*International Journal of Innovative Computing, Information and Control*
Volume **7**, Number **1**, January **2011**

ICIC International ©2011 ISSN 1349-4198
pp. 487–500

# HIERARCHICAL INFORMATION-THEORETIC CO-CLUSTERING FOR HIGH DIMENSIONAL DATA

Yuanyuan Wang[1], Yunming Ye[1,*], Xutao Li[1], Michael K. Ng[2]
and Joshua Huang[3]

[1]Department of Computer Science
Shenzhen Graduate School of Harbin Institute of Technology
HIT Campus of Shenzhen University Town, Shenzhen 518005, P. R. China
pinetreeyuan@gmail.com
*Corresponding author: yeyunming@hit.edu.cn

[2]Department of Mathematics
Hong Kong Baptist University
Hong Kong, P. R. China
mng@math.hkbu.edu.hk

[3]E-Business Technology Institute
The University of Hong Kong
Hong Kong, P. R. China
jhuang@eti.hku.hk

Received September 2009; revised March 2010

Abstract. *Hierarchical clustering is an important technique for hierarchical data exploration applications. However, most existing hierarchial methods are based on traditional one-side clustering, which is not effective for handling high dimensional data. In this paper, we develop a partitional hierarchical co-clustering framework and propose a Hierarchical Information-Theoretical Co-Clustering (HITCC) algorithm. The algorithm conducts a series of binary partitions of objects on a data set via the Information-Theoretical Co-Clustering (ITCC) procedure, and generates a hierarchical management of object clusters. Due to simultaneously clustering of features and objects in the process of building a cluster tree, the HITCC algorithm can identify subspace clusters at different-level abstractions and acquire good clustering hierarchies. Compared with the flat ITCC algorithm and six state-of-the-art hierarchical clustering algorithms on various data sets, the new algorithm demonstrated much better performance.*

**Keywords:** Hierarchical clustering, Co-clustering, Text clustering

## 1. Introduction.

Numerous high dimensional data have emerged recently and the data are still increasing at a high speed, for example, text data on the Internet, microarray data in biological research and transaction data in business. It is essential to arrange these data in a comprehensible manner for management and analysis. Hierarchical management is one of the most widely used manners, such as arranging documents in the form of a tree for user's navigation and search, and grouping microarray data in a hierarchical structure for visualization and analysis.

Hierarchical clustering is an important approach to hierarchical data exploration applications. Traditionally, hierarchial clustering is mainly based on agglomerative algorithms [1, 2, 3, 4]. However, these algorithms have high computational costs and are not effective for handling high dimensional data. Recently, various studies have shown that hierarchical clustering can be obtained by a sequence of partitions that are generated by partitional algorithms. These types of algorithms have low computational costs. Zhao