Chapter 16
Usage Profile Generation from Web Usage Data Using Hybrid Biclustering Algorithm

R. Rathipriya
Periyar University, India

K. Thangavel
Periyar University, India

J. Bagyamani
Government Arts College, Dharmapuri, India

ABSTRACT

Biclustering has the potential to make significant contributions in the fields of information retrieval, web mining, and so forth. In this paper, the authors analyze the complex association between users and pages of a web site by using a biclustering algorithm. This method automatically identifies the groups of users that show similar browsing patterns under a specific subset of the pages. In this paper, mutation operator from Genetic Algorithms is incorporated into the Binary Particle Swarm Optimization (BPSO) for biclustering of web usage data. This hybridization can increase the diversity of the population and help the particles effectively escape from the local optimum. It detects optimized user profile group according to coherent browsing behavior. Experiments are performed on a benchmark clickstream dataset to test the effectiveness of the proposed algorithm. The results show that the proposed algorithm has higher performance than existing PSO methods. The interpretation of this biclustering results are useful for marketing and sales strategies.

1. INTRODUCTION

With the speedy growth of the World Wide Web (WWW), the study of knowledge discovery in web, modeling and predicting the user's access on a web site has become very important .Web usage mining is a researching area that studies the mining of usage data from the web log files. Its objective is to mine web log files to discover relations among users regarding their browsing interest (Chakraborty & Maka, 2005).

From the business and application point of view, knowledge obtained from the Web usage patterns could be directly applied to efficiently manage activities related to e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Gov-

DOI: 10.4018/978-1-4666-3628-6.ch016

ernment, Digital Libraries, and so on (Abhraham & Ramos, 2003). Jespersen, Throhauge, and Bach Pedersen (2002) proposed a hybrid approach for analyzing the visitor click stream sequences.

A user profile (Chen & Shahabi, 2001) is a collection of personal information. The information is stored without adding further description or interpreting this information. It represents cognitive skills, intellectual abilities, intention of browsing, browsing styles, preferences and interactions with the pages of specific web sites. User profiling is the process that refers to construction of user profile via the extraction from a set of data and it is a fundamental task in web personalization.

In Martín-Bautista, Kraft, Vila, Chen, and Cruz (2004) and Martín-Bautista, Vila, and Escobar-Jeria (2008), two types of profiles are proposed. They are simple profiles which are represented by data extracted from the users' interest and the extended profiles containing the additional information about the user such as the age, the language level, location and others. Mobasher, Cooley, and Srivastava (1999) and Mobasher, Dai, Luo, Nakagawa, Sun, and Wiltshire (2000) proposed the web personalization system, which consists of offline tasks related to the mining if usage data and online process of automatic Web page customization based on the knowledge discovered. The LumberJack model proposed by Chi, Rosien, and Heer (2002) builds up user profiles by combining both clustering of user sessions and traditional statistical traffic analysis using k–means algorithm.

Li (2009) has attempted to provide an up-to-date survey of the rapidly growing area of Web session clustering and analyzed the shortcoming of traditional similarity measurement between web sessions. They proposed a framework of Web session clustering using sequence alignment in computational biology. Lee and Fu (2008) used hierarchical agglomerative clustering to cluster users' browsing behaviors. In this paper, an improved Two Levels of Prediction Model was presented to achieve higher hit ratio which did not suffer from the heterogeneity user's behavior. Labroche (2007) proposed a comparison of relational clustering algorithms on web usage data to characterize user access profiles. These methods only rely on numerical values that represents the distance or the dissimilarity between web user sessions to construct web user profiles.

In the literature, in order to obtain the user profiles from web usage data, clustering and association rules (Agrawal, Imielinski, & Swami, 1993) are applied frequently. User profiles derived from the clustering results can be utilized to guide strategies of marketing according to the groups (Krishnapuram, Joshi, & Nasraoui, 2001). The association rules discover associations and correlations among items where the presence of an item or group of them in a transaction implies the presences of other items (Agrawal et al., 1993). Association rules are used to identify the relations among visits of users with a certain navigational pattern to the web site. In Alam, Dobbie, and Riddle (2008), swarm intelligence based PSO-clustering algorithm for the clustering of Web user sessions is proposed, in which author claimed that PSO clustering approach performs better than the benchmark k-means clustering algorithm for clustering Web usage sessions.

In Rambharose and Nikov (2010), various Computational Intelligence (CI) models such as Fuzzy Systems, Genetic Algorithms, Neural Networks, Artificial Immune Systems, Particle Swarm Optimization, Ant Colony Optimization, Bee Colony Optimization and Wasp Colony Optimization for personalization of interactive web systems are reviewed and compared regarding their inception, functions, performance and application to personalization of interactive web systems. But, PSO was credited with good performance as compared to the other methods. In Premalatha and Natarajan (2010), the modification strategies are proposed in PSO using GA. Experiment results are examined with benchmark functions and results show that the proposed hybrid models outperform the standard PSO.