Application in Web Mining

Jin Tak Lee
Computer Engineering
Undergraduate Student, Class of 2017
Melissa Cragin, PhD, Executive Director, Midwest Big Data Hub
07.18.16

1. Abstract

The volume, velocity and variety of Data accumulated now is unprecedented in the history of data. As the flow of data has become inconceivable, managing such 'Big Data' has become one of the key elements in the modern society. With extracting and analyzing right information, one can make quick and optimized decisions. Thus, it was inevitable that Web mining has been evolved as well in order to mine useful information from Big Data. This paper deals with the study of Web Mining and how I used Web mining technique to search for data to build a database for Midwest Big Data Hub.

2. Introduction

Since the early 90's, there has been an astonishing growth in the Web, and the countless amounts of information and data are being published in the Web pages every minutes. With such abundance of information, Web mining has become the important technique to extract the useful information from web pages. Web mining can be specified into three types: Web content mining, Web structure mining and Web usage mining[1]. Web content mining is the process of extracting useful information from the contents of web documents. Web structure mining is the process of discovering structure information from the Web. Web Usage mining is the process of discovering meaningful patterns from data generated by client-server transactions on one or more Web localities[2]. This research mainly deals with Web content mining and how I applied it for the Midwest Big Data Hub project.

The structure of the report is the following: Section 2.1 briefly introduces Midwest Big Data Hub and how I attempted to use web mining for the project. section 3 analyzes the technique and tool of Web content mining in more detail. Section 3.4 discusses Web scraping and the principle behind it. Then Section 4 focuses on how my research uses web content mining technique to extract useful information for the Midwest Big Data Hub. Finally, section 5 and 6 discuss the findings and conclusion respectively.

2.1 Midwest Big Data Hub Project

As the volume of data has been increasing drastically on the Web over the last few decades, managing such Big Data has become the one of the core issues for many sectors ranging from business to government. In particular, NSF has announced four regional data hub across the contry to deal with Big Data related issues. Those BD Hubs constitute a "big data brain trust" that will conceive, plan, and support big data partneships and activities to address regional and national challenges [3]. The Midwest Big Data Hub which is lead by Prof. Cragin is one of the BD Hubs. The hub offers platform for other spokes on Midwest region to collaborate on areas such as agriculture, the food, energy, water nexus and smart cities [4].

Midwest Big Data hub is still at the initiative stage and needs to create a database of the available resources and services in the Midwest region. My project is to collect the metadata of such resources in the Midwest and sort it in the structured format such as spreadsheet. At first, I have been "human crawler", manually searching for data from each organizations' Web pages. Such search method is web mining in the most primitive manner. Although it is easy way for web mining, I've found the job somewhat tedious and *non-intellectually challenging*. My search sometimes took unnecessary time to browse through Web pages to find the exact information I need. Then, I've contemplated on a more efficient way to browse the Web automatically and extract the exact information I need. That is how I've come across the Web content mining technique.

3. Web Content Mining

3.1 Overview

Since the useful informtation can be mined throughout the web contents such as text, image, videos etc, web content mining has been an excellent technique for several applications. Web content mining not only helps to analyze trend or customer behavior but also examines the search result of search engine[5]. Even though the search engine such as Google returns the general result of the search, these engines are not always guaranteed to fetch the exact information that one needs. Furthermore, search for specific data through web pages can be quite hard tasks without web content mining if there is an immense amount of data to go through. Various types of data can be mined, but Web content mining becomes complicated when it has to mine unstructured, structured, semi structured and multimedia data[6].

3.2 Unstructured Text Mining

Content mining can be done on unstructured text[7]. Most web documents are actually text. Unstructured text extraction deals with text mining, information retrieval and natural language processing. The technique is mostly based on the machine learning and natural language processing to learn what to extract by manually feeding a mining tool with examples [8]. Pattern, the web mining tool used in this research, is also heavily based on machine learning. The Pattern module can be trained to link specific adjective to certain words. Like pattern, unstructured text extraction is not only limited to text itself but to concepts and relations of different entities. Another direction of research in this area is Web question-answering. Even though Web question-answering was first studied in information retrieval literature, it becomes very important in the Web as the Web offers the largest source of information. Web question-answering is extended to the Web by query transformation, query expansion and then selection[9].

3.3 Structured Text Mining

Structured Text Mining is probably the most widely studied topic of Web Content Mining. The reason that it is so popular is structured text often represent their host pages' essential information. Thus, extracting such data enables one to provide value added services such as comparative shopping and meta-search [10]. Furthermore, it is faster and more effectieve extraction than unstructured text mining since it is easier to traverse the web. Structured extraction can be done through several program approaches. The first approach is to write an extraction program for each Web pages based on recognized pattern. However this approach is not only tedious but can be manually intensive [11]. Wrapper induction or wrapper learning is the second approach. It is currently one of the main mining techniques. In this technque, users can set rules by manually labeling trained Web pages and learning program extracts data based on the rules [12]. Web crawling which is used in the research is considered as structured text mining as well.

3.4 Web Crawling

Web crawling is a program which browses Web pages in methodical and automated manner [13]. When user provides source web site to crawl, module called spider goes through every hyperlinks which can be found from the provided Web site. The fascinating aspect of spider module is that one can increase the speed of the web crawling by increasing the number of spiders which corresponds to the number of usable threads. Web scraper is a recent variant of Web crawlers. Web scraper parses certain types of information. I tried to crawl several linked pages from a source page and scrape simple data such as locations or phone numbers of institutions in Midwest region for this project. Web scraping technique includes HTTP programming, DOM parsing, and HTML parsing which can be done through tools such as Pattern, Scrapy, etc [14].

