any of these was much lower: ‘Kinnock’ (a former political party leader in the U.K.) occurred once in the New Zealand corpus but 1,411 times in the BNC corpus.

Tables 5 to 7 continue the trends identified above, but also include common abbreviations like ‘prev’ for ‘previous’ for navigation buttons. The word ‘Millimolar’ is the only non-computing term that could be construed as academic. Its frequency in the New Zealand corpus is due to its repeated inclusion in a large collection of CellML pages. The word ‘workarounds’ appears for its extensive use in computer documentation. Although it has no ostensible connection to computing, it is frequently found in pages describing ways to circumvent problems with software. In this context, it is commonly found in the phrase “known bugs and workarounds”.

Table 5. The ten highest relative frequency words between the BNC and Australian corpora

Word	Origin	BNC	Au	Au/BNC
jcu	James Cook University	1	255625	255625
prev*	Navigation link	1	155446	155446
monash	Monash University	2	206232	103116
uq	University of Queensland	1	78530	78530
constr*	Java programming language	1	49951	49951
www*	World Wide Web	2	87996	43998
tafe	Technical and Further Education, Australia	1	38494	38494
emacs	Text editor	4	109196	27299
trobe	La Trobe University	4	97945	24486
Workarounds*	Used in software documentation	1	24022	24022

* Word appears in more than one of Tables 5 to 7.

Table 6. The ten highest relative frequency words between the BNC and New Zealand corpora

Word	Source	BNC	NZ	NZ/BNC
umi	ProQuest UMI – commercial digital library	1	18090	18090
iwi	Maori word	1	4124	4124
tangata	Maori word	1	4022	4022
tonu	Maori word	1	2967	2967
millimolar	Chemical unit of measurement	1	2844	2844
prev*	Navigation link	1	2546	2546
php*	Dynamic Web page creation software	3	6751	2250
waka	Maori word	1	2241	2241
marae	Maori word	1	1935	1935
constr*	Java programming language	1	1807	1807

* Word appears in more than one of Tables 5 to 7.

Table 7. The ten highest relative frequency words between the BNC and U.K. corpora

Word	Origin	BNC	UK	UK/BNC
prev*	Navigation link	1	271562	271562
constr*	Java programming language	1	108347	108347
www*	World Wide Web	2	117798	58899
wk	Abbreviation for ‘week’	2	89068	44534
mmu	Manchester Metropolitan University	1	37668	37668
php*	Dynamic Web page creation software	3	100667	33555
const	Keyword used in programming languages	7	234010	33430
applet	Commonly used for Java programs in Web pages	2	65920	32960
perl	Programming language name	3	98247	32749
workarounds*	Used in software documentation	1	31735	31735

* Word appears in more than one of Tables 5 to 7.

Discussion

Basic finding will be discussed first, and then implications for clustering will be discussed in a separate section. As mentioned earlier, the findings have particular relevance for clustering because effective clustering is critical for many types of Web mining and Web intelligence (c.f., Poudat & Cleuziou, 2003).

Text Characteristics of English Language University Web Sites

One overarching finding can be drawn from all of the results: that academic Webs exhibit instances of anomalous behavior, seemingly from every perspective from which they can be measured. This confirms and extends similar conclusions based upon linking phenomena (Thelwall & Wilkinson, 2003; Thelwall, 2002) and demonstrates that attempts to provide simple mathematical models of the Web (e.g. Barabási & Albert, [1999]; Pennock, Flake, Lawrence, Glover & Giles, [2002] for link and page-based models) are unlikely to be convincing for academic Web spaces. For scientific Web intelligence, the widespread occurrence of different types of anomaly underscores the importance of data cleansing.

Question 1 From Figure 3 it can be seem that large numbers of Web pages contain few words and so would be unlikely to be susceptible to algorithms that attempt to ascertain their semantic content (e.g., topic clustering, identifying relationships between Web topics).

