IMPACT: International Journal of Research in Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 3, Issue 10, Oct 2015, 27-34
© Impact Journals

A SURVEY ON WEB MINING TOOLS

KALPANA WANI, ARCHANA SHIRKE & PARIMITA DAS
Assistant Professor, Department of IT, Fr. C. R. I. T Vashi, Navi Mumbai, Maharashtra, India

ABSTRACT

Aim of this paper is to study and analyze different Web Mining Tools and techniques used to mine the information from World Wide Web. This survey will provide the detail information of different web mining tool/techniques for Web Content Mining, Web Structure Mining and Web Usage Mining, as well as a comparative study of their advantages and disadvantages.

KEYWORDS: Search Agent, Personalized Web Agents, Web Usage Mining, Web Content Mining, Web Structure Mining

I. INTRODUCTION

The World Wide Web (WWW) is a vast resource of multiple types of information in varied formats which is very useful for the analysis of business progress, which is very important now a days to stand in the competition of business. Researchers are beginning to investigate human behavior in this distributed Web data warehouse and are trying to build models for understanding human behavior in virtual environments. Data mining, often called Web mining when applied to the Internet, is a process of extracting hidden predictive information and discovering meaningful patterns, profiles, and trends from large databases. Web mining is an iterative process of discovering knowledge and is proving to be a valuable strategy for understanding consumer and business activity on the Web. Basically there are three sub categories for mining web information. These sub categories are

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

A. Web Content Mining

The Web content mining refers to the discovery of useful information from web contents which include text, image, audio, video, etc. The mining of link structure aims at developing techniques to take advantage of the collective judgment of web page quality which is available in the form of hyperlinks that is web structure mining [2]. It includes extraction of structured data/information from web pages, identification, similarity and integration of data's with similar meaning, view extraction from online sources, and concept hierarchy, knowledge incorporation [1].

Web Content Mining Strategies: A. Web Content Mining Approaches:

Two approaches used in web content mining are Agent based approach and database approach. The three types of agents they are

- Intelligent search agents
- Information filtering/Categorizing agent,
- Personalized web agents

Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine information. Adapted web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well formed database containing schemas and attributes with defined domains.

Web content mining has the following approaches to mine data

- Unstructured text mining,
- Structured mining,
- Semi structured text mining, and
- Multimedia mining.

1. Unstructured Text Data Mining:

Most of the Web content data is of unstructured text data. Content mining requires application of data mining and text mining techniques [4]. The research around applying data mining techniques to unstructured text is termed Knowledge Discovery in Texts (KDT), or text data mining, or text mining. Some of the techniques used in text mining are

- Information Extraction,
- Topic Tracking
- Summarization
- Categorization
- Clustering
- Information Visualization

2. Structured Data Mining

The Structured data on the Web represents their host pages. Structured data is easier to extract when compared to unstructured texts. The techniques used for mining structured data are

- Web Crawler
- Wrapper Generation,
- Page content Mining.

3. Semi-Structured Data Mining

Semi-structured data evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real world objects without sending the application writer into contortions. HTML is a special case of such intra-document structure. The techniques used for semi structured data mining are

- Object Exchange Model (OEM),
- Top down Extraction
- Web Data Extraction language

4. Multimedia Data Mining

The techniques of Multimedia data mining are:

- SKICAT
- Color Histogram Matching
- Multimedia Miner
- Shot Boundary Detection.

B. Web Structure Mining

Web structure mining is based on the link structures with or without the description of links. Markov chain model can be used to categorize web pages and is useful to generate information such as similarity and relationship between different websites. The goal of web structure mining is to generate structured summary about websites and web pages. It uses treelike structure to analyze and describe HTML or XML. Some algorithms have been proposed to model the Web topology such as HITS, PageRank and improvements of HITS by adding content information to the links structure and by using outlier filtering . These models are mainly applied as a method to calculate the quality rank or relevancy of each Web page. Some examples are the Clever system and Google . Some other applications of the models include Web pages categorization and discovering micro communities on the Web.

