***I.J. Education and Management Engineering,* 2016, 5, 24-31**
Published Online September 2016 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2016.05.03

Available online at http://www.mecs-press.net/ijeme

An Intelligent Survey of Personalized Information Retrieval using Web Scraper

Bhaskar Ghosh Dastidar [a], Devanjan Banerjee [b], Subhabrata Sengupta [a,b,*]

[a] *Affiliate of Istitute of Engineering and Management, 130 Bishnupally,Kolkata 700093, City Kolkata,India.*
[b] *Affiliate of Istitute of Engineering and Management, Purbachal Police Line East, Burdwan-713103, India*
[ab] *Affiliate of Istitute of Engineering and Management, Salt Lake,Sector V,Kolkata 700091, India*

Abstract

In this paper we aim to do an intelligent background survey of Personalized Information Retrieval, a specialized and crucial subsection of Information Retrieval or IR. We have chosen the method of IR as Web Scraping, a technique that is extremely popular and is proven to have multi-domain usage.

Index Terms: Web-Scraping, Information Retrieval, IR, Personal Information Retrieval, Semantic, Web, WWW, Internet, Bots, Spider, Crawler, Jaunt**.**

1. Introduction

This paper deals with web scrapers and their use in Information retrieval with a focus on personalized information retrieval. Web scraping is a hot topic in today's perspective and it has multi faced applications. But two of the most important utilities of scraping are information retrieval for personal usage and for analytic purposes. We have dealt with the first type throughout this paper.

The contents of this paper have been majorly divided into parts. Introduction takes on the subject matter on a wider basis and gives a general view on all aspects of Information retrieval and Web scraping. Background Survey starts off with the evolution of the need of web scraping and introduces all older forms of web scraping and its methodologies and its drawbacks. An example of web scraping is also provided. Research Findings details out all a few of the popular existing web scrapers and its relevance in today's context. Finally we conclude in the last section with web scraping usage, problems and future scope.

1.1. Data on the internet and the need of web scraping

* Corresponding author.
E-mail address:

Data generation and growth rate of the same is an abrupt process these days and will grow exponentially with each passing day. Users on the internet can enjoy abundant services and information in e-commerce websites, electronic newspapers, blog & social networks.

Although this data is available for its consumption by users, quite an amount of time is spent retrieving this information and processing it. Moreover, the format of data in the form of HTML and other web languages are not suited for automated agents and computer programs. This has favored the research in several fields such as web scrapping. Web scraping, a process of extracting useful information from HTML pages, which is the main formatting tool of information on the WWW, implemented using any programming language and semantic annotation is the main target in this paper as a method of a better or efficient personal information extraction from the web. The pages that are being scraped in context may include/comprise of metadata or semantic markups and annotations, which can be used in the location of specific data snippets. If the annotations are embedded in the pages, the same way Microformat does, this technique can be viewed as a special case of DOM or SAX parsing. Another methodology ascertains that the organized annotations, stored in a semantic layer and separately from web pages, so the scraping job runs faster by retrieving schema information and important meta-data before scraping the actual web page. [7] The technique which makes it possible to add semantics, meaning, structure and a formal parity to unstructured textual documents is called Semantic Annotation, an important aspect in semantic information extraction. Through our project, we revisit, explore and discuss some information extraction techniques using web scraping and semantic annotation for the creation of.

1.2. Personalized Information Retrieval explained with an example

A busy person finds it very difficult to sit on the internet at the end of the day and browse for news articles, important headlines of the day, information specific to a particular domain of his choice, scores of his favorite sports matches, trending information at the moment and sentiments encircling that trend. While Google would be the ideal best friend in such a need, it is no doubt hectic visit the plethora of links that Google search results put forward for a specific topic. An intelligent web scraper will be the perfect tool in this situation. A web scraper, intelligent enough to scrape the internet occasionally (based on user's choice of an interval) and takes into account the user's query comprising various keywords according to his liking and domains of interest, for e.g. Cricket, Election, Mobile Phones, Sachin Tendulkar, Chicken Biriyani.

The web scraper will identify the keywords of importance as above and will run it's scraping job on some predetermined websites. It will map Cricket with sports section of any media website, Chicken Biriyani with food websites and Mobile Phones with E-Commerce websites.

