Chapter II
SeqPAM:
A Sequence Clustering Algorithm for Web Personalization

Pradeep Kumar
University of Hyderabad, India

Raju S. Bapi
University of Hyderabad, India

P. Radha Krishna
Institute for Development & Research in Banking Technology, India

ABSTRACT

With the growth in the number of Web users and necessity for making information available on the Web, the problem of Web personalization has become very critical and popular. Developers are trying to customize a Web site to the needs of specific users with the help of knowledge acquired from user navigational behavior. Since user page visits are intrinsically sequential in nature, efficient clustering algorithms for sequential data are needed. In this chapter, we introduce a similarity preserving function called sequence and set similarity measure S3M that captures both the order of occurrence of page visits as well as the content of pages. We conducted pilot experiments comparing the results of PAM, a standard clustering algorithm, with two similarity measures: Cosine and S3M. The goodness of the clusters resulting from both the measures was computed using a cluster validation technique based on average levensthein distance. Results on pilot dataset established the effectiveness of S3M for sequential data. Based on these results, we proposed a new clustering algorithm, SeqPAM for clustering sequential data. We tested the new algorithm on two datasets namely, cti and msnbc datasets. We provided recommendations for Web personalization based on the clusters obtained from SeqPAM for msnbc dataset.

INTRODUCTION

The wide spread evolution of global information infrastructure, especially based on Internet and the immense popularity of Web technology among people, have added to the number of consumers as well as disseminators of information. Until date, plenty of search engines are being developed,

however, researchers are trying to build more efficient search engines. Web site developers and Web mining researchers are trying to address the problem of average users in quickly finding what they are looking for from the vast and ever-increasing global information network.

One solution to meet the user requirements is to develop a system that personalizes the Web space. Personalizing the Web space means developing a strategy, which implicitly or explicitly captures the visitor's information on a particular Web site. With the help of this knowledge, the system should decide what information should be presented to the visitor and in what fashion.

Web personalization is an important task from the point of view of the user as well as from the application point of view. Web personalization helps organizations in developing customer-centric Web sites. For example, Web sites that display products and take orders are becoming common for many types of business. Organizations can thus present customized Web pages created in real time, on the fly, for a variety of users such as suppliers, retailers, and employees. The log data obtained from various sources such as proxy server and Web server helps in personalizing Web according to the interest and tastes of the user community. Personalized content enables organizations to form lasting and loyal relationships with customers by providing individualized information, offerings, and services. For example, if an end user visits the site, she would see pricing and information that is appropriate to her, while a re-seller would see a totally different set of price and shipping instructions. This kind of personalization can be effectively achieved by using Web mining approaches. Many existing commercial systems achieve personalization by capturing minimal declarative information provided by the user. In general, this information includes user interests and personal information about the user. Clustering of user page visits may help Web miners and Web developers in personalizing the Web sites better.

The Web personalization process can be divided into two phases: off-line and online (Mobasher, Dai, & Luo, 2002). The off-line phase consists of the data preparation tasks resulting in a user transaction file. The off-line phase of usage-based Web personalization can be further divided into two separate stages. The first stage is preprocessing of data and it includes data cleaning, filtering, and transaction identification. The second stage comprises application of mining techniques to discover usage patterns via methods such as association-rule mining and clustering. Once the mining tasks are accomplished in the off-line phase, the URL clusters and the frequent Web pages can be used by the online component of the architecture to provide dynamic recommendation to users.

This chapter addresses the following three main issues related to sequential access log data for Web personalization. Firstly, for Web personalization we adopt a new similarity metric S^3M proposed earlier (Kumar, Rao, Krishna, Bapi & Laha, 2005). Secondly, we compare the results of clusters obtained using the standard clustering algorithm, *Partition Around Medoid (PAM),* with two measures: *Cosine* and S^3M similarity measures. Based on the comparative results, we design a new partition-clustering algorithm called

Table 1. Table of notations

Symbol	Description
D	Dataset
N	Total number of item sets in D
k	Number of clusters
$\hat{t}_j$	Medoid of j^{th} cluster
t_{j_s}	s^{th} member of j^{th} cluster
$\|C_j\|$	Total number of items in the j^{th} cluster
τ	Tolerance on total benefit