A Study on Web Mining Tools

D.Nithya[1], Dr.S.Sivakumari[2]

[1]Assistant Professor, [2]Professor and Head

Department of Computer Science and Engineering

Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

Abstract— In the current scenario, millions of customers are accessing daily the internet and World Wide Web (WWW) to search the information and execute their requirements. Websites are a common platform to exchange the information between users. Web mining is one of the applications of Data mining techniques for extracting information from web data. The web mining process can be classified into three types: web content mining, web usage mining and web structure mining. Web content mining is the process of extracting or discovering useful information web pages. It includes image, audio, video and metadata. Web structure mining deals with the hyperlink structure of web. Web usage mining is the process of extracting information from web server logs. This paper is a study and analysis of different techniques and tools used in web mining for mining the information from internet.

Keywords— Web mining, Web content mining, Web usage mining and Web structure mining**.**

I. INTRODUCTION

The development of internet in today's world has generated a huge amount of data becomes very popular and its growth is very fast. It has the collection of text, images, videos and other form of data. To handle these huge volumes of data and extract meaningful information and knowledge, there is a need to develop some new techniques and tools. Data mining is a process of extracting useful information from the large data set, when it is applied to the web content is called a web mining. Web mining is the process of web based contents such as documents and links between web pages [1, 2]. The complete web mining is divided into four subtasks:

- Resource finding
- Information selection and preprocessing
- Generalization and
- Analysis

The aim of resource finding is to extract the information from the web documents. During the second task, extract/select the relevant information and filter the irrelevant information from the actual data. Generalization is used to discovery the general patterns by applying machine learning or data mining techniques. During analysis, the patterns are analyzed and verified [3, 4].

II. CLASSIFICATION OF WEB MINING

Web mining is an iterative process for fetching the facts from web data. Fig. 1.1 depicts the major classification of web mining.

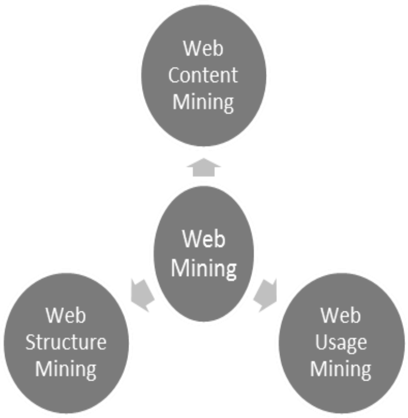

Fig 1: Classification of Web Mining

Structured mining is used when the data available in the tabular form (i.e,) consist of rows and columns. The data content in structured mining is fully structured. Semi-structured mining is used when the data is partially structured in the form of HTML tags. Un-structured mining is used when the data is un-structured which contains images, audio and video.

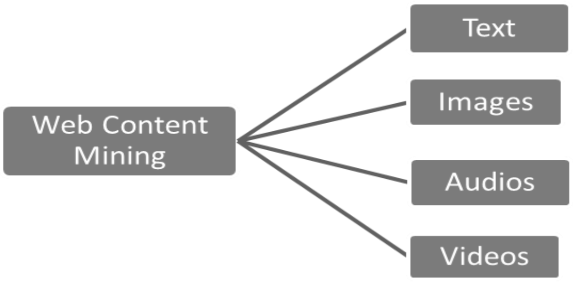

Fig 2: Web Content Mining

Web Content mining can be differentiated into two types of applications. They are

1. Agent based approach and
2. Database approach.

The first approach transforms the information from semi and unstructured data into structured data. The second approach follows the standard database querying mechanism and data mining applications to analyse the result [5].

a. *Web Content Mining Tools:*

The tools related to web content mining is available which can extract useful information from web pages. There are different types tools available for web content mining are:

i. Web Info Extractor:
Web Information Extractor is a powerful tool in web data mining, extracting web content and monitoring the web content update. It has the ability to extract the structured or unstructured data (includes text, images, video and audio) from web pages. No need to define complex rules, browse to the web page and click what to define the extracting content and run as you want or automatically [6].

ii. Screen-Scraper:
Screen-Scraper is a tool to extract and mine data from websites and provide to the user in a format they can use. It can easily invoke screen scraper from .NET, JAVA, PHP.

iii. Web Content Extractor:
Web Content Extractor is a highly accurate and effective tool for extracting data from web sites. This tool extracts the product data from online shopping, stock market, financial, song or movie information, helpful for extracting news from different news sites for reporter. This tool helps for business people to analyse the real estate data, market figures and pricing differences [6].

iv. Octoparse:
Octoparse is client-side software written in .NET for extracting information from websites. It is cloud based web crawling and web scraping software that helps to extract any web data without coding in real time. It can collect data from websites and sort the data into database.

v. Scrapy:
Scrapy is a free and open source software written in Python for extracting data from websites. It is application framework for extracting structured and crawling data used for applications like data mining and information processing.