It will then provide necessary information and a sentiment analysis perhaps of its extracted data and provide the data to the user in order. Thus, the whole process of personalized information retrieval is done in a jiffy by the web scraper and the work done by the user is reduced significantly. [3]

2. Related Work

Information Retrieval has come a long way. Before the internet came into existence and IR shifted its focus on web searches and kind, it was prevalent since as early as 1960 in commercial and intelligence applications. The speed and accuracy of the same have increased exponentially with processing power and storage capacity. Such development and advancement in the field of IR have also resulted in a rapid progressing of querying techniques from the manual library based approaches to a far digital end.

2.1. Evolution of data extraction techniques from the World Wide Web

When Tim Berners-Lee created the World Wide Web, the net content of information in web pages was very low. It didn’t need an automated IR system to retrieve and scrape data off them. But with the expansion of the internet, the need for the same was essential. Developments in this field were carried on in two specific directions: searching the content of the web page and also the links incoming to that specific web page. [21]

The reference based query expansion by Bradshaw, Scheinkman, and Hammond of Northwestern University's Intelligent Information lab is a brilliant expansion of early IR technique. Although this methodology suffers from a drawback of being limited to conceptually homogenous texts archives allows the documents to be indexed according to the way they are referenced in other articles. Bradshaw's famous and invaluable observation that "people rarely form queries of longer than three words" still holds true. Their assumption is searchers will submit queries which are not ambiguous based on which the index system works. The main technique used to yield high-quality search results was to index documents. The usage of references is a very strong way to index and more importantly rank the documents because the reference pair coincides with the document that contains the required information. [17]

Working on the drawback reference based query expansion by Bradshaw, Chau and Yeh proposed a technique which works with the heterogeneous document. The main idea behind it was the native Asian people would also be able to search for the query passed using their own language, this idea was known as multilingual query expansion. Chau and his sub-ordinates gave a work around so that the person in need of the desired information will browse through the directory or hierarchy of concepts that are generalized and changed to the person's native language. The information seeker clicks on the concept of interest through which he submits the query after which the system returns the result by showing the document based on the concept category.

The most common information retrieval is the ad-hoc querying where a query searches for a set of documents which are static. The commercial search engines like AltaVista and Google are known for using the above information retrieval technique. The search engines work on a huge database while searching for a particular keyword. The drawback with the above-mentioned technique is that the precision is very low. An example illustrating the drawback would be if we search for the query "Who is the president of India?" This information retrieval technique would return the documents where the keyword president is associated. However other unrelated documents about "How to campaign as a president?" or "Who is the president of INDIA TODAY" are also retrieved along with it. [10]

A promising technology for information retrieval is Agent Paradigm. The information retrieval technique with an agent-based approach means that the IR systems are more scalable, flexible, extensible and interoperable. Other statistical approaches like that of n-grams, semantic indexing are particularly interesting techniques with which the text objects are analyzed, they are also not dependent on the language of the text, they are resistant to misspellings and use many well known mathematical techniques for a natural language analysis.

To extract, categorize, rank and index the most legit information available on today's worldwide web to the user requesting the information, web scraping is a crucial tool. Web scrapers can also determine the relevance of the extracted information to the user. Personal information and that of a high quality to the user is the main aim of any web scraper. Further, it forwards any data analysis, web usage patterns, fine tune crawlers and spiders, comparative market researchers etc. [6]

2.2. Forms of web content mining and tools used in that context

In recent times, much research has gone into search queries having linguistic structure. Identifying noun forms from a user presented phrase and performing the IR procedure based on those noun forms eases the process and is more realistic in today's context. [15]

The major forms of web content mining are done in the following ways:

- Unstructured data mining
- Structured data mining
- Semi-structured data mining

- Multimedia data mining.

Under the subtopic of unstructured data mining comes the mining of web documents which this paper deals with. Personal Information retrieval is a challenging task when it comes to mining from web pages, primarily because of the complexity of all the HTML tags that may be present. Web scrapers help a great deal in simplifying the task.

Structured data mining plays a very crucial role in the development of crawlers or web spiders. Web crawlers crawl and visit web pages for the later processing of them by search engines to determine page quality and thereby a rank. Spiders use different graph algorithms to analyze the web and perform different tasks. [16]

The different above mentioned methodologies of web content mining puts to use a number of tools and services. Some of them are very prevalent while some are still scarcely put to use. [5]

Basic Scrapers are used to scrape web pages using keywords that are provided by the user. It targets a specific web page and specific HTML tags.

