Public Abstract

First Name:Michael

Middle Name:

Last Name:Phinney

Adviser's First Name:Chi-Ren

Adviser's Last Name:Shyu

Co-Adviser's First Name:

Co-Adviser's Last Name:

Graduation Term:SP 2017

Department:Computer Science

Degree:PhD

Title:Distributed Frequent Hierarchical Pattern Mining for Robust and Efficient Large-Scale Association Discovery

Frequent pattern mining is a classic data mining technique, generally applicable to a wide range of application domains, and a mature area of research. The fundamental challenge arises from the combinatorial nature of frequent itemsets, scaling exponentially with respect to the number of unique items. Apriori-based and FPTree-based algorithms have dominated the space thus far. Initial phases of this research relied on the Apriori algorithm and utilized a distributed computing environment; we proposed the Cartesian Scheduler to manage Apriori's candidate generation process. To address the limitation of bottom-up frequent pattern mining algorithms such as Apriori and FPGrowth, we propose the Frequent Hierarchical Pattern Tree (FHPTree): a tree structure and new frequent pattern mining paradigm. The classic problem is redefined as frequent hierarchical pattern mining where the goal is to detect frequent maximal pattern covers. Under the proposed paradigm, compressed representations of maximal patterns are mined using a top-down FHPTree traversal, FHPGrowth, which detects large patterns before their subsets, thus yielding significant reductions in computation time. The FHPTree memory footprint is small; the number of nodes in the structure scales linearly with respect to the number of unique items. Additionally, the FHPTree serves as a persistent, dynamic data structure to index frequent patterns and enable efficient search. When the search space is exponential, efficient targeted mining capabilities are paramount; this is one of the key contributions of the FHPTree. In this dissertation, will demonstrate the performance of FHPGrowth, achieving a 300x speed up over state-of-the-art maximal pattern mining algorithms and approximately a 2400x speedup when utilizing FHPGrowth in a distributed computing environment. In addition, we allude to future research opportunities, and suggest various modifications to further optimize the FHPTree and FHPGrowth. Moreover, the methods we offer will have an impact on other data mining research areas including contrast set mining as well as spatial and temporal mining.