A Survey on Web Usage Mining

Amrita Soni[1], Dr. Ashish Bansal[2]

M.E.(IT) 4th sem S.V.I.T.S. Indore [1], HOD of IT Department, S.V.I.T.S. Indore[2]

Email: amritasoni13@yahoo.in

ABSTRACT- Web data is expanding day by day; Extraction of useful knowledge from WWW data is considered as web mining. It is mainly concerned with 3 types, about the content (content mining), how should be the structure (structure mining), how and where and how much usage of web data (usage mining) has to be done. Web usage mining has many emerging implications as network traffic control and flow analysis, adaptive website management, personalization, creation of adaptive websites etc. In this paper we introduces a review of web usage mining techniques and its benefits.

Keywords: *WWW* (World Wide Web), web mining, content mining, structure mining, web data.

1. INTRODUCTION

Web mining is the use of Data mining techniques to extract information from web documents and services. Web mining is decomposed into three sub tasks. Resource finding, Generalization and Analysis [1].

"Web mining aims to **discover useful information and knowledge from the Web** hyperlink structure, page contents, and usage data."

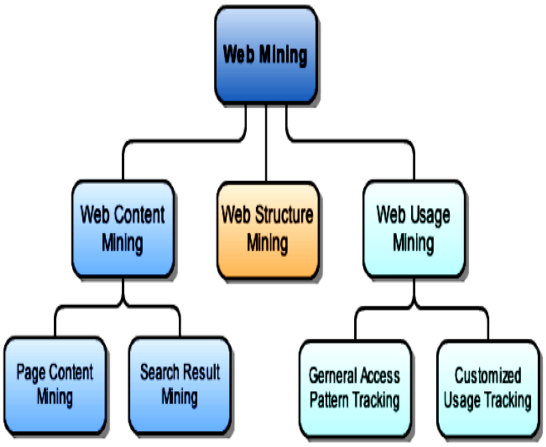

Figure 1: Types of Web Mining

Nowadays the usage of web resources and data are growing exponentially as numbers of users are increasing every day. Growing data storage and usage make the network (WWW) complex and may un-handle in future. So, few efficient techniques of web usage mining are useful for retrieving knowledge from huge data available on the web.

Web mining basically classifies in 3 major categories: content mining, structure mining and usage mining. These techniques are used depending upon what to mine from the web [2]. In this paper we focus mainly web usage mining.

2. DATA MINING VS. WEB MINING

2.1 DATA MINING

Data mining is also called **knowledge discovery in databases** (**KDD**). It is commonly defined as the process of discovering useful **patterns** or knowledge from data sources, e.g., databases, texts, images, the Web, etc. The patterns must be valid, potentially useful, and understandable.

Data mining is a multi-disciplinary field involving machine learning, statistics, databases, artificial intelligence, information retrieval, and visualization.

A data mining application usually starts with an understanding of the application domain by **data analysts** (**data miners**), who then identify suitable data sources and the target data. With the data, data mining can be performed, which is usually carried out in three main steps:

i. **Pre-processing**: The raw data is usually not suitable for mining due to various reasons. It may need to be cleaned in order to remove noises or abnormalities. The data may also be too large and/or involve many irrelevant attributes, which call for data reduction through sampling and attribute selection. Details about data pre-processing can be found in any standard data mining textbook.

ii. **Data mining**: The processed data is then fed to a data mining algorithm which will produce patterns or knowledge.

iii. **Post-processing**: In many applications, not all discovered patterns are useful. This step identifies those useful ones for applications. Various valuation and visualization techniques are used to make the decision.

2.2 WEB MINING

Web mining aims to discover useful information or knowledge from the **Web hyperlink structure**, **page content**, and **usage data**. Although Web mining uses many data mining techniques, as mentioned above it is not purely an application of traditional data mining due to the heterogeneity and semi-structured or unstructured nature of the Web data. Many new mining tasks and algorithms were invented in the past decade. Based on the primary kinds of data used in the mining process, Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

The **Web Mining** process is similar to the **Data Mining** process. The difference is usually in the data collection. In traditional data mining, the data is often already collected and stored in a data warehouse. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. We will devote a whole chapter on crawling. Once the data is collected, we go through the same three-step process: data pre-processing, Web data mining and post-processing. However, the techniques used for each step can be quite different from those used in traditional data mining.

3. TYPES OF WEB MINING

3.1 Web Content Mining

Web content mining is the process of extracting useful information from the contents of web documents. Content data is the collection of facts a web page is designed to contain. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to web content has been the most widely researched. Issues addressed in text mining include topic discovery and tracking, extracting association patterns, clustering of web documents and classification of web pages [4].

