論 文 内 容 の 要 旨

博士論文題目:

Computational Complexity of Finding Correlated Items in Market Basket Databases
(販売データベース中に存在する関連性の高いアイテムのマイニングに要する計算量)

氏　　名: 権 娟大

近年の計算機技術の進展により，大量のデータを格納したデータベース中のすべてのデータを解析することが可能になった．そのような大規模データベース中のすべてのデータを解析し，データベースの利用者にとって有益な情報を得ることをデータマイニングと呼ぶ．データマイニングにおいて最もよく研究されている問題の一つとして，大量の販売トランザクションを含む販売データベースから有意な相関規則を求める問題が挙げられる．有意な相関規則を求める問題は，まず関連性のあるアイテム（商品）の集合をすべて見つけ出し，次に得られた集合を用いて有意な相関規則を構成することにより解くことができる．本論文では相関規則を求める問題に関連したいくつかの問題について議論する．

与えられたアイテム集合中のアイテム間の関連性の度合を示す尺度としてサポートと呼ばれる概念が提案されている．与えられた閾値を越えるサポートをもつアイテム集合を頻出集合と呼ぶ．すべての頻出集合を求めるアルゴリズムが多数提案されているが，それらの多くは計算量について議論されておらず，その性能がベンチマーク上で実験的に評価されているだけである．本論文では頻出集合問題を与えられた大きさをもつ頻出集合が存在するか否かを判定する問題として形式的に定義し，その NP 完全性を示す．この結果から，すべての頻出集合（従って，有意な相関規則）を求めることは P=NP でない限りデータベースの大きさに対する多項式時間では不可能であることが導かれる．さらに，本論文ではすべての頻出集合が効率良く求められるデータベースの部分クラスを提案する．

また，サポートに関するいくつかの弱点が指摘されている．例えば，高いサポートをもつアイテムを多く含むようなアイテム集合に対しては，そのサポートもアイテム間の関連性の有無に関わらず高くなる傾向がある．本論文ではサポートに代わるいくつかの尺度を提案する．これらは

- 与えられたアイテム集合中の各アイテム間に関連性がないと仮定したときのサポートの期待値に対する実際のサポートの比率
- 与えられたアイテム集合中のどのアイテムも含まないようなトランザクションの割合

などを考慮して定義される．これらの尺度のもとで，あるアイテム集合中のアイテム間の関連性を示す値が与えられた閾値を越えるならば，そのアイテム集合を高共起度集合と呼ぶ．本論文では，高共起度集合問題を与えられた大きさをもつ高共起度集合が存在するか否かを判定する問題として形式的に定義し，どの尺度に対してもこの問題が NP 完全であることを示す．さらに，本論文ではすべての高共起度集合が効率良く求められるデータベースの部分クラスを提案する．

氏 名	権 娟大

(論文審査結果の要旨)

データマイニングの対象の一つとして，販売データベースの解析が挙げられる．販売データベースにおいて同時によく買われる商品の組合せ（頻出集合と呼ぶ）を発見するマイニングアルゴリズムが種々提案されたが，その性能評価はテストデータ上でしか行われておらず，漸近的挙動は不明であった．本論文は，頻出集合のマイニングに要する計算量について形式的に議論している．また，頻出集合モデルでは，共起性が高いとは言えない商品の組合せが頻出集合になり得るという問題点に関して，頻出集合に代わるモデルをいくつか提案している．本論文の主な成果は次のように要約される．

(1) データマイニングにおいて重要な基本的問題である頻出集合問題が NP 完全であることを示している．ここで，頻出集合問題とは，販売データベース D，実数 r $(0 < r < 1)$，正整数 k が与えられたとき，r 以上の率で同時に買われる商品の部分集合で，要素数が k 以上のものが存在するかどうかを判定する問題である．上の NP 完全性の結果から，頻出集合を効率良く求めることは一般には困難であることが分かる．さらに，顧客の現実的な挙動を表す制約条件下で，頻出集合を求める多項式時間マイニングアルゴリズムを提案している．

(2) 頻出集合モデルでは，商品の集合の共起性の高さを実際の出現頻度に基づいて評価している．これに対し，本論文で提案されている共起度と呼ばれる評価基準では，商品の集合の実際の出現頻度をその期待値と比較することにより，関連性の高さを相対的に評価している．さらに，高共起度集合を求める問題が NP 困難であることを示し，上と同様の制約条件下で，与えられた要素数をもつすべての高共起度集合が多項式時間で求められることを示している．

以上のように，本論文は，関連性の高い商品の集合を求める計算量について形式的に議論するとともに，それらが多項式時間で求められるための現実的な条件を提案しており，データマイニングの分野において，学術上，実用上寄与するところが多い．よって，本論文は博士（工学）の学位論文として充分に価値のあるものと認める．