ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

Web Scraping of Social Networks

Renita Crystal Pereira, Vanitha T

Dept of M.Sc Software Technology, AIMIT, St. Aloysius College (Autonomous), Mangalore, Karnataka, India

ABSTRACT: Web scraping is a process of extracting data from the internet through various methods. The internet is a universally accessible resource for millions of people. As the usage of internet has commonly increased in everywhere there is highly growth in competition between the organizations in their business. Web scraping helps in automation of web through the use in various techniques. The usage of web scraping have been in many fields and beneficial for weather data monitoring, website change detection, research, web data integration, contact scraping and online price comparison.. In this paper we will go through the tools and techniques used in scraping and its impact on the social networks.

KEYWORDS: web scraping; social networks; data extraction, scraping tools; techniques.

I. INTRODUCTION

Through web scraping services unstructured data are converted into structured data which can be stored and verified in a centralized data bank. The aim is to collect, store and analyse data. The data analysis is very much needed in a society to extract any information and transforming it into a format helpful to interpret. Thus, web scraping services have a direct influence on the outcome which is needed from the data collection. Web data extraction is the process of transforming the useful content on websites into valuable business assets. There are several web extracting software that has emerged in the market which helps to address this problem. The software aids in extracting structured content from a web page and exposes the required services as APIs and makes it useable for further processing. It is necessary to know the available technologies in the market today. The available technologies that are related may be in different languages written such as java, python, php etc. The benefits of this are beyond the limitations of the users. Since there is rise in new online business through internet this has an adverse effect on the consumers as well. Online marketing analyst use web scraping methods to grab some information from other competitors such as emails, targeted keywords and links and also traffic source. The scraping techniques are used for personal as well as commercial usage. All the techniques available has its own pros and cons to overcome this there is need to have a clear idea on the usage of these techniques in social networking.

II. LITERATURE REVIEW

The related work on the web scraping techniques involves [1] in this paper the various aspects such as different web scraping semantic levels and the sentimental approach has been included. This paper gives the study on human opinion mining where screen scraping plays the major role. The most common available tools and techniques are been used by many users which are free and easy to use. [2] uses the algorithm to explain XPath using Tree edit distance matching algorithm. The problem of computing the tree edit distance between trees is a variation of the classic string edit distance problem for extracting data. The authors of [7] have a study on the techniques of web content mining and relate web scraping tools that are available. As these paper consists many topics under data mining which gives the clear idea on the different available techniques under data mining and we can compare this with web scraping.

The author of [15] in this paper the author has calculated the total estimated value of web scrapers in industries which has a survey as 68% is used for e-commerce and social media, rest of which is digital news publishing and online directories. As the increase in use of internet more and of the users tend to use social networking sites .The papers[4], has adopted the web scraping techniques in the web advertising field which also explains the collaborative filtering ways of web scraping with preferred implementation ads. As the scraping has an usage in adverting field this has become important in learning various methods.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Special Issue 7, October 2015

III. OVERVIEW

The overview on web scraping and use in the networking process is explained in this paper.

Websites are More Important Than APIs

The biggest one is that site owners generally care way more about maintaining their public-facing visitor website than they do about their structured data feeds. We've seen it very publicly with Twitter clamping down on their developer ecosystem, and I've seen it multiple times in my projects where APIs change or feeds move without warning. Sometimes it's deliberate, but most of the time these sorts of problems happen because no one at the organization really cares or maintains the structured data. If it goes offline or gets horribly mangled, no one really notices.

No Rate-Limiting

Another thing to think about is that the concept of rate-limiting is virtually non-existent for public websites. Aside from the occasional catches on sign up pages, most businesses generally don't build a lot of defences against automated access.

Anonymous Access

There are also fewer ways for the website's administrators to track your behaviour, which can be useful if you want gather data more privately. With APIs, you often have to register to get a key and then send along that key with every request. But with simple HTTP requests, you're basically anonymous besides your IP address and cookies.

The Data's Already in Your Face

Web scraping is also universally available There is no need to have to wait for a site to open up an API or even contact anyone at the organization. Just spend some time browsing the site until you find the data you need and figure out some basic access patterns.

