The Application of Sampling to the Design of Structural Analysis Web Crawlers

Mandlenkosi Victor Gwetu

Business Information Systems Department, University of Venda
Thohoyandou, Limpopo, South Africa

Abstract

The growth of the World Wide Web (WWW) has seen it evolve into a rich information resource. It is constantly traversed with the aid of crawlers so as to harvest web content. When collecting data, crawlers have the potential of causing service denial to web servers. This paper proposes the application of sampling as a selection strategy in the design of structural analysis web crawlers. This has the benefit of alleviating the problems of bandwidth costs to web servers whilst retaining the quality of the data that is mined by crawlers. The initial results of this study are promising and are presented in this paper.

Keywords: web crawler, sampling, web server, denial of service attacks.

1. Introduction

At its inception, the internet was a resource for academic purpose [1]. It has now grown to be a network offering services such as ecommerce, social networking and search engines on the WWW. As the WWW has grown it has increasingly become a target of data mining mainly for its content. The focus of this paper is agents called web crawlers whose primary goal is traversing the WWW in search of data for processing. Web crawlers normally explore the web in search of either textual content or structural data. They are used by search engines as well as for research purposes. Search engines collect web content from documents on the WWW in order to assist users in finding the content they search for. Research on the WWW is warranted by a growing need to analyze the content and the structure of the web in order to highlight trends and tendencies. Examples of the use of structural data from websites include surveys on the most popular web server and link analysis. This paper focuses only on web crawlers used for structural analysis.

Web sites benefit from web crawlers visiting their sites because they help market their existence. The problem however is the potential of web crawlers causing the denial of services to these websites. A web crawler can put strain on a website that is on a server with low bandwidth such that it can no longer reliably service requests from users on the internet. Fig. 1 illustrates a potential set up in which a web crawler requiring high bandwidth traverses hosts that are on networks with low bandwidth. The robots exclusion protocol [2] was proposed as a guide to crawler designers to know the parts of a website that the site owners do not want to be crawled. The protocol however does not specify issues such as speed of access and service denial.

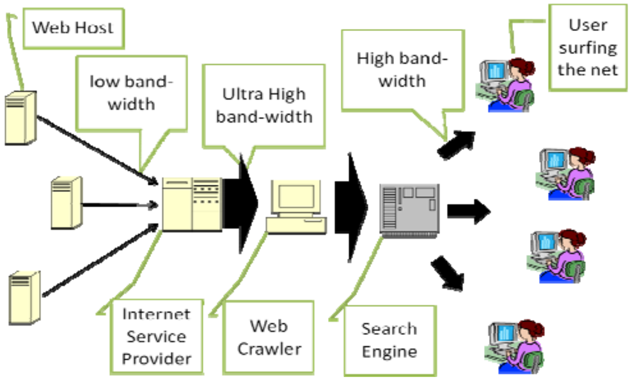

Fig. 1 Bandwidth congestion

This paper chronicles some of the different research that has done on web crawlers and then proposes the application of sampling to web crawling. The benefits of this research include the reduction of denial of service attacks on web servers in areas of low bandwidth, especially in developing countries.

In order to design an effective full scale web crawler for use on the WWW, huge computational resources are required due to the vast number of pages to crawl. This research will enable the design of web crawlers that can function on less computational resources and thus open up the area of web crawler implementation to researchers with limited resources.

Due to the larger number of pages to crawl, web crawlers often run for long periods of time. The use of sampling will mean fewer pages are visited therefore web crawlers will have shorter run times. The research problem to be addressed by this paper is the improvement of the design of web crawlers so as to reduce service denial attacks, computational resources and download times. All this needs to be done in a manner that does not compromise the quality of data that is mined by the web crawlers.

2. Literature review

Due to the popularity of search engines on the web, the role of web crawlers has become all the more critical in order to facilitate accurate and relevant results. A significant amount of research has thus been done in this area. This research mainly centers on the design of web crawlers as well as the implications of their use.

One of the pioneering studies in this area was the work of Lawrence Page and Sergey Brin in the initial stages of the design of the Google search engine. They stated that "running a crawler which connects to more than half a million servers, and generates tens of millions of log entries generates a fair amount of email and phone calls" [3]. This is because their web crawler focused on computational power and relied on the robots exclusion protocol to determine which sites to crawl. They also state that some people "do not know about the robots exclusion protocol, and think their page should be protected from indexing". The existence of the robots exclusion protocol does not translate to its use by all web sites. As such, an approach that is sensitive to web server performance and band width scarcity is needed to protect those ignorant of the robots exclusion protocol.

Eichmann has recommended the following general advice, "the pace and frequency of information acquisition should be appropriate for the capacity of the server and the network connections lying between the agent and that server" [4]. Mike Thelwall also encourages web crawler designers to look beyond legal implications and consider ethical issues [5].This has motivated the current study to find a solution that facilitates this ethic.

The concept of the application of sampling to web pages has been considered before, albeit to web archiving. Jared Lyle proposed using purposive, systematic and random sampling methodologies in Web archiving [6]. His research tested sampling as a viable means of appraising electronic records, for a lasting and useful present and future. Random samples were used to appraise over four million Umich.edu related documents and results showed that random sampling of personal documents gave a representative sample. The researcher's proposed research extends this idea to web crawling and seeks to find out whether sampled pages will be representative of a whole website. The focus in this paper is also unique in that it majors on structural components of a website as opposed to its content.

The study done by David Gibson et al [7] recognized that web pages contain a combination of unique content and template material. Template material is present across multiple pages and is used primarily for formatting, navigation and branding. They then studied the nature, evolution, and prevalence of these templates on the web. The results showed that about forty to fifty percent of the material on the web is template and that the contents of the templates such as text and links are growing at a rate of between six and eight percent per year. They crawled websites in order to study the prevalence of templates where as this study uses their work as a premise. We argue that if templates are as prevalent as up to 50% of the material on the web then trawling a complete website should contain recurring features. If a sample of websites is appropriately selected then it should be possible to capture the attributes of a website using fewer web pages.

