CHAPTER 1

INTRODUCTION

This chapter consists of five (5) sections that elucidate a brief overview of Web Structure Mining and Web Content Mining on the web forums. There are Section 1.1 Background on Web Crawling, Section 1.2 Problems, Section 1.3 Objectives, Section 1.4 Research Questions and Section 1.5 Contribution.

1.1 Current Landscape and Background

In the past 20 years, World Wide Web (WWW) continues to grow at an overwhelming rate until it has become an indispensable tool in human daily life especially web forums. With just a few finger taps and mouse clicks, the topic-related information on the screen will be displayed in a few seconds. Over time, data obtained from the Internet has evolved into big data, increasing the web complexity. Especially for web forums where web forum has become an essential resource on the Web nowadays due to its massive and rich information contributed by millions of Internet users in the world every day. For instance, one may search information from Quora, Reddit and etc. Therefore, web mining especially crawling data from web forums can provide a better understanding of the world trends and things that people are concerned about and interested in from time to time. However, web mining is not an easy task in which the process of data extraction is time-consuming and pretty tedious due to its in-depth link structures and a large number of duplicated pages. Hence, web mining has become a crucial approach for extracting meaningful information from the Internet.

Web mining is the process of using data mining techniques to extract and discover valuable information. Web mining can be further categorized into three (3) categories which are web structure mining, web content mining and web usage mining. Web structure mining is to retrieve previously unrevealed relations between web pages. It can be applied to sitemaps for linking one's business information and enable navigation and cluster information. The desired information can be fetched through content mining. Web content mining refers to the process of extraction of valuable information from the contents contained in the web pages. Web usage mining provides the information regarding the clients' access to the web page from users web

used logs, which record each activity made by every client while surfing. However, I will only mainly focus on web structure mining in this paper. Therefore, we propose a web forum crawler called FoCUS (Forum Crawler Under Supervision), a supervised web-scale forum crawler that exploits the organizational characteristics of the Web forum sites and simulates the human behaviour of the visiting web forums.

1.2 Problems

Nowadays, Web forum consists of massive and rich information contributed by millions of Internet users in the world every day. Therefore, crawling data from Web forums can provide a better understanding of the world trends and things that people are concerned about and interested in from time to time. However, crawling data from web forums is not easy due to its in-depth link structures and a large number of duplicated pages. Besides, we should realize that the most useful information is contained in the deeper levels of the web forum site. So, using the traditional crawling method (traditional breadth-first crawling (TBFC)) cannot crawl deeply enough in the web forum site due to it needs to avoid falling into Spider Trap (a structural issue within a website that causes crawlers to stuck in one part of the website and never finishes crawling irrelevant URLs) (Guo et al., 2006). Consequently, information extracted from the site is lesser because information gained is a small part of the whole information contained in the site.

On top of that, most of the modern websites now are dynamic that doesn't have a standard web structure. Collecting data in a machine-understandable format can be a challenge due to the lack of uniformity. For example, a webpage can be created using CSS, HTML, JavaScript, PHP and etc. The data extraction process becomes challenging when web crawlers need large-scale structured data. The problem gets worse when web crawlers need to extract data from thousands of websites relating to a specific schema (Quantzig, 2018). As a result, most crawlers are dependent on web layout where each website requires a separate crawler program which will incur high development and maintenance cost for the team. Worse still, a minor change of web tags of the website usually requires constant and continuous maintenance works which are incurring a lot of cost and time too.

1.3 Objectives

The objective of this research is to tackle the issues stated in **section 1.2** which the implemented system is able to crawl most of the forums' **index, thread and page-flipping URLs** that in this world even though the structure of the forums present in the world are not uniform. The system will also help to store all the crawled results into a CSV file with well-labeled the type of each of the links so that the user can easily view the results by opening the file with common software application like Microsoft Excel or Google Spreadsheet.

1.4 Research Questions

1. How to build a good Web Structure Mining technique for web forums?
2. Can FoCUS crawl multiple web forums that are having different DOM tree structure?
3. Can FoCUS effectively detect and crawl all index URLs?
4. Can FoCUS effectively detect and crawl all thread URLs?
5. Can FoCUS effectively detect and crawl all page-flipping URLs?
6. Can people easily use the FoCUS crawler to crawl URLs in web forums?

1.5 Contributions

Most programmers, data scientists, or data engineers are overwhelmed by regular web structure changes as they need to maintain or update their web crawlers from time to time in order to keep the data collection pipeline operational and clean. Furthermore, understanding the structures of many different web forums in the world is not an easy task as not every programmer, software developer or data scientist is expert in the area of web structure mining and web content mining. In addition, many web page sources now have incredibly complicated HTML structure due to the advancement of web page designs and regular updates. This will lead to the increase of difficulty of crawling data from web forums and many efforts need to be carried out in order to crawl a web forum. Hence, this solution of FoCUS provides an approach for detecting index, thread and page-flipping URLs in a web forum. After done learning on 1 web forum, the user can use the regular expressions learned which are regular expression of index, thread and page-flipping URLs to crawl other web forums that have the similar link structure to the forum that the user used for learning just now. By doing this, the user can easily crawl on many web forums in the

world by just inputting the web forum's URL to the FoCUS crawler. Besides, people also no longer need to rely on the web forum to publish or release its official public API for accessing the data on the particular web forum. There are some advantages of FoCUS Crawler where it is easy to use and can conduct large-scale crawling as long as the web forums are having similar link structure to the one that user trained.

1.6 Chapter Summary

In a nutshell, this first chapter is an introduction that gives a clear explanation to the readers about the background and current landscape of web mining as well as the problems of web crawling in todays' world and the objective to implement FoCUS. Other than that, there are 6 research questions listed in this chapter and the contribution of FoCUS web structure mining technique are elucidated as well in this chapter.