

Methodologies for Crawler Based Web Surveys

Mike Thelwall¹

*School of Computing, University of Wolverhampton, Wulfruna Street, Wolverhampton
WV1 1SB, UK
m.thelwall@wlv.ac.uk*

Abstract

There have been many attempts to study the content of the web, either through human or automatic agents. Five different previously used web survey methodologies are described and analysed, each justifiable in its own right, but a simple experiment is presented that demonstrates concrete differences between them. The concept of crawling the web also bears further inspection, including the scope of the pages to crawl, the method used to access and index each page, and the algorithm for the identification of duplicate pages. The issues involved here will be well-known to many computer scientists but, with the increasing use of crawlers and search engines in other disciplines, they now require a public discussion in the wider research community. This paper concludes that any scientific attempt to crawl the web must make available the parameters under which it is operating so that researchers can, in principle, replicate experiments or be aware of and take into account differences between methodologies. A new hybrid random page selection methodology is also introduced.

Keywords

Web surveys, search engine, indexing, random walks

1 Introduction

The current importance of the web has spawned many attempts to analyse sets of web pages in order to derive general conclusions about its properties and patterns of use (Crowston and Williams, 1997; Hooi-Im et al, 1998; Ingwerson, 1998; Lawrence and Giles, 1998a; Miller and Mather, 1998; Henzinger et al., 1999; Koehler, 1999; Lawrence and Giles, 1999a; Smith, 1999; Snyder and Rosenbaum, 1999; Henzinger *et al.*, 2000; Thelwall, 2000a; Thelwall, 2000b; Thelwall, 2000d). The enormous size of the web means that any study can only use a fraction of it. Moreover, the complexity of automatic web crawling means that the outcome of any exercise that directly or indirectly involves an automatic downloading component is dependent upon a series of technical decisions about its operation. This paper addresses the issue of the variability of results of web surveys as a result of methodological decisions and seeks to verify that differences exist, using the simplest of statistics about a web site: the number of its pages that can be crawled.

¹ Thelwall, M. (2002). Methodologies for Crawler Based Web Surveys, *Internet Research*, 12(2), 124-138.

The initial decision to be made in any survey is how to decide upon which sites or pages to study. Various different methodologies have been used, a common one being the selection of web sites directly or indirectly from existing search engines. The extensive literature on the use of search engines for information retrieval is an indication of their importance here (Ardo and Lundberg, 1998; Byers, 1998; Hersovici *et al.*, 1998; Kamiya, 1998; Kirsch, 1998; Miller and Bharat, 1998; Schwartz, 1998; Snyder and Rosenbaum, 1998; Ben-Shaul *et al.* 1999; Dean and Henzinger, 1999; Gordon and Patak, 1999; Hawking, and Craswell, 1999; Ho and Goh, 1999; Jenkins *et al.*, 1999; Kumar *et al.*, 1999; Lawrence and Giles, 1999b; Overmeer, 1999; Rafiei and Mendelzon, 2000; Scime, 2000; Spink *et al.*, 2000). Many surveys have, however, used direct methods for the selection of sites using the systematic checking of legal domain names or the random testing of Internet Protocol (IP) addresses, the unique identifier that each web server must possess. Commercial search engines only index a fraction of the web and thus surveys using, or seeded by, search engines are potentially operating on a biased subset of the web. At the other extreme, surveys finding web pages directly through domain names or random IP addresses will also index sites that, although accessible through the web, are not intended for a general audience. The Lawrence and Giles (1999b) survey used a technique to minimise this, however. Random IP address crawlers will also not be able to find some domains which share an IP address with others in a process known as virtual hosting (Fielding, 1999). Domain name based surveys, which test ranges of legal domain names for use, can also suffer from the former problem, but it may be that the sites with relatively standard domain names that this methodology is restricted to are more likely to be for public consumption. Such surveys suffer from a problem that IP based ones do not, an inherent mathematical inability to be genuinely randomly chosen from any complete domain space (Thelwall, 2000b). With the various types of survey all having some limitations, an important question is whether they are likely to give demonstrably different results in response to research questions. This also begs the question of exactly which pages should be counted as part of the web in any given survey, in particular whether to use an inclusive definition or to attempt to exclude pages not intended for general access. It is important to consider the algorithm by which web sites are crawled. The function of the web crawler, spider or robot is to fetch web pages, archive information about them in a database and to extract and follow links in the pages retrieved. The database is then normally processed by a separate program or suite of programs to extract meaningful information, perhaps in the form of a searchable index for a search engine site. Although the task of a web crawler is conceptually simple and deals only with quantifiable electronic entities, in fact its practicalities make its algorithm a complex creation. The obstacles that have to be overcome in the engineering of a crawler will be well known by their designers, but are not well documented. Academic research on web crawlers has tended to focus on the product of the web crawler, or its unique features, rather than on the crawler itself. The situation is further complicated by the fact that the workings of many major search engines are commercial secrets. With the increasing use of the results of web crawlers outside computer science there is, however, a real need for an explicit discussion of the factors involved. Some articles illustrating this necessity are given below.

One number-critical activity is to estimate the total number of web pages in existence (Bray, 1996; Bharat and Broder, 1998; Lawrence and Giles, 1999a) or the proportion with certain properties (Bar-Yossef, 2000; Kishi *et al.* 2000; Henzinger *et al.*, 2000; Thelwall, 2000a; Thelwall, 2000d). Since the web is too large to survey in entirety, such attempts must be based upon a clearly defined subset of the web, ideally coinciding with accepted definitions (see below) as far as is feasible, although it is believed to be impossible to

precisely quantify differences in the context of practically infinite size of the Web, described below. An important question in this context is how to define that subset and the extent to which the results would be different for alternative realistic definitions. Another numerical exercise is the attempt to count links to academic web sites and to use them in an attempt to evaluate factors connected to the research impact of the institution (Ingwerson, 1998; Smith, 1999; Snyder and Rosenbaum, 1999; Thelwall, 2000b), with recent work showing some progress in this direction (Thelwall, 2001a). Here it is particularly important to avoid counting links from a page more than once, and so identifying duplicate pages is essential. But this task is troublesome too, because, for example, a page containing a text-based hit counter will always be different when reloaded.

This paper seeks to clarify and place fully in the public domain the issues involved in web crawling. It uses a web crawler designed to test the effect of varying the parameters in a crawl and presents some results from partial web crawls.

