

A COMPLETE PREPROCESSING METHODOLOGY FOR WUM

**CHANDRASHEKAR HC¹, NEELAMBIKE S² & VEERAGANGADHARA SWAMY TM³
& B. S. THEERTHARAJU**

^{1,2}Research Scholar, Department of CSE, VTU, Karnataka, India

³Assistant Professor, Department of ISE, VTU, Karnataka, India

⁴Associate Professor, Department of ISE, VTU, Karnataka, India

ABSTRACT

The popularity of Web resulted in heavy traffic in the Internet. This intense increase in internet traffic has caused significant increase in the user perceived latency. In order to reduce this, WUM is found to be potentially useful. The exponential growth of the Web in terms of Web sites and their users has generated huge amount of data related to the user's interactions with the Web sites. This data is recorded in the Web access log files of Web servers and usually referred as Web Usage Data (WUD). In this paper, a comprehensive preprocessing methodology is proposed as a prerequisite and first stage for Web mining applications that has four steps:

Data Collection, Data Cleaning, Identification of Users & Sessions, and finally the Data Aggregation. Here an attempt is made to reduce the quantity of the WUD and thereby improve the quality of WUD for effective mining of usage patterns. Several heuristics have been proposed for cleaning the WUD which is then aggregated and recorded in the relational database. Experimental results show that the proposed methodologies reduce the volume of the WUD effectively.

KEYWORDS: Web Usage Mining, Data Preprocessing, Server Logs, Users and User Sessions