Available online at www.sciencedirect.com
ScienceDirect
Data & Knowledge Engineering 65 (2008) 512–543
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Sequence-based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-means algorithm

Sungjune Park [a,*], Nallan C. Suresh [b], Bong-Keun Jeong [a]

[a] *The University of North Carolina at Charlotte, Business Information Systems and Operations Management, The Belk College of Business, 9201 University City Blvd, Charlotte, NC 28223, United States*
[b] *Department of Operations Management and Strategy, School of Management, State University of New York, Buffalo, NY 14260, United States*

Received 5 June 2007; received in revised form 18 December 2007; accepted 28 January 2008
Available online 6 February 2008

Abstract

We develop a general sequence-based clustering method by proposing new sequence representation schemes in association with Markov models. The resulting sequence representations allow for calculation of vector-based distances (dissimilarities) between Web user sessions and thus can be used as inputs of various clustering algorithms. We develop an evaluation framework in which the performances of the algorithms are compared in terms of whether the clusters (groups of Web users who follow the same Markov process) are correctly identified using a replicated clustering approach. A series of experiments is conducted to investigate whether clustering performance is affected by different sequence representations and different distance measures as well as by other factors such as number of actual Web user clusters, number of Web pages, similarity between clusters, minimum session length, number of user sessions, and number of clusters to form. A new, fuzzy ART-enhanced K-means algorithm is also developed and its superior performance is demonstrated.
Published by Elsevier B.V.

Keywords: Web usage mining; Clustering methods; Simulation; Artificial intelligence; Markov chain

1. Introduction

With the abundance of information available on the World Wide Web (WWW), the issue of how to extract useful knowledge from the Web has gained significant attention among researchers in data mining and knowledge discovery areas. According to a survey conducted by Computer Industry Almanac, number of online users worldwide exceeded one billion in 2005 – up from only 45 million in 1995, and the number is expected to increase to two billion by 2011 [1]. Jupiter Research projects that online retail sales will continue to rise to a level of around $144 billion in 2010, when the Internet will influence nearly half of total retail sales [2].

* Corresponding author. Tel.: +1 704 687 7628; fax: +1 704 687 6330.
E-mail address: supark@uncc.edu (S. Park).

0169-023X/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.datak.2008.01.002

From a business point of view, increasing demand for Web services such as e-commerce, e-banking, and e-CRM has changed the way the Web is being used. In today's competitive business environment, Web services have become an absolute necessity for organizations, not only to distribute and collect information but also to discover useful patterns from collected data. Knowledge gained from collected data can then be used to develop business strategies. Many organizations have begun implementing value-added services on the Web to gain competitive advantage and create loyal customers. In order to make the Web more user-friendly for individuals and create long-term relationships with them, companies now realize that providing personalized products and services is crucial. This type of personalization can be achieved through a variety of methods focused on discovering each individual's needs [3]. Web mining enables extraction of such knowledge for personalization and improved Web services.

Web mining refers to the effort of Knowledge Discovery in Data (KDD) from the web. It can be defined as the process of applying data mining techniques to extract useful knowledge from the huge amount of information available from the web. It is often categorized into three major areas [4,5]:

- *Web content mining*: mining of text, image, audio, video, metadata, and hypertexts in order to extract useful concepts and rules and summarize the content on the web.
- *Web structure mining*: mining of underlying link structures of the Web in order to categorize Web pages, measure similarities and reveal relationships between different Web sites.
- *Web usage mining*: mining of the data generated by the Web users' interactions with the web, including Web server access logs, user queries, and mouse-clicks in order to extract patterns and trends in Web users' behaviors.

This paper deals with Web usage mining, for which many data mining techniques such as statistical analysis, clustering, classification, association rules, sequential pattern discovery, and dependency modeling have been applied to Web server logs. Among these, clustering, which has been one of the most frequently used techniques, forms the focus of this study.

One important use of clustering in Web usage mining is aimed at finding groups which share common interests and behaviors by analyzing the data collected in Web servers. Recent Web usage mining studies conducted for this purpose attempt to incorporate sequence of page visits, although clustering based on frequency or Boolean representation of Web usage is still common in many studies as we will discuss later in the literature review section. However, clustering Web users based on sequences of Web navigation remains a relatively undeveloped area in that there are few guidelines on how to evaluate the performance of sequence-based clustering methods [6]. This study contributes to the topic of *sequence-based clustering for Web usage mining* in two aspects: (1) by developing a general methodology that allows for the use of any distance-based clustering methods in identifying Web user clusters based on sequence of page visits, and (2) by developing a systematic evaluation framework, based on replicated clustering, that can be used to compare the performance of various clustering methods. This study also develops an enhanced K-means clustering algorithm for sequence-based clustering by combining the strength of fuzzy ART neural network to address certain limitations of the K-means algorithm.

One research question addressed here is as follows: Can sequence-based clustering methods perform better than frequency-based? If so, what factors affect the performance of sequence-based clustering? A second research question involves the use of Markov model in Web usage mining, the efficacy of which has been stated repeatedly in past studies. For this reason, clustering based on a Markov model is considered in some Web usage mining studies [7,8]. However, whether the Web user clusters identified by cluster analysis are indeed Web user clusters that share common Markov process is an open question.

In order to address these research questions, we first develop a general sequence-based clustering method by proposing new sequence representation schemes in association with Markov models. The resulting sequence representations allow for calculation of vector-based distances (dissimilarities) between Web user sessions and are used as inputs of various clustering algorithms. Two clustering algorithms are selected to conduct various experiments: (1) K-means algorithm and (2) a hybrid algorithm referred to here as "FAK" that first utilizes fuzzy adaptive resonance theory (fuzzy ART) to determine initial cluster centroids for a K-means algorithm. We then investigate how well the clusters, or groups of Web users who follow the same Markov process, are