2 Page Selection Methodologies

There have been several different approaches used to choose sites or sets of web pages for a survey that attempts to select with some degree of randomness. One experimental survey used the random URL feature that was previously available in AltaVista (Miller and Mather, 1998) and another uses an unspecified method (Turau, 1998), but most have described a controlled attempt at selection. The following approaches are search engine based, or have similar theoretical limitations.

- Using a copy of the whole database of a large search engine (Broder *et al.* 2000).
- Using a systematic approach to select links from the directory structure of a search engine (Hooi-Im *et al.*, 1998; Callaghan and Pie, 1998, Cockburn and Wilson, 1996; Koehler, 1999; Crowston and Williams, 2000).
- Selecting pages from a random walk seeded with pages from a predefined large initial set (Henzinger *et al.*, 1999; Bar-Yossef *et al.*, 2000; Rusmevichientong *et al.*, 2001).

Another possible approach is to use a metacrawler (e.g. Lawrence and Giles, 1998b) to combine the results of several search engines, but this has the same theoretical limitations, albeit of a lesser magnitude. There are also two strategies that have been used to obtain unmediated access to the web.

- Selecting web sites by random choices of IP addresses (Lawrence and Giles, 1999a; Kishi *et al.*, 2000).
- Selecting web sites by systematic searching of a domain name space (Thelwall, 2000a; Thelwall, 2000c; Thelwall, 2000d).

All of these methods have strengths and weaknesses, excluding certain types of web sites. Search-engine based surveys exclude all pages neither in the crawler's database nor linked to by a chain of pages starting with one in the crawler's database. This, although restrictive, can be justified in terms of the selection including the most 'findable' pages on the web, a concept that can be precisely defined, for example in terms of random walks (Henzinger *et al.*, 2000). It is known, however, that commercial search engines cover only a fraction of the web, and a fraction known to be biased by page interlinking (Lawrence & Giles, 1999a). Commercial search engines use various unpublished algorithms to decide which pages to crawl and which to keep (Mauldin, 1997) and so their use is problematical from an accountability point of view. Another example of a type of site likely to be underrepresented in a search engine is one where most or all content is not indexed because it is in a non-HTML form such as Java, Macromedia Flash, or images of text. The

direct methods that bypass search engines and test the use of legal names or addresses are also problematic for both technical and non-technical reasons. Many web sites are not intended to be used publicly, but are not protected from general access because they contain no sensitive information. This includes default self-description web sites that are automatically generated by web servers when they are installed and sites for Internet access to various electronic devices. These sites may be found by direct searching even though it could be argued that they do not really form part of the public web. Lawrence and Giles (1999a) attempted to circumvent this problem by using a database of regular expressions to weed out known sites of this kind, but the methodology is unable to identify individual web sites created by humans that are not intended for general use. There are also more complex types of sites that can cause problems for certain crawling approaches. For example, large sites which rent or give space to individual users without giving them domain names or separate IP addresses, such as `geocities.yahoo.com`, will effectively shield the subsites from crawlers that do not follow links from external sites unless the host site maintains a crawlable directory of all hosted subsite home pages. It should also be mentioned that some sites contain a combination of public and private pages and that a crawl of these may well cover the public pages and retrieve an access denial error message for the remainder. A different technical problem is that with the advent of the virtual server in the HyperText Transfer Protocol (HTTP) 1.1 (Fielding *et al.*, 1999), one IP address can be used for multiple domain names. A survey indexing only using IP addresses will only get one of these sites. Web site hosting agencies may, for example, host tens of thousands of domain names with a few IP addresses, making this a potentially serious problem. Although it is possible to query a Domain Name System server in order to find out the names of all of its virtual host sites using a single IP address, this request may be refused (Albitz and Liu, 1992), and so it is not possible to guarantee to be able to have access to this information. Hence IP based searches cannot find all possible domain names. Name based surveys crawl for sites in batches by name pattern, for example checking for sites at the domain names `www.a.com` to `www.zzzz.com`. These are useful for surveying particular domain areas and do not suffer from the IP problem, but this approach cannot be made genuinely random. The reason is that the size of the domain name space means that any survey of randomly generated domain names from the space of all possible domain names is statistically almost certain not to find any at all in a reasonable amount of time. In the case cited above it is likely, but as yet unproven, that the short domain names, for which the approach is feasible, are unrepresentative of the domain space as a whole. For example these web sites are likely to be older than the average site, having 'got to the shorter names first'.

A further method of surveying web pages in order to obtain a relatively indiscriminate selection is the random walk. This starts with a large initial collection of web pages and proceeds in a series of steps, at each one making a decision whether to follow a link from the current page or to jump to another from the known set. The random walk approach and those based upon search engines differ from the IP and domain based survey techniques in the details of site coverage. The latter surveys the publicly indexable pages on a site, essentially all pages findable by following links from the home page, whereas a search engine based survey can potentially cover a larger variety of pages on sites. The additional pages can be found by following links from other sites or using, directly or indirectly, information about URLs that are not linked to but have been registered in the search engine used. The term 'publicly accessible web' has been used in the context of random walks to describe the area of the web reachable by a finite sequence of links from the starting set of pages, although pages only reachable by a large sequence

of links are in practice unlikely to be included any given random walk survey (Henzinger *et al.*, 2000). In fact, the set of pages reachable by random walks is not essentially different from that indexable by a search engine, see Kirsch (1998), for an example. The main difference is that a search engine may attempt to index most pages found, rather than just a random sample.

In order to compare the two main types of survey, IP based and search engine based, an IP based crawl was made and then its results were analysed in relation to AltaVista. The results were then factored with the type of domain name used to see if there were clear differences in coverage for the publicly indexable set. The IP based crawl reached a selection of 1407 sites, chosen by repeatedly testing the class A, B and C Internet Protocol addresses at random and retaining only those linked to a web server. This gave a random selection of mainly commercial sites, encompassing all continents except Antarctica. The crawler used and its design parameters are discussed later in this paper. The figures given below are from counting all file types on the site and after excluding multiple pages with identical HTML.

2.1 Naming styles for sites identifiable from IP addresses

Some web sites are publicly available but not intended for general use. It was conjectured that those for public use were more likely to have a standard type of domain name, whereas others would be more likely to have extra prefixes indicating their function or subordinate status within a domain. For example mail.wlv.ac.uk is a valid web address, but the main external web site is at www.wlv.ac.uk.