Web content mining refers to investigative approaches the user with a to provide structured overview of existing web sites available

Figure2: Web Content Mining (Extraction of knowledge from Website content)[3]

- Deals with the analysis of the content of web pages
- The aim is to facilitate the search for information on the Internet
- Classification and grouping of online documents or finding documents for keywords
- Text mining methods are employed
- Agent-based approach
- Database-based approach.

3.2 Web Structure Mining

The structure of a typical web graph consists of web pages as nodes, and hyperlinks as edges connecting related pages. Web structure mining is the process of discovering structure information from the web. This can be further divided into two kinds based on the kind of structure information used[4].

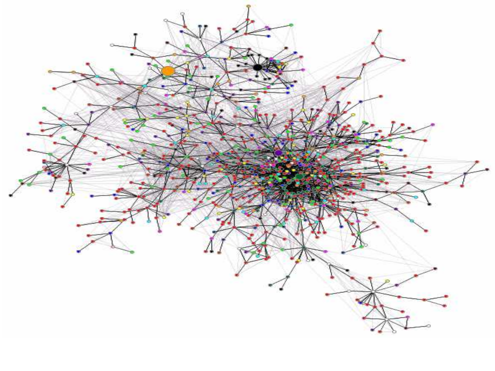

Figure3: Web Structure Mining (Extraction of knowledge from hyperlink structures) [3]

i. **Hyperlinks:** A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an *intra-document hyperlink*, and a hyperlink that connects two different pages is called an *inter-document hyperlink*. There has been a significant body of work on hyperlink analysis, of which Desikan, Srivastava, Kumar, and Tan (2002) provide an up-to-date survey.

ii. **Document Structure:** In addition, the content within a Web page can also be organized in a tree structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents (Wang and Liu 1998; Moh, Lim, and Ng 2000).

3.3 Data Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications (Srivastava, Cooley, Deshpande, and Tan 2000).

Usage data captures the identity or origin of web users along with their browsing behavior at a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

Web Server Data User logs are collected by the web server and typically include IP address, page reference and access time.

Application Server Data Commercial application servers such as Weblogic, 1, 2 StoryServer,3 have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

Application Level Data New kinds of events can be defined in an application, and logging can be turned on for them — generating histories of these events. It must be noted, however, that any end applications require a combination of one or more of the techniques applied in the above the categories.

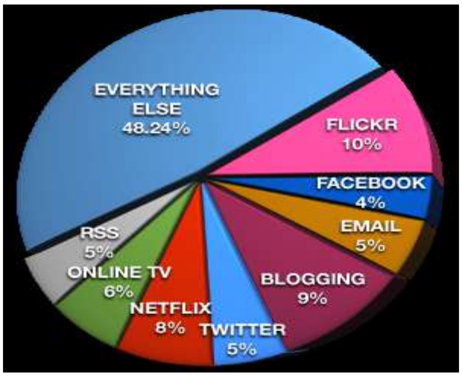

Figure4: Data Usage Mining (Extraction of knowledge from user behavior)[3]

Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with *Web. M. Spiliopoulou* abstract the potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users.

There are no definite distinctions between the Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources, which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining [5].

4. CONCLUSION

In this paper, we have delineated three different types of web mining, namely web content mining, web structure mining and web usage mining. The development and application of Web mining techniques in the context of Web content, usage, and structure data will lead to tangible improvements in many Web applications, from search engines and Web agents to Web analytics and personalization.

Future efforts, investigating architectures and algorithms that can exploit and enable a more effective integration and mining of content, usage, and structure data from different sources promise to lead to the next generation of intelligent Web Applications.

REFERENCES

[1]. "Web Usage Data Clustering using Dbscan algorithm and Set similarities" international conference data storage and data engineering 2010.

[2]. ""A Survey on web usage mining with neural network and proposed solutions on several issues", journal of information, knowledge and research in computer engineering.

[3]. http://webdocs.cs.ualberta.ca/~zaiane/courses/cmput690/slides/Chapter9/sld009.htm,Zugegriffen am 10.04.2010.

[4]. "A Review on Web Mining", International Journal of Engineering Research and Technology (IJERT) Vol. 1 Issue 2, April – 2012 ISSN: 278-0181.

[5]. "Web Mining and Knowledge Discovery of Usage Patterns", Yan Wang, February, 2000.