Analysis of Click Stream Patterns Using Soft Biclustering Approaches

P. K. Nizar Banu, Narasu’s Sarathy Institute of Technology, India
H. Hannah Inbarani, Periyar University, India

ABSTRACT

As websites increase in complexity, locating needed information becomes a difficult task. Such difficulty is often related to the websites’ design but also ineffective and inefficient navigation processes. Research in web mining addresses this problem by applying techniques from data mining and machine learning to web data and documents. In this study, the authors examine web usage mining, applying data mining techniques to web server logs. Web usage mining has gained much attention as a potential approach to fulfill the requirement of web personalization. In this paper, the authors propose K-means biclustering, rough biclustering and fuzzy biclustering approaches to disclose the duality between users and pages by grouping them in both dimensions simultaneously. The simultaneous clustering of users and pages discovers biclusters that correspond to groups of users that exhibit highly correlated ratings on groups of pages. The results indicate that the fuzzy C-means biclustering algorithm best and is able to detect partial matching of preferences.

Keywords: Biclustering, Fuzzy Biclustering, Page Clustering, Rough Biclustering, User Clustering, Web Mining, Web Usage Mining

INTRODUCTION

The World Wide Web is a popular and interactive medium to disseminate information today. The web is huge, diverse, and dynamic and thus raises the scalability, multimedia data, and temporal issues respectively. Due to these situations, we are currently drowning in information and facing information overload (Maes, 1994; Chakraborty, Dom, Gibson, Kleinberg, Kumar, & Raghavan, 1999). Therefore, web mining becomes an active and popular research field. Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services (Nasraoui, Rojas, & Cardona, 2006; Etzioni, 1996; Zamir & Etzioni, 1998). Based on the several research studies we can broadly classify web mining into three domains such as web content mining, web structure mining and web usage mining. Web Content Mining is the process of extracting knowledge from the content of the web documents and their descriptions. Web Structure Mining is the process of inferring knowledge from the structure of data.
Web usage mining is the process of applying data mining techniques to the discovery of behavior patterns based on web data, for various applications. Usage patterns extracted from web data can be applied to a wide range of applications such as web personalization, system improvement, site modification, business intelligence discovery, usage characterization, and so on. The overall process of web usage mining is generally divided into two main tasks; data preparation and pattern discovery. The data preparation tasks build a server session file where each session is a sequence of requests of different types made by single user during a single visit to a site (Liu, Chen, & Song, 2002; Yan, Jacobsen, Garcia-Molina, & Dayal, 1996). Pattern discovery converge the algorithms and techniques from several research areas, such as data mining, machine learning, statistics, and pattern recognition. In Web Usage Mining research existing approaches are single sided approaches which discover either user cluster (Xiao, Zhang, Jia, & Li, 2001) or page cluster (Smith & Ng, 2003).

Recently, web usage mining techniques have been widely applied for discovering interesting and frequent user navigation patterns from Web server logs (Liu & Keselji, 2007). Most of these efforts have proposed using various data mining or machine learning techniques to model and understand web user activity. Of the existed methods, some are non-sequential, such as association rule mining and clustering; and some are sequential, such as sequential or navigational pattern mining (Perkowitz & Etzioni, 2000).

By analyzing the characteristics of the clusters, web designers may understand the users better and may provide more suitable, customized services to the users. In web usage mining, clustering algorithms can be used in two ways: usage clickstream clusters (Xiao et al., 2001) and page clickstream clusters (Song & Shepperd, 2006). User clustering and Page clustering algorithms cannot detect partial matching of preferences because their similarity measures consider the entire set of users or pages respectively. The simultaneous clustering of users and pages discovers biclusters which correspond to group of users which exhibit highly correlated ratings on groups of pages. In this paper, three Biclustering algorithms are proposed to extract user profiles by integrating user clustering and page clustering techniques. Therefore, this research explores a new user profiling method integrating user clustering and page clustering techniques.

Furthermore User clustering and page clustering algorithms cannot detect partial matching of preferences because their similarity measures consider the entire set of users or pages respectively.

The paper is organized as follows. The first section describes user and page clustering approaches for web usage mining. A brief introduction to the biclustering technique is presented. The proposed Biclustering algorithms to perform user and page clustering simultaneously is described. The experimental results are then presented and analyzed.

BACKGROUND

Regarding web usage clustering algorithm, there exist two main approaches: user clustering which clusters users based on their similarity and page clustering which creates page clusters based on similarity between pages. However both user and page clustering are one sided approaches, in the sense, that they examine similarities either only between users or only between pages separately. This way, they ignore the clear duality that exists between users and pages. This research paper proposes a method for integrating user clustering and page clustering techniques. Traditional clustering approaches generate partitions; in a partition, each pattern belongs to one and only one cluster. Hence, the clusters in a hard clustering are disjoint. The rough clusters resulting from the proposed algorithm provide interpretations of different navigation orientations of users present in the sessions without having to fit each object into only one group. Such descriptions can help web miners to identify potential and
Sharing Knowledge When it Cannot be Made Explicit: The Case of Product Lifecycle Management Systems

The Effect of Total Quality Management on Innovation Performance with Mediator Role of Organizational Learning: (Case Study: ABBAND Mazandaran)