3. Methodology

To investigate the viability of the use of sampling in web crawler design, a simple crawler that utilizes sampling is to be designed. A sample of 200 websites is to be used to test the performance of the crawler. This sample will be collected from web directories from different countries. One web directory per country and two countries are to be selected from each continent. In the end 20 websites will be sampled from each of the chosen directories. Stratified sampling is to be used in selecting the sites to ensure that the samples are representative of different sub-domains such as .ac and .co. The overall sample would then be crawled to collect data for analysis from each page in the site. The same will be done for a sample of the pages on the site and the results will then be compared to see if sampling gave results that are similar to those of the full crawl.

The functionality of a prototype of the crawler is illustrated in Fig. 1. The crawler only analyzes pages for the number of images, forms and external links. In addition to using sampling the crawler respects the robots exclusion protocol by not visiting all the Uniform Resource Locators (URLs) contained in the robots.txt file. A delay of one second is intentionally observed between page crawls as a means of further reducing the likelihood of using up valuable web server bandwidth.

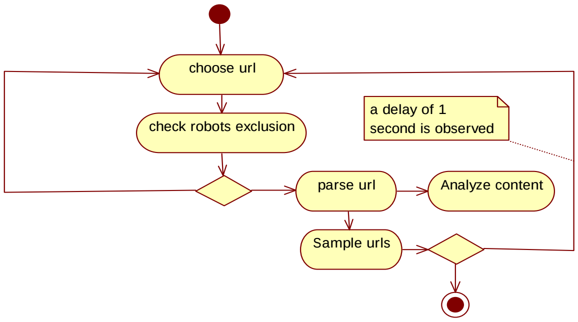

Fig. 2 The design of a simple crawler that uses sampling

4. Results

The crawler prototype simply samples alternate URLs in the order they appear on each web page. A test was carried out using ten websites randomly selected from a South African web directory. The selected sites all consisted of less than 20 pages and were equally divided between .co and .ac sub-domains. This pretest was carried out to assess the viability of the methodology. The results are shown in Fig. 3. Even though the sample is small and not representative of all sites on the WWW the results are promising. They show that the structural data collected from a full crawl was very similar to the data from a partial or sampled crawl.

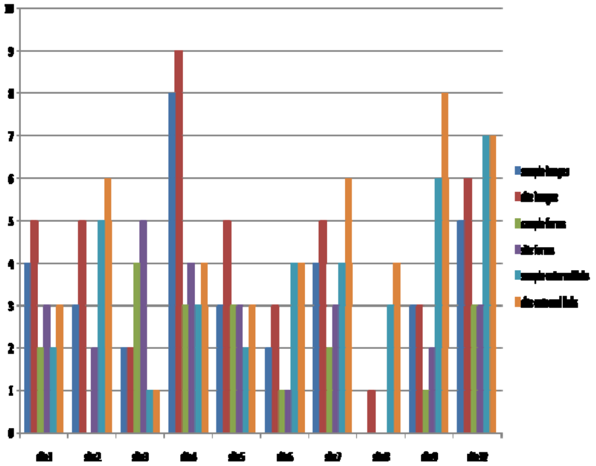

Fig. 3 Web crawl results

4. Conclusions

There is a lot of duplication on the web and thus it may not be necessary to traverse the whole web. The use of sampling will mean fewer pages are visited and therefore web crawlers will have shorter run times. The problem addressed by this research is the improvement of the design of web crawlers so as to reduce computational resources and download times during crawl times. To investigate the viability of the use of sampling in web crawler design as a means of achieving this, a simple crawler that utilizes structural content sampling was designed.

A major sample of 200 websites is to be used to test the performance of the crawler. A minor sample of 20 websites will be selected from the major sample using stratified sampling. Both samples will then be crawled to collect data for analysis. The major sample will go through a full crawl while the minor sample will undergo a partial crawl. The results of a pretest of this exercise show that the use of sampling in crawling gave similar results as the full crawl.

Further work needs to be done to introduce multi-threading in the crawler so as to enable it to crawl more than one site at a time. Also investigation into different sampling methods needs to be carried out in order to improve the results.

References

[1] P. J. Dennin, A. Hearn, and C. W. Kern, History and overview of csnet, ACM SIGCOMM, 1983, pp. 138-45

[2] M. Koster, A Standard for Robot Exclusion, 1994, Last accessed 7 June 2009 from: http://www.robotstxt.org/wc/norobots.html

[3] S. Brin, and L. Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, 2009, Last accessed 7 June 2009 from: http://infolab.stanford.edu/~backrub/google.html

[4] D. Eichmann, Ethical Web Agents, WWW conference, Chicago, IL, 17-20 October 1994, pp. 3-13.

[5] M. Thelwall, Web Crawling Ethics Revisited: Cost, Privacy and Denial of Service, Last accessed 7 June 2009 from: http://www.scit.wlv.ac.uk/~cm1993/papers/Web_Crawling_Ethics_preprint.doc

[6] J. Lyle, Sampling the Umich.edu Domain, 4th International Web Archiving Workshop, 2004

[7] D. Gibson, K. Punera, and A. Tomkins, The Volume and Evolution of Web Page Templates, WWW 2005, Chiba, Japan, 10-14 May 2005.

M. V. Gwetu has an MSc in computer science from the National University of Science and Technology (2004, Zimbabwe) and is currently working as a lecturer at the University of Venda. His research interests include search engines and web crawlers as well as programming languages. He is a Sun Microsystems certified java programmer as well as a web component developer.