摘 要

論文名稱：超大型數據集的高效用模式探勘近似隨機演算法

頁數：

校所別：國立臺北科技大學 電子工程系 碩士班

畢業時間：一百一十一學年度 第二學期

學位：碩士

研究生：呂彥旻

指導教授：黃柏鈞 博士

關鍵詞： 高效用模式探勘、超大型數據集、近似演算法

高效用模式挖掘是資料探勘中常見的一種方法，其主要目的在項目集中尋找數據庫中最高效用的關聯規則。該方法考慮項目的數量和不同項目的重要性，並相對於傳統頻繁模式挖掘方法，在處理現實世界的大規模資料庫時提供更高效、更準確的結果。相較於過去的演算法，它們未考慮到整個數據集可能存放於輔助儲存設備，而非主記憶體。這一點非常重要，通常輔助儲存設備會性能會遠低於主記憶體。為了減少輔助儲存設備所造成的性能瓶頸，因此我們提出了一種高效用模式探勘近似隨機演算法框架（PAHUPMA），它只需要掃描整個數據集的一小部分即可以發現高效用模式。我們對高效用模式挖掘方法與其他演算法進行了一系列的分析和性能評估，結果顯示我們提出的方法更能夠符合現實環境。

ABSTRACT

Title: A Probabilistic Approximate High-utility Pattern Mining Algorithm for Ultra Large Scale Datasets

Pages:

School: National Taipei University of Technology

Department: Department of Electronic Engineering

Date: August, 2023

Degree: Master

Researcher: Yan-Min Lu

Advisor: Po-Chun Huang, Ph.D.

Keywords: High utility pattern mining, Ultra large scale datasets, Approximate algorithm

Efficient pattern mining is a widely used technique in data exploration, aimed at discovering the most valuable association rules within a database for a given set of items. This approach considers both the number of items and the significance of different elements. It delivers more efficient and precise outcomes when handling large databases containing real-world data, surpassing traditional frequent pattern mining methods. Unlike previous algorithms, they fail to consider the scenario where the complete dataset is stored in secondary storage instead of main memory. This aspect is highly significant since the performance of secondary storage is typically significantly lower than that of primary memory. To mitigate the performance bottleneck arising from auxiliary storage devices, we propose a framework called the probabilistic approximate high-utility pattern mining algorithm (PAHUPMA), which requires scanning only a small fraction of the complete dataset to uncover efficient usage patterns. We have conducted a series of comprehensive analyses and performance evaluations comparing the efficient pattern mining method with other algorithms. The results demonstrate that our proposed method is better suited for real-world environments.