GPGPUによる飽和・高価値アイテム集合マイニング

栗山　裕介

要旨

データマイニングとは，データベース中の大量のデータから，必要な情報や有用な知識を抽出する技術の総称である．対象となる情報の形式もアイテム集合，系列データ，時系列と多種多様であり，それぞれ多くの研究が行われてきた．アイテム集合マイニングはデータベース中に頻繁に出現するアイテム集合を抽出する．しかし，頻度による条件を低く設定した場合に膨大なパターンが発見される．そこで，同じ支持度の中で最大のパターンである同値類のみを発見する飽和アイテム集合マイニングが提案された．極大な頻出パターンのみを発見することで，頻出パターンの持つ情報を失わずに不要な頻出パターンを捨て，解を絞ることが出来る．一方，アイテム集合マイニングの別の拡張として高価値アイテム集合マイニングが提案された．アイテム集合マイニングでは，頻度をもとに頻出パターンを発見してきたが，現実の社会では価値が存在する．高価値アイテム集合マイニングでは，アイテムの価値に着目してパターンマイニングを行う．これにより，頻度は少なくとも，価値の高いパターンを発見することが出来る．しかし，こちらもアイテム集合マイニング同様に，価値による条件を低く設定した場合に膨大なパターンが発見される．そこで，高価値アイテム集合マイニングと飽和アイテム集合マイニングをあわせた飽和・高価値アイテム集合マイニングが提案されている．飽和・高価値アイテム集合マイニングとは，価値による条件を満たした上で，頻度による飽和の関係を満たすパターンを発見する分析手法である．本研究では，飽和・高価値アイテム集合マイニングのアルゴリズムを，GPGPUを用いて実装する．GPGPUによる実装は，パターンの評価値・上界値・右，左の枝刈り条件の確認の際に利用しており，それぞれトランザクションによる並列化をしている．評価値と上界値では，パターンに対する各トランザクションの値を並列演算している．右，左の枝刈り条件の確認では，現在のパターンと対象のアイテムを各トランザクションが保持しているかのマッチングをとるのに並列演算をしている．評価実験では，様々なパラメタ設定のもと，中国の小売業のデータ10万件と，Twitterから抽出したネガティブ語のデータ3千万件を対象に，パターン抽出を行った．実験の目的は，デー タの大規模化である．結果として，3千万件まで実際に実行し結果を得ることが出来た．