International Journal of Innovative
Computing, Information and Control
Volume **6**, Number **5**, May **2010**

ICIC International ©2010 ISSN 1349-4198
pp. 2171–2179

HCLUWIN: AN ALGORITHM FOR CLUSTERING HETEROGENEOUS DATA STREAMS OVER SLIDING WINDOWS

Jiadong Ren[1,2], Changzhen Hu[2] and Ruiqing Ma[1]

[1]College of Information Science and Engineering
Yanshan University
Qinhuangdao, P. R. China
mrq153101@163.com

[2]School of Computer Science and Technology
Beijing Institute of Technology
Beijing, P. R. China
jdren@ysu.edu.cn

Received November 2008; revised May 2009

Abstract. *Many applications in web usage mining, such as business intelligence and usage characterization, require effective and efficient techniques to discover the users with similar usage patterns and the web pages with correlate contents in the physical world. Clustering click streams can help to achieve the goal. Despite the high processing rate, the existing methods for clustering click streams over sliding widows suffer from the missing of categorical attributes in click stream data. In this paper, we present HCluWin, an approach for clustering heterogeneous data streams which contain both continuous attributes and categorical attributes over sliding windows. A Heterogeneous Temporal Cluster Feature (HTCF) is introduced to monitor the distribution statistics of heterogeneous data points. Based on this structure, Exponential Histogram of Heterogeneous Cluster Feature (EHHCF) is presented. Simultaneously, a new similarity measure between two heterogeneous objects is proposed. Experimental results show that the clustering quality of HCluWin is higher than CluWin and the stream processing rate of HCluWin is higher than HCluStream.*

Keywords: Data stream, Clustering, Heterogeneous attribute, Sliding windows

1. **Introduction.** Web usage mining is the process of applying data mining techniques to the discovery of usage patterns from web data [1]. In the web environment, a group of consecutive page views sent by users constitutes a click-stream. In order to get useful patterns for the business intelligence and the usage characterization, clustering analysis over data streams is applied in the web usage mining.

Clustering is an important and traditional method in data mining. Clustering can be used to address many problems. For example, clustering can be used to improve classification accuracy [2]; a proximity clustering approach is used in the DNA computing [3]. Recently, clustering has been applied in data streams mining. Most algorithms on data streams clustering can only manipulate continuous attributes. For example, in CluStream [4], the final clusters are generated by using modified k-means algorithm which is only applied in the case where the mean of data points makes sense. HPStream [5] is proposed to cluster high dimensional data streams. HPStream introduces a fading cluster structure and the projection based clustering methodology. This fading cluster structure describes the cluster feature of data points of continuous attributes. Beringer considers the parallel streams clustering problem and has developed an efficient online version of the classical k-means clustering algorithm in [6]. In addition, a few algorithms can only