Automation Anywhere (AA) is an automated web extraction tool to scrape web pages and use it for data mining. [2] Web Info Extractor (WIE) is a tool to extract information from web pages and manage data for content analysis. It can extract both structured and unstructured data.

3. Research Findings

The web scrapping process does not always receive a positive intimation. The question that arises is whether the process of web scrapping is legal or whether it complies with the terms and conditions of each and every website? So in this section various research findings about the web scrapping process has been discussed.

The initial history of web scrapping was not very positive as the concept of web scrapping was not approved legally. The enforceability of these terms is unclear. When we duplicate the original expression it is considered illegal, however in USA the concept that we can duplicate the information elsewhere is pardonable.

According to the courts in US they consider computer to be personal assets thus the persons using scrapers are responsible for using or accessing the computers of other persons in an unauthorized way. For example, the case of eBay v. Bidder's Edge, this involved the trespassing of information from the eBay website by using bots. Bots are basically software which are used to perform various functions like scrapping data, placing bids, clicking ads any many more. This case involved the automatic placing of bids, known as auction sniping. The Bidder's Edge used the bots to scrape the data from the eBay site and displayed it on their own website. This further led to the involvement of auction bids automatically. To authenticate the trespassing activity, the users should show a way to defend their intention.

An example of web scrapping in the travel industry is the case of Hipmunk. Hipmunk was used to scrape data, collect the price information and other vital statistics. The information was obtained before it could be received by the online travel agents and suppliers with an actual partnership agreement. However, this activity was stopped immediately when they were asked to.

The well known example of screen scrapping was with the American Airlines the firm involved in this case goes with the name FareChase. American Airlines prevented the firm from using softwares to access their confidential information after taking legal help from the courts. In March 2003 AA alleged that FareChase was using their software to gather information and trespass their web servers without their consent. However, both the parties ended up in a mutual settlement dropping the charges against each other by the end of June.

Some of the fraud bots detect the loop holes in our system. Taking advantage of the vulnerabilities like ads, pop ups etc. The bots can obtain the valuable information from the user. To secure his credentials the user can perform various steps like traffic monitoring, blacklisting, rate limiting and anti-spam protection.

Web Scrapping is still considered as a very important technology all though so many legal cases are still going on against it. FMiner is very important software which extracts information from the web and displays information in a readable format for the users. The software's main implementation idea is to obtain the information from the URL of the websites. FMiner is considered as the best software available for extraction of information from the web. The web scraping software is designed to scan through set sites, and scrape set data at your requirement. The FMiner software extracts information and stores the data in various formats so that the users can access them easily. The formats used by FMiner are .csv, access, Excel or even an SQL server. [22] The diagrammatic representation of FMiner is provided in Fig 1.