Question 2 The existence of high frequency unknown words marks out all of the Web corpora from the BNC, as does their higher overall percentage of unknown words. It is likely that high frequency unknown words are not present in the BNC because of its creation strategy, which deliberately excluded any very long text. In contrast the Web corpora are entire collections and can therefore contain many pages written about the same topic, possibly repeatedly using words that are highly specific to that topic. The relatively low percentage of known words in the Web corpora, particularly for the U.K. and Australia, is striking. Note, however, that there is a scaling effect present: the New Zealand corpus is most comparable to the BNC in terms of size. If the BNC (or New Zealand corpus) was scaled up by a factor of 10 to approximately match the UK and Australia corpora then the proportion of unknown words could be expected to increase, since low frequency known words would presumably be more likely to have their frequencies increased by the corpus extension than low frequency unknown words. As a result, the analysis of word types that are unknown is important to help interpret these results.

Question 3 In the classification exercise, it was observed that pages with many low frequency words come from three sources: foreign language pages; pages with many computer variable names; and highly specialist academic sources. Figures 5 to 10 show that low frequency words are not predominantly spelling mistakes, but are a range of types, including foreign words and proper nouns. At lower frequencies, errors occur evenly across the different corpora, but two of the Web sources also include higher frequency error words. Academic words occur in the BNC, partly because some of its texts are academic, and partly because some medical terms occur in non-academic texts. Non-words are unique to the Web and, although present in all Webs, were particularly evident in the U.K. because of several collections of pages of protein sequences. The high frequency of some non-words is a surprising finding. For example, the protein sequence ‘allgsqfgqg’, which appears in the U.K. academic Web 1000 times, is a sequence that is common to many proteins. Unknown computing words are almost unique to the Web corpora, occurring in similar frequencies in each of them. Foreign words are significantly more evident on the Web than in the BNC. This reflects the occasional inclusion of foreign word or phrases in predominantly English BNC texts, in contrast to the inclusion of whole collections of non-English documents on the Web. Proper nouns, although occurring almost equally in the BNC and Web corpora and ostensibly not academic terms, are an interesting case because they can be useful markers of theory. In sociology, entire theories are marked primarily by their allegiance to individual people such as Marx, Durkheim or Derrida. The existence of these names in a text might be a good word frequency indicator of its content. Yet some personal names are relatively common: there are chemists and physicists called Marx, and there must be Smiths in every field, so personal names will not always be good indicators of text content. Geographic names are also ostensibly not academic terms, except perhaps in the context of regional studies (e.g., American Studies, European Studies). In subjects including Environmental Sciences and Geography, particular place names can be synonymous with theories and phenomena around which subfields are based. Examples are the names of locations in which important archaeological artifacts/unusual geographic phenomena/unique mineral deposits/unique cultural practices are found.

Question 4 The tables of high total and relative frequency words showed clear patterns. First, the name and abbreviations of universities and departments occur frequently in their Web sites. Second, generic Web-related words (e.g., url, homepage, Website) also occur frequently. Third, the names of specific Web programs for providing online information (e.g., ProQuest) are common because they generate large numbers of pages either containing credit lines or links to other related pages created by the software. Fourth, computing related terms were evident, some directly connected to the Web. Perhaps the strangest high frequency word was 'workaround', used in a highly specific computing context, but not on the face of it an inherently technical word. It is infrequently used away from the Web. The results also show that New Zealand university Web sites contain a considerable body of work in Maori. In the U.K., the minority language of Welsh also appears frequently in Welsh university Web sites, but not frequently enough to appear in the tables. This underlines the multilingual nature of university Web sites.

It seems likely that much of the above discussion will apply to the other two large developed English speaking countries: Canada and the U.S.A. Probably in both cases minority languages (French, Spanish) will make a bigger impact than in Australia and the U.K. In mainland Europe, the extensive use of English in academic Webs (Thelwall, Tang & Price, 2003) makes them bilingual or multilingual, and an analysis of the university Web sites of any mainland European country would need to separate out the pages written in different languages in order to get useful results. Different techniques would also be needed for countries with non-ASCII languages. For the rest of the world (e.g., Africa, South America and Central America) multilingualism is likely to be common and Web sites sizes often smaller. Future scientific Web intelligence research will need to take language factors into account, except perhaps for the two relatively monolingual countries of the U.K. and Australia.

