International Journal of Innovative
Computing, Information and Control
Volume **7**, Number **5(B)**, May **2011**

 ISSN 1349-4198
pp. **2809–2819**

A NOVEL TWO-STAGE SCHEME BUILT-UPON CLUSTERING FOR SEQUENTIAL PATTERN MINING

Qianwen Yang, Jisong Kou, Fuzan Chen* and Minqiang Li

School of Management
Tianjin University
No. 92, Weijin Road, Nankai District, Tianjin, P. R. China
qwjackie@gmail.com; { jskou; mqli }@tju.edu.cn
*Corresponding author: fzchen@tju.edu.cn

Received January 2010; revised April 2010

Abstract. *A novel two-stage scheme for mining sequential patterns is proposed. It clusters the sequences into several groups in the first phase. A n-tuple data structure is designed to represent sequences and reduce the dimensionality. A more understandable and accurate method for measuring similarities SMCS among the above sequences is presented, which captures more specific information about sequences so that the similarity is computed more accurately. In the second phase, stratograms are employed to visualize the patterns. Stratogram provides more information, such as frequency of the sequences, which helps discover and extract significant patterns. The efficacy of the proposed method is verified on one real life dataset and one synthetic dataset, and experimental results show that the proposed method has advantages in accuracy, expressivity and comprehensibility in comparison with conventional approaches.*
Keywords: Sequential pattern mining, Clustering, Similarity measure, Stratogram

1. **Introduction.** Mining sequential patterns from a large database is an important research topic in data mining [1]. It has been applied in applications such as customer shopping sequences, web clickstreams, biological sequences, disease treatments and so on. Lots of studies have contributed to the mining of sequential patterns. On one hand, many efficient algorithms were designed to find sequential patterns in a sequence database [2-10]. Another research topic is to apply the mining techniques to special data types, such as the web [11] and biological data [12]. This paper aims at designing an efficient strategy for sequential patterns mining, especially in the web data.

A sequence database consists of sequences of ordered elements or events, recorded with or without a concrete notion of time. Sequential pattern mining is to discover subsequences or frequently occurring ordered events in a sequence database. Some of the sequential pattern mining algorithms are the Apriori-like [2,3], i.e., based on the Apriori property proposed in association mining. The representative Apriori-like algorithm was GSP (Generalized Sequential Patterns) proposed by Srikant and Agrawal [1]. It adopted the candidate generation and test procedure. The algorithm made multiple passes over the total data set.

In view of the huge time and space complexities of Apriori-like sequential pattern mining method, more effective and faster algorithms for mining sequential patterns were proposed [4-8]. SPADE was the Apriori-based sequential pattern mining algorithm for vertical data format. It transferred the sequence database into vertical data format that is indexed with the sequence identifier and event identifier of the items. The FreeSpan algorithm [9] and PrefixSpan [10] algorithm mined sequential patterns by pattern growth and did not require candidate generation.