A publicly accessible database of UK university website links and a discussion of the need for human intervention in web crawling

Mike Thelwall
School of Computing and Information Technology, University of Wolverhampton,
Wulfruna Street, Wolverhampton, WV1 1SB, UK.
m.thelwall@wlv.ac.uk

Abstract

This paper describes and gives access to a database of the link structures of 109 UK university and higher education college websites, as created by a specialist information science web crawler in June and July of 2001. With the increasing interest in web links by information and computer scientists this is an attempt to make available raw data for research that is not reliant upon the opaque techniques of commercial search engines. Basic tools for querying are also provided. The key issues concerning running an accurate web crawler are also discussed. Access is also given to the normally hidden crawler stop list with the aim of making the crawl process more transparent. The necessity of having such a list is discussed, with the conclusion that fully automatic crawling is not socially or empirically desirable because of the existence of database-generated areas of the web and the proliferation of the phenomenon of mirroring.

Keywords: Web links, Web Impact Factor, search engines, web crawlers.

Foreword

This paper follows the lead of Rousseau and Rousseau (2000) in being a hybrid of a traditional scholarly article and a resource for academics, but has shifted the balance further towards providing the resource. Although it is not normal practice to publish an electronic resource in an (electronic) journal, it is believed that in this particular case it may be of use to the Cybermetrics research community, and that its publication in the eponymous journal will allow for public review, including criticism, and wider dissemination.

Introduction

Individual hyperlinks are not only the glue that holds web sites together but also the means for authors to refer to other sites and for surfers to visit them. Computer programs also use the same links to retrieve large numbers of pages in a process known as crawling or spidering. The end result of this may be a search engine database, another type of navigation aid, data for research, or even a spam email list.

Collectively, web links can be used to extract more information such as the existence of highly interlinked communities of page authors, and the identities of the most popular link target choices. This latter idea underpin's Google's highly

successful ranking algorithm (Brin & Page, 1998). The analogy between citations and hyperlinks has not gone unnoticed by computer scientists and has lead to bibliometric speculations by information scientists, including the creation of a metric, the Web Impact Factor, based upon link counts (Ingwersen, 1998). Recent results correlating a web link metric with research ratings for UK universities (Thelwall, 2002a) confirm that this is a promising area of research. As with the much older field of citation analysis, many questions must be asked and answered about the reasons for web link creation and their distribution before any derived metrics can have their validity assessed with a degree of confidence.

Information about web links over large web sites can be obtained by building a web crawler, but for those without access to this type of program the results of someone else's may be used. The easiest way is to use the advanced search feature of a search engine such as HotBot or AltaVista, but this approach has problems (Bar-Ilan, 1999; Rousseau, 1999; Smith, 1999; Snyder & Rosenbaum, 1999; Thelwall, 2000; Bar-Ilan, 2001a). The fundamental problem with this is a lack of knowledge of how the engine's crawler covers the web and how it extracts results from its database. Indeed AltaVista, which appears to be the most reliable and with the largest coverage, uses approximation techniques that return counterintuitive results. These include a request for more results leading to less, refining a search but getting more matches, and repeating the same query and retrieving a completely different answer. These anomalies have not stopped AltaVista's use in statistically significant link metrics (Thelwall, 2002b), yet they mitigate against its unguarded use as a primary data source. A second method of obtaining link data is to use the Internet Archive (archive.org). This is a far-sighted attempt to create an archive of databases from a commercial crawler and other sources that can be used by interested scientists, without charge. The current disadvantages of this are the source of the data and the difficulty for non unix programmers in writing code to extract the desired information.

Bar-Ilan (2001b) has called for the information science community to have its own web crawler so that all can have access to transparently collected data without clogging the Internet with numerous personal crawlers. This paper seeks to provide an interim stopgap measure in the form of a database of the link structure of the UK university web. This is claimed to be the ideal source for such an exercise for several reasons: academic use of the web is more mature than other sectors; academic web sites provide an opportunity for close comparison with academic articles; the size of the body of UK universities is manageable, yet large enough for statistical measures; the UK has an official research assessment exercise that can be used to assign scores to individual universities. More information and statistics are available from the Times Higher Education Supplement (Mayfield University Consultants, 2001).

