A Memory Efficient Algorithm with Enhance Preprocessing Technique for Web Usage Mining

Nisarg Pathak¹, Viral Shah², and Chandramohan Ajmeera³

¹ Narsinhbhai Institute of Computer Studies and Management, Sarva Vidyalaya Campus, Kadi - 382 715, Gujarat, India  
nisarg.pathak@gmail.com

² SAP BI-BO, Accenture India, Mumbai, India  
viral54@gmail.com

³ Department of Computer Science and Engineering, Osmania University, Hyderabad, Telangana, India  
cm.ajmeera@gmail.com

Abstract. Huge amount of data is generated daily by billions of web users. The usage pattern of the web data could be very prized to the company in the field of understanding consumer behavior. Web usage mining includes three phases namely preprocessing, pattern discovery and pattern analysis. The focus of this paper is to establish an algorithm for pattern discovery based on the association between the users accessed web pages. We have proposed a complete preprocessing methodology to identify the distinct users. The foundation of the algorithm is to find the frequently accessed web pages. The biggest constrain for mining web usage patterns are computation overhead and memory overhead. The performance evaluation of algorithm shows that our algorithm is efficient and scalable.

Keywords: Web Usage Mining, Clustering, Association rules mining, Server Log, User Identification.

1 Introduction

Web Mining is applications of data mining techniques on web server log data to automatically retrieves, extract and analyze information. To discover the knowledge regarding website usage it is essential to mine the web usage files generated on web servers [¹], [²]. The web mining is roughly categories in three different horizons as follows:

Web Content Mining. Mining the information contain in the web page i.e. text, image, audio or video data in the web.

Web Structure Mining. It is mining the website’s hyperlink structure. It uses the graph theory to analyze the graph structure of web.

Web Usage Mining. It is used to discover the usage pattern of web surfing, to understand the surfing patterns. It uses the secondary data retrieve from the interaction of users to the web server while surfing the website. The
user interaction generates *registration data, user session, cookies, user URL request, & mouse clicks* etc... The web usage mining is three step process:

– Preprocessing: It cleans web log data by removing log entries that are not needed for the mining process, data integration, identify users, & sessions.

– Pattern discovery: Different data mining methods alike path analysis, association rule, sequential patterns, cluster & classification are applied to discover patterns.

– Pattern analysis: The patterns are analyzed using various tools to filter out the potentially useful rules/patterns [9].

The discovered rules and patterns can then be used for improving the system performance for making modifications to the web site. The web usage mining is an aid to improve the website, to attract visitors, or to give regular users a personalized and adaptive services.

### 1.1 Pattern Discovery Phase

Association rule mining is an important and fundamental task of data mining to find the correlation among data items [10]. We have proposed the algorithm here where we have tried to apply the *Clustered Association Rule Mining* to the web usage data. We have obtain the association among web pages that are frequently appear together in user’s session. At the end of the algorithm execution we have the result like $A.html, B.html \rightarrow C.html$. The interpretation of the result is *The user who have visited page A and page B, it is quite possible that during the same session user will visit the page C.*

The information contained in association rules can be used to learn about website visitor behavior patterns, enhance website structure making it more effective for the visitors, or improve web marketing campaigns [3] [4]. The web page association is differs from the general association mining on transactional databases. The web usage data contains a large number of tightly correlated items due to the link structure of a website [5]. The web pages in the same user sessions gives *hard association rules* which may not be the potentially useful [6]. Some of the proposed association rule interestingness measures are all-confidence [7], collective strength [8], conviction and lift.

### 1.2 Web Log Data

It’s the data which points the information about website visitor activity. Any user query is register as a log information in server whenever a user request for any page, text, image etc...

**Location of Web Log File:** Web log file can be located either from *Web server logs* or *Web proxy server* or can be from *Client’s browser*.

**Type of Web Log File:** Server logs are also found in different types. They are *Access log file, Error log file, Agent log file and Referrer log file*.

**Web Log File Format:** The file is simple plain text file. The display of log files data are either in *W3C Extended format* or *NCSA common and IIS format*. 