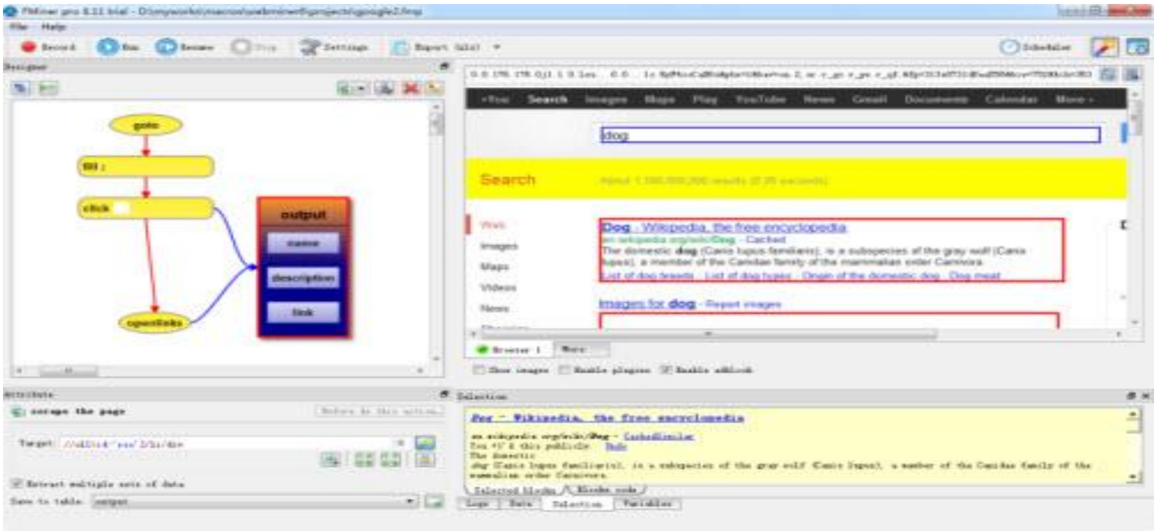

Fig.1. The FMiner Interface

Another well known application which is an extension of the Firefox browser is the ChickenFoot. It adds functions which are used for data extraction to the JavaScript and we can run them directly on the browser itself. The well known feature for which ChickenFoot is famous is the concept of “Embedding” thus it can interact with the JavaScript easily. The distribution of this software can also be carried out with ease. The only drawback of using this application is its slow execution time which makes it very difficult to use. The main reason for the slow execution speed is the fact that it runs in the browser making it slow and unfeasible to use. ChickenFoot extracts information from the web using the command find() .[18]

Piggy Bank is also an extension of the Firefox browser which is used for reducing the gap between what we have right now and the semantic web. The users of this software send scripts along with the regular expression to extract the data from the web by using the URLs of the desired websites. The system then displays the scrapped information in a semantic way after visiting the desired web page. The idea used in the above case is a beautiful approach as it is spread across the globe however; the community is not developed yet. [21]

Sifter is built using the concept of PiggyBank web scrapper. It scrapes the data from any desired web page automatically. However, the scope of this scrapper is very limited as it looks for the largest group of outgoing links in the desired web page.

Some other very frequently used tools used by students, IT professionals are as follows:

- Diff bot
- Scrappy
- Selenium data scraping
- Apache Camel
- Archive.is
- Jaxer
- Import.io

In the field of journalism too the journalists prefer using the web scrappers to extract statistical information. The online marketers collect chunks of information about the contact details. They prefer scrapping of data for search engine optimization. This would also reduce the manual labour of the online marketers who would otherwise waste a lot of time in gathering the required information. The developers also need various scraping tools so that they can fill their applications with huge amount of variable data.

Various site indexers use the bots as well. The search engines like Google, Yahoo, Bing use the site indexers as well for crawling the website and displaying the content according to their priorities so that the users can access the information easily.

Since our work mainly deals with personalized data retrieval through data scraping, we focus on the ways of data extraction from web pages present.

They are as follows:

- HTTP Programming
- HTML Parsers
- DOM and SAX parsers
- Open Source web scraping libraries.
- Semantic annotation recognizing
- NLP recognition of keywords

4. Conclusions

The Internet is filled with publicly available data. Information can range from sports scores, weather forecasts, financial results, etc. This data can help answer questions we have, by testing the hypothesis.

The main uses of Web Scraping comprise of the super set of and not limited to the domain of online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration. [4]

Some Business Use Cases

- Gathering data from multiple sources for market analysis and Lead generation
- Research
- Data Integration
- Helps monitoring of competitor's inventory information
- Stock prices
- Order status from e-commerce portals

Acknowledgements

We hereby acknowledge our sincere thanks to Prof. Dr. Debika Bhattacharya, head of the department, Computer Science and Engineering.

We would also like to thank Prof. Subhabharta Sengupta, our teacher, mentor and guide without whom we would have never learnt about Web Scraping and Information Retrieval. Without the constant support and encouragement, we would have never been able to realize this project.

References

[1] Rahul Dhawan, Murda Shukla, Priyanka Puvar, Bhagirath Prajapati, "A Novel Approach to Web Scraping Technology", International Journal of Advanced Research in Computer Science and Software Engineering.

[2] David Martinez, Richard Baron Penman, Timothy Baldwin, “Web Scraping Made Simple with Site Scraper”.

[3] Jose ́ Ignacio Fernandez-Villamor, Jacobo Blasco-Garc ́ıa, Carlos A'. Iglesias, Mercedes Garijo, "A Semantic scraping model for web resources-Applying Linked Data to Web Page Screen Scraping".

[4] Rushabh A. Patel, Mansi Patel, “A SURVEY ON INFORMATION RETRIEVAL FROM WEB USING WEB SCRAPING TECHNIQUE”, IJIRT | Volume 1 Issue 6 | ISSN: 2349-6002.

[5] Amit Sheth, Clemens Bertram, David Avant, Brian Hammond, Krysztof Kochut, and Yashodhan Wake, Coquette. “Managing Semantic content for the web”, IEEE INTERNET COMPUTING,1089-7801/02/$17.00 ©2002 IEEE.