Implications for Academic Web Clustering

Question 1 Because of the large proportion of pages with too little text to be clustered (e.g., less than 10 words), algorithms may need to be designed to avoid pages with low word counts, perhaps filtering them out as part of data cleansing. This may not be necessary for domain clustering, because small pages may be combined with others.

Questions 2 & 3 The cosine similarity measure relies upon similar documents tending to use similar words, but documents discussing the same topic can use different words (e.g., synonyms). There are sophisticated techniques to avoid this problem (Deerwester, Dumais, Furnas *et al*., 1990; Santamaria, Gonzalo & Verdejo, 2003), but these do not correct error words. Two approaches for dealing with errors are to use automatic spelling correction (e.g., Golding, & Roth, 1999), or to remove all low frequency words from the vocabulary. The results above show, however, that most low frequency words are not errors and so their removal risks loosing information. Automatic spelling correction will be time-consuming and perhaps difficult, given the number of non-words that are not spelling mistakes. Error words do not have to be a critical problem, however, because low frequency words could be expected to occur at random and so, statistically, should not have a big effect on clustering.

Specialist academic words, evident in all corpora, are not necessarily useful for clustering, especially if they are rare even within the discipline. For example, in chemistry every molecule has its own unique name, and simple word frequencies may not be sufficient to identify two groups of researchers, because even if the researchers are studying similar molecules, the molecules will still have different names. It is not clear whether academic words will help clustering.

Proper nouns are an interesting case because of their potential to yield useful topic-related information, as discussed above. Manual filtering may needed exclude proper nouns that cause problems, such as the names of two authors operating in different fields, or geographic names that are frequently used throughout universities in a region (e.g., 'London'). Alternatively, a blanket approach may need to be adopted: banning all words identified as proper nouns, identifying them through capitalization.

In the Web corpora there are document types with many low frequency words. These include foreign language texts, computer program listings, specialist academic pages, and biological compound information pages. Clustering such pages would be inherently difficult because, using the cosine measure, they would appear distant from all other documents. A possible solution is to remove pages with many low frequency words from the corpus before clustering. An alternative approach is to separate out the different types of low frequency word pages, and apply different rules to them.

Foreign language pages produce pressure to cluster by language in addition to more frequently containing low frequency words that do not help clustering. This would be acceptable if foreign language pages were found predominantly in departments teaching the language, but this is probably not the case: translations are also present in academic Webs. Hence, it would probably help clustering to remove all foreign language pages. This needs to be further investigated for Welsh and New Zealand universities, with their extensive use of minority languages.

The variable names in computer program listings are relatively arbitrary. If all computer science departments contained large collections of different variable names then from a word frequency point of view they would all contain many unique and low frequency words and so the cosine measure would assess them as different to each other and everything else. Biological compound information pages should be excluded, when possible, both because of the problem caused by the many low frequency non-words, but also because of the occurrence of some high frequency non-words which would tend to make the pages similar only to other biological compound pages with the same common subsequence.

Question 4 The existence of high frequency words that are peripheral to a document's meaning, such as Web page creation software names, is a potential problem. These would not normally be a obstacle to clustering Web pages, since they would tend to occur one per page, but if clustering whole domains then the combined individual page word frequencies could cause a problem. The frequent occurrence of university and department names will tend to make pages from the same university appear similar with the cosine measure and is a problem for clustering. There is a case, therefore, when using domain clustering to manually identify unwanted high frequency words. This may be possible to automate, perhaps by choosing words that are high frequency in the Web corpus but not in the BNC. An alternative blanket approach would be to adopt a standard information retrieval technique and use a heuristic such as discarding the top third and bottom third of words based upon frequency in the corpus (c.f., Luhn, 1958; Salton & McGill, 1983). A disadvantage of this is that frequency anomalies can occur in the remaining middle third of words. Note that not all high frequency words are necessarily an issue. For example, generic Web-related words should not be a problem for clustering if they occur uniformly across sites.

