Available online at www.sciencedirect.com
SCIENCE DIRECT®
Information Processing and Management 42 (2006) 1471–1483
INFORMATION PROCESSING & MANAGEMENT
www.elsevier.com/locate/infoproman

Automated Web issue analysis: A nurse prescribing case study

Mike Thelwall [a,*], Saheeda Thelwall [b,1], Ruth Fairclough [a,2]

[a] School of Computing and Information Technology, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK
[b] School of Health, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1SB, UK

Received 16 March 2006; accepted 16 March 2006
Available online 16 May 2006

Abstract

Web issue analysis, a new automated technique designed to rapidly give timely management intelligence about a topic from an automated large-scale analysis of relevant pages from the Web, is introduced and demonstrated. The technique includes hyperlink and URL analysis to identify common direct and indirect sources of Web information. In addition, text analysis through natural language processing techniques is used identify relevant common nouns and noun phrases. A case study approach is taken, applying Web issue analysis to the topic of nurse prescribing. The results are presented in descriptive form and a qualitative analysis is used to argue that new information has been found. The nurse prescribing results demonstrate interesting new findings, such as the parochial nature of the topic in the UK, an apparent absence of similar concepts internationally, at least in the English-speaking world, and a significant concern with mental health issues. These demonstrate that automated Web issue analysis is capable of quickly delivering new insights into a problem. General limitations are that the success of Web issue analysis is dependant upon the particular topic chosen and the ability to find a phrase that accurately captures the topic and is not used in other contexts, as well as being language-specific.

Keywords: Web; Automated Web issue analysis; Link analysis; Nurse prescribing; Medical informatics

1. Introduction

Healthcare information and healthcare initiatives typically need to be communicated to large professional bodies such as doctors, nurses and health managers. Health-related information can be produced by a wide variety of people including academics, doctors, government spokespersons, and in some cases, non-medical people. The Web is a popular publication medium for a wide variety of health information (Zeng et al., 2004), of varying quality and accuracy (Bernstam, Shelton, Walji, & Meric-Bernstam, 2005), and is being

* Corresponding author. Tel.: +44 1902 321470; fax: +44 1902 321478.
E-mail addresses: m.thelwall@wlv.ac.uk (M. Thelwall), s.thelwall@wlv.ac.uk (S. Thelwall), r.fairclough@wlv.ac.uk (R. Fairclough).
[1] Tel.: +44 1902 328713; fax: +44 1902 321478.
[2] Tel.: +44 1902 321000; fax: +44 1902 321478.

0306-4573/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2006.03.011

increasingly seen as central to information provision within the health services (Murphy et al., 2004), including in the role of keeping practitioners up to date with current guidelines. The Web also seems to be a vehicle for an increased internationalisation of medical education (Hovenga, 2004). For those responsible for any aspect of healthcare information, Web publishing is a problem because of the conflicting messages it can give (Burd, Chiu, & McNaught, 2004), and hence there is a need to gain insights into what healthcare information is published for any given topic in order to decide how to respond to it. Other researchers have tackled the problem of variable quality Internet information by evaluating metrics for predicting health Web site quality (Currò et al., 2004; Hernández-Borges et al., 2003). This is useful from the perspective of deciding which sites to use or recommend, but does not help managers identify and respond to unwanted information, particularly when it comes from an unexpected source, such as a medical article in an online newspaper.

Previous researchers have developed a variety of methods designed to identify aspects of online communities or topics, although these have tended to either rely upon simple link analyses (Garrido & Halavais, 2003; Park, 2003; Tang & Thelwall, 2003) or to be very labour intensive (Foot, Schneider, Dougherty, Xenos, & Larsen, 2003; Weare & Lin, 2000). In computer science, various forms of Web mining have been developed to extract information from Web pages or log files (Chakrabarti, 2003; Kosala & Blockeel, 2000), but these have typically not been designed to be applied to wider social issues, with the closest perhaps being community identification (Flake, Lawrence, Giles, & Coetzee, 2000) and topic clustering (Chakrabarti, Joshi, Punera, & Pennock, 2002). Topic identification and tracking is also a recognised task within computer science and computational linguistics with online variants following a long tradition of offline research, primarily through the TREC conferences (e.g., Chakrabarti, VanDen Berg, & Dom, 1999; e.g., Clifton, Cooley, & Rennie, 2004; Ozmutlu & Cavdur, 2005). This task is more narrowly focussed than issue analysis (as described below), however, with a typical application being the identification and categorisation of news stories. Issue tracking, the task of identifying the scope of a broad social issue and tracking it, has a pedigree from before the Web as a specific social science task, triggered by the pioneering study of Lancaster and Lee (1985), who tracked research related to acid rain over time in several databases. A more recent example is Wormell's (2000) analysis of topics related to the Danish welfare state, a study that was able to take advantage of the availability of multiple different sources of electronic information. In bibliometrics, the mapping of papers or authors in an attempt to describe areas of science is an established practice (e.g., Leydesdorff, 1989; Small, 1973; White & Griffith, 1982). In this paper we apply Web issue analysis (Thelwall, Vann, & Fairclough, in press) to systematically identify all issues relevant to any selected health topic, at least those issues that are reflected on the Web. In essence, the method starts with one or more topic descriptions, such as 'nurse prescribing', and downloads all Web pages (via Google) that allude to the topic. These Web pages are then used for a range of types of link analysis. The pages are then processed to extract their noun phrases and a frequency table is produced giving the number of sites containing the noun or noun phrase. Nouns and noun phrases are much better indicators of topic discussed in a document than individual words since they can be complete concept representations. Site frequencies are reasonable indicators of the popularity of topics and are better than raw frequency counts or page based frequency counts because Web sites are often highly repetitive, duplicating content in many or all site pages (Thelwall, 2002), which is made easy by database driven Web site technology (Dørup, Hansen, Ribe, & Larsen, 2002). In Web issue analysis, the set of nouns and noun phrases extracted from topic-relevant pages are the candidate *topic-relevant issues*. The site frequency counts of noun phrases are suggestive indicators of their *topic-relevant popularity*. The table of topic-relevant issues and popularities is described as the *Web environment* of the topic in the belief that researchers and information managers can gain useful topic-relevant insights from its Web environment.