TOOLS

Some of the powerful webs scraping tools available are:

Tool Name	Availability
Uipath	free
Import.io	free
Kimono:	free
Screen scraper	free
Bixo	free
Darcy Ripper	free
DEiXTo	free
Pattern	free
Web Mining Services	commercial
80legs	commercial
FMiner	commercial
iWeb	commercial
Mozenda	commercial
Scrapy	commercial
TheWebMiner	commercial
Visual Web Ripper	commercial
WebSundew	commercial
Darcy Ripper	commercial
DEiXTo	commercial
Pattern	commercial
Web Mining Services	commercial
80legs	commercial
FMiner	commercial
Helium Scraper	commercial

Screenshots of some tools:

Web scraping Social networking and scraping

The social media deals with sharing data... As Screen scraping has been there for past years there are chances of using the information negatively by many victims too. As a result many companies have secured themselves by adding privacies polices where the scrapers might have difficulty in accessing the information. The intellectual property rights would help the business websites to be on the safer side from attackers. Sometimes even privacies features are not effective. As there is huge data available there is a difficulty to maintain such data. Now the complexity of using the internet has increased for various purposes which in turn allows to have complexity in programming too. The responses to these might be html or they might be json, in rare cases they will be xml or something else. The new technologies come up with many new features that makes the user more secure in having the information on the internet.

The usage of internet and social networking sites increases day by day such as twitter, facebook, linked-in and many more, the information of the users also been in high volume throughout the internet accessible from anywhere. This will also have an advantage for hackers to steal information. As data is in huge amount throughout the internet scraping involves various steps in extracting the required information. To scrape is to build a chain from the relatively unformed mass of online data to formatted information, and along this chain relatively raw textual data is progressively stripped of its useless elements and formatted so as to produce a well-ordered, useable data set. Social networking are essential in business point of view where the idea of increasing profit comes into existence. As in case of online shopping will also help the users to get easy shopping and save time as well. In other side there is benefit on promoting the business and gaining profit out of it.

As well web scraping may be referred as collecting information and extracting it and making use of it some point of view whereas it is a process of involving business intelligence and having profit out it through managing the data effectively. Today business intelligence plays a major role in industry there is need in having rise of web scraping tools and technologies.

IV.CONCLUSION

This paper has an overview on the web scraping and tools and techniques facing many challenges as the extraction of the data are not that easy. These techniques ensure that the gathered information is accurate, reliable and having higher confidentiality as the data present is in huge amount which is difficult to manage and maintain. Even though techniques useful there are some challenges faced that may be such as the high volume of web scraping can cause regulatory damage to the pages. Scale of measure the scales of the web scraper can differ with the units of measure of the source file thus making it somewhat hard for the interpretation of the data. The Level of source complexity in case if the information being extracted is very complicated web scraping will also be paralyzed.

REFERENCES

1. Jos´e Ignacio Fern´andez-Villamor, JacoboBlasco-Garc´ıa, Carlos ´A. Iglesias, Mercedes GarijoDepartamento de Ingnier´ıa de SistemasTelem´aticos, Universidad Polit´ecnica de Madrid, Spain jifv@dit.upm.es, j.blasco@alumnos.upm.es, cif@dit.upm.es, mga@dit.upm.es
2. Web Data Extraction, Applications and Techniques: A SurveyEmilio Ferraraa, Pasquale De Meob, Giacomo Fiumarac, Robert Baumgartnerd
3. WebSelF: A Web Scraping Framework Jakob Thomsen1, Erik Ernst1 , Claus Brabrand2 , and Michael Schwartzbach
4. Exploiting web scraping in a collaborative filtering- based approach to web advertising
5. Eloisa Vargiu1, 2, Mirko Urru1
6. Dipartimento di Matematica e Informatica, Università di Cagliari, Italy. 2. Barcelona Digital Technology Centre, Spain
7. Faustina Johnson and Santosh Kumar Gupta.Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
8. Cohen and Fan. Learning page-independent heuristics for extracting data from web pages. CN, 31(11-16), 1999.
9. R. Baumgartner, K. Fro¨schl, M. Hronsky, M. P¨ottler, and N. Walchhofer. Semantic online tourism market monitoring. Proc. 17th ENTER eTourism International Conference, 2010.
10. R. Baumgartner, W. Gatterbauer, and G. Gottlob. Web data extraction system. Encyclopedia of Database Systems, pages 3465–3471, 2009.
11. K. Kaiser and S. Miksch. Information extraction. a survey. Technical report, E188 - Institut fu¨r Softwaretechnik und Interaktive Systeme; Technische Universita¨t Wien, 2005
12. Faustina Johnson and Santosh Kumar Gupta.Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012