The 1407 sites checked were split into two groups according to the domain name obtained from a reverse DNS lookup of the IP address. A domain name was counted as 'standard' if there was only one name before either a generic Top Level Domain (gTLD) or a recognised Second Level Domain subcategory of a national TLD (e.g. co, ac, sch, plc, edu, com, net: the exact selection used varies by country), or if there was a 'www.' followed by one name and one of these two types of ending. The results are shown in table I. Note that the median is a more appropriate measure than the mean, which is unreliable for web page counts and particularly unreliable for relatively small counts.

Table I The distribution of standard named site sizes in the survey.

	Number	Median	Mean
All Sites	1407	6	111
Standard Named Sites	856	10	92.1
Non-standard Names	551	1	140

The difference between the page size distributions for the standard and non-standard name endings was examined with the non-parametric Mann-Whitney U test and was found to be significant at the 0.01% level, giving clear evidence that the two domains are different in character. A domain name based survey that only surveyed standard-named sites by first finding their IP address would, therefore, expect to get different results than one using all sites found on IP addresses.

2.2 AltaVista's coverage of sites identifiable from IP addresses

The sites found by the IP address search were checked in AltaVista to see if any of the pages had been indexed. This search engine was chosen for its reliably reported relatively large coverage at the time (Lawrence and Giles, 1999a) and its advanced syntax that allows direct domain specific searching. Since the survey was completed it appears that Google has built much greater web coverage, but AltaVista may be more stable for this kind of search (Thelwall, 2001b). The results are shown in table II.

Table II A comparison of the survey site sizes in AltaVista and the sites outside AltaVista.

	Indexed in AltaVista			Not indexed in AltaVista		
	Number	Median	Mean	Number	Median	Mean
All Sites	582	16	138	825	2	92.1
Standard Named Sites	534	16	133	322	3	24.9
Non-standard Names	48	10	195	503	1	135

As can be seen from the table, the vast majority, 91.8%, of sites indexed by AltaVista had standard domain names, but it indexed only a small majority, 62.4%, of sites with standard names. The impact of name type on AltaVista indexing was tested to be highly significant. It is perhaps not surprising that AltaVista tended to index the larger sites: the median size of standard site indexed was 16 pages, whereas the median size of standard site not indexed was only 3 pages. Care is needed when analysing web site page counts because the data is not normally distributed and therefore significance tests on the *mean* are not possible. The problem is essentially that a minority of web sites are known to be very large. For example, although the mean size in this survey was small, a university web site could be expected to have somewhere near 100,000 and would, therefore, dramatically increase the mean size of a small sample of web sites. For this reason conclusions about general properties of a section of the web based on numbers small enough not to include a reasonable proportion of the larger web sites, such as that in Kishi *et al.* (2000), based upon 85 with no large sites, are without merit. It is, nevertheless, possible to use non-parametric statistics to test for differences in distribution of data, even based upon relatively small numbers. The difference in page size distributions between sites covered by AltaVista and those not was found to be highly significant at the 0.01% level, using the Mann-Whitney test, indicating that AltaVista's coverage is significantly different from that of an 'unvetted' IP address crawl.

2.3 Short and longer standard domain names

A comparison was made between all the standard names found and the 42 standard names in which the identifying part of the name was short, meaning up to four characters long, as used in Thelwall (2000a). The results are shown in table III with the median being again the more useful measure of central tendency.

Table III A comparison of short with standard domain name sites

	Number	Median	Mean
Standard domain name	856	10	92
Long standard domain name	814	10	83
Short standard domain name	42	32.5	270

The difference in the distribution of site counts for these sites was tested and found to be significant at the 0.01% level, using the Mann-Whitney test.

2.4 Summary

This experiment has provided evidence that all of the methods of selecting sites cover areas of the Internet which have demonstrably different properties, even when considering the most basic statistic of page counts. The fact that the property checked was the size of the publicly indexable set rather than any realisation of the concept of the publicly accessible pages does not affect this conclusion.

3 Obstacles to Reliable Automatic Indexing of Web Sites

Once a collection of sites has been selected for study, consideration must be given to the process by which they are crawled and indexed. Any functioning web crawler is the product of a series of decisions concerning how to process the pages downloaded and how to ensure comprehensive coverage of the selected area. In principle, all that is needed is a program to download web pages and to extract their links for subsequent downloading, but there are many potential pitfalls in this process. One elementary problem is that many web pages contain errors in the HTML syntax. A common mistake, for example, is to omit the closing quotes in a link reference. In this case it would be reasonable for the program parsing the page to terminate the linked URL at the tag end character '>' or an end of line character instead, effectively guessing at an automatic correction of the mistake. Other common errors are to include too many or too few slashes in URLs, for example `http:/` or `http:///`. The crawling software can be programmed to automatically correct such errors or to only accept valid links and the decision about which method to adopt will effect the results of the crawler. The point here is that the small-scale technical details of a crawling algorithm can have an impact on its results, but such issues do not normally merit discussion in publications because they are essentially peripheral to the design of an effective crawler. Three more fundamental issues are discussed below, and all are summarised in table IV.

Table IV A summary of obstacles to reliable automatic indexing of web sites

The existence of errors in web pages and the use of different strategies for automatically correcting them
The phenomenon of different URLs pointing to the same web page. <ul style="list-style-type: none"> • Multiple home page names • Virtual hosting allowing subsites to appear as complete sites under a different domain name • File copying between servers, including entire sites and subsites
The difficulty in constructing an algorithm to identify duplicate pages <ul style="list-style-type: none"> • The technical difficulty for huge collections of pages • The existence of pages with variable components, such as hit counters
The existence of effectively infinite collections of web pages created automatically by servers in response to URL requests, recursively embedding new URLs in the generated pages

3.1 Non-Unique URLs

A single web based resource may be accessible by multiple URLs. A web survey must therefore either take steps to avoid duplication or state that it is allowing it. Page duplication can occur in four ways: by the use of duplicate domain names, for example `microsoft.com` and `www.microsoft.com`; by a server recognising several URLs as referring to a single file; by the file being copied from one location to another; and by more than one server having access to the file. There is a formal mechanism in the HyperText Transfer Protocol, the redirection header (Fielding *et al.*, 1999), which is used to notify some kinds of duplication, but in the cases discussed below this does not necessarily happen.

