Chapter 42
Web Usage Mining and the Challenge of Big Data:
A Review of Emerging Tools and Techniques

Abubakr Gafar Abdalla
University of Khartoum, Sudan

Tarig Mohamed Ahmed
University of Khartoum, Sudan

Mohamed Elhassan Seliaman
King Faisal University, Saudi Arabia

ABSTRACT

The web is a rich data mining source which is dynamic and fast growing, providing great opportunities which are often not exploited. Web data represent a real challenge to traditional data mining techniques due to its huge amount and the unstructured nature. Web logs contain information about the interactions between visitors and the website. Analyzing these logs provides insights into visitors' behavior, usage patterns, and trends. Web usage mining, also known as web log mining, is the process of applying data mining techniques to discover useful information hidden in web server's logs. Web logs are primarily used by Web administrators to know how much traffic they get and to detect broken links and other types of errors. Web usage mining extracts useful information that can be beneficial to a number of application areas such as: web personalization, website restructuring, system performance improvement, and business intelligence. The Web usage mining process involves three main phases: pre-processing, pattern discovery, and pattern analysis. Various preprocessing techniques have been proposed to extract information from log files and group primitive data items into meaningful, lighter level abstractions that are suitable for mining, usually in forms of visitors' sessions. Major data mining techniques in web usage mining pattern discovery are: clustering, association analysis, classification, and sequential patterns discovery. This chapter discusses the process of web usage mining, its procedure, methods, and patterns discovery techniques. The chapter also presents a practical example using real web log data.

DOI: 10.4018/978-1-4666-9840-6.ch042

INTRODUCTION

The explosive growth of the internet and the substantial amount of information being generated daily has turned the web into a huge information store. The relationships between the data available online are often not exploited.Web mining analyzes web data to help create a more useful environment in which users and organizations manage information in more intelligent ways. (Srivastava, Cooley, Desphande, & Tan, 2000).

The internet has become an important medium to conduct business transactions. Therefore the application of data mining techniques in the web has become increasingly important to organizations to extract useful knowledge that can be utilized in many ways such as improving the web system performance, restructuring website design, providing personalized web pages, and deriving business intelligence. Web data mining methods have strong practical applications in E-Systems and form the basis for marketing and e-commerce activities. It can be used to provide fast and efficient services to customers as well as building intelligent web sites for businesses. Data mining in e-business is considered to be a very promising research area.

Web data mining deals with different type of data, which is semi-structured or even unstructured, called web data. Web data, can be divided into three categories: content data, structure data, and usage data. This type of data differentiates web mining from data mining.

Web data represent a new challenge to traditional data mining algorithms that work with structured data. The nature of the web data which is less structured, and the rapid growth of information being generated daily, it has become necessary for users to utilize automated tools in order to find the required information. There are several commercial web analysis tools but most of them provide explicit statistics without real knowledge. These tools are also considered slow, inflexible, and provide only limited features. While some tools are being developed that using data mining techniques, but the research still in its first stages and faces real challenges such as large storage requirements and scalability problems (Rana, 2012).

The main objectives of this chapter are:

1. To extensively review the web usage mining methods and types;
2. To identify the mean web usage mining challenges due to the Big Data phenomena;
3. To describe the Big Data solutions for web usage mining;
4. To evaluate the different emerging methodologies and implementation tools for Big Data web usage mining.

This chapter discusses the web usage mining process, also known as web log mining, is a three-phase process: pre-processing, pattern discovery, and pattern analysis. There are many data sources for web usage mining, among all; the web server's log file is the most widely used source of information. This chapter will also cover the following major techniques in web usage mining pattern discovery in relation to Big Data: