高效率常見超集合探勘演算法之研究

摘要

過去對於探勘常見項目集的研究僅限於找出資料庫中交易紀錄的子集合，在這篇論文中，我們提出一個新的探勘主題：常見超集合探勘。常見超集合意指它包含資料庫中各筆紀錄的筆數多於最小門檻值，而原本用來探勘常見子集合的演算法並無法直接套用，因此我們以補集合的角度，提出了三個快速的演算法來解決這個新的問題。首先為 Apriori-C：此為使用先廣後深搜尋的演算法，並且以掃描資料庫的方式來決定具有相同長度之候選超集合的支持度，第二個方法是 Eclat-C：此為採用先深後廣搜尋的演算法，並且搭配交集法來計算候選超集合的支持度，最後是 DCT：此方法可利用過去常見子集合探勘的演算法來進行探勘，如此可以省下開發新系統的成本。

常見超集合的探勘可以應用在電子化的遠距學習系統，生物資訊及工作排程的問題上。尤其在線上學習系統，我們可以利用常見超集合來代表一群學生的學習行為，並且藉以預測學生的學習成就，使得老師可以及時發現學生的學習迷失等行為；此外，透過常見超集合的探勘，我們也可以為學生推薦個人化的課程，以達到因材施教的教學目標。

在實驗的部份，我們比較了各演算法的效率，並且分別改變實驗資料庫的下列四種變因：交易資料的筆數、每筆交易資料的平均長度、資料庫中項目的總數和最小門檻值。在最後的分析當中，可以清楚地看出我們提出的各種方法皆十分有效率並且具有可延伸性。

Efficient Algorithms for the Discovery of Frequent Superset

Abstract

The algorithms for the discovery of frequent itemset have been investigated widely. These frequent itemsets are subsets of database. In this thesis, we propose a novel mining task: mining frequent superset from the database of itemsets that is useful in bioinformatics, E-learning systems, jobshop scheduling, and so on. A frequent superset means that the number of transactions contained in it is not less than minimum support threshold. Intuitively, according to the Apriori algorithm, the level-wise discovering starts from 1-itemset, 2-itemset, and so forth. However, such steps cannot utilize the property of Apriori to reduce search space, because if an itemset is not frequent, its superset maybe frequent. In order to solve this problem, we propose three methods. The first is the Apriori-based approach, called Apriori-C. The second is the Eclat-based approach, called Eclat-C, which is a depth-first approach. The last is the proposed data complement technique (DCT) that we utilize original frequent itemset mining approach to discover frequent superset.

The experimental studies compare the performance of the proposed three methods by considering the effect of the number of transactions, the average length of transactions, the number of different items, and minimum support. The analysis shows that the proposed algorithms are time efficient and scalable.