CHAPTER 2

LITERATURE REVIEW

In the past 20 years, there have been many researchers conducting research in the field of web mining, which is an area of data mining techniques used to discover and extract useful information available on the internet. It just likes the concept of extracting informative data available on web pages over the internet (Kumar and Singh, 2016). Besides, web mining can be further divided into three main categories based on the type of data, namely web content mining, web structure mining, and web usage mining (Mughal, 2018). So, in this paper, there will be 2 sections in this chapter where section 1 will be discussing web structure mining while section 2 will be discussing web content mining.

2.1 Web Structure Mining

Web structure mining is one of the core techniques in web mining that deals with document structure and hyperlink structure. A common web's structure consists of web pages as nodes and hyperlinks as edges connecting to other nodes (web pages). A web page's content can also be arranged or structured in a tree-structured way using various XML and HTML tags (Srinath, 2017). As a result, web structure mining basically shows the structural summary of a website based on the document and hyperlinks structure where it identifies the relationship between all connected web pages of a website (Mughal, 2018).

2.1.1 PageRank

The PageRank algorithm was developed by Sergey Brin and Lawrence Page during their PhD studies at Stanford University in 1998 (Brin and Page, 1998). Even the current most well-known search engine in the world, Google, was also formed by the PageRank algorithm. So, the PageRank algorithm is actually an algorithm that often ranks pages, it will make the different pages that are linked to a particular web page calculates or describes the importance of that page. The calculated links are known as backlinks. If a backlink is produced from an important page then the weightage of this link will be higher than those whose links are coming from non-important pages (Kumar P and Kumar Singh, 2010). In short, this algorithm calculates the importance of a web page by counting the quality and number of pages link in which the more important or relevant web pages are more likely to be linked by other web pages.

2.1.2 Web Content Structure

Gu et al. (2002) have proposed this web content structure technique that can facilitate automatic web page adaptation. This technique simulates how a user understands the web layout structure when he or she browses a page such as objects' size, position, colour, background and etc. It tries to represent the author's presentation intention by identifying the logical relationship of web content based on visual layout information. In the beginning, it will undergo a detection process to detect the physical structure of the particular web content in order to find out the presentation scheme. So, the detection process can help to identify how should a web page be divided and subdivided into smaller parts (objects). After that, according to the web visual representation, the whole page is divided into blocks through projection and the adjacent blocks are merged if they are visually similar. This dividing and merging process will keep repeating until the layout structure of the whole page is constructed. In short, this technique utilizes the projection-based algorithm to segment a web page into blocks first then only further divides the blocks into sub-blocks or merged if they are visually similar.

2.2 Web Content Mining

Web content mining is all about extracting useful information from the contents within websites. The contents may contain video, text, audio, images, and structured records such as lists and tables, which are all aimed to provide the user with a better and simpler understanding of the information delivered from the web (Srivastava, Desikan and Kumar, 2005). Besides, web content are made up of unstructured, semi-structured, and multimedia data in most cases as most of the web contents are text-based which are unstructured data. So, web content mining somehow is related to text mining also. Meanwhile, those pictures and videos in some websites especially E-Commerce websites like Shopee and Lazada as well as retailing websites like Adidas and Nike which all are multimedia data (Navadiya and Patel, 2012). Representation of semistructured data is in the form of tags (such as HTML, XML) especially HTML tags since most of the web pages are built using HTML (Mughal, 2018).

2.2.1 Information Extraction

This is one of the most common and popular web content mining techniques for unstructured data which also can be applied to text mining as well. This technique can help to extract information from unstructured data that is present on web pattern matching is used where it traces the keywords and phrases and then finds out the connection of keywords within texts. It will identify the extraction of attributes, entities, and their relationships from semi-structured or unstructured texts (Mebrahtu and Srinivasulu, 2017). When a large volume of text is there then this technique is very useful. Whatever information is extracted using this technique is then transformed into a more structured form where the data will be

stored in a database for future retrieval. Precision and recall processes are being applied to evaluate the efficacy and relevance of the outcomes (Rai, 2019).

2.2.2 Web Crawler

Crawler is a computer program developed by software engineers or programmers that is able to traverse through the hypertext structure (e.g. HTML) of any web page. Consequently, anyone can use web crawlers to collect information from the web. Besides, search engines like Google, Bing also use crawlers often to collect information about what is available on public web pages (Saini and Mohan Pandey, 2015). However, there are 2 different kinds of crawlers which are internal and external web crawlers. Internal web crawlers are crawlers that crawl through internal pages of the website which are returned by external crawlers while external crawlers are crawlers that crawl through unknown websites (Leela Mary, Silambarasan and Phil Scholar, 2017).

2.3 Chapter Summary and Evaluation

To sum it all up, we have covered up what is web structure mining and web content mining and also the techniques that can be applied for each of them, which are PageRank algorithm and Web Content Structure approach for web structure mining while Information Extraction and Web Crawler for web content mining. We can apply the techniques mentioned in the web structure mining section to understand and identify the structure of any websites first and then use the techniques mentioned in web content mining to extract the data from the websites.