Low-Load Server Crawler: Design and Evaluation

Katsuko T. Nakahira
Nagaoka University of Technology
1603-1 Kamitomiokamachi, Nagaoka
Niigata, Japan
katsuko@vos.nagaokaut.ac.jp

Tetsuya Hoshino
Nagaoka University of Technology
1603-1 Kamitomiokamachi, Nagaoka
Niigata, Japan
065365@mis.nagaokaut.ac.jp

Yoshiki Mikami
Nagaoka University of Technology
1603-1 Kamitomiokamachi, Nagaoka
Niigata, Japan
mikami@kjs.nagaokaut.ac.jp

ABSTRACT

This paper proposes a method of crawling Web servers connected to the Internet without imposing a high processing load. We are using the crawler for a field survey of the digital divide, including the ability to connect to the network. Rather than employing normal Web "page" crawling algorithm, which usually collect all pages found on the target server, we have developed "server" crawling algorithm, which collect only minimum pages from the same server and achieved low-load and high-speed crawling of servers.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Search process; K.4.1 [**Public Policy Issues**]: Trans-border data flow

General Terms

Design, Experimentation

Keywords

Global Digital Divide, Server crawler

1. INTRODUCTION

Broadband access shows an increased penetration every year in advanced countries according to OECD statistics [1]. However, the penetration still remains low in developing countries. In order to conduct a field survey of this digital divide, including the ability to connect to the network, authors are crawling of all Web servers connected to the Internet under the specified country code domains [2]. When crawling servers in developing countries with immature broadband networks, it is important that the traffic generated by crawler itself does not impose a heavy load on the surveyed servers. So far not much work has been done on crawler algorithm which are aware of traffic workload. A work by Shaozhi Ye proposed a workload-aware web crawling algorithm, which enables crawler to adapt its crawling speed based on server workload detection [3]. But it still doesn't decrease total workload to the sever. This paper reports on a low-load crawling method, which traces links in Web sites down to only a certain depth levels in order to reduce the traffic load imposed by crawling without deteriorating the coverage in terms of the number of servers crawled.

WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. FIELD SURVEY METHOD

In this paper, we define the set of pages that can be crawled within the same domain from the starting URL as a" site tree". We define the number of consecutive links traced from the starting URL as the" depth level". First, an HTML file is collected from the starting URL, the URL where crawling is to be started. The links included in the file are classified into internal URLs and external URLs (these will be discussed later) depending on their domain names. Thereafter, HTML files linked by internal URLs are collected recursively. This is repeated until the specified depth level has been reached for each site tree. Thus, link information is collected semi-comprehensively.

2.1 Definitions of internal URL and external URL

An internal URL is one that belongs to the same domain as the starting URL or to one of its subdomains. External URLs are those that are not internal URLs. The domain of any external URL is different from the domain of the starting URL, which suggests that their servers are also different. Since an external URL is outside the scope of the site tree of the starting URL, external URLs are not crawled recursively. They are used as new starting URLs.

2.2 Constraints on query URLs

Query URLs are URLs with query characters attached. An unlimited number of these may be generated, and this can hamper semi-comprehensive crawling. Our method prevents an increase in the number of pages subjected to crawling by limiting the number of crawls for one script to one. F or example, When" /index.php?id=2 " has been crawled, no more URLs in" /index.php " are crawled.

3. SURVEY RESULTS

To reduce the traffic load due to crawling, our method limits the scope of crawling by limiting the crawling depth. In order to find the optimum depth level, we carried out crawling and examined the relation between the depth level and the number of external URLs collected. The number of starting URLs used in this survey was about 3400. These URLs have

ccTLDs of African countries. The maximum depth level of crawling was 20.

