A Graph-Based Approach for Mining Closed Frequent Patterns

以图形理論為基礎挖掘封閉常用資料集的方法

Lee-Wen Huang

Dept. of Computer Science and Engineering
National Sun Yat-Sen University
&
Department of Management Information Systems
Far East University

huanglw@gmail.com

Ye-In Chang

Dept. of Computer Science and Engineering
National Sun Yat-Sen University

changyi@cse.nsysu.edu.tw

Abstract

Data Mining means a process of nontrivial extraction of implicit, previously and potentially useful information from data in databases. Mining closed frequent patterns is a further work of mining association rules, which aims to find the set of necessary subsets of frequent patterns that could be representative of all frequent patterns. In this paper, we design a graph-based approach, considering the character of data, to mine the closed frequent patterns efficiently. Two features of market basket analysis are considered – the number of items is large; the number of associated items for each item is small. Combining the cut-point method and the clique concept, the new algorithm can find the closed frequent patterns efficiently. The simulation results show that the new algorithm outperforms the FP-Growth algorithm in the execution time and the space of storage.

Keywords: Association Rule, Closed frequent patterns, Data Mining, Graphic-based mining

1. Introduction

In recent years, with the computers and information industries growing more and more rapidly, varies data around us become more complexity and huge. Commercial behavior, scientific statistics, natural phenomena and DNA projects are some examples to produce lots of data every day. In the past, we may put these data in drawers or databases to extract some information efficiently. However, as the amount of data grows, it becomes very difficult to determine the useful data. Generally speaking, there are four kinds of problems that we have met [2,3,7,9,13]. First, there are too much information for us to digest. Second, it is difficult to recognize the reality of the information. Third, it is hard to guarantee the security of these information. Finally, we cannot deal with different forms of information easily because of much data. In order to solve these four problems, people start to think how to find useful knowledge without overwhelming by information flooding. Although recent database management systems can provide quick insertion, deletion, query and statistic, they cannot detect the relations and rules between data. So, there comes a research named data mining which is used to help making decisions and information retrieval. Data mining is one kind of analysis techniques which are data driven. This is different from Online Analytical Processing (OLAP) as well as the ad-hoc query and reporting approach, both of which are hypothesis driven [5]. A hypothesis is formulated by the query types. That is, if a pattern cannot be formulated, we cannot use the hypothesis to find this pattern. However, since data mining is data driven, it is not important whether we know the rules or query types, the