6ZE-04

相関ルールにおける論理的パーティションに関する研究

青山 聡† 杜 小勇‡ 石井 直宏†
†名古屋工業大学 ‡中国人民大学

1 はじめに

近年、大規模データベースが構築されるなかでデータマイニングへの期待は一層強まっている。データマイニングアルゴリズムでは大規模データベースに対する計算量は膨大なため、高速で効率的なアルゴリズムの開発が重要な課題の一つである。

本研究では、データマイニングの一つの手法である相関ルール (Association Rule) 獲得のパーティションアルゴリズムに注目し、アルゴリズムに論理的パーティション (Logical Partition) という概念を導入することによって効率化を図った。

2 パーティションアルゴリズムの概要

相関ルールは $X \Rightarrow Y$ という形で表現され、この場合はアイテム集合 (商品の集まり) X が購入されるならばアイテム集合 Y も同時に購入されることを表す確率的なルールである。相関ルールではアイテム集合がデータベース中でどのぐらい出現するかという割合を支持度という値で表している。相関ルールを獲得する問題点はユーザの定義した閾値 (最小支持度) を超える支持度を持つすべてのアイテム集合 (これをラージアイテム集合と呼ぶ) を求めることに帰着できる。ラージアイテム集合を求めるアルゴリズムは数多く提案されている。その一つにパーティションアルゴリズムがある。

パーティションアルゴリズムではデータベースをパーティションと呼ばれる単位に区切り、二つのフェイズでそれぞれ一度づつデータベースを走査することでラージアイテム集合を求めることができる。最初のフェイズではそれぞれのパーティションからその大きさに合わせたローカルな最小支持度を超えるアイテム集合を求める。これらはラージアイテム集合になりうるので候補アイテム集合と呼び、すべてまとめたものをグローバル候補アイテム集合と呼ぶ。次のフェイズでは先に求められたグローバル候補アイテム集合に含まれるアイテム集合に対しデータベース全体での支持度を特定する。そして、最小支持度を越える支持度を持つラージアイテム集合を求める。

3 アルゴリズムの問題点と解決

パーティションアルゴリズムの問題点はデータスキューの存在するデータベースに対してパフォーマンスが低下することである。データスキューとは一時的に連続して同じようなデータが出現することで、これによりデータベースに偏りが生じる。この場合最初のフェイズで実際にはラージアイテム集合にならない間違った候補アイテム集合を数多く発見してしまう。そのため余分な計算量が発生しパフォーマンスが低下してしまう。

そこで本研究では論理的パーティションという概念を導入した。論理的パーティションとは、連続した幾つかのパーティションを仮想的に一つにまとめた大きなパーティションである。

論理的パーティションの大きさを 3 とした場合を図 1 に示す。

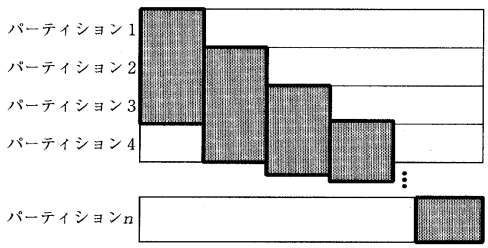

図 1: 論理的パーティション

従来のアルゴリズムではパーティションはそれぞれ独立して扱っていた。しかし、幾つかのパーティションをまとめた論理的パーティションを用いることで従来より大きな範囲で調べることができる。そのため局所的に出現するような間違った候補アイテム集合を減らすことができる。

論理的パーティションを用いた場合最初のフェイズでは一度求めた候補アイテム集合を続くパーティションでも調べるため従来の場合より余分に計算量がかかってしまう。しかし、論理的パーティション内でのそれぞれの候補アイテム集合の出現回数を保存しておくことで、次のフェイズで全体での割合を求める際に従来より計算量を減らすことができる。

4 実験

本研究ではまず論理的パーティションの最適な大きさを調べる実験を行った。パーティションの数を様々に変えてどれくらいの大きさが最適であるか調べた。本実験

The Logical Partition for Association Rule
Satoshi Aoyama† Xiaoyong Du‡ Naohiro Ishii†
†Nagoya Institute of Technology
‡The Renmin University of China