In this paper, Web issue analysis is applied to a specific case study to demonstrate its capabilities for providing management information in a national context. The medical field chosen is nurse prescribing in the UK. The objective of the case study is to investigate whether an automated Web issue analysis can produce useful information about the context of Web publishing for nurse prescribing.

2. Nurse prescribing background

In the UK, recent years have seen a Department of Health initiative to train a proportion of nurses to prescribe a range of medicines. Legislation was passed in 1992 to give prescriptive powers to district nurses and

health visitors so that they could legally prescribe from a restricted formulary (the Nurse Prescribers' Formulary). The government announced in May 2001 that prescriptive authority would be extended to additional nurse roles within both primary and secondary care. Nurses can prescribe both as 'extended' (the Extended Nurse Prescribers' Formulary) and 'supplementary' prescribers (the whole of the British National Formulary when they enter into a voluntary partnership with an independent prescriber). The aims of extending nurse prescribing were to provide patients with quicker access to medicine and enable nurses to use their skills appropriately. This initiative has, so far, been well received by staff and patients, having mainly positive effects, but since it is in its early stages, policy makers need to keep a close watch on how it is developing and identify any new issues that may arise (Latter & Courtenay, 2004). Nurse prescribing is a health issue that is particularly well suited to an Internet analysis because the dispersed and relatively isolated nature of practitioners, combined with the need for ongoing support for prescribers, makes the Web a natural tool for the provision of information during and after the initial formal training period (Smith, 2004).

The practice of nurse prescribing is an international one, with the US being prominent (Hales & Dignam, 2002). The introduction of prescriptive authority around the world has taken time to develop and establish changes. In the US only advanced practice nurses (i.e. registered nurses with advanced knowledge and skills) are allowed to prescribe. In all 50 states there are varying levels of prescriptive authority, requirements, standards and practice (Phillips, 2005; Ploncynski, Oldenburg, & Buck, 2003), in contrast to the more uniform approach in the UK (Mullay, Mason, & Frogatt, 2003). In Sweden, nurse prescribing has met with severe resistance from the general practitioners and nurses have received little support (Willhelmsson & Foldevi, 2003). New Zealand has had legislation in place to allow prescriptive authority for both nurses and health care professionals since 1998, and this has taken time to develop and establish, with a focus on international developments in USA, UK and Sweden examining what lessons could be learnt and how issues could be best addressed. The approach was taken to build strong relationships with stakeholders and flexible policy and legal arrangements that can respond to change (Hughes & Lockyer, 2004).

3. Methods

3.1. Design of the study

The study is designed to produce three different types of information about nurse prescribing from HTML Web pages.

1. URLs of Web pages containing the phrase 'nurse prescribing' (henceforth: 'nurse prescribing pages').
2. URLs of pages linked from by the above pages (outlinks).
3. Noun phrases in nurse prescribing pages.

The motivating belief for collecting these three types of information is that

1. URLs may give useful information about the types and geographic locations of organisations publishing nurse prescribing-related information online.
2. URLs of links in nurse prescribing pages may indicate where nurse prescribing theory or practice is drawn from, by analogy with citations.
3. Nouns and noun phrases in nurse prescribing pages may indicate to the topics that are most relevant to nurse prescribing.

Descriptive approaches are used to summarise aspects of each of the above three types of data.

3.2. Data processing

The data collection and processing stages are illustrated in Fig. 1 and are described in detail below.

Collecting URLs of Web pages containing "nurse prescribing": The Google API was used to obtain from Google Web pages containing the phrase "nurse prescribing". We used Google searches of international

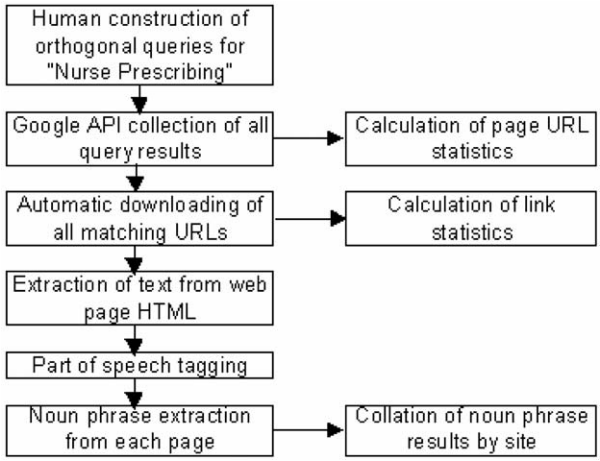

Fig. 1. The sequence of operations to obtain the text and link data.

nursing sites, particularly in the US, to look for other English phrases describing the same concept but were not able to find any, so all of the results are based upon the single phrase "nurse prescribing". The Google Applications Programming Interface, or API (Google, 2005), is a software tool that can be used to automatically send 1000 queries to Google per day. Each query returns up to 10 matches. It is possible to request up to 100 pages of 10 results for a single query, giving a total of 1000 matches. There were more than 1000 matches for "nurse prescribing" so a series of queries was needed so that each query gave less than 1000 results, and the sum total of all the queries would be all Google's pages containing "nurse prescribing". This was achieved by identifying a set of 8 "approximately orthogonal" words. This is a set of words, each of which occurs in approximately half of all pages containing "nurse prescribing", and the words are orthogonal in the sense that any two of the words would split the set of pages into four approximately equal quarters. A query was constructed for each possible combination of inclusion and exclusion of words. These queries were submitted to Google, via its API, and the URLs of all matching pages saved to a file. Google's logic is not perfect (Bar-Ilan, 2004; Mettrop & Nieuwenhuysen, 2001; Rousseau, 1999) and some duplicate URLs were produced by this method. These were automatically identified and removed.