A server can recognise different URLs for the same file for many different reasons. There are some common server default settings that automatically duplicate URLs, such as responding to a request for a directory by searching for a file with a standard name in the directory, often `index.html`, and returning it without flagging a change of name with a redirection command. The file `index.html` could also be accessed directly by name, giving two URLs for the same file. Another source of duplicate URLs is by the use of virtual servers. With HTTP 1.1, servers sitting on a single IP address can respond to different domain names by using a separate base file path for each. If one file path is in the domain referenced by another then this gives rise to duplicate names. For example `cba.scit.wlv.ac.uk` is a virtual domain containing a small subset of the pages of `www.scit.wlv.ac.uk`. In this case `http://cba.scit.wlv.ac.uk/` points to the same directory structure as `http://www.scit.wlv.ac.uk/~cm1993/cba/`. A further type of duplication occurs when a directory or file has been given duplicate names at the operating system level.

There are likewise situations when different servers have access to the same files, which can result even in different domain names IP addresses and paths referring to a single computer file. This can occur when there is more than one server running on a single computer, perhaps one as a backup. It can also occur when several computers have access to a common, or partially shared, file storage system. This could be the case when an old computer is allowed to continue running after its replacement is functional or in situations where multiple servers are needed to satisfy the demand for access to a web site. In the latter case incoming requests for a standard URL may be automatically redirected to a different domain name used by a less busy server.

Replication is also an issue because files can be copied from one location to another for various reasons, for example if the information that they contain is relevant in more than one context. This copying could be of a single file, or part of a systematic duplication scheme. Some files of widespread use are systematically copied to multiple servers around the globe in a process known as mirroring. Entire websites can also be mirrored, as occurs with the tu cows software repository website for example. It is also possible for websites to be partially copied, such as to provide an area of common content for the localised national websites of multinational companies. A further variation occurs when websites contain copies of numerous files around the globe as a repository. For example the UK Mirror Service (www.mirror.ac.uk) contains thousands of computer-related files and web sites copied from the rest of the Internet in order to provide fast efficient access to UK higher education.

An attempt to identify duplicate files is not straightforward. If a set of files were downloaded then a simple check for duplication would be to compare files in pairs, in principle discarding one of the duplicates when a match was found. If two HTML pages were found to be identical then a further check could be made on any relative path names used, as these could cause identical files in different directory locations to be effectively different web pages. For example two pages both consisting of a relative image reference such as `` could actually be referring to different images with the same name if the pages were in different directories. To ascertain whether these two pages were actually identical in function, the two files called `next.htm` would also have to be compared, creating a potentially recursive checking situation. It may be that a large number of pages referenced by the initial pages checked need to be compared before deciding on whether they are duplicates. This level of detail would be needed to differentiate between a virtual server based upon a subset of the main server but containing some of its own unique pages, and a virtual server which simply duplicates a subset of the main server. This level of checking would be computationally extremely expensive and so it is believed that most or all web crawlers adopt a compromise approach. The spider described by Heydon and Najork (1999) uses a literal HTML check, ignoring the possibility that pages with identical HTML could be different. In fact, in order to make the comparison process fast enough, in this example numbers calculated from the HTML are compared instead, which gives rise to the possibility that non-identical pages are occasionally regarded as duplicates. In contrast, one experimental 'archiver' does attempt to compare the raw HTML of pages, and allows for small changes such as in a page access count or date (Jackson and Burden, 1999), describing such duplicates as 'essentially equal'. Relative path issues are not taken into account, however.

It is ultimately impossible to verify whether two separate HTML files are conceptually the same or not, because identical files could be contextually different. A simple example of this is that two pages with identical HTML consisting of the message 'My home page will appear here shortly' would be conveying different information if they were in the default directories of two different people. It is believed that no crawlers attempt to distinguish between conceptually different pages with identical HTML, an apparently difficult task with little benefit for search engine users.

3.2 Non-Unique Pages

A web server may return different pages for the same URL, making the comparison of pages with different URLs for duplicate checking non-trivial. One simple way in which this can occur is by the server inserting text into a page as it is delivered, for example a

page access count or the current date and time could be inserted using Active Server Pages technology or Server Side Includes. This would result in a continually changing page with one URL, which could still be conceptually the same page. It is in principle impossible to tell in a HTML document whether any text is a relatively permanently fixture or whether it has been inserted automatically by the server. To cope with this situation an arbitrary rule must be adopted and specified. For example the possibility of server insertions in a page could be ignored and two pages compared as described in the previous section. Alternatively, pages could be compared and if found to be, say, 95% similar with the difference in a small area, then the pages could be accepted as being the same. It has been claimed that some web sites have pages that have deliberately continuously varying content with the specific purpose of foiling automatic crawlers (Ockenden, 2000). The online auction sites mentioned in this article were the target of automatic crawlers and would lose advertising revenue by the consequent reduction in the number of human visits. It is also increasingly common that web sites are personalised for individual users either at the request of the user (Manber *et al.*, 2000) or by a dynamic process on the server that is opaque to the visitor (Perkowitz and Etzioni, 2000).

In some cases there may also be two or more pages assigned to an URL with the server deciding which one to send depending on the exact nature of the request. This occurs, for example, when a site contains versions of its pages designed for particular browsers. An HTTP request for a page normally contains a description of the browser used, which makes this possible. The results of a survey are therefore possibly dependant on the exact request HTTP header sent. It would also be possible to survey with two or more headers, comparing the differences. Some sites, such as Lycos, attempt to identify the geographical location of the visitor and to deliver the appropriate regional variation of the page. The variation may be a minor one, such as changing the advert banner to a locally relevant one, or a major change such as redirection to a completely different national site.

Some web servers send an extra piece of information with web pages, the Etag (Fielding *et al.*, 1999). This was designed to be used by caching services in order to verify whether a page has changed since it was last downloaded. This cannot be used to aid checking in indexing, however, because common implementations of this do not guarantee that different Etags represent different web pages, only that the same Etag for a single page requested at different times means that the page has not changed significantly.