Conclusions

We have provided background information about word use in three English language academic Webs in addition to specific data cleansing suggestions for document clustering. Four general conclusions can now be drawn that are useful foundations for scientific Web intelligence. First, the data confirmed the tendency for statistics obtained from academic Web spaces to obey natural laws, but with exceptions. Based upon this, we conjecture that s*tatistics obtained by counting any identifiable components of Web pages in national systems of university Web sites will display simple mathematical patterns, but with anomalies.* It will be interesting to see if there are quantifiable aspects of academic Web pages that do not obey the conjecture. Even if there are, the knowledge that strong simple patterns, but with anomalies, are likely for any kind of data in academic Web sites forms a useful starting point for the design of any new Web mining algorithm. In particular, those who ignore the existence of anomalies, and do not compensate for them (e.g., by filtering them out), are not likely to get good results. Second, the Web sites of English speaking countries are not exclusively written in English, but contain pages in a wide variety of different languages. This is another complicating factor for any algorithm based upon word frequencies. Third, many high frequency words, even proper nouns, can be poor indicators of content. University and departmental names are possible sources of problems for text-based topic clustering. Logical solutions are to filter out manually identified problematic names or to automatically filter out all high frequency words. Fourth, many pages are empty of text, or almost empty, and these may need to be taken excluded from analyses.

References

Aguillo, I. F. (1998). STM information on the Web and the development of new Internet R&D databases and indicators. Online Information 98: Proceedings, 239-243.

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

Aston, G., & Burnard, L. (1998). The BNC Handbook. Edinburgh: Edinburgh University Press.

Atkinson, K. (2003). Spell Checking Oriented Word Lists (SCOWL), revision 5, January 3, 2003. Retrieved September 3, 2003, from http://wordlist.sourceforge.net/

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern Information Retrieval. Wokingham, UK: Addison-Wesley.

Bailey, P., Craswell, N., & Hawking, D. (2003). Engineering a multi-purpose test collection for Web retrieval experiments. Information Processing & Management, 39(6), 853-871.

Baldi, P., Frasconi, P., & Smyth, P. (2003). Modelling the Internet and the Web. Chichester, UK: Wiley.

Barabási, A.L. & Albert, R. (1999). Emergence of scaling in random networks. Science, 286, 509-512.

Blair, I.V., Urland, G.R., & Ma, J.E. (2002). Using Internet search engines to estimate word frequency. Behavior Research Methods, Instruments, & Computers, 34(2), 286-90

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

Börner, K., Chen, C., & Boyack, K. (2003). Visualizing knowledge domains. *Annual Review of Information Science & Technology*, 37, 179-255.

Boynton, J., Glanville, J., McDaid, D., & Lefebvre, C. (1998). Identifying systematic reviews in MEDLINE: developing an objective approach to search strategy design. Journal of Information Science, 24(3), 137-154.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. & Wiener, J. (2000). Graph structure in the Web. Computer Networks, 33(1-6), 309-320.

Burnard, L. (Ed.) (1995). Users' reference guide to the British National Corpus. Oxford: Oxford University Computing Services.

Cothey, V. (2005, to appear). Web-crawling reliability. Journal of the American Society for Information Science and Technology.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407.

Ghani, R., Jones, R, & Mladenić, D. (2001). Mining the web to create minority language corpora. Proceedings of the tenth international conference on Information and knowledge management, 279-286.

Golding, A. R., & Roth, D. (1999). A winnow-based approach to context-sensitive spelling correction. Machine Learning, 34(1-3), 107-130.

Heimeriks, G., Hörlesberger, M., & van den Besselaar, P. (2003). Mapping communication and collaboration in heterogeneous research networks. Scientometrics, 58(2), 391-413.

Heyer, G., Quasthoff, U., & Wolff, C. (2002). Automatic analysis of large text corpora - A contribution to structuring WEB communities. Lecture Notes in Computer Science, 2346, 15-26.

Jain, A., Murty, & Flynn, (1999). Data clustering: a review. ACM Computing Surveys, 31(3), 264-323.

Keller, F., & Lapata, M. (2003). Using the Web to obtain frequencies for unseen bigrams. Computational Linguistics, 29 (3), 459-484.