Page URL distribution statistics: In order to see which types of sites contain the phrase "nurse prescribing", the set of URLS reported by Google was processed and summarised by domain name. The domain name of an URL is normally the portion after the initial http:// and before the first subsequent slash. In most cases the collection of URLs sharing a common domain name form a coherent 'site', although there may be any number of pages in a 'site'. A different heuristic for identifying sites may also be used: including all pages with domain names with the same ending (e.g. the 'site' wlv.ac.uk would also include www.wlv.ac.uk and www.scit.wlv.ac.uk). There are some exceptions, such as www.geocities.com, which contains over a million individual sites. Nevertheless, equating domain names with sites seems a reasonable approximation for the URL data and this kind of approach is in common use in commercial Web server log file analysis software. Nevertheless, this is an assumption that needs to be assessed in practice by the relevance of the results produced. A ranked list of domains and the number of URLs associated with each domain was calculated to give an illustration of the most productive sites for nurse prescribing.

Domain names can usefully be assigned to generic site types in some cases, mainly based upon the structure of the domain name. The *top-level domain* (TLD) of a domain name, normally the segment of letters following the final dot, exists in the three main varieties.

- National TLDs are normally administered by a country and intended to signify affiliation with that country. In practice, however, there are exceptions, such as .tv being widely used for television, and .fr being used to signify predominantly French-language sites.

- Specific TLDs (e.g., edu, mil, gov) have their use restricted, again with some exceptions, to a specific type of user (e.g., US education, military and government, respectively).
- Generic TLDs (e.g. com, org, info) are widely used for many purposes (despite their initially prescribed remit).

A ranked list of TLDs provides some evidence of the origins of the pages in the data set but the usage exceptions discussed above and the generic TLDs' unknown purposes combine to make the ranked list suggestive of page distribution rather than definitive.

Some country codes are subdivided by second-level domain, effectively creating both specific and generic second level domains. The second level domains can give useful information about the origins of the pages in some cases. For example, nhs.uk pages are from the UK National Health Service (NHS) whereas .co.uk pages tend to be UK companies, although, like .com, this ending is widely used. Note that most European countries (e.g., France) do not use a second level domain naming system. The terminology STLD is used to describe sites grouped by second-level domain where such a convention exists, and otherwise grouped by TLD. A ranked list of STLDs provides a more fine-grained description of site origins than a TLD ranked list and is particularly useful when a significant number of URLs originate in countries using the second-level naming system.

Identifying Web pages containing "nurse prescribing": The file of URLs reported by Google as containing "nurse prescribing" was filtered to remove all non-HTML Web pages, and the remainder downloaded using the program SocSciBot (Thelwall, 2004).

Outlink statistics: All URLs returned by the Google searches were downloaded. The hyperlinks in each page were extracted. For each page, all links to other pages within the same site were discarded because these site self-links (Björneborn & Ingwersen, 2004) tend to be for navigational purposes and are less significant than links to other sites, which presumably tend to be more deliberately chosen (Smith, 1999). The remaining 'site outlinks' (Björneborn & Ingwersen, 2004) seem likely to indicate pages that the authors of the nurse prescribing pages thought relevant to their topic. Presumably the network formed by the interlinking pages centres on nurse prescribing and its structure will cast some light on the nurse prescribing Web environment. Presumably also, the most frequently linked to pages from this set tend to be most useful to the nurse prescribing topic, and so it would be useful to identify the most frequent link URLs. By extension, the most frequently linked to domains, TLDs and STLDs may give information about the distribution of links.

Text extraction and collation: A program was written to process each downloaded Web page and extract its text, discarding its HTML tags. This produced a set of files of plain text, one for each Web page. These files were then collated by site, so that the text of all pages within a single site was stored in a single site-based file. Sites were identified by domain name, using the second level domain name or third-level domain as appropriate. For example, in the .edu domain, the second level domain identifies university Web sites (e.g., washington.edu) whereas for .uk the third level domain identifies the site (e.g., oxford.ac.uk).

Part of speech tagging: The free program Lingua::En::Tagger 0.11 (Coburn, 2005) was used to tag each individual text file for parts of speech. This involves automatically assessing each word and predicting its correct part of speech, using a dictionary and a mathematical model based upon the word and surrounding text (Mitkov, 2003).

Noun and noun phrase extraction: Lingua::En::Tagger 0.11 (Coburn, 2005) was used to extract (a) a list of nouns and (b) a maximal list of noun phrases from each tagged file. Noun phrases are extracts from sentences that contain nouns and are judged by an artificial intelligence algorithm to form a coherent phrase. Nouns and noun phrases were both listed in a separate file for each Web page. Nouns are not necessarily single words, but could be consecutive words describing a single object, for example Public Health. The noun phrase list was calculated maximally in the sense of (i) identifying the longest possible phrases containing each noun and then (ii) for each identified noun phrase adding any shorter noun phrases contained in the longer phrase.

Noun and noun phrase collation: A single list was created of all noun phrases, recording the number of sites in which they appeared.

4. Results

The Google API searches returned 6772 URLs from 1619 domains. After downloading these URLs and excluding errors and non-HTML pages, there were a total of 1217 Web sites containing some text, although the smallest contained only a few words.

4.1. Page URL statistics

The results show that a clear majority of pages containing the phrase "nurse prescribing" come from the UK domain. Presumably most of the com and org results are also UK sites. Notable for its virtual absence is the edu domain, indicating that nurse prescribing is not an academic topic in the US, either because the concept is not widely recognised or because other words are used to describe this activity. The former is more likely: the key concept in the US seems to be that of the advanced nurse practitioner, with prescribing not being a core part of the role since some states have not allowed it (Rosseter, 2000). Moreover, there did not seem to be an equivalent phrase in the US Web documents that discussed prescribing for nurses. In this context, however, the small but significant percentage of US government gov pages is interesting. Even though nurse prescribing is an English phrase, the absence of significant pages in non-English nations points to an absence of the concept from the rest of European academic Web at least, unless an alternative phrase is used, since Web publishing in English is widespread in European universities (about 50% of pages are English) (Thelwall, Tang, & Price, 2003).