3.3 Infinite Collections of Pages

Theoretically infinite sets of pages can occur when a server automatically generates pages in response to a request. A common example of a server-generated page is the results returned by a search engine in response to keywords submitted. This is a benign example from an indexing point of view, but in other cases a server does create a theoretically infinite set of pages. This can happen, for example, when the server embeds additional information, useful to itself, in the links of a page. This could take the form of a request identification string, a technique often used for systems that require a user logon to access a database, but potentially also available for publicly accessible pages. This additional information could have the syntax of data at the end of an URL and because of this problem, Henzinger *et al.* (2000) prune URLs that contain such information. The data could also, however, be embedded in a manner making it indistinguishable from a normal directory structure, and would therefore be impossible for web crawlers to directly identify. If there is a loop of links involving pages with this feature then a theoretically infinite collection of pages can be created. This situation can also occur as a deliberate act of

sabotage on web crawlers, in something known as a ‘spider trap’. One use of the spider trap is as a defence mechanism to keep spiders out of a site in order to protect email addresses from being found and used for unsolicited mailings. It can be, in practice, impossible to distinguish an infinite collection from a large finite collection and so human intervention or an heuristic is needed for a crawler that attempts to index entire sites. The result of this would be a decision to ignore all links from a set of pages identified as a potential cause of trouble.

4 Executing a Survey

Some web surveys will be of a well-defined and easily indexable collection of web pages and will not have a problem in defining the area of study. For others an exact specification is an issue and the method of identifying relevant pages should be available. The exact definition of a web page is known to be problematic (Pitkow, 1998; Haas and Grams, 2000; Thelwall, 2000b) with conflicting definitions given in different places (Berners-Lee, 1993; Boutell, 1996; PC Webopedia, 1999). In any case a precise definition of the types of resources surveyed should be available. For example the survey could include just HTML pages or all files of any type accessible over the Internet using HTTP. It will also be relevant to know the IP port number(s) used for the survey. The official port of the web is 80, but other ports such as 8080 can be and are used. The survey may choose to operate only on port 80, on port 80 unless explicitly stated in the referencing URL, or may follow an algorithm to check likely ports if a request on port 80 fails.

4.1 The Concept of Publicly Indexable Pages

As discussed previously, a publicly indexable crawl starts at the home page of a site and systematically finds other pages by following links. The method will, however, only find pages linked to directly or indirectly by the home page. On some sites this might be all of the pages, but on others it will not. It is not possible to guarantee to find all pages on a site without having access to its file system because some pages might not be linked to any other web pages on the Internet and because a combinatorial search for legal URLs is mathematically not practical.

The definition of publicly indexable does not normally include JavaScript programs, a further restriction of its usefulness. Web pages can contain links to other pages that can only be found by executing JavaScript statements. The simplest is `document.location.href=filename.htm`, but URLs can also be constructed at run-time making the task of identifying them more complex than simple text processing. As a practical consideration, JavaScript programs are normally ignored in an indexing attempt and so sites using a lot of script-generated links that are not duplicated in the HTML may be less effectively indexed than others. Search engines have an advantage here, being able to use user submitted URLs to extend their coverage of sites.

4.2 The Statement of Coverage

The following list is a set of guidelines for the kind of information that other researchers should be entitled to request information about when a survey has been published. It is not reasonable to expect this kind of detailed information to clutter all such publications, but researchers should be prepared to divulge such information upon request.

- The dates and duration of the attempt and the location of the crawler program.

- The types of resources surveyed, including port details.
- The method of crawling. This could be a starting point and an attempt to follow all subsequent links in the chosen area. A depth restriction may also be declared, such as following only two levels of links.
- The HTTP header sent.
- The method of identifying and counting duplicate URLs for pages.
- Whether JavaScript links were followed, and if so the conditions under which they will be identified.
- Whether Java or any other application links were followed, and if so the conditions under which they will be identified.
- Any methods used to rectify incorrect links.
- A list of pages that were omitted, or pages the links of which were ignored, and a general description of why.
- The timeout period for non-responding servers - especially important if access speed is an issue or if the survey includes busy and quiet times for Internet traffic.
- The method of dealing with servers that did not respond, and a list or tally of servers that were omitted due to not responding.

5 Crawling Parameters

A web crawler was constructed that was programmed to be able to use a variety of different strategies to cover an area of the web. The web crawler was different in specification from that associated with a commercial search engine in that its primary design goal was to be able to ensure comprehensive and error-free coverage of individual web sites. This objective was possible because it would not be required to crawl a significant proportion of the web and would, therefore, not need to be optimised for efficiency. For example, it would not need heuristics to select which known URLs were the most important to crawl (Cho *et al.*, 1998; Cho and Garica-Molina, 2000) and in the identification of duplicates (Heydon and Najork, 1999). As a result of the design priority, the crawler underwent extensive testing of its ability to crawl general web sites, and specially constructed bug-ridden sites were also used to test its correct functioning in unusual circumstances. Over two million pages were crawled, in all, during this phase of the project. The testing process was built around human analysis of comprehensive error logs. When the crawler encountered an unexpected occurrence, essentially something other than a syntactically correct, complete HTML document it would record details in the log. A human would then analyse the log, recreate the problem, and then make a decision about how to deal with the situation, in most cases recoding of the program to implement the solution. Logged errors were a mixture of events that had not been anticipated and errors in web pages, Internet transfers or web page header information. To illustrate this process, an early log reported a malformed URL. Human inspection then revealed a syntax error in the HTML with missing end quotes in an anchor tag. A decision was made to recode the crawler to search for and accept the end of tag marker as the end of the URL. The same error would not, therefore, need to be logged again. The errors found ranged from relatively simple ones like this, to a bizarre error where a certain very rare character string was being rejected by the database software used and had to be specially intercepted and altered before saving. Possible transfer errors resulting in incomplete pages were identified by the absence of all of the tags `</body>`, `</html>` and `</frameset>`. When this occurrence was identified the page was repeatedly downloaded until either a closing tag was found or

successive downloads produced the same page, in which case an HTML error in a correctly transferred page was inferred. Some events were left in the logs to aid testing of the correct functioning of the duplicate checking process. Duplicate pages identified were logged whenever the identified duplication did not conform to a standard pattern, for example a root directory and file index.html in the same directory. They were also logged when the checking was not for similar, but not identical pages and the pages identified as essentially duplicate but not identical. An additional check of correct functioning was made by extensive comparison with results from AltaVista advanced searches of the same sites. All discrepancies were followed up and accounted for.

In order to produce statistically realistic results, a random sample of the web had to be identified, and the IP based crawl method was used. The purpose of choosing a large random selection was to estimate the significance of the parameters tested for the whole web. Each attempt at crawling was repeated with the lists of pages obtained combined and duplicates discarded. The duplication was an attempt to eliminate variations due to temporary factors, such as a server being temporarily unavailable. This had a significant impact on the results because one of the largest sites was unobtainable for some of the duration of the survey. In each case the only statistic used was the number of pages downloaded.

The crawler operated from July 14, 2000 to July 31, 2000. The standard settings used were as follows.

