International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(4), pp.212-217
DOI: http://dx.doi.org/10.21172/1.74.028
e-ISSN:2278-621X

A SURVEY- WEB MINING TOOLS AND TECHNIQUE

Prof. Prerak Thakkar[1], Prof. Gopi Bhatt[2], Prof. Anirudh Kurtkoti[3], Prof. Siddharth Shah[4], Prof. Chinmay Joshi[5]

Abstract:This paper is a study and analysis of the different Web mining tools and techniques, which is used for mining the information from WWW (World Wide Web). This paper content the different techniques with their benefits and drawbacks for Web Content Mining, Web Structure Mining and Web Usage Mining.

Keywords: Web Mining, WWW, Web Content Mining, Web Structure Mining, Web Usage Mining

I. INTRODUCTION

The World Wide Web is a very huge source of information with the different formats like text, image, audio, video, etc. Now a days WWW is very useful to identify the human interest, which is useful for the business prospective and current scenario of the world. Data mining is a process of extracting useful information from the large dataset, when it is applied to the web content is called a web mining. The complete web mining procedure is divided in to four sub parts [1].

1. Resource finding (Retrieving Web documents): The aim of this process is to collect the data from the web such as e-papers, tweets, user comments etc.
2. Selecting information and Pre-Processing: In this process we will select the relevant information and filter irrelevant data from the actual data collected in Resource finding by using pre-processing.
3. Generalization: This procedure will be used to discover the general patterns by applying machine learning or data mining techniques.
4. Analysis: finally in this stage we will interpret and verify the patterns discovered form the generalization.

[1] *Department of Computer Engineering, A. D. Patel Institute of Technology, Anand, Gujarat.*
[2] *Department of Computer Engineering, A. D. Patel Institute of Technology, Anand, Gujarat.*
[3] *Department of Computer Engineering, A. D. Patel Institute of Technology, Anand, Gujarat.*
[4] *Department of Computer Engineering, A. D. Patel Institute of Technology, Anand, Gujarat.*
[5] *Department of Computer Engineering, A. D. Patel Institute of Technology, Anand, Gujarat.*

II. WEB MINING

Web mining is an iterative process for fetching the facts from the web data. Basically there are three sub categories for mining web information. These sub categories are:

A. Web Content Mining
B. Web Structure Mining
C. Wed Usage Mining

A. Web Content Mining:

Web Content Mining is a process of fetching useful patent from the data available on Web like text content or multimedia content (Images, audio and video) [2]. First we will collect all these data from the different source of Web then we will apply the appropriate technique based on the type of data and analysis result will be generated.

Web Content Mining has mainly three approaches based on type of data

i. Structured mining
ii. Semi-Structured mining
iii. Un-Structured mining

i. Structured mining:

This approach is used when data is fully structured. The data which is available in tabular form which consist of specific rows and columns are known as the fully structured data.
The techniques used for structured data are:

- Web Crawler
- Wrapper Generation
- Page Content Mining

ii. Semi-Structured Mining:

This approach is used when data is partially structured. The data which contains HTML tags are known as the Semi-structured data.
The techniques used for structured data are:

- OEM (Object Exchange Model)
- Top-Down extraction

iii. Un-Structured mining:

This approach is used when data is un-structured. Images, audio, video etc. are un-structured data.
The techniques used for structured data are:

- Multimedia Miner
- Color Histogram Matching
- Shot Boundary Detection

B. Web Structured mining :

Web Structure Mining discovers the link structure of the websites [3]. The goal of the Web Structural Mining is to discover link structure summary of the hyper-links from the web

pages. Web Structural Mining analyse the web pages and identify the structure of the hyperlinks between webpages to find the relationship between different web sites. This analysis is used for the integration and comparison of web page schemas.

C. Web Usage Mining:

Web usage mining is based on the usage of the web user. This includes the techniques which can be used to analyse the behaviour of the web user. It can also be used to predict user access pattern from one or more web server. It contain four processing stages:

i. Data Collection
ii. Data Pre-processing
iii. Data Clustering
iv. Pattern discovery and Analysis

i. Data Collection:

Data collection is the process of collection relevant information and usage pattern trends, which is useful to improve performance and management of the web servers. The main source of data collection is the web log records.

ii. Data pre-processing:

The main aim of the data per-processing is to find the useful information form the data collection. First step in data pre-processing is to remove the irrelevant and noisy data by using data cleaning methods. Then data will be arranged based on the recently accessed data having high index and least recently used data having low index. Inappropriate data may cause data pre-processing very difficult.

iii. Data Clustering:

In data clustering similar objects are a grouped together in a single class called cluster. Then a label is assign to each cluster, through which the cluster is identified.

iv. Pattern discovery and Analysis:

Using this pattern discovery and analysis, important and relevant information can be easily predicted based on data analysis. Web data include from the different sources like web servers, proxy servers, logs, cookies, etc. analysis of these data can help to understand the user behaviour and also used to improve the design of web structure.

III. WEB MINING TOOL

There are different tolls available for the web mining.

A. Web Content Mining [2]:

In the simple words web content mining is nothing but to collect the data available on websites. This is very tedious and time consuming task. To deal with such a process the tools related to web content mining is available which can extract the useful information from the web. Different type of web content mining tools are:

i. Web Info Extractor:

This tool is useful in mining web data, extracting web content, and monitoring content update. It has the ability to extract the structured and unstructured data [4]

ii. Mozenda [5]:

To fetch the web data easily and to manage it efficiently Mozenda is useful tool. Mozenda collect and organize web data in the most efficient and cost effective way possible. Its cloud-based architecture enables rapid deployment, ease of use, and scalability.

iii. Screen-Scraper [6]:

Screen-Scraper mining allows mining the content from the SQL Database and the SQL Servers, which is useful to collect data from SQL for web content mining. One can easily invoke the Screen-Scraper from .NET, JAVA, PHP. It works on Windows, Linux, MAC OS.

iv. Web Content Extractor:

It is most powerful and easy tool for data mining and data retrieval form the internet. This tool is able to extract the data from the online shopping sites, real estate sites, financial sites, etc. in HTML, XML, SQL Script format.