The STLD results are useful for the subdivision of UK pages, surprisingly showing a dominance of commercial domain sites. Tables 1 and 2 point to a cause: Web sites of journals and magazines. This also seems to account for the com domain's prominence. The US government .gov domain is an apparent anomaly as a significant non-UK source. An investigation into the source of URLs from this domain found them all to come from PubMed, (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed) being mainly online journal articles or abstracts written by British authors. Hence the .gov results should not be associated with US origins.

From a methods perspective, note that the third column of Table 1 is similar to Table 2, which is almost redundant (see also Tables 3 and 4) indicating that for this issue multiple domain names for a single site are

Table 1
Most common TLDs, STLDs and sites for pages containing "nurse prescribing"

Pages	%	Source TLD	Pages	%	Source STLD	Pages	%	Source site
3573	61	uk	1175	20	co.uk	233	4	nurse-prescriber.co.uk
1161	20	com	1161	20	com	198	3	nih.gov
307	5	org	905	16	ac.uk	176	3	publications.parliament.uk
200	3	gov	503	9	nhs.uk	88	2	scot.nhs.uk
113	2	net	499	9	org.uk	87	1	rdlearning.org.uk
97	2	nz	307	5	org	86	1	nephronline.org
44	1	ca	292	5	gov.uk	83	1	dh.gov.uk
44	1	ie	200	3	gov	79	1	internurse.com
30	1	info	185	3	parliament.uk	74	1	nmap.ac.uk
28	0	au	113	2	net	72	1	scotland.gov.uk
26	0	pl	44	1	ca	67	1	pjonline.com
25	0	edu	44	1	ie	66	1	pharmj.com
23	0	I.P.	39	1	co.nz	58	1	findarticles.com
23	0	de	30	1	info	56	1	wales.nhs.uk
15	0	kr	26	0	pl	55	1	best-book-price.co.uk
14	0	ro	25	0	edu	52	1	luton.ac.uk
13	0	biz	23	0	I.P.	50	1	bmjjournals.com
13	0	ch	23	0	de	46	1	sandwell.gov.uk
13	0	us	22	0	org.nz	45	1	the-stationery-office.co.uk
12	0	jp	20	0	govt.nz	44	1	ppa.org.uk

Table 2
Most common domains for pages containing "nurse prescribing"

Count	%	Domain
233	4	http://www.nurse-prescriber.co.uk
198	3	http://www.ncbi.nlm.nih.gov
176	3	http://www.publications.parliament.uk
87	1	http://www.rdlearning.org.uk
86	1	http://www.nephronline.org
83	1	http://www.dh.gov.uk
78	1	http://www.internurse.com
74	1	http://nmap.ac.uk
72	1	http://www.scotland.gov.uk
67	1	http://www.pjonline.com
66	1	http://www.pharmj.com
64	1	http://www.show.scot.nhs.uk
58	1	http://www.findarticles.com
55	1	http://www.best-book-price.co.uk
52	1	http://www.luton.ac.uk
52	1	http://www.wales.nhs.uk
46	1	http://www.webwell.sandwell.gov.uk
45	1	http://www.parliament.the-stationery-office.co.uk
44	1	http://www.ppa.org.uk
43	1	http://www.pharmafocus.com

Table 3
Most linked-to TLDs from pages containing "nurse prescribing"

Inlink count	%	Target TLD	Inlink count	%	Target STLD	Inlink count	%	Target site
25,477	51	uk	18,837	29	com	1222	5	www.parliament.uk
18,837	29	com	7176	15	co.uk	1889	4	doh.gov.uk
3890	6	org	4966	11	ac.uk	1431	1	dh.gov.uk
1369	4	net	6008	9	gov.uk	192	1	w3.org
1264	2	de	2736	7	nhs.uk	442	1	amazon.com
227	1	nz	3890	6	org	470	1	amazon.co.uk
223	1	tv	1323	5	parliament.uk	265	1	adtech.de
1188	1	edu	1369	4	net	221	1	parliamentlive.tv
339	1	au	3203	4	org.uk	68	1	adobe.com
3197	1	gov	1264	2	de	128	1	ingenta.com
163	0	ch	223	1	tv	597	1	bmjjournals.com
177	0	I.P.	1188	1	edu	180	1	www.nhs.uk
240	0	info	3197	1	gov	149	1	nursingtimes.net
137	0	jp	163	0	ch	249	1	manchester.ac.uk
58	0	es	177	0	I.P.	93	1	inpharm.com
444	0	ca	42	0	govt.nz	770	1	scot.nhs.uk
55	0	no	240	0	info	136	1	doubleclick.net
151	0	ie	107	0	com.au	839	1	bbc.co.uk
38	0	us	72	0	co.nz	107	1	nhsdirect.nhs.uk
55	0	ru	59	0	ac.nz	24	1	celcat.com

not a significant factor. This would probably not have been true for a more purely academic topic, for example, because many universities make extensive use of multiple domain names (Thelwall, 2002).

