Top-Down Approach to Frequent Pattern Mining

Yan Xie
University of Illinois at Chicago
yxie8@uic.edu

Philip S. Yu*
University of Illinois at Chicago
psyu@uic.edu

Extended abstract:

Frequent pattern mining is considered to be one of the fundamental techniques in data mining, where its applications extend to many mining areas, including but not limited to association rule mining, classification and clustering. Over the past decade, frequent pattern mining has been extensively studied by many data mining researchers, leading to a large number of publications and algorithms for mining frequent patterns and/or other extended objectives.

Frequent pattern mining is essentially a challenging problem due to the huge number of item combinations that can exist. Based on the property that all subpatterns of a frequent pattern must be frequent as well, the Apriori principle significantly reduces the size of the candidate pattern set, and leads to great improvement of mining performance. Various algorithms building on this very same principle have been proposed to further improve the computational efficiency. For example, one representative method is FP-growth which deploys a divide-and-conquer approach to avoid candidate generations, and another is Eclat where mining is performed with data presented in vertical data format.

So far, most of the existing pattern mining algorithms mine the complete set of frequent patterns in a transaction database. However, when the length of the frequent patterns becomes long, none of these methods can produce results within a reasonable amount of time, since all possible combinations of frequent items are exponential in size. The introduction of closed or maximal frequent patterns can alleviate the result size issue to a certain extent, but still there are many cases where the frequent pattern set is just too large, even if we only target closed or maximal frequent patterns.

The above dilemma is intrinsic to the frequent pattern mining problem, if one wants to get the complete frequent pattern set. However, as we observe in many applications, mining tasks in practice usually attach greater importance to patterns with large sizes. For example, in bioinformatics, long sequences are usually of much higher significance compared to those short ones. Thus, if the complete frequent pattern set is prohibitively large and only a small number of long ones are of practical interests, it becomes

* corresponding author. Tel: (312) 996-0498

problematic to have the mining algorithm grow patterns from short to long, as it may get stuck at the intermediate stages.

Motivated by the above analysis, we design a new approach that can directly mine those interesting long patterns without going through the agonizing process of growing an explosive intermediate set. In particular, we are interested in solving the following problem: For a transaction database, can we efficiently get its top-k maximal frequent patterns? Compared to the state-of-the-art bottom-up approaches, we explore the opposite direction and propose a top-down mining strategy here. The main idea is as follows. We find that items and their co-occurrence relationships can be summarized by a so-called pattern graph, where frequent patterns will show up as cliques; this makes cliques promising frequent pattern candidates. Now, the problem of finding long maximal frequent patterns can be transformed into the problem of detecting large maximal cliques: Starting from the pattern graph derived from the transaction database, we can find a list of maximal cliques with very large sizes, where each maximal clique is one that cannot be further enlarged by adding more vertices onto it. The essence of this step is to bypass the bottom-up pattern growth procedure since we have directly jumped to the top of the search space. Although it is not guaranteed that each maximal clique c will represent a true maximum frequent pattern, due to the filtering done by the pattern graph, it is likely that the set of items in c is only slightly different from some true maximum frequent pattern p. To this end, a separate refinement step is needed to effectively transform c into p. Finally, the set of large cliques will give rise to a set of large maximum frequent patterns, and we can follow this strategy to find the top-k maximal frequent patterns.

Keywords: Frequent Pattern, Pattern Graph, Data Mining