Data Mining Homework 2 Proposal

0256026 陳上奎

Problem definition

Association rules是data mining的重要課題，用於從大量數據之中挖掘出有價值的數據項之間的相關關係，藉由association rules可以減少大量無用而雜亂的數據，將其轉化成易於觀察理解的資料關係。

和上次作業使用相同的input source，即movie dataset，並可以自己輸入support S, confidence C, top-k K等參數，而輸出則是top-k association rules，而輸出結果的排序則是依support, confidence, lexicography來做排序。

Related works

找尋association rules可以依方法分為兩類

1. 產生candidate
   1. Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994.
   2. Fournier-Viger, Philippe, Cheng-Wei Wu, and Vincent S. Tseng. "Mining top-k association rules." *Advances in Artificial Intelligence*. Springer Berlin Heidelberg, 2012. 61-73.

基本上這部分的works都是先找出所有或是對現有的candidate去作拓展，這些方法在尋找candidate後必須對整個dataset去做搜尋看其是否滿足其minsupport，故會較花時間。

1. 不產生candidate
   1. Han, Jiawei, Jian Pei, and Yiwen Yin. "Mining frequent patterns without candidate generation." *ACM SIGMOD Record*. Vol. 29. No. 2. ACM, 2000.
   2. Han, Jiawei, et al. "Mining top-k frequent closed patterns without minimum support." *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002.

不產生candidate的方式有許多種，在這邊主要以FP-Growth為參考方式，並去尋找在top-k的情形下作pruning的可能性。

Proposed methods

雖然是要找top-k association rules，可是要找到top-k前至多須找出所有的frequent pattern，在對這些找到的frequent pattern所找到的association rules作top-k。

從這個流程可以知道要在短時間內找到所有須被考慮的frequent pattern，而利用Apriori演算法會對整個dataset掃過多次，因為每當你要去驗證其是否有超過minsupport時，必定要去對所有dataset作搜尋來做驗證。因此會希望以實作FP-Growth來減少產生candidate並去做test的時間。

由上面可以歸納出流程：

1. Parsing dataset movies:

因為上次作業到最後才發現parsing的結果可能會missing一部分的data，因此會對讀檔和parsing檔案的部分作重新修正。

1. 使用FP-Growth
   1. 建出FP-tree
   2. 利用FP-tree找出frequent patterns
   3. 並利用frequent patterns找出top-k association rules

會使用FP-Growth除了可以減少搜尋dataset的時間之外，因為FP-Growth的特性，即使原先的dataset在大也可以另用分割的方式分成兩個FP-tree在做合併，因為FP-tree的大小會比dataset小很多。