- All files linked to from HTML files were downloaded, including all non-HTML resources, with the links checked being standard links, client-side image maps and automatic redirection header fields. The standard web port, 80, was used and for domain names starting with 'www.', other domain names differing only in the initial set of characters up to the first full stop were considered to be subdomains of the same site. Domains with names extending the main name were also considered to be subdomains and included in the count. Requests without a response were timed out after 180 seconds.
- Anchor tags with missing ending quotes were automatically corrected to terminate at the end of tag marker '>' or at the end of the current line, whichever was earlier.
- Any page generating an error during its transfer was downloaded a second time. Any HTML page without a </BODY>, </HTML> or </FRAMESET> tag was also downloaded a second time in case of unflagged errors.
- A 'weak' duplication check was used which checked whether the files in the same directory were at least 95% identical. Differences were only allowed to occur in one continuous section of the text file, and were not allowed to include any links or embedded resource names. The contents of embedded resources were not checked, allowing pages with graphical counters to be recognised as duplicates. File sizes were allowed to differ by up to 3 bytes.
- The request header used was that of Netscape 4.7.
- Document counts were based upon the number of files, and not the number of screens, for example a frameset page with three frames would count as four documents.
- Two pages on separate sites had to be manually excluded in order to avoid infinite looping.

In the testing that follows, only the differences from the above configuration will be noted in each case. In order to guarantee a common set of files to download, a main crawl was first completed using the above parameters, but without duplicate checking. Subsequent crawls were then made from this database rather than the web, with the exception of crawls to check the effect of different request headers.

5.1 Web Definitions Comparison

The number of pages obtained including all linked documents was compared to the number including just HTML documents. The results are shown in table V. The arithmetic mean is a useful measure in this case since the same set of sites are being compared, but the median is unchanged at 6 for all the categories of pages in tables V and VI. No statistical tests of significance are necessary because the differences are concrete.

Table V A comparison of definitions of the web

	All documents	HTML pages only
Mean number of pages per site	111.0 (129%)	85.9 (100%)

Only 30 (19%) of the sites included non-HTML pages, but the amount varied in those that had. One site, for example, had 44 image files and only one web page (2%).

5.2 Tests for File Duplication

The results for different checks for file duplication are shown in tables VI and VII.

Table VI Total web page duplicates found using different tests
*Areas of the site had to be manually excluded

	Mean number of pages
No checking	88.7* (103%)
Check for identical HTML	85.9 (100%)
Check for weakly similar HTML	85.4 (99%)

Table VII Total web sites with at least one of the following types of duplicates

	HTML	Percent
Any different	439	31.2%
Identical HTML	410	29.1%
Weakly similar HTML	83	5.9%

The most common case of a duplicate file name was of a standard name such as 'index.htm' or 'default.htm' duplicating with the URL of the root of the directory. In most sites there were no other kind of file duplications. Two sites in particular appeared to be much bigger when not checking for duplicates. They were both using a server executable to redirect URLs without using the HTTP redirect mechanism. The facility used was effectively a server-side image map to back up a client side image map for those browsers not supporting it. Users with a modern browser would not have seen the link because it would have been hidden by the client-side image map. There were many single pages on

the site that were referred to in nine different ways because of the referencing technique used.

5.3 Request Headers

No differences were found in the number of pages in any site after comparing crawling with a Netscape 4.7 request, an Internet Explorer 3 request and an Internet Explorer 5 request. In fact no differences of any type were detected, so it must be the case that it is rare for a web site to make use of this information to customise web pages.

5.4 Summary

The differences between crawling methodologies found in the results appear to be relatively small in the context of a continually changing and expanding web, but it is not appropriate to conduct statistical tests on the data in order to be more specific. One problem (of many) is that the character of the largest web sites on the web, none of which were represented in this survey, could be different from that of smaller web sites, and, therefore calculations based upon the smaller may be completely misleading. The results should, therefore, be taken as illustrative of the problems that exist rather than indicative of the likely range of variation.

The issue of page duplication is less pressing for random walks, which do not attempt to index entire sites, but only to visit pages at random. Indeed, it is true to a limited extent that duplicate pages do not make a difference to a random walk, since it is user behaviour that is being modelled and, therefore, if there is a link to each of two duplicate pages then the probability that one of the two is visited is the same as if the links were switched to point to the same page. It would, however, make a difference in the related Google algorithm (Brin and Page, 1998). The problem lies with sites where a page links to itself with a different name, for example <http://www.a.com/> linking to <http://www.a.com/home.htm>, which most human users would identify and ignore, but the more complex occurrences of this phenomenon would be difficult to detect and correct automatically.

6 Discussion

Of the two experiments described in this paper, the first provides evidence that the different page selection methodologies may yield different results for the same question, and the second illustrates arguments given to show that crawler design parameters can affect survey results. Unfortunately, there is no clear answer to either problem because it is impossible to select pages at random from the web and it seems unreasonable, in the face of an ever-changing web, to argue for a single, agreed and unambiguous algorithm for web crawling. One anonymous reviewer of this paper has made the sensible suggestion that a public “crawler testbed” could be developed to help different designers to standardise their procedures or benchmark their algorithms, providing a basis for comparability. This would need to be a collaborative and ongoing project that allowed new web idiosyncrasies to be added, when discovered.

The essential difference between publicly indexable crawls and random walks or search engine based crawls is that the former are an attempt to be exhaustive in coverage of a sample of web sites, whereas the latter attempt to take into account a model of surfing behaviour in order to favour pages more likely to be visited by human users, and therefore

to reflect more the web as experienced in reality. The random walk and search engine-based approaches do not, however, take into account that many sites on the web are publicised by other means, such as print media, and that, therefore, users are also likely to visit sites that are not findable by following links. It thus seems desirable to combine the approaches to develop a more sophisticated model of user behaviour to include the possibility of a user jumping to a site based on external information. The logical way of implementing this would be to modify the random walk algorithm so that at each step, the possibility of jumping to a site obtained from an IP or domain name based search would be added. Experimentation would then be needed to ascertain whether this did give rise to an acceptable model of surfer behaviour, and as to the best value for the probability of such a jump. The advantage of this composite approach would be the larger theoretical coverage of the web in addition to the enhanced model of user behaviour.