[6] Malik, S.K., Rizvi, “Information Extraction Using Web Usage Mining” IEEE, Xplore Digital Library.

[7] Chang, C., Kayed, M., Girgis, M., and Shaalan, K. (2006), “A survey of web information extraction systems”, IEEE Transactions on Knowledge and Data Engineering.

[8] Hogue, A. (2005), “Thresher: Automating the unwrapping of semantic content from the world wide web”, In Proceedings of the Fourteenth International World Wide Web Conference, pages 86–95. ACM Press.

[9] R. Cooley, B. Mobasher, and J. Srivastava, “Web Mining: Information and Pattern Discovery on the World Wide Web”, IEEE 1997.

[10] Brijendra Singh, Hemant Kumar Singh," WEB DATA MINING RESEARCH: A SURVEY", IEEE2010.

[11] Kai Zhong Zhang And Dennis Shasha,” Simple Fast Algorithms For Editing Distance Between The Trees And Related Problems”.

[12] Kuo - Chung Tai, “The Tree-To-Tree Correction Problem”. ACM 1979.

[13] Real World Application of Web Scraping http://imacros.net/overview/data-extraction

[14] Fuhr, N. and Grojohann, K. ‘XIRQL: An extension of XQL for information retrieval.’ In Proceedings of the ACM SIGIR 2000 Workshop on XML and Information Retrieval.

[15] Golbeck, J., Parsia, B., and Hendler, J. ‘Trust networks on the Semantic Web.’ To appear in the Proceedings of Cooperative Intelligent Agents 2003, August 27-29, Helsinki, Finland.

[16] Mayfield, J., McNamee, P. and Piatko, C. ‘The JHU/APL HAIRCUT system at TREC-8.’ The Eighth Text Retrieval Conference (TREC-8), pages 445-452, November 1999.

[17] Shah, U., Finin, T., Joshi, A., Cost, R. S. and Mayfield, J. ‘Information Retrieval on the Semantic Web.’ 10th International Conference on Information and Knowledge Management, November 2002.

[18] Kopena, J., and Regli, W., ‘DAMLJessKB: A tool for reasoning with the Semantic Web.' IEEE Intelligent Systems 18(3), May/June 2003.

[19] Bar-Yossef, Z., Kanza, Y., Kogan, Y., Nutt, W. and Sagiv, Y.. ‘Quest: Querying semantically tagged documents on the World Wide Web.’ In Proc. of the 4th Workshop on Next Generation Information Technologies and Systems, volume NGITS'99, Zikhron-Yaakov (Israel), July 1999.

[20] Abiteboul, S., Quass, D., McHugh, J. Widom, J. and Wiener, J. ‘The Lorel query language for semistructured data.’International Journal on Digital Libraries 1, pages 68-88, April 1997.

[21] Arocena, G. and Mendelzon, A. ‘WebOQL: Restructuring documents, databases and webs.’ In International Conference on Data Engineering, pages 24-33. IEEE Computer Society, 1998.

[22] Berners-Lee, T., and Fischetti, M. Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor. Harper, San Francisco. 1999.

Authors' Profiles

Bhaskar Ghosh Dastidar was born in Kolkata, West Bengal in 1993 and is pursuing his B.Tech in Computer Science and Engineering from Institute of Engineering and Management, Kolkata. He is in his final year of Engineering and will pass out in the month of June, 2016.

Devanjan Banerjee was born in Bardhaman, West Bengal in 1995 and is pursuing his B.Tech in Computer Science and Engineering from Institute of Engineering and Management, Kolkata. He is in his final year of Engineering and will pass out in the month of June, 2016.

Subhabrata Sengupta was born in Kolkata, West Bengal, 1984, earned Mtech degree in the field of Information Technology in 2011 from Jadavpur University. He is now an Assistant Professor at Institute of Engineering and Management in the department of Computer Science and Engineering. His current research interests include Information Retrieval.

How to cite this paper: Bhaskar Ghosh Dastidar, Devanjan Banerjee, Subhabrata Sengupta,"An Intelligent Survey of Personalized Information Retrieval using Web Scraper", International Journal of Education and Management Engineering(IJEME), Vol.6, No.5, pp.24-31, 2016.DOI: 10.5815/ijeme.2016.05.03