Tools are also provided to enable link-based queries of the database to be conducted. These allow summary statistics or complete listings to be extracted for all pages with target URLs or domain names containing specified text. This mimics and extends the functionality of AltaVista's link: command.

The Crawler

A web crawler works by recursively, requesting web pages, extracting URLs from the HTML and repeating the process with each new URL. A check may be performed before requesting a new page to see whether it is in the site of area being covered. Behind this simple description lie many choices that must be made. The functioning

of the crawler as a distributed system has already been described in Thelwall (2001a), but a recap of the essential features is given here.

Web Page Qualification

Web pages are requested over the Internet using a mechanism known as the HyperText Transfer Protocol (HTTP). To avoid unnecessary downloading of non-HTML pages not containing links, before requesting it, the same protocol is used to ask for identifying information (the HTTP HEAD request). If this does not identify the page as an HTML document then it is ignored.

URLs are taken as qualifying for a given site if the domain name part ends in known domain name of the site being crawled. Thus, www.scit.wlv.ac.uk/index.html and www.wolverhampton.ac.uk/lib/ would both qualify for Wolverhampton University (wlv.ac.uk and wolverhampton.ac.uk).

Web Link Extraction

URLs can be indicated in web pages in many ways. A minor technical issue is that they can be specified in full or partially only, to be interpreted relative to the URL of the current page.

URLs can be specified in the HTML in three ways:

- The anchor tag, including a client-side image map
- A meta tag in the head of the HTML indicating redirection to an alternative page
- The frame tag
- Outside the HTML URLs can be accessed in the browser in three further ways
- Through HTIP header redirection requests
- Through embedded programs running through the browser, such as JavaScript, Java, Shockwave and Flash.
- Through non-HTML document types with a hyperlinking feature for example MS Word or PDF documents accessed on the web
- Through server side image maps.

The crawler supports all of the above except for the last three. It is not possible to easily extract URLs from embedded programs since these may be built by the code itself when running and so a decision was made to ignore all such links. This would mean that a site using this kind of technology without the backup of HTML links would not be covered completely.

Non-HTML documents can still be web pages if they are delivered by a web server and have an URL. The decision not to attempt to extract links from any non-HTML documents was essentially a practical one of not having the resources to cope with the increase in complexity for supporting types less simple than HTML.

Server side image maps are pictures in web pages that form an URL from the coordinates of the point that a user clicks on. A program on the server returns a page based upon the coordinates sent. Although it is possible to crawl every possible pair of coordinates for any server side image map found, it is impractical to do this.

Server Programs

Some web pages have their URLs formed by user actions on the previous page. This is the case when a form is filled in. A simple example of this, used on one university home page, is to have a drop-down list of choices instead of a selection of links. The information chosen can be sent to the server either as part of the URL (a ‘GET’ request) or as information sent with the URL, but not forming part of it (a

'POST' request). In the latter case many different pages can have the same URL. The same mechanism can send information typed by the user, for example a search query. Although pages with user selection components but not text boxes, using the URL encoded method could be deduced from the HTML, a policy decision was made not to attempt to do this.

Dealing with Errors

One aspect of a web crawler that is often not of interest to those other than its programmers is its algorithms for dealing with errors. There are essentially two kinds: Internet transfer errors and HTML errors. The former case encompasses all events that can prevent a complete web page from being received by the crawler, whereas the latter includes all mistakes in the web page designs.

The most common Internet transfer error occurs when an URL is sent for a web page that does not exist. This can be identified when the HTTP header include the code 404 for 'file not found'. A simple error message may also be sent in an accompanying HTML page. Such pages are flagged as missing by the crawler in its database. This is an example of an error with a definitive cause, but other errors are impossible to fully diagnose automatically. One example of this occurs when a web server is switched off. A request for one of its pages will not be returned with an error message: it will just not be answered. The reason for a request not being answered within the default time period (60 seconds) could be that the server is offline permanently or temporarily, or even that its part of the Internet is particularly busy.