Kilgarriff, A. (2003). BNC database and word frequency lists. Retrieved December 2, 2003, from http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/bnc-readme.html

Leydesdorff, L., & Curran, M. (2000). Mapping university-industry-government relations on the Internet: the construction of indicators for a knowledge-based economy. Cybermetrics, 4(1). Retrieved November 20, 2003, from http://www.cindoc.csic.es/cybermetrics/articles/v4i1p2.html

Li, W. (1992). Random texts exhibit Zipf's-Law-like word frequency distribution. IEEE Transactions on Information Theory, 38(6), 1842-1845.

Lindsay, R. K., & Gordon, M. D. (1999). Literature-based discovery by lexical statistics. Journal of the American Society for Information Science, 50(7), 574-587.

Luhn, H.P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2, 159-165.

Hand, P., Mannila, D., & Smyth, H. (2001). Principles of data mining. Boston: MA: MIT Press.

McEnery, T. & Wilson, A. (2001). Corpus linguistics. Edinburgh: Edinburgh University Press.

Menasalvas, E., Segovia, J., & Szczepaniak, P.S. (Eds.) (2003). Advances in Web Intelligence. Berlin: Springer.

Middleton, I., McConnell, M. & Davidson, G. (1999). Presenting a model for the structure and content of a university World Wide Web site. Journal of Information Science, 25(3), 219-227.

Miller, G.A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41.

Mitkov, R. (Ed.) (2003). The Oxford handbook of computational linguistics. Oxford, UK: Oxford University Press.

Neuendorf, K. (2002). The content analysis guidebook. California: Sage.

Oakes, M. (1998). Statistics for corpus linguistics. Edinburgh: Edinburgh University Press.

Pennock, D., Flake, G., Lawrence, S., Glover, E., & Giles, C.L. (2002). Winners don't take all: Characterizing the competition for links on the Web. Proceedings of the National Academy of Sciences, 99(8), 5207-5211.

Porter, M. (1980). An algorithm for suffix stripping. Program, 14 (3), 130-137.

Poudat, C. & Cleuziou, G. (2003). Genre and domain processing in an Information Retrieval perspective. Lecture Notes in Computer Science, 2722, 399-402.

Resnik, P., & Smith, N.A. (2003). The Web as a parallel corpus. Computational Linguistics, 29(3), 349-380.

Rousseau, R. (1997). Sitations: An exploratory study. Cybermetrics, 1(1). Retrieved October 16, 2003, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Rousseau, R. (1999). Daily times series of common single word searches in AltaVista and NorthernLight. Cybermetrics, 2-3. Retrieved October 16, 2003, from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

Salton, G. & McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Sanderson, M., & Van Rijsbergen, C. J. (1999). No good? The impact on retrieval effectiveness of skewed frequency distributions. ACM Transactions on Information Systems, 17(4), 440-465.

Santamaria, C., Gonzalo, J., & Verdejo, F. (2003). Automatic association of Web directories with word senses. Computational Linguistics, 29(3), 485-502.

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

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

Thelwall, M. & Wilkinson, D. (2003). Graph structure in three national academic Webs: Power laws with anomalies. Journal of the American Society for Information Science and Technology, 54(8), 706-712.

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

Thelwall, M. (2002). 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. (2003). A free database of university Web links: Data collection issues. Cybermetrics. Retrieved October 16, 2003, from http://www.cindoc.csic.es/cybermetrics/articles/v6i1p2.html.

Thelwall, M. (2004, to appear). Scientific Web Intelligence: Finding relationships in university webs. *Communications of the ACM.*

Thelwall, M. (2005, to appear). Data cleansing and validation for Multiple Site Link Structure Analysis, In: Scime, A. (Ed.), Web Mining: Applications and Techniques. Hershey, PA: Idea Group.

Vilensky, B. (1998). Can analysis of word frequency distinguish between writings of different authors? Physica-A, 231(4), 705-711.

Weeber, M., Vos, R., & Baayen, R.H. (2000). Extracting the lowest-frequency words: pitfalls and possibilities. Computational Linguistics, 26(3), 301-17.

Zipf, G.K. (1949). Human behavior and the principle of least effort: an introduction to human ecology. Cambridge, MA: Addison-Wesley.