C. Web Usage Mining

The Web usage mining is also known as Web Log mining, which is used to analyze the behavior of website users. This focuses on technique that can be used to predict the user behavior while user interacts with the web. It also uses the secondary data on the web where the activity involves automatic discovery of user access patterns from one or more web servers. It contains four processing stages including

- Data collection
- Preprocessing
- Pattern discovery and Analysis

Data Collection

The data collection is the discovery of hidden information and usage pattern trends, which could aid the Web managers for improving the management, performance and controlling of the Web servers.

Data Preprocessing

The selection of useful data is an important task in the data pre-processing stage. The data's were selected in each data type to generate the cluster models for finding web user access and server usage patterns. The removal of irrelevant and noisy data is an initial step in this task. The most recently accessed data were indexed with higher value of 'time index' while the least recently accessed data were placed at the bottom with lowest value . This becomes the critical step to obtain more precise analysis result due to time dependence characteristics of Web usage data.

Data Clustering

The method of clustering is broadly used in different projects by researchers for finding the usage patterns or user profiles. The clustering algorithms become the most mining method in websites and the cluster objects include user groups (to describe user actions) and web pages.

Pattern Discovery and Analysis

Using this pattern discovery and pattern analysis, relevant and useful information can be easily predicted based on data analysis and Graph. Web usages data includes data from web server access, proxy server and browser logs, user profiles, sessions or transactions, queries, registration data, cookies, bookmark data, mouse clicks and scrolls or any other data as result of interaction. Analysis of web access logs for web sites can help understand the user behavior and also its web structure, thus improving the design of this massive collection of resources. There are two tendencies in Web Usage Mining driven by the applications of the discoveries: General Access Pattern Tracking and Customized Usage Tracking .

II. WEB MINING TOOLS

As we have seen web mining having sub categories as Web Content Mining, Web Structure Mining, Web Usage Mining. For these different categories there are different tools available for mining the data. We will see tools of these different categories one by one.

A. Web Content Mining Tool [2]

i) Web Info Extractor

This tool is helpful in mining web data, extracting web content, and monitoring content update. Thorny template rules are not required to be defined.

ii) Mozenda

To extract web data easily and to manage it affordably Mozenda is useful. With Mozenda, users can set up agents that regularly extract, store and circulate data to several destinations. Once information is in the Mozenda systems users can repurpose, format, and mash up the data to be used in other online/offline applications or as intelligence.

iii) Screen-Scraper

Screen-scraper allows mining the content from the web, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

iv) Web Content Extractor

Most powerful and easy to use data extraction tool for web scraping, data mining or data retrieval from the internet is Web Content Extractor.

v) Automation Anywhere7

Automation anywhere is a web data extraction tool used for retrieving web data effortlessly, screen scrape from web pages or use it for web mining.

Commonalities and Differences between the Above Tools

- **Commonalities**

  All the tools automate the business task and retrieve the web data in an efficient way.

- **Differences**

  - Screen-scrapper needs prior knowledge of proxy server and some knowledge of HTML and HTTP where as other tools do not require any such knowledge and it need Internet connection to run.
  - Automation-Anywhere 7 allows recording of actions this facility is not provided in the other tools.
  - Though we have setup file, Mozenda will not allow us to install without Internet connection, thi is not the case with other tools.

B. Web Structure Mining[3]

There are some possible tasks of link mining which are applicable in Web structure mining and are described as follows:

1. Link-Based Classification

Is the most recent upgrade of a classic data mining task to linked Domains. The task is to focus on the prediction of the category of a web page, based on words that occur on the page, links between pages, anchor text, html tags and other possible attributes found on the web page.

2. Link-Based Cluster Analysis

The goal in cluster analysis is to find naturally occurring sub-classes. The data is segmented into groups, where similar objects are grouped together, and dissimilar objects are grouped into different groups. Different than the previous task, link-based cluster analysis is unsupervised and can be used to discover hidden patterns from data.

3. Link Type

There are a wide range of tasks concerning the prediction of the existence of links, such as predicting the type of link between two entities, or predicting the purpose of a link.

4. Link Strength

Links could be associated with weights.