This tool helps the businessmen to analyse the market figures and pricing differences. It is also helpful for news reporter to extract various news from the different news sites.[4]

v. Automation anywhere:

Automation anywhere is data retrieval tool used for fetching data easily from web pages and used for web mining. This tool is used for the automation of business process.

Comparative analysis based on features of different tool: [2]

Tool	Features			
	Prior knowledge Required	Handle structured data	Handle un-structured data	User friendly
Web Info Extractor	-	√	√	Easy
Mozenda	-	√	√	Easy
Screen-Scraper	√	√	√	Complex
Web Content Extractor	-	√	-	Complex
Automation anywhere	-	√	√	Easy

B. Web Structured Mining [7]:

The data on the web are in from of web pages. Web pages are linked with other webpages through which we can identify the cardinality of web pages. Some tasks of link mining are used in web structure mining, which are as follows:

i. Link based classification:

This is a very efficient method to link the webpage content. In this task, it will predict the category of the web-page based on the data content of the web page, links between the web pages html tags, etc. It is supervised analysis.

ii. Link based cluster analysis:

The motive of this analysis is to find the naturally occurring sub-classes. In this task the data is divided in to a groups, called a cluster, where the similar type of data will be placed in the same cluster. It is unsupervised analysis.

iii. Link type:

This task is used to predict the existence of link between two pages including its type and purpose of that link.

iv. Link strength:

Links could be associated with the weights.

v. Link cardinality:

The main task is to predict the number of links between the objects.

Uses of Web structure mining:

- Easy to predict which page will be added to which collection.
- Finding related pages.
- Identify the duplicate websites.
- Used to rank the user's query.

C. Web Usage Mining [8]:

Tasks done in the web usage mining are:

i. Data Pre-processing:

The data which is collected from the log files is noisy data. Before using that data it must be clean, this processes is known as data pre-processing. This process contain four phases: data cleaning, session reconstruction, content and structure information retrieval and data abstraction.

ii. Usage Pre-processing:

It is very difficult task because of the incomplete and inconsistent data in server log. Only IP address, agent and server side click stream are available which leads too many problem like multiple IP/single server session, multiple agent/single user, etc.

iii. Content pre-processing:

Content pre-processing focuses on to convert un-structured and semi-structured data in to from which is used for mining.

iv. Pattern Discovery:

It focuses on the different techniques to recognize the different pattern. Discovery of desired patterns and to extract understandable knowledge from them is a challenging task.

v. Pattern Analysis:

This task identifies the interesting patterns and uninteresting patterns from the patterns discovered from the total pattern discovered in pattern discovery phase.

Tools for web usage mining:

- **Data Pre-processing tools:**

o Data Preparator: It performs data cleaning and transformation on data collected from the web.

- Sumatra TT: It is a platform independent tool for data transformation. It support Rapid application Development.
- Lisp Miner: performs data pre-processing by analysing click stream and data collected.
- Speed tracer: It mines the web server log and reconstruct the navigational path for session identification.

- **Pattern discovery tools:**

- SEWEBAR-CMS: Provides interaction between data analyst and domain expert to perform discovery of pattern.
- i-Miner: Discover data cluster by using Fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis.
- Argunaut: Develop the pattern of useful data by using sequence of various rule.
- MiDas(Mining Internet Data for Associative Sequences): Discover marketing based navigational pattern from log files.

- **Pattern Analysis tools:**

- Webalizer: GNU GPL licence based and produces web pages after analysing pattern.
- Naviz: Visualization tool that combines 2-D graph of visitor access and grouping of related pages. It describes the pattern of user navigation on the web.
- WebViz: Analyze the patterns and provide them in the from of graphical patterns.
- Web Miner: Maines the useful pattern and provides the user specific infromation.
- Stratdyn: Enhances WUM and provides visualization of patterns.

IV. CONCLUSION

This paper describe the deatils of Web content mining, web structure mining and web usage mining incudeing its techniqes and tools with is all the features. so this will be useful for researcher to take decision which tool is better to use for analysis.

REFERENCES

[1] H. Blockeel, R. Kosala, “Web mining research: A survey,” ACM SIGKDD Explorations, Vol. 2 No. 1, pp. 1-15, June 2000.

[2] V. Bharanipriya & V. Kamakshi Prasad, Web Content Mining tools: A Comparative Study in International Journal of Information Technology and Knowledge Management January-June 2011, Volume 4, No. 1, pp. 211-215.

[3] B. Singh, H.K. Singh, “Web data Mining Research”, in the Proceedings of 2010 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, Dec. 2010.

[4] Arvind Kumar Sharma, P.C. Gupta, “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, October 2012

[5] www.mozeda.com

[6] www.screen-scraper.com

[7] Preeti Chopra, Md. Ataullah, a Survey on Improving the Efficiency of Different Web Structure Mining Algorithms in International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-3, February 2013.

[8] Kamika Chaudhary, Santosh Kumar Gupta, Web Usage Mining Tools & Techniques: A Survey in International Journal of Scientific & Engineering Research, Volume 4, Issue 6,June-2013 1762 ISSN 2229-5518