4.2. Link URL statistics

Link URLs continue the theme of UK dominance of the phrase "nurse prescribing", but with a slightly increased internationalism (Table 1). By analogy with citations (Borgman & Furner, 2002), which have been theorised as tending to reflect the transfer or use of prior knowledge (Merton, 1973), the distribution of link

Table 4
Most linked-to domains from pages containing "nurse prescribing"

Inlink count	%	Target domain
1222	5	www.parliament.uk
1346	4	www.doh.gov.uk
1431	1	www.dh.gov.uk
469	1	www.amazon.co.uk
265	1	adserver.adtech.de
375	1	www.amazon.com
221	1	www.parliamentlive.tv
68	1	www.adobe.com
128	1	www.ingenta.com
180	1	www.nhs.uk
121	1	www.nursingtimes.net
93	1	www.inpharm.com
107	1	www.nhsdirect.nhs.uk
24	1	www.celcat.com
24	1	www.venuesearch.co.uk
136	1	www.mahealthcareevents.co.uk
195	1	www.manchester.ac.uk
40	1	www.virco-hosting.com
85	1	jigsaw.w3.org
65	1	ad.uk.doubleclick.net

URLs may reflect the countries or domains that influence nurse prescribing practice. Alternatively, they may also reflect, at least in part, the location of resources related to nurse prescribing, or the wider social or organisational connections of those that discuss nurse prescribing (Harries, Wilkinson, Price, Fairclough, & Thelwall, 2004; Wilkinson, Harries, Thelwall, & Price, 2003). This is consistent with the hypothesis that nurse prescribing is a very UK-centred topic in all regards, at least on the Web, assuming the generic TLDs to be UK-dominated because of the absence of a strong showing for any non-UK national TLD. It would be interesting to see if this is reflected in scholarly literature citations. The STLD results echo the page URL co.uk dominance within the UK, but the most linked to sites and domains (Table 4) emphasise the importance of the UK government for nurse prescribing, unusual for an academic topic (Stuart & Thelwall, 2005). Of note is the relatively low showing of the natural 'owner' of the nurse prescribing topic, nhs.uk, perhaps because much internet-type activity occurs through the huge virtual private network NHSnet (NHSnet, 2005), which is not part of the Internet. The NHS is the UK's government-run national health service system, with most of its services being free or subsidised (www.nhs.uk/England/AboutTheNhs/Default.cmsx). The minor role of nhs.uk perhaps also reflects the fact that nurse prescribing is primarily an internal NHS matter, and so there is little need to publish information about it to the public.

Table 5 includes some general Web resources, such as Adobe Acrobat, as highly linked to pages, but it is difficult to effectively interpret this table because some of the URLs are present as a result of repeated links across the pages of a single site. This is the case for the parliament.uk links, for example.

The overall results (Tables 1–5) are consistent with nurse prescribing being driven by UK journals and professional magazines, strongly supported by UK universities and drawing upon these two sources and the UK government. Web publishing is natural in academia, in contrast, and the legal implications and regulations for nurse prescribing are a natural reason for government publishing on the topic.

4.3. Interlinking

Fig. 2 shows the interlinking of the top 50 sites using as raw data the links from nurse prescribing pages. The link structure is relatively sparse despite the connectivity of the four relatively peripheral sites (google, msn, doubleclick, adobe). This suggests that the practice of linking is relatively unimportant to nurse prescribing on the Web and that nurse prescribing could not be considered a strongly nationally structured online topic.

Table 5
Most linked-to URLs from pages containing "nurse prescribing"

Inlink count	%	Target URL
91	1	http://www.doh.gov.uk/nurseprescribing/pomlist.htm
231	1	http://www.parliament.uk
221	1	http://www.parliament.uk/site_information/contact_us.cfm
221	1	http://www.parliament.uk/glossary/glossary.cfm
221	1	http://www.parliamentlive.tv
221	1	http://www.parliament.uk/index.cfm
221	1	http://www.parliament.uk/index/index.cfm
37	1	http://www.ingenta.com/corporate
103	1	http://www.nhsdirect.nhs.uk
24	1	http://www.celcat.com/webpub.html
24	1	http://www.venuesearch.co.uk/healthcare_events/homepage.htm
83	1	http://www.mahealthcareevents.co.uk
80	1	http://www.quaybooks.com
83	1	http://auth.athensams.net/?ath_returl=%22http:/www.swetswise.com/public/login.do%22&ath_dspid=SWETS
79	1	http://www.mastertravel.co.uk
39	1	http://www.adobe.com/products/acrobat/readstep2.html
19	1	http://forum.snitz.com
82	0	http://www.nhs.uk
23	0	http://www.psncdatabase.co.uk
77	0	http://www.nursingtimes.net

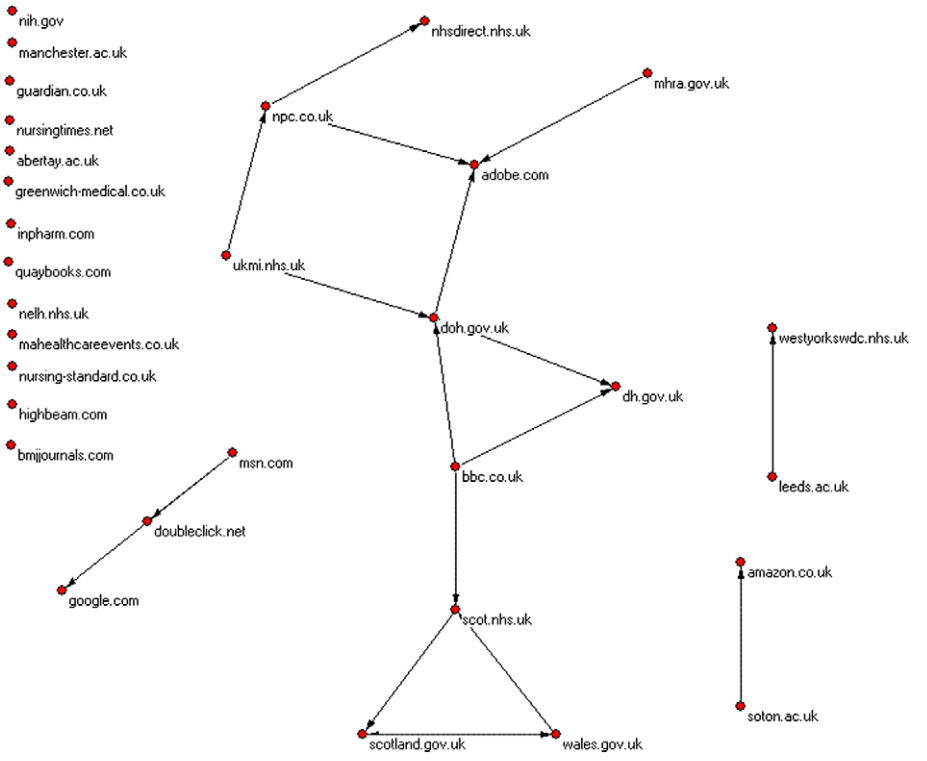

Fig. 2. Interlinking of the top 50 nurse prescribing sites, based upon their nurse prescribing pages.

