A Probabilistic Approximate High-utility Pattern Mining Algorithm for Ultra Abstract

超大型數據集的高效用模式探勘近似隨機演算法

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 HUMP. We believe that the outstanding performance of PAHUPMA can effectively extend the spectrum of applications of high-utility pattern mining.