3.1 Relation between the number of pages collected and the number of external URLs

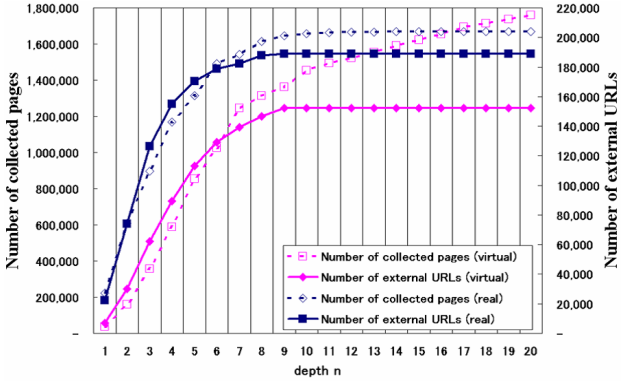

Figure 1: Relation between number of collected pages and number of external URLs

Figure 1 shows the cumulative number of pages collected and the cumulative number of external URLs collected on the vertical axis, against the depth level of sites on the horizontal axis. Rhombuses indicate the total number of pages collected, while squares indicate the number of external URLs. The dotted lines show values obtained in an experiment in a virtual Web space while the solid lines show measured values resulting from crawling in a real Web space starting with the same seed URL as that in the virtual Web space. In the case of the virtual Web space, the gradient of the curve for the cumulative number of pages collected changes at depth level 7 but does not level off even at depth level 20. However, the cumulative number of external URLs collected in the virtual Web space levels off at depth level 9. These trends more or less hold in the measured values as well.

3.2 Distribution of the number of site trees for different maximum depth levels

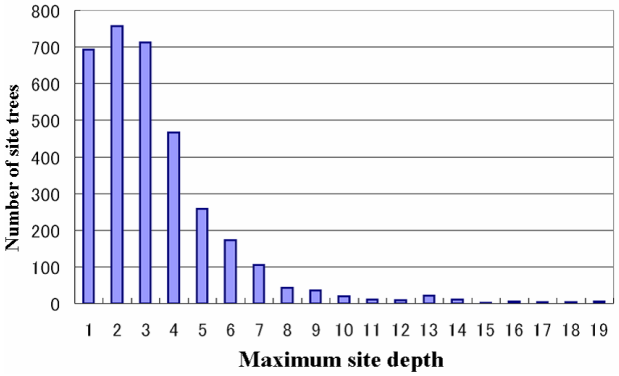

Figure 2: Distribution of number of site trees for different depth levels

Figure 2 shows the distribution of the number of site trees on the vertical axis against the maximum depth level applying for those sites on the horizontal axis. The number of site trees has already become small at depth level 8 and decreases rapidly thereafter.

3.3 Relation between the traffic load and the number of external URLs collected

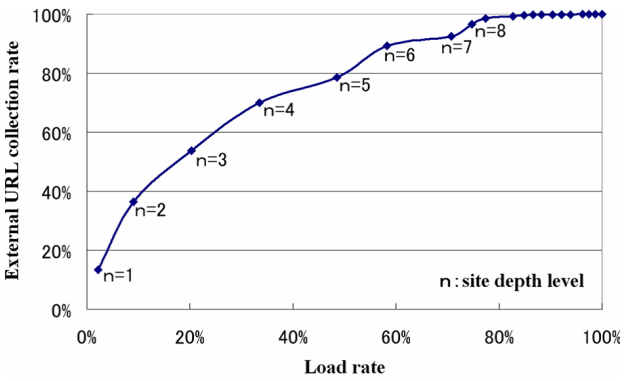

Figure 3: Relation between external URL collection rate and load rate

Figure 3 shows the cumulative percentage of the number of URLs that can be collected on the vertical axis, and the required load rate on the horizontal axis. The 100% load rate is the load applying for depth level 20. The chart shows that a reduction of the load by 20% can be achieved without affecting the number of external URLs that can be collected.

4. CONCLUSIONS

It goes without saying that there is a tradeoff between the processing load of crawling and the number of URLs that can be captured. An optimal balance between the load and the capture rate can be found for a specific network environment by varying the depth level as a parameter. Figure 3 provides a rule of thumb for this. A semi-comprehensive crawling that limits the crawl depth to level 8 can reduce the processing load of crawling by 20%. If we can be satisfied with an external URL collection rate of 80%, the processing load can be nearly halved. Figure 3 enables us to select an optimal crawling strategy for the specific network environment of the servers under consideration and for the specific purpose of crawling. While Figure 3 indicates the capture rate when the crawling depth level is limited, it also provides an indication of the situation if we attempt to capture all URLs.

5. REFERENCES

[1] *OECD Broadband Statistics to June 2006*. OECD, 2006.

[2] Katsuko T. Nakahira et. al. Geographic Location of Web Servers under African Domains. *The 15th International World Wide Web Conference*, 2006.

[3] Shaozhi Ye et. al. Workload-Aware Web Crawling and Server Workload Detection. *Asia Pacific Advanced Network 2004*, July 2004.