4.4. Text analysis

Many of the most frequent nouns were general words, such as nurse, health and university. The top 1000 nouns were inspected, and it was found that many were not relevant: 35 related to time or date; 21 were university-related terms; 88 were geographic descriptions; and 14 were Internet-specific words. After excluding these types of noun, the left of Table 6 gives a more interesting set. The right of Table 6 gives the top 30 noun phrases, excluding single word nouns. Since these were almost all topic-relevant, an unedited list is presented. Note that multiple-word "nouns" are included as nouns, typically being sets of consecutive capitalised words. Confusingly, counts of nouns that are also noun phrases are different: this is because heuristics are used to extract both.

The noun and noun phrase tables are exploratory tools, designed to give stakeholders a broad overview of the range of topics discussed on the Web in relation to nurse prescribing. For example, from Table 6 (rank 35/rank 4) a lesson may be drawn that mental health is widely seen as important for nurse prescribing. This is probably not caused by a focus on mental health nurses, who have not been at the forefront of nurse prescribing (Nolan, Bradley, & Carr, 2004) and hence reflects important concerns of general nurses. General nurses may be able to recognise the signs and symptoms of someone suffering from mental illness and also help undiagnosed patients to access mental health services: two important reasons for them to need information about this issue since one in four people in the UK experience serious mental distress at some point in their life (Camm, 2005). In the US every state already has prescriptive authority for mental health nurses (Gournay, 2002).

Table 6
The top 30 medical-related nouns, and top 30 noun phrases most frequently identified in nurse prescribing domains

Rank	Domains	Noun	Rank	Domains	Noun phrase
1	843	Health	1	458	Nurse prescribing
2	695	Nursing	2	381	Health care
3	605	Prescribing	3	359	Primary care
4	594	Care	4	296	Mental health
13	380	nhs	5	258	Health professionals
15	365	Community	6	258	Royal college
16	363	Management	7	246	Health services
17	358	Medical	8	244	Patient care
28	315	Development	9	239	Public health
32	299	Services	10	233	Health visitors
35	290	Mental health	11	211	Department of health
37	287	Medicine	12	211	Health promotion
40	276	gp	13	209	General practice
41	276	Health care	14	207	Nursing practice
43	268	Service	15	206	Clinical practice
47	261	Midwifery	16	200	Health service
48	254	Royal college	17	195	Nurse practitioner
50	252	gps	18	194	Palliative care
51	250	Children	19	186	Older people
53	247	Healthcare	20	185	Clinical governance
55	231	Hospital	21	180	Nurse prescribers
56	229	Assessment	22	173	Community nursing
61	217	Public health	23	171	Professional development
62	217	Medicines	24	170	Nurse practitioners
63	216	Public	25	169	Practice nurses
65	213	Family	26	167	Health nursing
66	213	Trust	27	166	Evidence-based
67	212	Drug	28	159	Young people
70	199	Society	29	155	District nurses
71	197	Pharmacy	30	154	Web site

Interestingly, the noun and noun phrase results do not overlap to the extent that might be expected. This overlap suggests that both should continue to be used. Logically, noun phrases ought to be better because they are more specific (e.g. 'University of Cambridge' rather than 'University' and 'Cambridge') but the high frequency of the word 'university' is also in this case a useful indicator that many universities are discussed, so the generality of nouns can sometimes be an advantage. Nevertheless, there seems to be more scope for automating the noun phrase analysis because its results did not need to be manually filtered.

It is probably not possible to give a definitive list of the unexpected items in Table 6: presumably different stake holders will see different sections of the picture and the table will serve to illuminate sections in which they are less well engaged.

5. Discussion

As discussed in Section 3, all the data should be viewed in the knowledge of the limitations of its origins. The documents included are those that are (a) publicly available on the Web and (b) indexed in Google. Point (a) is a purpose of the study, but should not be forgotten, and the omission of invisible Web pages (Ru & Horowitz, 2005), and presumably many in NHSnet, is a serious concern. Viewing Web documents as a subset of all documents about the topic, the large number of academic pages is not necessarily an indicator of more academic than commercial interest, but of a greater tendency for academics to publish on the Web (Lawrence & Giles, 1999), related to a culture of openness and inter-organisational information sharing (Whitley, 2000). An important corollary, however, is that all the results (e.g. concerning national differences) will be skewed towards academic pages. Point (b) is an additional limitation because search engines have biases, for example better coverage of documents in developed nations (Vaughan & Thelwall, 2004).

The generation of summary statistics through URL processing, for example domain names, TLDs and STLDs, is another limitation because of the underlying assumptions discussed above concerning the origins of the pages (e.g. .uk pages tend to be of UK origin). In one case, for pages from the .gov domain, the assumption was checked and found not to be correct. It seems reasonable, however, to use these aggregation methods and then to either manually check any apparent anomalies that they produce, or, if significant time is available, to check all of the results.

The approach used in this paper, giving the top results in each category, is probably most appropriate for categories where the top few results formed the vast majority of the data (e.g. Table 1) but is less satisfactory when the top results form a small percentage of the data, and possibly a misleading one. A more accurate approach in the latter case would be to take a random sample of results and report a classification of these using content analysis (Weare & Lin, 2000). This approach was not adopted, however, because the ultimate goal of this research is to develop tools that can automate the extraction and processing of data. Hence it is necessary to avoid any method that requires extensive human work. Consequently, however, the tables should be treated as suggestive evidence rather than conclusive proof of any patterns found.