の評価値としてコストは次の二つの計算量を合わせたものを考えた。

- 最初のフェイズで他のパーティションで候補アイテム集合を調べた回数
- 次のフェイズでパーティション毎に候補アイテム集合の出現回数を調べた回数

前者は従来に比べて余分な計算量を表す。後者は従来に比べ軽減された計算量を表す。これらを合わせて考えて計算量が少なくすむ大きさが適したものであるとする。実験結果を図2、3に示す。この時のパーティション数は10とした。

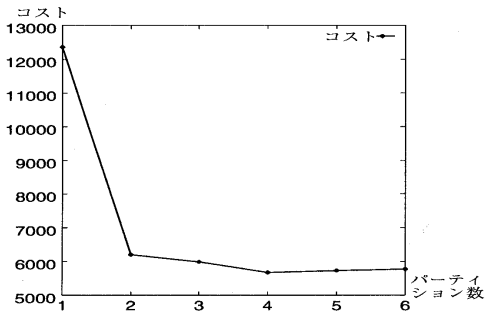

図 2: データスキューのない場合

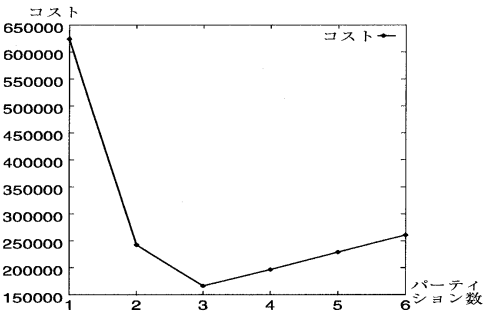

図 3: データスキューのある場合

図ではパーティション数1の場合が従来の場合を表している。

次に物理的なパーティションの数と論理的パーティションの大きさの関係について調べた。データベースを様々な大きさのパーティションにし、その際の最適な論理的パーティションの大きさを調べた。その結果を図4に示す。

5 まとめ

データスキューのない場合、論理的パーティションの大きさはどの大きさをとってもあまり変わらないことが分かった。データスキューの存在するデータの場合は論

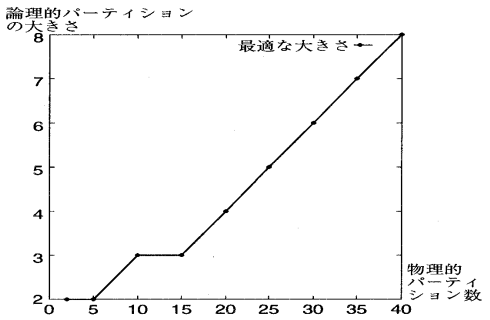

図 4: 物理的パーティションの数と論理的パーティションの関係

理的パーティションの大きさとしてある最適なパーティションの大きさが存在することが分かった。大きさをあまり大きくすると間違った候補アイテム集合に対する処理が増え余分な計算量が発生しパフォーマンスは低下した。データベース中にデータスキューが存在しているかはあらかじめ分からないため実際に論理的パーティションを用いる場合大きさとしてはパーティション2、3個分の大きさをとっておくことが妥当であると考える。

また、データベースを分割する物理的パーティション数と最適な論理パーティション数との間には線形な関係が成り立っていた。一つのデータに対してある一定の大きさの最適な論理的パーティションが存在することが分かった。

参考文献

[1] Ashok Savasere,Edward Omiecinski and Shamkant Navathe. An Efficient Algorithm for Mining Association Rules in Large Databases. In *Proceedings of the 21th VLDB Conference Zurich,Swizerland,1995*

[2] Rakesh Agrawal and Ramakrishnan Strikant. Fast Alogrithms for Mining Association Rules. In *Proceedings of the 20th VLDB Conference Santiago,Chile,1994*

[3] Rakesh Agrawal,Tomasz Imielinski and Arun Swami. Mining Association Rules between Set of Items in Large Database. In *Proceedings of the 1993 ACM SIGMOD Conference Washington DC,USA,May 1993*

[4] M.J.Zaki,S.Parthasarathy,M.Ogihara and W.Li. New Algorithms for Fast Discovery of Association Rules.