A second combination approach suggests itself based upon the experiments, to conduct an IP-based survey, but to only include sites that have standard names. The advantage of this is that the coverage would be of sites shown to be more likely to be indexed in AltaVista, and, therefore, more likely to be designed for public consumption. This seems to be a less reliable method than the approach described above because of the existence of public web sites without standard names, and the existing use of regular expressions for a similar purpose.

Information scientists, and others analysing and measuring the web need to be aware that differences do exist in crawling and site selection methodologies, differences that will affect the results in uncontrolled ways, but that there are no simple answers to this problem. One conclusion is, however, possible: small sample sizes, particularly ones that are too small to give a reasonable sample of the larger sites in the domain surveyed, can give completely misleading results unless accompanied by a careful statistical analysis. In the context of an already enormous and expanding web the challenge is to make sense of the vast amount of information available, in a controlled manner.

7 References

- Albitz, P. and Liu, C. (1992), *DNS and BIND*, O'Reilly & Associates Inc., Sebastopol, CA.
- Ardo, A. and Lundberg, S. (1998), "A regional distributed WWW search and indexing service - the DESIRE way", *Computer Networks and ISDN Systems*, Vol. 30 No.1-7, pp.173-83.
- Bar-Yossef, Z., Berg, A., Chien, S., Fakcharoenphol, J. and Weitz, D., (2000), "Approximating Aggregate Queries about Web Pages via Random Walks", *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, pp. 535-541.
- Ben-Shaul, I., Herscovici, M., Jacovi, M., Maarek, Y. S., Pelleg, D., Shtalhaim, M., Soroka, V. and Ur, S. (1999), "Adding support for dynamic and focused search with Fetuccino", *Computer Networks*, Vol. 31, pp. 1653-1665.
- Berners-Lee, T. (1993), "WorldWide Web Seminar", <http://www.w3.org/Talks/General/Concepts.html> accessed 22 October 1999.
- Bharat, K. and Broder, A. (1998), "A technique for measuring the relative size and overlap of public Web search engines", *Computer Networks and ISDN Systems*, Vol 30, pp. 379-388.
- Boutell, T. (1996), "What are WWW, hypertext and hypermedia?", <http://www.boutell.com/faq/oldfaq/htext.htm> Accessed 22 October 1999.
- Bray, T. (1996). "Measuring the Web", *Computer Networks and ISDN Systems*. Vol. 28 No. 7-11, pp.993-1005.

- Brin, S. and Page, L. (1998). "The Anatomy of a large scale hypertextual web search engine", *Computer Networks and ISDN Systems*, Vol. 30, pp. 107-117.
- Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A. and Wiener, J. (2000), "Graph structure in the web", *Computer Networks*, Vol. 33 No. 1-6, pp. 309-320.
- Byers, D. (1998), "Full-text indexing of non-textual resources", *Computer Networks and ISDN Systems*, Vol. 30 No. 1-7, pp. 141-8.
- Callaghan, J. and Pie, A. (1998) "Business use of Internet Web sites-could do better!", *British Telecommunications Engineering*, Vol. 17 No. 1, pp. 56-65.
- Cho, J. and Garcia-Molina, H. (2000), "The Evolution of the Web and Implications for an Incremental Crawler", *Proceedings of the 26th VLDB Conference*, Cairo, Egypt, pp. 200-209.
- Cho, J., Garcia-Molina, H. and Page, L. (1998), "Efficient crawling through URL ordering", *Computer Networks and ISDN Systems*, Vol.30 No. 1-7, pp. 161-72.
- Cockburn, C. and Wilson, T. D. (1996) 'Business use of the World-Wide Web', *International Journal of Information Management*, Vol. 16 No. 2, pp. 83-102.
- Crowston, K. and Williams, M. (2000). "Reproduced and emergent genres of communication in the world wide web", *Information Society*, Vol.16 No.3, pp.201-15.
- Dean, J. and Henzinger, M. R. (1999), "Finding Related Pages in the World Wide Web", *Computer Networks*. Vol. 31 No. 11-16, pp. 1467-79.
- Gordon, M. and Patak, P. (1999), "Finding Information on the World Wide Web: the retrieval effectiveness of Search Engines", *Information Processing and Management*, Vol. 35, pp. 141-180.
- Fielding, R., Irvine, U. C., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P. and Berners-Lee, T. (1999), "Hypertext Transfer Protocol -- HTTP/1.1", <ftp://ftp.isi.edu/in-notes/rfc2616.txt>, Accessed 12 December 1999.
- Haas, S. W. and Grams, E. S. (2000). "Readers, authors and page structure: A discussion of four questions arising from a content analysis of web pages", *Journal of the American Society for Information Science*, Vol. 51 No. 2, pp. 181-192.
- Hawking, D. and Craswell N. (1999), "Results and Challenges in Web Search Evaluation", *Computer Networks*, Vol.31 No. 11-16, pp. 1321-30.
- Henzinger, M.R., Heydon, A., Mitzenmacher M. and Najork, M. (1999), "Measuring Index Quality using random walks on the Web", *Computer Networks and ISDN Systems*, Vol. 31 pp. 1291-1303.
- Henzinger, M.R., Heydon, A., Mitzenmacher M. and Najork, M. (2000), "On near-uniform URL sampling", *Computer Networks*, Vol. 33 No 1-6, pp. 295-308.
- Hersovici, M., Jacovi, M., Maarek, Y. S., Pelleg, D., Shtalhaim, M. and Ur, S. (1998), "The shark-search algorithm - An application: tailored web site mapping", *Computer Networks and ISDN Systems*, Vol.30 No.1-7, pp. 317-26.
- Heydon, A. and Najork, M. (1999), "Mercator: A scalable, extensible Web crawler", *World Wide Web*, Vol. 2, pp. 219-229.
- Ho, C. and Goh, A. (1999), "Jamaica: a World Wide Web profiler", *Internet Research: Electronic Networking Applications & Policy*, Vol. 9 No. 2, pp. 129-139.
- Hooi-Im, N., Pan, Y. J. and Wilson, T. D. (1998), "Business Use of The World Wide Web: A Report on Further Investigations", *International Journal of Information Management*, Vol. 18, pp. 291-314.
- Ingwersen, P. (1998), "Web Impact Factors", *Journal of Documentation*, Vol. 54, pp. 236-243.