The methods used are sensitive to the actual phrase chosen, "nurse prescribing". There may be other English phrases that describe the same concept, particularly outside of the UK, but we were unable to find any before the data collection or in the results, so it seems likely that no such phrases are in common use in academia in the English speaking world. In particular, the results point to nurse prescribing being less significant as a concept in the US because no alternative phrase was found suggesting that the nurse practitioner status dominates and that prescribing is seen as a secondary issue. Of course, in non-English speaking countries there may well be phrases equivalent to "nurse prescribing". This is a concern that was less significant for the previous Web issue analysis study (Thelwall et al., in press) of the impact and spread of a United Nations initiative which was associated with a specific phrase.

A generic automated technique to improve the quality of the results would be to develop algorithms that processed the contents of each page, discarding segments that were identified as likely to be irrelevant (e.g., Miles-Board, Carr, & Hall, 2002). These irrelevant segments would probably include advertising and navigation bars. An alternative would be to only process only text and links that were close to the identified phrase. The former technique would probably be unsuitable for PDF analyses. The natural language processing algorithms used are another part that could potentially be improved, for example customising the algorithms to recognise medical terminology.

6. Conclusions

The Web analysis was able to identify a number of interesting facts. Whilst many would probably serve to confirm stakeholders' suspicions, others (e.g. mental health, the UK focus, the minor nhs.uk role, the disconnectedness of nurse prescribing Web sites) may present surprises. Overall, then, the results should help give managers an evidence-based map of online nurse prescribing information, as well as suggesting avenues for further exploration. It is important that the results of a Web issue analysis should be interpreted by a human expert and not taken at face value, however, in case false connections have been identified because of anomalous publishing practices. The Web approach seems promising because it can give useful and timely information, can be almost fully automated and uses a free data source. The quality of information could be improved, when full automation is not necessary, with additional manual filtering to remove anomalies, or to classify random samples. Future similar studies will give the additional benefit of allowing comparisons between different topics that will help identify anomalies.

References

Bar-Ilan, J. (2004). Search engine ability to cope with the changing Web. In M. Levene & A. Poulovasilis (Eds.), *Web Dynamics*. Berlin: Springer-Verlag.

Bernstam, E. V., Shelton, D. M., Walji, M., & Meric-Bernstam, F. (2005). Instruments to assess the quality of health information on the world wide Web: What can our patients actually use? *International Journal of Medical Informatics 74*(1), 13–19.

Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for Webometrics. *Journal of the American Society for Information Science and Technology, 55*(14), 1216–1227.

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

Burd, A., Chiu, T., & McNaught, C. (2004). Screening internet Websites for educational potential in undergraduate medical education. *Medical Informatics and The Internet in Medicine, 29*(3–4), 185–197.

Camm, J. (2005). Early intervention and mental health. *Community Practitioner, 78*(4), 124–125.

Chakrabarti, S. (2003). *Mining the Web: Analysis of hypertext and semi structured data*. New York: Morgan Kaufmann.

Chakrabarti, S., VanDen Berg, M., & Dom, B. (1999). *Focused crawling: A new approach to topic-specific Web resource discovery*. Paper presented at the 8th International World Wide Web Conference.

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

Clifton, C., Cooley, R., & Rennie, J. (2004). Topcat: Data mining for topic identification in a text corpus. *IEEE Transactions on Knowledge and Data Engineering, 16*(8), 949–964.

Coburn, A. (2005). Lingua:En:Tagger—part-of-speech tagger for English natural language processing. Available from http://search.cpan.org/dist/Lingua-EN-Tagger/Tagger.pm.

Currò, V., Buonuomo, P. S., Onesimo, R., Rose, P. D., Vituzzi, A., Di Tanna, G. L., et al. (2004). A quality evaluation methodology of health Web-pages for non-professionals. *Medical Informatics and The Internet in Medicine, 29*(2), 95–107.

Dørup, J., Hansen, M. S., Ribe, L. R., & Larsen, K. W. (2002). A comparison of technologies for database-driven Websites for medical education. *Medical Informatics and The Internet in Medicine, 27*(4), 281–289.

Flake, G. W., Lawrence, S., Giles, C. L., & Coetzee, F. M. (2000). *Efficient identification of Web communities*. Paper presented at the 6th International Conference on Knowledge Discovery and Data Mining, New York.

Foot, K. A., Schneider, S. M., Dougherty, M., Xenos, M., & Larsen, E. (2003). Analyzing linking practices: Candidate sites in the 2002 us electoral Web sphere. *Journal of Computer Mediated Communication, 8*(4). Available from http://www.ascusc.org/jcmc/vol8/issue4/foot.html.

Garrido, M., & Halavais, A. (2003). Mapping networks of support for the Zapatista movement: Applying social network analysis to study contemporary social movements. In M. McCaughey & M. Ayers (Eds.), *Cyberactivism: Online activism in theory and practice* (pp. 165–184). London: Routledge.

Google (2005). Google Web APIs (beta). Available from http://www.google.com/apis/.

Gournay, K. (2002). Prescribing: The great debate. *Nursing Standard, 17*(9), 22.

Hales, A., & Dignam, D. (2002). Nurse prescribing: Lessons from the US. *Nursing New Zealand, 8*(10), 12–15.

Harries, G., Wilkinson, D., Price, E., Fairclough, R., & Thelwall, M. (2004). Hyperlinks as a data source for science mapping. *Journal of Information Science, 30*(5), 436–447.

Hernández-Borges, A., Macías-Cervi, P., Gaspar-Guardado, A., Arcaya, M. L. T.-Á. D., Ruíz-Rabaza, A., & Jiménez-Sosa, A. (2003). User preference as quality markers of paediatric Web sites. *Medical Informatics and the Internet in Medicine, 28*(3), 183–194.