This type of error can be dealt with by a large crawler by fetching other pages simultaneously, so that delays in responses to requests to one server do not greatly slow down the overall crawl rate (known as multithreading). This is not practical for an individual site crawler that requests only one page at a time (for politeness), because long delays can result when many URLs reference an unresponsive server. The technique used is to put URLs with inconclusive errors at the back of the downloading queue so that they can be tried later after several tries the server is listed as unavailable and its subsequent pages ignored.

HTML is an official language with agreed rules for elements including links. This does not stop designers from making errors in their pages, and some are so common that web browsers automatically correct them. One example is forgetting to close quotes at the end of a tag. The crawler does not stick rigidly to official HTML, but attempts to correct as many errors as possible.

Human Intervention

It is impossible to run a web crawler fully automatically if its purpose is to cover an entire site. This is because web servers can create new pages upon request, including links to further new pages. A simple example of this is an online calendar linked to a database of activities. The address of each page in this example could be an encoded version of the date and it could contain links to the next day. This situation occurred at one university, with the crawler eventually requesting pages from 2030 before human intervention stopped it. During the crawling process each log was periodically checked to ensure that no infinite crawling was occurring. When this check was made, a further check was conducted for large sets of pages without external links, for example daily server access logs. Both of these were added to a banned list of areas that the crawler is instructed to ignore. See appendix 1 for a full list. In the banned list were also included any identified large mirror sites where the hosting university was not the creator of the content. The banned list has both a

practical function: to ensure smooth fast crawling, and a courteous purpose: to avoid unnecessarily using the target servers. Maintaining the list was hugely expensive in terms of time. Each of the entries in the list involved tracking the problem source from the database and then visiting several pages on the web site to identify the scope of the problem and the consequent correct entry in the banned list file. Commercial search engines may well employ heuristics to perform this identification (Bharat *et al*., 2000), but this option was not available since the crawl was not to be of a significant proportion of the entire web.

An important point to make is that the use of the banned list changes the external link count for a site because of the banning of mirror sites. A commercial web crawler may be expected to take measures to avoid crawling multiple versions of very large mirror sites, but the inevitable partial implementation of such a policy does create a source of uncertainty in their results. This is not completely avoided in the database published here because it, too relies on human intervention to identify the problem. In principle, mirror sites could be left in the results, but these would cause problems in subsequent link analyses. The reason is that it is common practice to put a link to the home page of the creating organisation on all or many of the mirrored pages. This creates a large anomaly when counting links. Banning these pages is intellectually justifiable, however, on the grounds that the work is not produced by the hosting institution.

Special Cases of Crawls

Two sites constitute special cases: the universities of Liverpool and Cambridge. Liverpool bans all robots from crawling its main site except those belonging to the major search engines. The crawl submitted here starts at an individual departmental web site instead of the main one, and may well miss important parts of the site, in addition to the main domain, through not finding links to any pages in them.

The Cambridge site was crawled normally except for the varsity.cam.ac.uk subsite. This is an online magazine which maintains a database of past publications accessible online through URLs. It was decided to crawl only the current issue of this site and not to allow large portions of it to be covered through a few links to a back issues, since it is essentially an archive resource. This was achieved by crawling Cambridge with the site banned and then crawling the subsite from the home page, concatenating the results.

The Web, the Publicly Indexable Web and Crawler Coverage

For the average web user, “The World-Wide Web” probably means everything on the Internet that they can access through their web browser. This would include non-web services, in particular FTP (File Transfer Protocol), that may integrate into their browser, but also document types that are not created using the official language of the web, HTML, including Adobe’s Portable Document Format, pictures and video files. A more technical definition would be all documents that are publicly available on the Internet through requests encoded using the official protocol of the web (the HyperText Transfer Protocol, HTTP, as observable in the http:// that starts web URLs).

The web as perceived by a user is different to that indexed by robots, including those building search engine databases. There are several reasons for this.