III. WEB STRUCTURE MINING

Web structure mining is otherwise called as link mining. It is to deal with structure of hyperlink within web pages itself. Based on the hyperlinks, web structure mining will classify the web pages and generate the information [7]. Web structure mining classified into two types as shown in fig. 3.

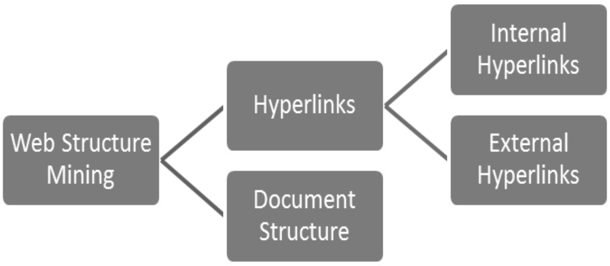

Fig. 3. Web Structure Mining

In web structure mining, hyperlinks can be classified into (i) Internal Hyperlinks and (ii) External Hyperlinks. Internal hyperlinks which connects the web page to another web page in the same website. External Hyperlink which connects the web pages to another web pages in different website [8]. In Document Structure the web pages are described in tree like structure.

a. *Web Structure Mining Tools:*

The tools related to Web structure mining is a process to discover the relationship between web pages linked by information. There are different types of tools available for web structure mining are:

i. HITS Algorithm:

Hyperlink-Induced Topic Search (HITS) is a link analysis algorithm that rates Web pages also known as hubs and authorities. The step in HITS algorithm is to retrieve the most relevant pages. This set is called as root set can be obtained by taking top pages and base set is generated by supplementing the root set with all web pages [9].

ii. Page Rank Algorithm:

Page Rank is an algorithm used by Google search to rank the websites in their search engines. It is a link analysis algorithm and works by counting the number and quality of links to a web page [10].

IV. WEB USAGE MINING

Web usage mining is based on the techniques that could predict the pattern of the user while the user interacts with web. It is otherwise called as web log mining [11]. It collects data from web log records to discover the patterns of web pages. Web log records are unformatted text file which contain data like User name, date, time, IP address, status code etc. whenever the user interacts with website, the information are recorded and maintained in web servers. Web usage mining includes web logs and application logs as shown in fig 4.Web logs maintain data like user browsing history. Application logs business transaction and are stored in application server [12].

Web usage mining consists of three phases:

1. Pre-processing
2. Pattern discovery
3. Pattern analysis

In web usage mining the first step is pre-processing, the noisy and useless data in web usage log file is cleaned and transformed so that the size can be reduced [12]. Second step, pattern discovery the cleaned and transformed log file is used to discover patterns [13]. Third step, Pattern Analysis in which discovered patterns are further analysed to generate more useful and related information to the user [14].

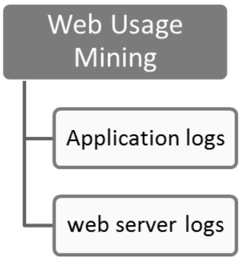

Fig 4: Web Usage Mining

a. Web Usage Mining Tools:

i.R:
R is an open source programming language and also software environment for graphics and statistical computing. It has been made up of languages like C, FORTRAN, Python, Ruby, Perl etc. It compiles and runs on UNIX and Windows platform.

ii. Oracle Data Mining:
Oracle Data Mining is a Data mining software developed by Oracle. It is implemented by Oracle Database Kernel and mining models. Database helps you to predict customer behaviour, customer details and identify best customers. The functions of Oracle Data mining can mine data tables, schema, transactional data, structured and unstructured data.

iii. Tableau:
Tableau is one of the business intelligence tool for analysing the data. It allows user to create and transform data into interactive and variations called dashboards. Data will be represented by graphs and charts. Tableau is used by businesses, researches and many government organizations for visually analysing the data.

iv. Speed Tracer:
Speed Tracer is one of the web usage mining and analysing tool. Speed Tracer tool helps to analyse and debug critical issues in web applications. It is a part of Google web Toolkit. It uses the information like IP address, Timestamp, URL address and session identification [15,16].

