Chapter 6
Discover Patterns from Web-Based Dataset

Raghvendra Kumar
LNCT College, India

Priyanka Pandey
LNCT College, India

Prasant Kumar Pattnaik
KIIT University, India

ABSTRACT

The Web can be defined as a depot of varied range of information present in the form of millions of websites dispersed around us. Often users find it difficult to locate the appropriate information fulfilling their needs with the abundant number of websites in the Web. Hence multiple research work has been conducted in the field of Web Mining so as to present any information matching the user's needs. The application of data mining techniques on web usage, web content or web structure data to find out useful data like users' way in patterns and website utility statistics on a whole can be defined as Web mining. The main cause behind development of such websites was to personalize the substance of a website on user's preference. New methods are developed to deal with a Web site using a link hierarchy and a conceptual link hierarchy respectively on the basis of how users have used the Web site link structure.

INTRODUCTION

Navigation and search can be considered as the two primary models to find appropriate data on the web. Most Web users often use the web browser to find their way through a Web site either by beginning through home page or a Web page found through a search engine or linked from another Web site. This is then succeeded by hyperlinks relevant in the starting page and the subsequent pages, until they reach the appropriate information. Usually the search option provided on the Web site is used to speed up searching of data by author Agarwal, C. C. (2004). In case of Web site that pertaining huge number of Web pages and hyperlinks between them, such maneuvers do not justify users need to piece of information. None the less, the fleetness in the information explosion in the Web is not without consequences.

DOI: 10.4018/978-1-5225-1877-8.ch006

Users have many websites that are static in nature and provide the purpose of relaying information to get hold of correct data to quench their needs with no official governance on websites. Static websites have certain advantages and disadvantages. The advantages can be counted as follows that, they are cheap and easy to create and often developed to supply the developers need rather than the user. The biggest disadvantage of the websites is the information contained in them is fixed upon publishing and cannot be changed until the developer decides to publish a newer version of the website which needs professional skill hence increasing the maintenance cost. Furthermore, different users will have dissimilar preferences on the piece of data in a website. Thus increasing the necessity of smart websites to meet user expectations to find appropriate information.

The application of data mining techniques on web resources to find patterns in the Web is called as web mining by author Altingovde, I. S. and Ulusoy, O. (2004). Web resources available are mainly of three kinds, namely usage data, content data and structure data. Web Mining can be categorized into three main parts first is web usage mining; second one is web content mining and last is web structure mining. A common source of usage data is web server logs which are textual data brought together by servers and are enriched with information on users browsing behavior and website usage statistics. It is a huge warehouse of users' activities in a website tracked by the web server. Web usage mining applies data mining techniques on this repository to discover useful knowledge about user behavior in a website. This understanding of web applications can be used to find a way through the website and signify popular links to users. Web content data consists of document's textual information which is highly unstructured and varying in websites. On the other hand, web logs have a fixed set of data fields that are utilized to interpret the data, however no such accurate indication enclosure is offered to represent a document's content. Data mining techniques suggested by author Ansari, S., Kohavi, R., Mason, L. and Zheng, Z. (2001), when applied to the content of a document to ascertain information such as topical relations between documents in a website is via means of web content mining. This piece of information can further be utilized to generate a topical hierarchy of the documents in a website so as to recognize and identify documents equivalent to user's need. Web structure data means hyperlinks interconnecting the group of documents in a website where each a number of outgoing and incoming hyperlinks have. A link between two documents in a website suggests that the documents may be related and may contain relevant information. Data mining techniques are implemented on web structure mining in the network of hyperlinks structure to extort information utilized for various purposes, for example web crawling.

The most primitive move towards intelligent websites was customization suggested by author BBC(2005) and Berendt, B., Mobasher, B., Nakagawa, M. and Spiliopoulou, M. (2002)., which involved altering the interface and contents of a website to go with a user's need. Categorization is done by asking the users to choose from a set of predefined interest categories by manually filling up a form e.g.: Yahoo. The drawback of this approach is that it is time consuming and needs to keep user profiles up-to-date, else they will remain static. Website personalization addresses this restraint via automated user profiling which is understanding of user profiling methods to secure a user's predilection in a website. On the other hand Personalization utilizes web mining techniques to involuntarily build a user's profile from web logs to convert information into a user's document. This was followed by recommender systems gives an idea by author Brickell, J., Dhillon, I. S. and Modha, D. S. (2007), which are web applications that make use of web usage mining techniques on web logs to come up with personalized recommendations on appealing information in a website. These recommendations can be straight forward as signifying popular links in a website or may be complex approaches like proposing links to documents that are related to a user need. Bulk of the tribulations faced by Web users are the inability to find the appropriate