Protected Areas

Many areas of the web have access blocked to the general public. This is normally achieved in one of two ways: by password protection; and through only allowing requests from a specified set of computers, as identifiable by their Internet Protocol address. The latter can be used to give access to digital libraries to all computers in subscribing institutions, for example.

Databases and Dynamically Generated Pages

Web pages can be created by computer programs in response to user actions. The classic case of this is the HTML created by search engines in response to keyword queries. There are many other applications, however, including searchable online commercial products catalogues and searchable email directories. A robot could not index sites of this type.

Non-Standard Linked Pages

Some web pages contain pathways to access other pages that do not use the standard HTML linking mechanism. Some of these pathways are impossible for robots to discover, whereas others impose an additional computational overhead that the robot owner may choose to avoid. A page that was only linked to in a non-standard way would not be found by a robot unless it was already in its starting list of URLs to crawl (for example if it had been submitted to a search site by its owner). An example of non-HTML access method is the hyperlinks in PDF files. These are indexed by Google, but not by any other general commercial search engine known to the author at the time of writing. An example of an impossible to index type is JavaScript links. It is increasingly common to find web pages where links are implemented by JavaScript programs through clicking on pictures or normal links, or by selecting an option from a drop-down list. The indexing difficulty is that the program may build the URL to be accessed through a series of computational steps. It is fundamentally impossible to write a program to understand what another program is doing, and very difficult to even approximate this capability, and so it is believed that no search engines make a serious attempt to do this.

Banned Areas

The web has an official mechanism for web site owners to ban robots from visiting and indexing their site, the robots.txt convention. This is often used to keep crawlers away from internal company information that is not sensitive enough to password-protect, and also to avoid runaway crawling of dynamically-generated pages.

Overview

A robot starts with a seed set of known URLs and is able to visit URLs that it is allowed to access and that it can discover by examining the pages downloaded. It can only cover a subset of the web, but it is impossible to give an estimate of the fraction of the web that is indexable by robots because dynamically generated pages make web size indeterminate. The seed set may be just the home page of the site, or a list of URLs found by a previous crawl or by some other method.

The key question to be addressed by researchers using commercial search engines or web crawlers to study the web itself is whether the portion covered forms a meaningful entity for study. Robots can only visit pages that they know the URL of and so cannot even cover all of the web that they are allowed to find. Lawrence and

Giles (1999) estimated that commercial search engines covered no more than about 16% of the web that was possible to find through link extraction from HTML and given the home page of the site.

There are two arguments that the area cover by crawlers is meaningful and worthy of scholarly attention. Firstly, search engines are important information retrieval tools in themselves. Secondly, correlations established between link count measures and a non-Internet phenomenon (research ratings) for UK universities support the hypothesis that commercial search engines and independent crawlers coverage of this area of least is not too arbitrary to be worth investigating. In the longer term, the importance of search engines as a potential source of new visitors to find a site is a factor that is conducive to content being crawlable by robots.

Summary

It is hoped that the database and tools provided here will prove useful in forwarding the understanding of web link phenomena and of the limitations of attempts at comprehensive web crawling. In addition, it is hoped that the publication of the banned list can provide additional information to non-specialists about the restrictions on crawlers that are necessary to allow them to function effectively on large sites.

References

Bar-Ilan, J. (1999). "Search Engine Results over Time - A Case Study on Search Engine Stability." **Cybermetrics**, 2/3. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p1.html

Bar-Ilan, J. (2001a). "How much information the search engines disclose on links to a web page? – A case study of the 'Cybermetrics' home page." **Proceedings of the 8th international Comverence on Scientometrics and Informetrics**, 1, 63-73.

Bar-Ilan, J. (2001b). "Data collection methods on the Web for informetric purposes - A review and analysis." **Scientometrics**, 50(1), 7-32.

Bharat, K. Broder, A. Z.,Dean J. and Henzinger, M. R. (2000). "A comparison of techniques to find mirrored hosts on the WWW." IEEE Data Engineering Bulletin, 23(4), 21-26.

