Available online at www.sciencedirect.com

ScienceDirect

Procedia Engineering 69 (2014) 492 – 499

Procedia Engineering

www.elsevier.com/locate/procedia

24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013

# A Fast Distributed Focused-Web Crawling

Harry T Yani Achsan[a], Wahyu Catur Wibowo[b]*

[a]*Paramadina University, Jl. Gatot Subroto Kav. 97, Jakarta 12790, Indonesia*
[b]*University of Indonesia, Depok 16424, Indonesia*

**Abstract**

Mining data from a web database becomes more challenging in recent years due to the exploding size of data, the rising of dynamic web, and the increasing performance of web security. Mining data from a web database differs from mining data from web sites because it is intended to collect specific data from a single web site. Collecting a very large data in a limited time tends to be detected as a cyber attack and will be banned from connecting into the web server. To avoid the problem, this paper proposes a crawling method to mine web database faster and cheaper than conventional web crawlers. The method used is to run hundreds of threads from a single web crawler in a single computer and to distribute the threads into hundreds or thousands publicly available proxy servers. This web crawler strategy highly increases the speed of mining and is more secure than using single thread of web crawler.

*Keywords:* focused web crawler; proxy; multi thread; web database

## 1. Introduction

Every search engine in the internet must have at least one web crawler. The other names of web crawler are web spider and bot. The main task of a web crawler is to crawl every web pages fed to it. It will retrieve the contents of web pages crawled, and parse it to get the data and hyperlinks. It then continues to crawl the found hyperlinks. The parser sends data to the indexer and saves it into the database. A search into a search engine actually does not search into the real web site, but it searches the search engine database. The output of search engines is a list of snapshots of web pages with its hyperlinks. The user should open web pages he needs by clicking the hyperlinks.

* Corresponding author. Tel.: +62 818 0854 0094; fax: +62 21 799 3375.
*E-mail address:* harry.achsan@paramadina.ac.id

doi:10.1016/j.proeng.2014.03.017

Since the output of search engines is a list of snapshot of web pages, it is impossible to find any specific structured data using search engines. For example, we cannot gather all conversation made by someone in a forum, or friends of friend data in a social network. Using general search engines like Bing, Hotbot, Google, or Yahoo, we cannot collect all book data with Library of Congress control number between 89211901 and 89211999.We have to use a special web crawler to collect specific and structured data called Focused Web Crawler.

Focused web crawler, sometimes called vertical or specific web crawler, is a tool for mining specific data from web databases. The data mined are structured or semi-structured because it is retrieved from database in web sites such as databases from social networks, forums, blogs, online libraries, online stores, or any web sites that use database to display their information. If we can collect the data from a web site, then we can retrieve the information and discover the knowledge contained in it.

Collecting specific data from a web site is trickier than gathering contents of web pages, because there is sometimes no hyperlinks to/among each entity that makes regular crawler unable find it. For instance, if we need data of all books about automation in Library of Congress web site (http://www.loc.gov), we usually do it using search engines by supplying a key words “site:loc.gov automation book”. Google and Yahoo gave no matched result in the first 300 of its results. On the other hand, the search facility in the home page of Library of Congress web site gave 806 matched results, but we have to retrieve41 web pages of its result set to get the data. Focused web crawler also gave 806 results with 100% matched and automatically put the data into its database.

A focused web crawler has to run carefully. Some web sites have a monitoring tool to watch every visitor. The tool can differentiate between human and bot by using special algorithms and can monitor data transfer rate for each visitor. It can also ban user’s IP address if the user violate any of web site restrictions. Web site restriction can be found in the file robots.txt in the root directory of web site. It consists of all directories and web pages forbidden to be crawled. If the monitoring tool detects any bot or crawler opening any of restricted web pages, it will ban the crawler IP address from accessing its web site. If the crawler originates from a university network, the consequence is fatal, no one can access the web site anymore.

The non existence of robots.txt file does not mean that there is no restriction. The monitoring tool can still watch the visitor behavior. It will check every visitor data transfer rate and compare it to the maximum data transfer rate allowed. It also monitors the duration a visitor interacts with the web site continuously, and check the size of the data transferred. All visitor behavior will be compared to fair usage. Any breach to fair usage will have consequences. These restrictions make focused web crawler do their job very slowly. It can take weeks or months to collect tens or hundreds of thousands entities.

The aim of this paper is to develop algorithms for fast focused web crawler that can run safely. It will be achieved by using multi-threaded programming and distributed access via proxy servers. This paper will also show how to retrieve pairs of IP address and port of public proxy servers and how to crawl nicely.

## 2. Related works

Focused web crawler plays an important role in information society. It is used to crawl social networks [1], to crawl forums [2], to crawl web pages in specific language [3], to browse offline, to mirror web site [5], to generate web site map [6], and to develop Business Intelligence [7]. Many people need it, and some people give the software for free [8-9] and free to try [4-6]. Unfortunately, not all of focused web crawler’s behaviour is in accordance with netiquette [10], resulting in many web sites implement user agent monitoring tool [11] to reject impolite crawler. To honour netiquette, focused web crawlers have to be improved.

Improvement of focused web crawler has been done by improving its strategies. Several strategies used by focused web crawler in the last decade has been reviewed and compared [12]. In the recent years, some researchers optimized the precision of focused web crawling results by implementing Bayesian classification [13-16], ontology [17], similarity [18], relevant topic [19], and Genetic Algorithm [34]. The more precision of a crawler makes it less web page visited, less data transfer rate, and more polite. Since many of web pages implement dynamic content, the content which are displayed different from the HTML source code, a number of improvements has been developed [13, 20, 21, 35-37] to overcome the problem.

One of the most important problems to overcome is to increase the speed of crawling politely. A multi-threaded crawler is proposed [22], which can speed up crawling, however it is detected as an impolite crawler if used to