6F－7

分類階層を考慮した相関ルールの並列抽出処理における負荷分散手法の評価

新谷 隆彦　喜連川 優
東京大学 生産技術研究所

1 はじめに

データマイニングで得られる情報の代表的なものに相関ルールがあり、我々はその抽出処理性能向上を目的とした並列処理方式の研究を進めてきた [1]。また、データはその特徴により分類階層化されている場合が多く、これを考慮することにより更に一般的なルールの抽出が可能となる。我々は分類階層を考慮した相関ルール抽出の並列処理方式を提案してきたが、従来の手法では処理負荷の偏りの影響が大きかった [2, 3]。

本稿では、従来の並列処理方式の問題点である処理負荷の偏りを低減させる手法を提案し、実際の分散メモリ型並列計算機上に実装し、その性能評価を行う。

2 分類階層を考慮した相関ルール

データの分類階層構造 $\mathcal{T}$ は図 1 に示すような木構造を有し、その要素をアイテム $\mathcal{I}$ と呼ぶ。$\mathcal{T}$ の辺はアイテム間の階層を示し、アイテム x から y への矢がある場合、x を y の上位アイテム ($ancestor$(x)) と呼ぶ。トランザクションデータベースを $\mathcal{D}=\{t_1, t_2, \ldots, t_n\}(t_i \subseteq \mathcal{I})$ とし、各要素 t_i をトランザクションと呼ぶ。長さ k のアイテム集合とは k 個のアイテムの組合せを指す。アイテム集合 X の支持度 $Sup(X)$ は $\mathcal{D}$ 全体に対して X を含むトランザクションの割合を表す。また、アイテム集合 X がトランザクション t のアイテムまたはそれらの上位アイテムで構成される場合、t は X を含むと表現する。分類階層を考慮した相関ルール (一般化相関ルール) は $X \Rightarrow Y$ で表現され、$X, Y \subset \mathcal{I}$, $X \cap Y = \phi$、Y は X の上位アイテムを含まない。一般化相関ルールは支持度、確信度の 2 つの値によりルールの有意性を示す。ここで、Y が X の上位アイテムであるルール $X \Rightarrow ancestor(X)$ の確信度は常に 100%であり、冗長なルールとなる。一般化相関ルール $X \Rightarrow Y$ の支持度 $Sup(X \Rightarrow Y)$ は $\mathcal{D}$ 全体に対し X と Y を共に含むトランザクションの割合 $Sup(X \cup Y)$ により、また、確信度 $Conf(X \Rightarrow Y)$ は $\mathcal{D}$ の中で X を含むトランザクションのうち、X と Y を共に含むトランザクションの割合 $Sup(X \cup Y)/Sup(X)$ により定義される。

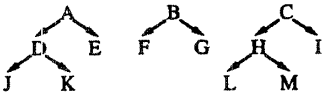

図 1: データの階層構造

一般化相関ルール抽出問題はユーザにより指定された最小支持度と最小確信度を満足する全てのルールを見出すことに相当する。その処理は、まず最小支持度を満足するアイテム集合 (ラージアイテム集合) を生成し、それらを用いて最小確信度を満足するルールを導出する。ラージアイテム集合生成処理はアイテム数、トランザクション数が多い場合に高負荷となり、この効率化が研究の中心となっている。

3 並列処理方式

はじめに、基本的な並列処理方式として分散メモリ型並列計算機環境をモデルとした並列処理方式 H-HPA(Hierarchical Hash Partitioned Apriori)[2] を示す。以下に、長さ k のラージアイテム集合を求める処理 (パス k) を示す。

1. 候補アイテム集合の作成　長さ $(k-1)$ のラージアイテム集合から k 個のアイテムの組合せ (長さ k の候補アイテム集合) を作成し、分類階層の上下関係であるアイテムの組合せを含むものを除去する。残った候補アイテム集合に対して、アイテムをその最上位アイテムに置き換えたものにハッシュ関数を適用し、ハッシュ値に対応するノードの識別子を求め、自分のノードの識別子と等しい場合に主記憶上のハッシュ表に保持する。

2. 支持回数の数え上げ　ローカルディスクからトランザクションデータベースを読み出し、各トランザクションのアイテムを最下位ラージアイテムに置き換え、k 個のアイテムの組合せを作成し、"1" と同一のハッシュ関数を適用し、対応するノードを求める。各ノードに対応する最下位アイテムを送信する。
同時に他ノードから送信されたアイテム集合から長さ k のアイテム集合を作成し、対応する候補アイテム集合とその上位候補アイテム集合の支持回数を 1 増加させる。

3. ラージアイテム集合の決定　全トランザクションの処理が終了した時点で、ノード毎にラージアイテム集合を決定し、他ノードへ放送する。

H-HPA では分類階層を考慮し、階層の上下関係である候補アイテム集合が同一のノードに割り当てられるように候補アイテム集合をノード間にハッシュ分割する。つまり、候補アイテム集合を木の組合せでまとめ、その木の組み合わせ単位でノード間に割り当てる。

Performance evaluation of Load Balancing for Parallel Mining Association Rules with Classification Hierarchy
Takahiko Shintani and Masaru Kitsuregawa
The University of Tokyo, Institute of Industrial Science
Roppongi 7-22-1, Minato-ku, Tokyo 106, Japan

Parameter	Value
Number of transactions	3200000
Average size of the transactions	10
Number of items	30000
Number of roots	30
Number of levels	5-6
Fanout	5

