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

在一個潛在龐大的資料集中，探索重要的模式通常具有相當大的價值。此類問題中最基本的形式是頻繁模式的挖掘，它的目的是發現在資料集中出現次數最多的項目集合。雖然已經有很多演算法來解決頻繁模式挖掘的問題，但是現實的環境通常更加複雜。例如：假設在數據集中，每一個項目都可能具有不同的價值，因此有了高效用模式挖掘，目的在於發現數據集中具有最高效用的項目集合。由於高效用模式不具有反單調的特性，因此高效用模式的子集不一定保證也是高效用模式。而缺乏反單調性特性，使得在龐大的數據集中挖掘高效用模式比頻繁模式更加困難。

在這項研究中，我們觀察到在現實生活中通常不需要進行確定性和精確性的高效用模式挖掘。相反，我們通常只需發現一些具有高概率是高效用模式的項目集，而不一定是真正具有最高效用的項目集。透過放寬這些要求，我們提出一種近似高效用模式挖掘演算法（PAHUPMA），他可以在給定的數據集中發現高效用模式，而無須掃描整個數據集。因此，高效用模式挖掘所產生的 IO 流量大幅減少，而顯著提升挖掘的性能。根據我們初步的實驗結果，PAHUPMA 所花費的只是現有演算法（如 HUI-Miner 和 HUPM）的一小部分。我們相信，PAHUPMA的優異性能能夠有效擴展高效用模式挖掘的應用範疇。

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, Probabilistic algorithms, Approximate algorithms

In a potentially huge dataset, it is often useful to discover the important patterns. The simplest form of such problems is the frequent pattern mining, which asks to discover the itemsets that appear for most times in the dataset. While there have been quite some efficient algorithms for solving the frequent pattern mining problem, more sophisticated forms of the problems are often of greater interest. For example, given the utility of each item in the dataset, the high-utility pattern mining problem asks to discover the itemsets with the highest total utility in the dataset. As high-utility patterns do not exhibit the anti-monotonicity property, with which a subset of a high-utility pattern is guaranteed to be a high-utility pattern as well. Without the anti-monotonicity, the mining of high-utility patterns in a huge dataset becomes much more difficult than that of frequent patterns.

In this work, we observed that the deterministic and exact high-utility pattern mining is often unnecessary in practice. Instead, it often suffices to discover some, but not all, patterns that have a high, but might not 100%, probability to have a high, but possibly not the highest total utility, in the given dataset. By relaxing such requirements, we present a probabilistic approximate high-utility pattern mining algorithm (PAHUPMA), which can discover the high-utility patterns in a given dataset without scanning through the whole dataset. As a result, the IO traffic generated for high-utility pattern mining is significantly reduced, which remarkably enhance the mining performance. According to our preliminary experimental results, the latency of PAHUPMA is only remnants of that of the existing deterministic algorithms, such as HUI-Miner and HUPM. We believe that the outstanding performance of PAHUPMA can effectively extend the spectrum of applications of high-utility pattern mining.