WEB USAGE MINING: USER NAVIGATIONAL PATTERNS EXTRACTION FROM WEB LOGS

Thilaga Rani
M.Phil Scholar
Dr.S.N.S Rajalakshmi College of Arts & Science
Coimbatore

Abstract

Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files. In this paper, we define the notion of a “user session” as being a temporally compact sequence of web accesses by a user. We also define a new distance measure between two web sessions that captures the organization of a web site. Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Web usage information mining could help to engage new customers, maintain current customers, track customers who are leaving web site.

Keywords: web usage mining, web mining.

Introduction

Usage information can be extracted to increase web server efficiency by perfecting and caching strategies. Based on several researches done in the area of web mining, we can broadly classify it into three domains: web content mining, web structure mining, and web usage mining. Personalization is a recent and informally-articulated notion, and deals with tailoring a user’s interaction with the web information space based on information about them. Web content mining is the process of extracting knowledge from web documents such as text and multimedia. Knowledge extraction from the structure of web and hyperlink references is called web structure mining. Web usage mining is the process of knowledge exploitation from the secondary data. By secondary data, we mean browser logs, user profiles, web server access logs, registration data, user sessions or transactions, cookies, user queries, mouse clicks and any other data that is the result of interaction with the Web. Mining typical user profiles from the vast amount of historical data stored in server or access logs is a possible approach to personalization that has been recently proposed. The goal of our web mining is to categorize these sessions. We define the notion of a “user session” as being a temporally compact sequence of web accesses by a user. In this paper, we propose a new method for extracting patterns from web logs based on ant clustering algorithm. We apply ant-based clustering for pattern discovery, other similar methods applied ant colony clustering to segregate visitors.

Methodology

The log file consists of image files (gif, jpg, bmp, jpeg…) and other unneeded resources like javascripts and errors. For user identification, we use IP address and session timeout of 30 minutes. So, a user with an IP address has 30 minutes to navigate in the web site. After a user's navigational sequence is extracted, it is displayed in a suitable format for the second step,
Pattern Discovery. We then classify the URLs of the web site into many groups and assign a number to each Group. Then, each user’s requested URL is substituted with its corresponding number. The output of this step is a file that consists of records, each record representing a navigational sequence of users in numbers. The log data from accesses to the server during a period of days was used. After filtering out irrelevant entries, the data was segmented into sessions. The maximum elapsed time between two consecutive accesses in the same session was set to &’ minutes.

Conclusion

In this paper, we have presented a new approach for automatic discovery of user session profiles in web log data. We defined the notion of a “user session” as being a temporally compact sequence of web accesses by a user. A new similarity measure to analyze session profiles is presented which captures both the individual URLs in a profile as well as the structure of the site. We have proposed a new method to extract navigational patterns from web logs.

References