Brin, S. and Page, L. (1998). "The Anatomy of a large scale hypertextual web search engine." **Computer Networks and ISDN Systems**, 30(1-7), 107-117.

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

Lawrence, S. and Giles, C. L. (1999). "Accessibility of information on the web, **Nature**, 400, 107-109.

Mayfield University Consultants, (2001). "League Tables 2001", **The Times Higher Education Supplement**, May 18, T2-T3.

Rousseau, B. & Rousseau, R. (2000). "LOTKA: A program to fit a power law distribution to observed frequency data." **Cybermetrics,** 4.

http://www.cindoc.csic.es/cybermetrics/articles/v4i1p4.html

Rousseau, R., (1999). "Daily time series of common single word searches in AltaVista and NorthernLight." **Cybermetrics**, 2/3. http://www.cindoc.csic.es/cybermetrics/articles/v2i1p2.html

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

Snyder, H. and Rosenbaum, H. (1999). "Can search engines be used for web-link analysis? A critical review." **Journal of Documentation**, 55(4), 375-384.

Thelwall, M. (2000). "Web Impact Factors and search engine coverage." **Journal of Documentation**, 56(2), 185-189.

Thelwall, M. (2001a, to appear). "A Web Crawler Design for Data Mining." **Journal of Information Science** 27(5).

Thelwall, M. (2002a, to appear). "Extracting macroscopic information from web links." **Journal of the American Society for Information Science and Technology**.

Thelwall, M. (2002b, to appear). "Sources of links for Web Impact Factor calculations." **Journal of Documentation**.

Appendix 1: The data file format

The data is in a simple structure in a plain text file. The format was chosen to allow it to be loaded easily into text editors as well as desktop spreadsheet and database packages, if desired.

Each line of the file describes either a link in a web page or the (partial) URL of a page crawled. All URLs are shortened by the removal of the initial http:// and domain names starting with www. have the initial www removed. The purpose of this is to slightly reduce the file size.

Lines beginning with a tab character represent links in web pages, the rest represent the pages crawled. The list of links above a page are those extracted from it. Following each page crawled is a tab character and a code number, with the following interpretation.

1: A normally downloaded web page, with its links on proceeding lines.
2: A non-definitive error was encountered downloading the page, (normally an unreachable server) and the attempt to download the page has been abandoned.
3: A HTTP redirection header, with the redirected page on the previous line (the full URL is always given for the new page).
4: [Not used.]
5: An error downloading the web page that could not be resolved (such as "file not found", or the page is of a non-HTML type.

The following section of code shows the web page with URL http://www.comp.glam.ac.uk/teaching redirected to http://www.comp.glam.ac.uk/teaching/ followed by an aborted web downloaded attempt of http://www2.glam.ac.uk/sot/doms/Research/cgrowth/expnotes.pdf due to its non-HTML type (although from the log it could also have been due to a download error since both are recorded as type 5), followed by the page http://www.comp.glam.ac.uk/pages/staff/tdhutchings/chapter8.html which had 2 URLs extracted from it.

```
      http://www.comp.glam.ac.uk/teaching/
.comp.glam.ac.uk/teaching     3
www2.glam.ac.uk/sot/doms/Research/cgrowth/expnotes.pdf      5
      .comp.glam.ac.uk/pages/staff/tdhutchings/Default.htm
      .comp.glam.ac.uk/pages/staff/tdhutchings/chapter8/sld010.htm
.comp.glam.ac.uk/pages/staff/tdhutchings/chapter8.html      1
```

Note that redirections are conducted transparently in most browsers so the type 3 event would not normally be noticed by a web surfer, unless they noticed the URL change automatically in the browser.

Appendix 2: The banned list

The banned list file banned.ini contains a list of partial URLs of banned areas for each web site for which it was considered necessary. The format is almost identical to that of the official robots.txt agreement. Essentially if any of the listed partial URLs match the start of an URL, it is banned and will not be crawled. The bracketed expression at the top of a list identifies the university that it applies to and the URL shortening convention is automatically applied.