5. Link Cardinality

The main task here is to predict the number of links between objects. There are some uses of web structure mining like it is:

- Used to rank the user's query
- Deciding what page will be added to the collection
- Page categorization
- Finding related pages
- Finding duplicated web sites
- To find out similarity between them Many authors have proposed web structure mining algorithms like: Page rank algorithm, weighted page rank algorithm, Hyper Induced Topic search algorithm, weighted Topic sensitive page rank algorithm. In next section we will explain these algorithms in detail.

C. Web Usage Mining[1]

There are different task to be carry out in Web Usage Mining that are given below and different tools are used for that work..

i) Data Preprocessing

The first step of Web Usage Mining is preprocessing of data stored in web logs as it is noisy in nature. The process of converting usage, content and structure information into data abstraction is described in preprocessing. The processing of preprocessing consists of four phases: data cleaning, session reconstruction, content and structure information retrieval and data abstraction.

ii) Usage Preprocessing

This is considered as most difficult task of web usage mining because of presence of incomplete and inconsistent data in server log. Only IP address, agent and server side click stream are available to identify users and server sessions which faces many problems like single IP address/multiple server sessions, multiple IP address/single server session, multiple IP address/single user and multiple agent/single user. Usage preprocessing also encountered the problem of inferring cached page references.

Iii) Content Preprocessing

Content preprocessing concerned with transforming unstructured and semi structured documents into the forms that are suitable for web usage mining.

iv) Discovery Pattern Discovery

It focuses on to uncover patterns from the abstractions produced as a result of preprocessing phase. It focuses on applying various methods and techniques developed from several fields such as data mining, machine learning, statistics and pattern recognition. Discovery of desired patterns and to extract understandable knowledge from them is a challenging task. This section explains some of algorithms

v) Pattern Analysis

The last step of web usage mining process is pattern analysis. This phase separates the interesting and uninteresting patterns from the overall patterns discovered during pattern discovery phase.

Table 1: Tools Used in Various Stages of Web Usage Mining

Tools	Features
Data Pre-Processing Tools	
Data Preparator	Performs cleaning, extraction and transformation of data before pattern discovery.
Sumatra TT	Platform independent data transformation tool. Based on Sumatra script and support Rapid application Development
Lisp Miner	Performs data pre-processing by analysing the click stream and data collected.
Speed Tracer	Mines web server logs and reconstruct the user navigational path for session identification
Pattern Discovery Tools	
Sewebar-Cms	Provides interaction between data analyst and domain expert to perform discovery of patterns. Helps in selection of rules among various rules in association rule mining [34].
i-Miner	Discover data cluster by using fuzzy clustering algorithm and fuzzy inference system for pattern discovery and analysis [33]
Argunaut	Develop the patterns of useful data by using sequence of various rules.
MiDas(Mining Internet Data for Associative Sequences)	Discover marketing based navigational pattern from log files. It applies more features to traditional sequential method.
Pattern Analysis Tools	
Webalizer	GNU GPL license based and produces web pages after analyzing patterns.
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 provides them in the form of graphical patterns
Web Miner	Mines the useful patterns and provides the user specific information
Stratdyn	Enhances WUM and provides visualization of patterns

III. CONCLUSIONS

This paper described several tool/techniques for Web Content Mining, Web Structure Mining and Web Usage Mining. We analyzed their strengths and limitations and provide comparison among them. So we can say that this paper may be used as a reference by researchers when deciding which tool/techniques are suitable.

IV. REFERENCES

1. 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.
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. 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.
4. Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, International Journal of Information Technology & Decision Making. Vol.7, No. 4, pp. 683-720. World Scientific Publishing Company (2008).
5. Darshna Navadiya, Roshni Patel, Web Content Mining Techniques-A Comprehensive Survey, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 10, December- 2012 ISSN: 2278-0181.
6. Mozenda, http://www.mozenda.com/web-mining-software Viewed 18 February 2013.
7. Web Content Extractor help. WCE, http://www.newprosoft.com/webcontent-extractor.htm Viewed 18 February 2013. [31] Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.
8. Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.