Chapter 9
Rule Optimization of Web-Logs Data Using Evolutionary Technique

Manish Kumar
IIIT, Allahabad, India

Sumit Kumar
IVY Comptech Pvt. Ltd., India

ABSTRACT

Web usage mining can extract useful information from Weblogs to discover user access patterns of Web pages. Web usage mining itself can be classified further depending on the kind of usage data. This may consider Web server data, application server data, or application level data. Web server data corresponds to the user logs that are collected at Web servers. Some of the typical data collected at Web server are the URL requested, the IP address from which the request originated, and timestamp. Weblog data is required to be cleaned, condensed, and transformed in order to retrieve and analyze significant and useful information. This chapter analyzes access frequent patterns by applying the FP-growth algorithm, which is further optimized by using Genetic Algorithm (GA) and fuzzy logic.

1. INTRODUCTION

Web mining is a data mining task to discover and retrieve useful information from large dataset. Web mining can be divided into: Web usage mining, Web content mining and Web structure mining. Web usage mining is a process of extracting useful information from Web-server logs i.e. user's history. Web content mining is the process to discover useful information from text, image, audio or video data and Web structure mining is the process to analyze the connection and node structure of a Web site. The phases involved in Web usage mining are data preprocessing, pattern discovery and pattern analysis. Preprocessing phase involves removal of unusual data like sound, image, graphics files and several server error codes. Pattern discovery extracts useful patterns from user sessions applying association rule mining and FP-growth (Han and Kamber, 2006). FP-growth algorithm is used for

DOI: 10.4018/978-1-4666-7456-1.ch009

generating association rules. The two important approaches for the optimization of the association rule: genetic algorithm (Agrawal, Lad and Manish, 2004; Pardasani, Parveen and Virendra, 2010) is applied with fuzzy logic (Jaisankar, Kannan and Veeramalai, 2010).

2. RELATED WORKS

Web usage mining consists of three phases: preprocessing, pattern discovery, and pattern analysis. B. Santhosh and Rukmani(2010) worked on 'Implementation of Web Usage Mining by using Apriori and FP-growth Algorithms'. Authors used Apriori algorithm to generate association rules that identifies the usage pattern of the client for a particular website. The output of the system is in term of memory usage and speed of producing association rules. Iyakutty and Sujatha (2010) proposed a new framework for Web usage data clustering for user's session. Web clustering involves grouping of the similar object and dissimilar object in different group. The initial clusters are selected based on statistical model to allow the iterative algorithm to converge to a better local minima and improving cluster quality using genetic algorithm based refinement. The method is scalable and can be coupled with a scalable clustering algorithm to address the large-scale clustering problems in Web mining. Biwei Li and Cunlai Chai (2010) presented a GA-based method to derive the fuzzy sets from a set of given transactions. Genetic algorithms provide efficient search algorithms to select a model, from mixed media data, based on preference criterion and objective function. It combines the strengths of rough set theory and genetic algorithm. Arslan et al. (2006) proposed method to find sequential accesses from weblog files using genetic algorithm. Weblog transaction, whether completed or not, is recorded and stored unstructured. Analyzing these log files is one of the important research areas of Web mining. Gyenesei (2000) presented methods for mining fuzzy quantitative association rules; namely without normalization and with normalization. The results showed that the numbers of large itemset and interesting rules found by fuzzy method are larger than the discrete method (Agrawal and Srikant, 1999). Hadzic and Hecker, (2011) presented an approach where a tree structured data is converted into flat represenattion for preserving the structural and attribute value information, thus enabling a wider range of data mining and analysis techniques. Luan et al. (2012) introduced an association rule algorithm for Web log mining that reduces the search range and avoids the problem of combinatorial explosion. Nithya and Sumathi (2012) focused on data cleaning by removing the noise. Weber et al. (2012) suggested an approach where data blogs can be used to visualize political issues covering various sub issues. In Mele (2013), Author focused on improving the search engine performance by using static caching and recommending the interesting Web pages, articles and blogs.

3. OBJECTIVE

Web contains large amount of incredible information. Though it is tough to deal with vast information with user's perspective, Web service provider's perspective and business analyst's perspective because of its high complexities. Web service providers want to predict the user's behavior to design the website according to user's perspective and also to reduce the traffic load. Analysis can be done on the user's history from weblog patterns to retrieve useful information. This information can be used in different forms and places in e-business, website designing, market campaigns, measuring the success of marketing efforts, customer-company behavior and many more applications.