Chapter 6

Identification of Optimal Web Page Set based on Web Usage using Biclustering Optimization Techniques

R. Rathipriya
Periyar University, India

ABSTRACT

The primary objective of this chapter is to propose Biclustering Optimization Techniques (BOT) to identify the optimal web pages from web usage data. Bio-inspired optimization techniques like Firefly algorithm and its variant are used as optimization tool to generate optimal usage profile from the given web usage dataset. Finally, empirical study is conducted on the benchmark clickstream datasets like MSNBC, MSWEB and CTI and their results are analyzed to know the performance of the proposed biclustering optimization techniques with respect to optimization techniques available in the literature.

INTRODUCTION

The characterization of web users and their usage behaviors are the important issues in the design and maintenance of any web based recommendation systems and for targeted advertisement in the web pages. Understanding the interests and behaviors of web users serves as the base for many web usage data mining applications such as personalized search, recommendation, personalization, business decision, advertisement, targeting marketing, electronic commerce (E-commerce) and customer relationship management.

For past one decade, a lot of research is being done in the area of web usage mining based on the goals of the analyst and applications, various algorithms can be applied for cluster analysis. Very frequently, the clustering methods are used for grouping related web users of a web site, it typically partitions users according to their similar browsing behavior under all pages of a website. On the other hand, related web users do not necessarily express similar interest to every page in the website, or in other words, there are users who can be similar only for a subset of pages. On the contrary, it is also possible to discriminate
Identification of Optimal Web Page Set based on Web Usage

groups of pages by using different groups of users. From this point of view, clustering cannot only be addressed horizontally (users) or vertically (pages), but also in the two dimensions simultaneously. This approach is called biclustering, which is the one of the most popular data mining techniques that extracts browsing or usage patterns from a flood of web data by simultaneously clustering of rows(users) and columns (pages) of a web usage data. Owing to its computational complexity, it is constantly a matter of interest for machine learning researchers.

In this chapter, Biclustering Optimization Techniques (BOT) based biclustering approach is used to identify the optimal web page set of a web site from highly correlated user group over a subset of web pages (called bicluster).

RELATED WORK

More sophisticated systems and techniques for discovery and analysis of user browsing patterns are available in the literature.

All these works were concentrated in application of clustering and its variants for web usage mining in different perspective to identify the usage based clusters. These clusters may be user clusters or page clusters.

The major limitations of listed clustering above are:

1. The clustering algorithms described above are mainly manipulated on one dimension (i.e. either users or pages) of the web usage data only, rather than taking into account the correlation between web users and pages. Moreover, it is based on the assumption that all related users behave similarly across all set of pages of a website and vice versa. However, in most cases, the web clusters do often exist in the form of co-occurrence of pages and users, i.e. users from the same group are particularly interested in one subset of web pages.
2. The characteristic of the web usage data is that it is not necessary to include all the users or pages in the clusters. In fact, it may be more useful to identify a subset of pages where a subset of the users acts in a coherent manner.

The application of traditional clustering techniques for pattern (i.e. browsing pattern or usage pattern) discovery and data analysis poses significant problems. Consider a case of hundreds of web pages being viewed by the users. Typically, the pages browsed by one set of users will have a high level of mutual exclusivity when compared to pages browsed by other set of users. Discovering such similar local grouping of attributes (or pages) may be the key to uncover many interesting and useful browsing patterns that are not otherwise apparent. In-order to identify such browsing or usage patterns in terms of usage profile is possible by using biclustering of web usage data.

In Symeonidis. P et al. (2006), use biclustering approach to provide recommendation to the users based on the user and item similarity of neighborhood biclusters. Most of our pervious works were focused on the application of biclustering techniques for web usage data.

R. Rathipriya et al. (2011) introduced Binary Particle Swarm Optimization based biclustering to identify the global optimal web usage profile from the given web usage data. Similarly,