Abstract

Today internet has made the life of human dependent on it. Almost everything and anything can be searched on net. Web pages usually contain huge amount of information that may not interest the user, as it may not be the part of the main content of the web page. Web Usage Mining (WUM) is one of the main applications of data mining, artificial intelligence and so on to the web data and forecast the user’s visiting behaviors and obtains their interests by investigating the samples. Since WUM directly involves in applications, such as, e-commerce, e-learning, Web analytics, information retrieval etc. Weblog data is one of the major sources which contain all the information regarding the users visited links, browsing patterns, time spent on a particular page or link and this information can be used in several applications like adaptive web sites, modified services, customer summary, pre-fetching, generate attractive web sites etc. There are varieties of problems related with the existing web usage mining approaches. Existing web usage mining algorithms suffer from difficulty of practical applicability. This paper continues the line of research on Web access log analysis is to analyze the patterns of web site usage and the features of users behavior. It is the fact that the normal Log data is very noisy and unclear and it is vital to preprocess the log data for efficient web usage mining process. Preprocessing is the process comprises of three phases which includes data cleaning, user identification, and pattern discovery and pattern analysis. Log data is characteristically noisy
and unclear, so preprocessing is an essential process for effective mining process. In this paper, a novel pre-processing technique is proposed by removing local and global noise and web robots. Preprocessing is an important step since the Web architecture is very complex in nature and 80% of the mining process is done at this phase. Anonymous Microsoft Web Dataset and MSNBC.com Anonymous Web Dataset are used for evaluating the proposed preprocessing technique.

References


Index Terms

Computer Science
Information Sciences
Keywords
Preprocessing  Data Cleaning  Path Completion  Travel Path set  Content Path