Web Crawler and Big Data

Pengyu Li pl26

Introduction

Today, Big Data has been a buzz word lately. It seems like a startup is not sexy if it doesn't mention "big data" in their technology. But what exactly is big data? How to access big data? Why is big data important in your business? And how important is web crawling in the age of "big data"? The paper is trying to figure out answers to questions above, especially the relationship between web crawler and big data.

What is Big Data?

From its literal meaning, big data is the high volume of data. Typically, big data is generated by the interaction of business, government agencies, and the internet. In many papers and blogs which describe big data, big data often has four characteristics, Volume, Variety, Velocity and Veracity which is denoted by 4Vs. Volume is the most important part of big data. The size of data determines the value and potential insight.

Why big data is important?

The high volume of data often conceals certain patterns. Using machine learning models to build a predictive application or knowledge discovery application will dig out the potential value of data. Take Google, for instance, Google has been collecting data from billions of web pages using their own web crawler --- Google Bot. Using this huge amount of data with other data from search engine queries, Google successfully improve the effectiveness of Google Ads. By continuing to do so, Google has been generating more profits for themselves and their Ads customers, and also keeping their search users happy. Amazon is doing similar things as Google. Amazon collects user's action towards a certain category of items, and input them into their recommendation systems model. The model will automatically recommend items that users may have interests in. It really increases the profit of Amazon.

How important is the Web Crawler in the age of Big Data?

To build a data-driven product is not easy. Firstly, you should collect a significant amount of data before you launch your product. Secondly, you will find there are limited ways to get the desired amount of data.

1) Direct User Input (survey, search form, etc)
2) Third Party APIs (Facebook, Twitter)
3) Server Logs
4) Web Crawling

Among all the above sources, user input is too slow and unlikely builds enough data in short time; third party APIs are also limited and most of the companies may not share their data to public; server logs are also unrealistic if you are only a startup. So that, most data-driven companies crawl web pages to collect data because most of the data that these companies need are in the form of the web page with no API access. However, crawling web pages is also not an easy task. So that, the price of the ticket to big data club is very high.

Web Crawling is a technique applied in search engine optimization. Internet bot would collect web data by following links on the site map. The scope of browsing is very vast, the whole World-Wide-Web. But for your own web crawling task, you could focus your crawling scope on the site of interests. To build simple a web crawler is easy, but there are also many difficulties in both technical side and non-technical side.

You must obey the privacy protocol when scraping data from a third party website. Sometimes, the owner of a website may not want someone to access to their data. Even the owner hasn't mentioned terms of use of their data, scarping also could cause some potential legal issues. Do not be evil, just as Google's slogan says.

The Web is a sea of information with billions of web pages created every single day. Most of the data these web pages contain are unstructured and messy. Collecting and organizing these messy data is not easy. While creating any data-driven product, you will be spending almost 90% of your time collecting, cleaning and filtering data. And, when it comes to crawling web pages, you have to have good programming and database skills. None of the data in the web pages are conveniently handed to you. Sometimes, scraping web pages gets even worse when you have to scrape data from a difficult source such as a PDF file. But, ultimately you will even have to scrape data from these sources to benefit your business.

Modern web applications use AJAX heavily. Some web pages may be rendered by AJAX totally. Unfortunately, web crawler doesn't have an easy solution to handle AJAX. There are two regular ways to solve it. First one is using some automatic testing utilities like Selenium and PhantomJS. They would create an environment for executing Javascript code. On the other word, you could render the page by yourself, and wait for a server to send data that you want. The drawback of this solution is obvious, it is too heavy and slow. Rendering web pages cost too much computing resource, it is unlikely to extend your project on a large scale. The second way is mocking user's behavior, you need to hack into message queue when rendering the page, and find out which request calls server to send data you want. Then you need to simulate this request with a cookie. This solution is light-weight and fast, but not very easy for new hands.

Web crawling becomes more and more popular with the explosion of Big Data techniques. Many good frameworks for scraping have emerged and played very important role in today's crawling scenarios, such as Scrapy. Scrapy works in Python environment. The philosophy of it is treating data you want to scrape as an object. The advantage of this framework is it is fast and powerful, you write the rules to extract the data and let Scrapy do the rest. And it is extensible by design, plug new functionality easily without having to touch the core. Also, it is written in Python and runs on Linux, Windows, Mac and BSD.

In coming years, we will be hearing lots of news about Big Data and new data-driven startup companies. Companies with a better understanding of Big Data will be able to provide a better experience to their customers, find potential customers and generate more profits. Web crawling is a significant part of building most data-driven products. It requires lots of time, effort and money. Maybe we could find an integrated solution including easily scraping data from AJAX rendered pages for our business in the very near future.

References:

[1] https://en.wikipedia.org/wiki/Web_scraping

[2] https://www.grepsr.com/importance-of-web-crawling-in-the-age-of-big-data-2/

[3] https://doc.scrapy.org/