Pattern Decomposition Algorithm for Data Mining of Frequent Patterns

Tech ID: 20304 / UC Case 2003-359-0

SUMMARY

Researchers in the UCLA Department of Computer Science have developed an algorithm that speeds up the mining of frequent patterns (also known as frequent itemsets, FI) in large datasets. Such an efficient algorithm is crucial to many tasks in data mining. The method uses a pattern decomposition algorithm to significantly reduce the size of the dataset on each pass, making it more efficient to mine all frequent patterns in a large dataset.The invention can be implemented as a data analyzing tool that can be bonded with database products (Sql2000, DB2, Oracle, etc.), or data processing software (SAS, Matlab, etc). It can be used for analyzing large datasets like genome data, medical data, protein analysis data, human genome data, and full information security data.

BACKGROUND

A fundamental problem in data mining is finding frequent patterns in large datasets. This problem is even worse in datasets containing highly frequent, yet often meaningful patters (e.g., free text). Finding frequent patterns enables essential data mining tasks, such as discovering association relationships, determining correlations between data, and finding sequential patterns.Several different algorithms have been proposed to find all frequent patterns in a dataset. The Apriori algorithm is widely cited in the literature. It generates candidate sets to limit pattern counting to only patterns that can meet the minimum support requirement. However, this algorithm exhibits poor performance when frequent pattern sizes are large, due to combinatory explosion. The Pincher-Search and Max-Miner methods attempt to avoid this problem by outputting only maximal frequent itemsets (MFI). These methods have limited use in association rule mining because a complete set of rules cannot be extracted without support information of the subsets of those maximal frequent sets.FP-tree-based mining claims the best performance in the recent literature. It first builds a compressed data representation from a dataset and then all mining tasks are performed on the FP-tree, rather than on the dataset. However, FP-tree-based mining uses a complicated data structure and performance gains are sensitive to the support threshold.

INNOVATION

Dr. Chus innovation uses pattern decomposition (PD) to mine frequent patterns. PD transforms the dataset, similar to the FP-tree algorithm. However, unlike the FP-tree algorithm, PD does not pre-calculate the new data representation. Instead, the dataset is transformed only when the changes may shorten subsequent passes (e.g., decrease the number of data items to count). PD uses a bottom-up search to find frequent sets, and shrinks the dataset when new infrequent itemsets are discovered.This method provides three significant improvements: 1. The dataset can be significantly reduced in each pass by decomposing transactions into short itemsets and combining regular patterns together. 2. The algorithm does not need to generate candidate sets since the reduced dataset does not contain any infrequent patterns found before. 3. Using a reduced dataset greatly saves time for counting pattern occurrence.

STATE OF DEVELOPMENT

Empirical evaluation shows that Dr. Chus algorithm outperforms the Apriori algorithm by one order of magnitude and is faster than the FP-tree method.

CONTACT

UCLA Technology Development Group
ncd@tdg.ucla.edu
tel: 310.794.0558.

INVENTORS

- Chu, Wesley W.

OTHER INFORMATION

KEYWORDS

communications

CATEGORIZED AS

- **Computer**
  - Software

RELATED CASES

2003-359-0