Web User Session Clustering Using Modified K-Means Algorithm

G. Poornalatha[1] and Prakash S. Raghavendra[2]

Department of Information Technology,
National Institute of Technology Karnataka (NITK), Surathkal,
Mangalore, India
[1]poornalathag@yahoo.com, [2]srp@nitk.ac.in

Abstract. The proliferation of internet along with the attractiveness of the web in recent years has made web mining as the research area of great magnitude. Web mining essentially has many advantages which makes this technology attractive to researchers. The analysis of web user's navigational pattern within a web site can provide useful information for applications like, server performance enhancements, restructuring a web site, direct marketing in e-commerce etc. The navigation paths may be explored based on some similarity criteria, in order to get the useful inference about the usage of web. The objective of this paper is to propose an effective clustering technique to group users' sessions by modifying K-means algorithm and suggest a method to compute the distance between sessions based on similarity of their web access path, which takes care of the issue of the user sessions that are of variable length.

Keywords: web mining, clustering; K-means, Jaccard Index.

1 Introduction

Now the present generation is living in an information era. Moreover, the evolution of the internet along with the popularity of the web has made even an ordinary person to use the information available at his finger tips for various purposes. Web has been adopted as a critical communication and information medium by a majority of the population. Due to the rapid growth in the use of web the task of analyzing, understanding and producing useful information manually from a vast quantity of data available on the web is a very complicated and time consuming task. Thus, there is a requirement to develop techniques to get the valuable information, hidden in the web data, so as to improve the web performance.

This paper focuses on clustering web user sessions based on their navigation path which is of variable length. Clustering is a technique for grouping user sessions such that, within a single cluster the usage pattern is more similar while sessions in different groups are dissimilar. The knowledge discovered from the clustering may be used to analyze the pattern of usage of the web site by the user, to recommend for restructuring of web site, to pre-fetch or cache the pages and predict the next page

A. Abraham et al. (Eds.): ACC 2011, Part II, CCIS 191, pp. 243–252, 2011.

visited by the user to reduce the latency etc. As a result, realizing user's navigation patterns on a web site is an important activity for browser to pre-fetch as well as the web site designer to take decisions on redesigning the site.

A number of clustering approaches have been proposed in the literature. For example, Federico et al. [1] present a survey of the developments in the area of web usage mining, where the view points on various techniques like association rules, clustering, sequence patterns etc. are given. Yunjuan et al. [2] suggest that the focus of web usage mining should be shifted from single user session to group of user sessions and applied clustering for identifying such cluster of similar sessions. They introduce an effective clustering technique using belief function based on Dempster-shafer's theory. Chaofeng Li et al. [3] presented an algorithm for clustering of web session based on increase of similarities. Here number of clusters is defined according to the knowledge of application fields and uses ROCK to decide the initial point for each cluster.

Dariusz Krol et al. [4] investigated on the internet system user behavior using cluster analysis. Here sessions are represented as vectors where each dimension represents a web page and stores the value of user interest in each page of a session. The sessions are clustered using Hard C-Means algorithm. Yongjian Fu et al. [5] proposed a generalization based clustering method which employs the attribute-oriented induction method to reduce the large dimensionality of data. Prakash S Raghavendra et al. [6] modeled user behavior as a vector of the time spent at each URL. The cosine of the vector is taken as the similarity/distance measure, instead of euclidean distance and modified the standard k-means algorithm accordingly. Jin-HuaXu et al. [7] presented vector analysis and k-means based algorithm for mining user clusters.

In the web usage domain, there are two kinds of interesting clusters to be discovered: usage clusters and page clusters. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user/s query or past history of information [8]. George Pallis et al. [9] assessed the quality of user session clusters in order to make inferences regarding the users' navigation behavior.

The studies have shown that the most commonly used partitioning-based clustering algorithm, is the K-means algorithm, which is more suitable for large datasets. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. Euclidean distance is generally used as a metric. The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments.

In this paper, an effective method is proposed to compare variable length sessions and basic k-means algorithm is modified to get effective clusters, such that the initial centroid assignments will not have much impact on the clusters. Jaccard index is used to analyze the goodness of the clusters obtained, while [9] uses chi square test to validate the clusters obtained by using EM algorithm. The main contribution of this paper is to propose, improved way of comparing user sessions represented as vectors, that are of variable length inherently and employing Jaccard index for analyzing the