Here is a simple example that bans all URLs which start with any of the three identified strings.

```
[st-and.ac.uk]
benromach.dcs.st-and.ac.uk/Teaching/Java/JDK_Docs
www-hons.dcs.st-and.ac.uk/mirrors
star-www.st-and.ac.uk/starlink/stardocs
```

The URLs http://benromach.dcs.st-and.ac.uk/Teaching/Java/JDK_Docs/index.html and http://www-hons.dcs.st-and.ac.uk/mirrors/first/second.html would be banned, but the following would not because they do not contain all of the partial URL.

- http://benromach.dcs.st-and.ac.uk/Teaching/Java
- http://benromach.dcs.st-and.ac.uk/Teaching/Java/JDK_Doc/list.htm

Appendix 3: The affiliated programs

The user is warned that these programs may take a long time to complete their calculations: several hours or even days for some requests. The author apologises for the poor user interface and documentation provided, claiming in mitigation that the emphasis was on publishing them in a timely manner.

The link extractor

This program will process all of the files and list or summarise link information. The following options are supported.

- Listing the source and target file for each matching link, or returning counts of the number of matching source or target pages.
- Restricting the results to only external links or only internal links (based upon the domain name versions identified in the file domain_names.ini).
- Restricting the results to only URLs or to only domain names containing the specified text.

The purpose of this program is to allow researchers to calculate summary statistics for their own area of web link interest.

The sub-site extractor

This program splits one web link file into two parts. A text string must be entered and a link file selected. The program will then create two sub-files, one with all pages containing the text and one with the remainder. For example, to extract the link structure of the School of Computing web site from the University of Wolverhampton web site file, the text .scit.wlv.ac.uk would need to be entered and the original Wolverhampton web link file selected. The program can also be used to remove all web pages from a given area from a file, for example if it is discovered that some of the pages are from a previously unidentified link site.

Performing advanced queries

The tools provided do not give access to the flexibility of full Boolean querying as provided by AltaVista Advanced Search, for example. Many types of queries can be built using them, however, and some examples will be given to illustrate the process.

To find all links in Cambridge university that point to any academic web site in New Zealand

This information may be expressed in the AltaVista query

```
(host:cam.ac.uk OR host:cambridge.ac.uk) AND link:ac.nz
```

although AltaVista's link command is not powerful enough to restrict its choice of ac.nz to only the domain of the link.
To find information on this, the link extractor program can be used with the match text .ac.nz, the choice of "domain only" for the text and the selection of the Cambridge file only for processing.

To find all links in UK universities that point to any web site in Papua New Guinea

This information may be expressed in the AltaVista query

```
host:ac.uk. AND link:png
```

but AltaVista may return many erroneous results, links to png graphics files.
To find information on this, the link extractor program can be used with the match text .png, the choice of "domain only" for the text and the selection of all files for processing.

To find all links in West Midlands universities that point to any URLs containing the text "Java"

This information may be expressed in the following AltaVista query.

```
(host:wlv.ac.uk. OR host:wolverhampton.ac.uk OR host:bham.ac.uk OR
  host:birmingham.ac.uk OR host:aston.ac.uk OR host:uce.ac.uk OR
 host:warwick.ac.uk OR host:cov.ac.uk OR host:coventry.ac.uk) AND
                          link:Java
```

To get results from the database, move all of the Midlands universities' data files into a separate folder and point the program at this, with the match text "Java" and th option to search the entire URL.

Appendix 4: Access to the database and programs

The database is available at
`http://www.scit.wlv.ac.uk/~cm1993/cybermetrics/database/`
together with the necessary files and programs. It is provided as a very large (199Mb) and self-extracting zip file. Running this will expand it to 2.3Gb of plain text files, one for each institution covered. Each file is named after the domain name of the institution, and contains a numerical string that can be ignored. The names of the files should not be changed because the processing programs need to extract the domain names from the file names.

Many may find the large file download a problem and so the author undertakes, for a minimum of two years, to provide the files on a CD-ROM to bone-fide researchers upon receipt of an empty CD-ROM case and self-addressed packaging for it. Postage and the CD will be provided free of charge.