- Jackson, M. S. and Burden, J. P. H. (1999), "WWLib-TNG - New Directions in Search Engine Technology", *IEE Informatics Colloquium: Lost in the Web - Navigation on the Internet*, November, pp 10/1-10/8.
- Jenkins, C., Jackson, M., Burden, P. and Wallis, J. (1999), "Automatic RDF metadata generation for resource discovery", *Computer Networks*, Vol. 31 no. 15, pp 11-16.
- Kamiya, H., Ohta, K., Kato, N., Mansfield, G. and Nemoto, Y. (1998), "An Improved Content Search Engine - Usage of Network Configuration Information", *Proceedings of IEEE TENCON 98*, pp. 21-24.
- Kirsch, S. (1998), "Infoseek's experiences searching the Internet", *SIGIR Forum*, Vol. 32, pp. 3-7.
- Kishi, N., Ohmori, T., Sasazuka, S., Kondo, A., Mizutani, M. and Ogawa, T. (2000). Estimating web properties by using search engines and random crawlers. *Proceedings of INET2000, The 10th Annual Internet Society Conference*. Yokohama (Japan), http://www.isoc.org/inet2000/cdproceedings/2a/2a_3.htm, accessed 20 November, 2000.
- Koehler, W. (1999), "Classifying Web sites and Web pages: the use of metrics and URL characteristics as markers", *Journal of Librarianship and Information Science*, Vol. 31 No. 1, pp. 21-31.
- Kumar, R., Raghavan, P., Rajagopalan, S. and Tompkins A. (1999), "Trawling the web for emerging cyber-communities", *Computer Networks*, Vol.31 No. 11-16, pp. 1481-93.
- Lawrence, S. and Giles, C. L. (1998a), "Searching the World Wide Web", *Science*, Vol. 280, pp. 98-100.
- Lawrence, S. and Giles, C. L. (1998b), "Inquirus, the NECI meta search engine", *Seventh International World Wide Web Conference*, Brisbane, Australia, pp.99-105.
- Lawrence, S. and Giles, C. L. (1999a), "Accessibility of information on the web", *Nature*, Vol. 400, pp. 107-109
- Lawrence, S. and Giles, C. L. (1999b), "Searching the Web: General and Scientific Information Access", *IEEE Communications Magazine*, Vol. 37 No. 1, pp. 116-122.
- Manber, U., Patel, A. and Robison, J. (2000), "Experience Personalisation with Yahoo!", *Communications of the ACM*, Vol. 43 No. 8, pp. 35-39.
- Mauldin, M. L. (1997), "Lycos: design choices in an Internet search service", *IEEE Expert*, Vol. 12 No. 1, pp. 8-11.
- Miller, H. and Mather, R. (1998). The presentation of self in WWW home pages. IRISS 98 International Conference: 25-27 March 1998, Bristol, UK <http://www.sosig.ac.uk/iriss/papers/paper21.htm>, accessed 20 November, 2000.
- Miller, R. C. and Bharat, K. (1998), "SPHINX: A framework for creating personal, site-specific web crawlers", *Computer Networks and ISDN Systems*, Vol. 30 No.1-7, pp. 119-30.
- Ockenden, M. (2000), "Scraping the page", *PC Pro*, Vol. 71, pp. 272.
- Overmeer, M. A. J. C. (1999), "My personal search engine", *Computer Networks*, Vol. 31, pp. 2271-2279.
- PC Webopaedia: http://webopedia.internet.com/TERM/W/World_Wide_Web.html, accessed 22 Oct 1999.
- Perkowitz, M. and Etzioni, O. (2000), "Adaptive Web Sites", *Communications of the ACM*, Vol. 43 No. 8, pp. 152-158.
- Pitkow, J. E. (1998), "Summary of WWW characterizations", *Computer Networks and ISDN systems*, Vol. 30, pp. 551-558.

- Rafiei, D. and Mendelzon, A. O. (2000), "What is this page known for? Computing Web page reputations", *Computer Networks*, Vol.33 No.1-6, pp.823-835.
- Rusmevichientong, P., Pennock, D. M. Lawrence, S. and Giles, C. L. (2001). "Methods for Sampling Pages Uniformly from the World Wide Web", The 2001 AAAI Fall Symposium Series, November 2-4, North Falmouth, Massachusetts. <http://www-users.cs.york.ac.uk/~tw/fall/Proceedings/pennock.pdf>, accessed 21 December, 2001.
- Schwartz, C. (1998), "Web Search Engines", *Journal of the American Society for Information Science*, Vol. 49, pp. 973-982.
- Scime, A. (2000), "Learning from the World Wide Web using organizational profiles in information searches", *Informing Science*, Vol. 3 No. 3, pp. 135-43.
- Smith, A. G. (1999), "A tale of two web spaces: comparing sites using Web Impact Factors", *Journal of Documentation*, Vol. 55, pp. 577-592
- Snyder, H. and Rosenbaum, H. (1998), "How Public is the Web?: Robots, Access and Scholarly Communication", *Proceedings of the ASIS 98 Annual Meeting*, pp. 453-462.
- Snyder, H. and Rosenbaum, H. (1999), "Can search engines be used for web-link analysis? A critical review", *Journal of Documentation*, Vol. 55, pp. 375-384.
- Spink, A., Jensen, B.J. and Ozmultu, H. C. (2000), "Use of query reformulation and relevance feedback by Excite users", *Internet Research: Electronic Networking Applications and Policy*, Vol. 10 No. 4, pp. 317-28.
- Thelwall, M. (2000a), "Commercial Web Sites: Lost in Cyberspace?", *Internet Research*, Vol. 10 No. 2, pp. 150-159.
- Thelwall, M. (2000b), "Web Impact Factors and search engine coverage", *Journal of Documentation*, Vol. 56, pp. 185-189.
- Thelwall, M. (2000c), "Effective web sites for small and medium-sized enterprises", *Journal of Small Business and Enterprise Development*, Vol. 7 No. 2, pp.149-159.
- Thelwall, M. (2000d), "Who is Using the .co.uk Domain? Professional and Media adoption of the Web", *International Journal of Information Management*, Vol. 20 No. 6, pp. 441 - 453.
- Thelwall, M. (2001a), "Extracting macroscopic information from web links", *Journal of the American Society for Information Science and Technology*, Vol. 52 No. 13, pp. 1157-1168.
- Thelwall, M. (2001b), The Responsiveness of Search Engine Indexes, *Cybermetrics*, Vol. 5 No. 1. <http://www.cindoc.csic.es/cybermetrics/articles/v5i1p1.html>, accessed 17 December, 2001.
- Turau, V. (1998), "What practices are being adopted on the web?", *Computer*, May, pp. 106-108.