Hovenga, E. J. S. (2004). Globalisation of health and medical informatics education—what are the issues? *International Journal of Medical Informatics, 73*(2), 101–109.

Hughes, F., & Lockyer, H. (2004). Evidence and engagement in the introduction of nurse prescribing in New Zealand. *Nurse Prescribing, 2*(3), 131–134.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2*(1), 1–15.
Lancaster, F. W., & Lee, J. l. (1985). Bibliometric techniques applied to issues management—a case-study. *Journal of the American Society for Information Science, 36*(6), 389–397.
Latter, S., & Courtenay, M. (2004). Effectiveness of nurse prescribing: A review of the literature. *Journal of Clinical Nursing, 13*(1), 26–32.
Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature, 400*, 107–109.
Leydesdorff, L. (1989). Words and co-words as indicators of intellectual organization. *Research Policy, 18*, 209–223.
Merton, R. K. (1973). *The sociology of science. Theoretical and empirical investigations*. Chicago: University of Chicago Press.
Mettrop, W., & Nieuwenhuysen, P. (2001). Internet search engines—fluctuations in document accessibility. *Journal of Documentation, 57*(5), 623–651.
Miles-Board, T., Carr, L., & Hall, W. (2002). Looking for linking: Associative links on the Web. In *Proceedings of ACM hypertext 2002*, pp. 76–77.
Mitkov, R. (2003). *The Oxford handbook of computational linguistics*. Oxford: Oxford University Press.
Mullay, S., Mason, C., & Frogatt, J. (2003). The progress of nurse prescribing in the United Kingdom. *Nurse Prescribing, 1*(3), 104–105.
Murphy, J., Stramer, K., Clamp, S., Grubb, P., Gosland, J., & Davis, S. (2004). Health informatics education for clinicians and managers—what's holding up progress? *International Journal of Medical Informatics, 73*(2), 205–213.
NHSnet (2005). Nhsnet—nhs information authority. Retrieved 29 September, 2005. Available from http://www.nhsia.nhs.uk/nhsnet/pages/.
Nolan, P., Bradley, E., & Carr, N. (2004). Nurse prescribing and the enhancement of mental of mental health services. *Nurse Prescriber, 1*(11). Available from http://www.nurse-prescriber.co.uk/Articles/NP_MentalHealthServices.htm.
Ozmutlu, S., & Cavdur, F. (2005). Neural network applications for automatic new topic identification. *Online Information Review, 29*(1), 34–53.
Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. *Connections, 25*(1), 49–61.
Phillips, B. J. (2005). A comprehensive look at the legislative issues affecting advanced nursing practice. *The Nurse Practitioner, 30*(2), 14–47.
Ploncynski, D., Oldenburg, N., & Buck, M. (2003). The past, present and future of nurse prescribing in the United States. *Nurse Prescribing, 1*(4), 170–174.
Rosseter, R. (2000). Nurse practitioners: The growing solution in health care delivery. Retrieved 29 September 2005. Available from http://www.aacn.nche.edu/Media/Backgrounders/npfact.htm.
Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and northernlight. *Cybermetrics, 2/3*. Available from http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html.
Ru, Y. B., & Horowitz, E. (2005). Indexing the invisible Web: A survey. *Online Information Review, 29*(3), 249–265.
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science, 24*(4), 265–269.
Smith, A. G. (1999). A tale of two Web spaces; comparing sites using Web impact factors. *Journal of Documentation, 55*(5), 577–592.
Smith, A. (2004). The use of the internet to support the education of nurse prescribers. *Nurse Prescribing, 2*(3), 127–130.
Stuart, D., & Thelwall, M. (2005). *What can university-to-government Web links tell us about a university's research productivity and the collaborations between universities and government?* Paper presented at the ISSI 2005.
Tang, R., & Thelwall, M. (2003). US academic departmental Web-site interlinking: Disciplinary differences. *Library and Information Science Research, 25*(4), 437–458.
Thelwall, M. (2002). Conceptualizing documentation on the Web: An evaluation of different heuristic-based models for counting links between university Web sites. *Journal of American Society for Information Science and Technology, 53*(12), 995–1005.
Thelwall, M. (2004). *Link analysis: An information science approach*. San Diego: Academic Press.
Thelwall, M., Vann, K., & Fairclough, R. (in press). Web issue analysis: An integrated water resource management case study. *Journal of the American Society for Information Science & Technology*.
Thelwall, M., Tang, R., & Price, E. (2003). Linguistic patterns of academic Web use in Western Europe. *Scientometrics, 56*(3), 417–432.
Vaughan, L., & Thelwall, M. (2004). Search engine coverage bias: Evidence and possible causes. *Information Processing & Management, 40*(4), 693–707.
Weare, C., & Lin, W. Y. (2000). Content analysis of the world wide Web-opportunities and challenges. *Social Science Computer Review, 18*(3), 272–292.
White, H. D., & Griffith, B. C. (1982). Author co-citation: A literature measure of intellectual structure. *Journal of American Society for Information Science, 32*(3), 163–172.
Whitley, R. (2000). *The intellectual and social organization of the sciences* (2nd ed.). Oxford: Oxford University Press.
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), 49–56.
Willhelmsson, S., & Foldevi, M. (2003). Exploring views on Swedish district nurses' prescribing—a focus group study in primary health care. *Journal of Clinical Nursing, 12*(5), 643–650.
Wormell, I. (2000). Critical aspects of the Danish welfare state—as revealed by issue tracking. *Scientometrics, 48*(2), 237–250.
Zeng, Q. T., Kogan, S., Plovnick, R. M., Crowell, J., Lacroix, E.-M., & Greenes, R. A. (2004). Positive attitudes and failed queries: An exploration of the conundrums of consumer health information retrieval. *International Journal of Medical Informatics, 73*(1), 45–55.