3.5 Pattern

Pattern which is used in the research is a fascinating Python package for web mining, natural language processing, machine learning and network analysis, with a focus on ease-of-use [15]. I have identified Pattern as the mining tool to test for conducting this task since it seems to be one of the easiest module to learn from the scrach. The syntax is straightforward, and function names and parameters are intuitive and self-exlplanatory. Pattern is well documented in general as well [16].

4. Method

Using Pattern, I planned to write a module that can sweep the Web and return more advanced search results than other search engines would. It turned out that I was too naïve. Web mining with pattern has turned out to be harder task than I originally expected it would since I had no prior knowledge of Web mining and machine learning. Pattern consists of several modules that can be chained together [17]. At first, I have looked for a module that would return me more detailed search results than the results that I would generally get by searching on search engines such as Google. For instance, I was looking for a type of module that would return me the urls of all the hpc centers in the Midwest region when I type in "HPC centers" and "Midwest" as the

keywords. However, I have found there isn't any module I can use to surpass the manual search on Google. That is when I realized the need to change the method of Web mining.

Since I couldn't get the outcome I wanted, I have changed the direction of the project. Instead of focusing on mining the entire Web for the given keywords, I have decided to use Pattern to mine the specific information from each organization's Web pages. I need to extract simple matadata such as contact information, location, leading organizations, etc of each institutes and services in Midwest region. Using web crawler, I have crawled the entire Web pages of each organizations' url and scraped some of metadata I need by using Pattern module.

5. Findings

I have mined 150 Web sites so far and extracted metadata both manually and by using Pattern. These data include basic information such as contact, location, leading organizations of various organizations in the Midwest Region. Even though manual Web scraping can be a lot quicker and more accurate than using Pattern at the moment, I have found the manual Web scraping is not useful in extraction of not apparent information such as funding agencies or target audience. Those information is sometimes not only hard to find but also can be time consuming to search for. However, Pattern can easily extract such information with the help of pattern search and the machine learning modules.

6. Conclusion

In this paper, I have introduced Web content mining and how I used Pattern to extract metadata from urls to create spreadsheet for Midwest Big Data Hub. Relatively easy usage of Pattern let me scrape simple information such as contact and locations from the url of the resources in the Midwest. However, I have seen more potential of Pattern for Web mining through machine learning application. In the future, I hope to train machine learning modules such as k-NN classifier on certain adjectives to scrape a bit more intricate information such as purpose or target audiences of the resources in the Midwest.

7. References

[1] "Big Data What it is and why it matters." Sas. n.p. n.d. Web. Date Accessed 21 July 2016

[2] Kalil Tom, Jim Kurose and Zhao Fen. "Big Announcements in Big Data." The WHITE HOUSE. 4 November 2015. Web. 22 July 2016

[3] Kalil Tom, Jim Kurose and Zhao Fen. "Big Announcements in Big Data." The WHITE HOUSE. 4 November 2015. Web. 22 July 2016

[4] Faustina Johnson, Santosh Kumar Gupta. "Web Content Mining Techniques: A Survey" *International Journal of Computer Applications Volume*. Volume 47-No.11. Web. June 2012

[5] Jaideep Shrivastava. “Web Mining: Accomplishments & Future Directions.” University of Minnesota. Web. 29 July 2016

[6] Faustina Johnson, Santosh Kumar Gupta. “Web Content Mining Techniques: A Survey” *International Journal of Computer Applications Volume*. Volume 47-No.11. Web. June 2012

[7] Faustina Johnson, Santosh Kumar Gupta. “Web Content Mining Techniques: A Survey” *International Journal of Computer Applications Volume*. Volume 47-No.11. Web. June 2012

[8] Faustina Johnson, Santosh Kumar Gupta. “Web Content Mining Techniques: A Survey” *International Journal of Computer Applications Volume*. Volume 47-No.11. Web. June 2012

[9] Faustina Johnson, Santosh Kumar Gupta. “Web Content Mining Techniques: A Survey” *International Journal of Computer Applications Volume*. Volume 47-No.11. Web. June 2012

[10] Bing Liu, Keven Chen-Chuan Chang. “Editorial: Special Issue on Web Content Mining” *ACM SIGKDD Explorations Newsletter. Volume 6 Issue 2* : 1 - 4. Web. December 2004

[10] Bing Liu, Keven Chen-Chuan Chang. “Editorial: Special Issue on Web Content Mining” *ACM SIGKDD Explorations Newsletter. Volume 6 Issue 2* : 1 - 4. Web. December 2004.

[10] Bing Liu, Keven Chen-Chuan Chang. “Editorial: Special Issue on Web Content Mining” *ACM SIGKDD Explorations Newsletter. Volume 6 Issue 2* : 1 - 4. Web. December 2004

[10] Bing Liu, Keven Chen-Chuan Chang. “Editorial: Special Issue on Web Content Mining” *ACM SIGKDD Explorations Newsletter. Volume 6 Issue 2* : 1 - 4. Web. December 2004

[10] Bing Liu, Keven Chen-Chuan Chang. “Editorial: Special Issue on Web Content Mining” *ACM SIGKDD Explorations Newsletter. Volume 6 Issue 2* : 1 - 4. Web. December 2004

[15] Eloisa Vargiu, Mirko Urru. “Exploiting web scraping in a collaborative filtering-based approach to web advertising” Artificial Intelligence Research. Vol.2, No.1. Web. 2013

[16] Eloisa Vargiu, Mirko Urru. “Exploiting web scraping in a collaborative filtering-based approach to web advertising” Artificial Intelligence Research. Vol.2, No.1. Web. 2013

[17] Tom De smedt, Walter Daelemans. “Pattern for Python” Journal of Machine Learning Research 13. N .pag. Web. 2012