表 1: データセットのパラメータ

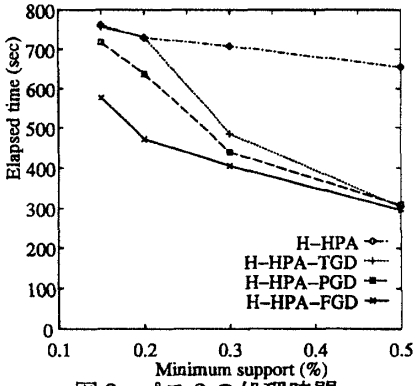

図 2: パス 2 の処理時間

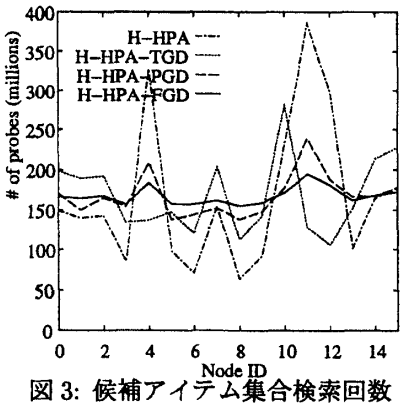

図 3: 候補アイテム集合検索回数

4 負荷分散手法

H-HPA は木の組み合わせ単位で候補アイテム集合をノード間に割り当てるため、均等な負荷バランスを実現することが困難である。また、トランザクションデータに偏りがある、つまり、極端に多くのトランザクションに含まれるアイテム集合が存在する場合、そのアイテム集合が割り当てられたノードに負荷が集中し、処理の偏りが生じる。本節ではノード間の処理負荷の偏りを低減させる手法を示す。本手法では支持度の高いアイテムからなる候補アイテム集合を全ノードに複製し、ノード毎に支持回数を数え上げ、各パスの最後に全ノードでの支持回数の総和を求め、ラージアイテム集合を決定する。他の候補アイテム集合は H-HPA と同様に処理する。

4.1 H-HPA with Tree Grain Duplicate : H-HPA-TGD

H-HPA-TGD では木の頂点にアイテムの支持度が高い木の組合せを見つけ出し、木の組合せの単位で候補アイテム集合を全ノードに複製する。

4.2 H-HPA with Path Grain Duplicate : H-HPA-PGD

H-HPA-PGD では支持度の高い最下位アイテムで構成される候補アイテム集合とその上位候補アイテム集合のパス単位で候補アイテム集合を全ノードに複製する。

4.3 H-HPA with Fine Grain Duplicate : H-HPA-FGD

H-HPA-PGD は H-HPA-TGD よりも細粒度の負荷制御が可能であるが、複製する候補アイテム集合を最下位アイテムで判断するため、最下位アイテムの支持度は低いが、その上位の中間アイテムの支持度が高いものが存在する場合、十分な負荷制御が実現できない。

H-HPA-FGD では支持度の高い中間アイテムからなる候補アイテム集合とその上位候補アイテム集合を全ノードに複製する。つまり、H-HPA-PGD のパスから部分的に候補アイテム集合を選び出すことになるため、不必要な候補アイテム集合の複製を回避し、更に細粒度の負荷制御が可能となる。

5 性能評価

提案した手法を IBM 社製分散メモリ型並列計算機 SP-2 上に実装し、性能測定を行った。本性能測定では 16 台のノード (RS/6000) が HPS(ハイパフォーマンス・スイッチ) と呼ばれる高速ネットワークを介して接続された構成を使用した。また、各ノードにはローカルディスクが接続されている。評価には、小売業における購買トランザクションを模倣して作成したデータセットを用いた。各パラメータを表 1 に示す。

図 2 に最小支持度を変化させた場合のパス 2 での処理時間を示す。また、図 3 に各ノードでの候補アイテム集合のハッシュ表の検索回数を示す。ここで、最小支持度を 0.3%とした。以上の実験では候補アイテム集合数の多いパス 2 についてのみ測定し、トランザクションデータファイルはノード間にほぼ均等になるように分割して、各ノードのローカルディスクに割り当てた。

結果から、H-HPA-FGD が他の手法よりも優れていることがわかる。H-HPA-FGD は他の手法よりも細粒度な単位で候補アイテム集合の複製を行うため、メモリ空間の余裕が少ない場合にも効果を発揮できる。また、余分な候補アイテム集合の複製を行わないため、ノード間の負荷の偏りを効果的に低減できている。

6 まとめ

本稿では分類階層を考慮した相関ルール抽出の並列処理方式の負荷分散手法を提案し、実際の並列計算機上への実装を行った。提案する手法によりノード間の負荷の偏りを低減することが可能であることを示した。

参考文献

[1] T.Shintani and M.Kitsuregawa. Hash based parallel algorithms for mining association rules. In *Proc. of 4th Int. Conf. on Parallel and Distributed Information Systems*, pp. 19–30, December 1996.

[2] 新谷隆彦, 喜連川優. 並列相関関係抽出処理における通信負荷削減方式. 第 54 回全国大会 2R–2, pp. 249–250. 情報処理学会, March 1997.

[3] 新谷隆彦, 喜連川優. 分類階層を考慮した相関ルール抽出の並列処理方式における負荷制御手法. 電子情報通信学会データ工学研究会 データベースワークショップ (DE97–44). 信学技法 Vol.97 No.161, pp. 103–108, July 1997.