VI. CONCLUSION

This paper describes about web content mining, web structure mining and web usage mining including its tools. Internet and websites provide rich platform for searching the information. Most of the websites are complex and larger in their size and structure. Therefore, it is mandatory to develop tools for websites.

REFERENCES

[1] R. Kosala, H. Blockeel.2000, "Web Mining Research: A Survey," In SIGKDD Explorations, ACM Press, pp.1-15.

[2] Md. Zahid Hasan, Khawja Jakaria Ahmad Chisty and Nur-E-Zaman Ayshik, 2012, "Research Challenges in Web Data Mining", International Journal of Computer Science and Telecommunications, Volume 3, Issue 7.

[3] Singh, Brijendra, and Hemant Kumar Singh, 2010, "Web data mining research: A survey", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC).

[4] Kavita Sharma, Gulshan Shrivastava, Vikas Kumar, 2011, "Web Mining: Today and Tomorrow", Proceedings of 3rd International Conference on Electronics Computer Technology, pp. 399 - 403.

[5] B. Singh, H.K. Singh, 2010, "Web data Mining Research", IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10.

[6] Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi, 2013, "Overview of Web Content Mining Tools", International Journal of Engineering and Science (IJES), Vol.2, ISSN: 2319 – 1813 ISBN: 2319 – 1805.

[7] Suvarn Sharma, Amit Bhagat, 2016, "Data Preprocessing Algorithm for Web Structure Mining",IEEE Fifth International Conference on Eco-Friendly Computing and Communication Systems, pp. 94-98.

[8] M. D. Costa and Z. Gong, 2005, "Web structure mining: an introduction", IEEE International Conference on Information Acquisition, pp. 590–595.

[9] Weiming Yang, 2016, "An Improved HITS Algorithm Based on Analysis of Web Page Links and Web Content Similarity", International Conference on Cyberworlds, IEEE, pp.147-150.

[10] Ashish Jain, Rajeev Sharma, Gireesh Dixit and Varsha Tomar, 2013,"Page Ranking Algorithms in Web Mining, Limitations of Existing methods and a New Method for Indexing Web Pages" IEEE International Conference on Communication Systems and Network Technologies.

[11] Lya Hulliyyatus Suadaa, 2014, "A Survey on Web Usage Mining Techniques and Applications", IEEE International Conference on Information Technology Systems and Innovation, PP 24-27,ISBN: 978-1-4799-6526-7.

[12] Theint Theint Aye, 2011, "Web Log Cleaning for Mining of Web Usage Patterns", Proceedings of 3rd International Conference on Computer Research and Development, Volume 2, pp. 490 - 494.

[13] A. Bhargav and M. Bhargav, 2014, "Pattern discovery and users classification through web usage mining," in IEEE International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), pp. 632–636.

[14] Dilip Singh Sisodia, Shrish Verma, 2012, "Web Usage Pattern Analysis through Web Logs: A Review", Proceedings of 2012 International Joint Conference on Computer Science and Software Engineering, pp. 49 - 53.

[15] Arvind Kumar Sharma, P.C. Gupta, 2012, "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining", in International Journal of Advanced Research in Computer Engineering &Technology (IJARCET) Volume 1, Issue 8.

[16] Chhavi, R 2012, 'A Study of Web Usage Mining Research Tool', International Journal of Advanced Networking and Applications, vol. 3, no.6, pp.1422-1429.

D.Nithya received B.E degree Computer Science and Engineering from Avinashilingam University in 2008 and obtained her Master degree in Computer Science and Engineering from Avinashilingam University, Coimbatore in the year 2010. At present she is an assistant professor in Department of Computer Science and Engineering, Faculty of Engineering, Avinashlingam University, Coimbatore, India since 2010. She is currently working toward Ph.D degree in Computer Science and Engineering from Avinashiingam University. Her research area is Web Mining.