様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ８ 日現在

機関番号：１２６１２
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００１６４

研究課題名（和文）集約バスケットからのデータマイニング手法の研究

研究課題名（英文）A Study on Data Mining from Aggregated Basket

研究代表者
沼尾　雅之（Numao, Masayuki）

電気通信大学・情報理工学（系）研究科・教授

研究者番号：９０５０８８２１

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：データマイニングの代表的な手法であるバスケット分析は，相関マイニングとも呼ばれ，どの商品が他の商品と一緒に買われるかのパターンを購買データから見つける手法である．既存アルゴリズムでは，複数のバスケットが集約されている場合などを扱えず，例えば，家族全員の買い物を１つにまとめたバスケットを分析することができなかった．
本研究では，集約バスケットを定義し，それを従来のバスケットに復元して商品間の相関を求める手法を提案し，実際に，復元アルゴリズムを開発し，相関分析システムを実装し，製造業の実際のデータによって実験・評価した．その結果，有効性が確認され，製造業，流通業に拡大できることが示された．

研究成果の概要（英文）：Basket analysis is a well-known data mining method to find the pattern of which item is purchased often with other items from point-of-sales (POS) data. In existing algorithm, however, the input basket is restricted to a set of items. Since the number of items in the set is 1 each at the most, it cannot deal with the aggregated basket where multiple baskets are merged into 1 basket. Thus it cannot deal with the basket of family items, or data collected in a certain period.
In this study, we defined the aggregated basket, proposed how to reconstruct the baskets from the aggregated basket to find the association pattern. We also developed the association mining system form the aggregated baskets and evaluated it by using the actual manufacturing data. The result shows that the proposed method is very useful to find the new patterns which cannot be extracted by the existing method, and it can be applied to manufacturing and distribution industry.

研究分野：知能情報学

キーワード：知識発見とデータマニング　バスケット分析　ビッグデータ　相関分析

１．研究開始当初の背景

データベースから知識を発見するデータマイニングは，ビッグデータ解析の主要な方法である．バスケット分析は，相関分析とも呼ばれ，複数のアイテムの間の共起関係を抽出するものである．バスケット分析の代表的な例としては，スーパーマーケットにおける購買データから，複数の商品が組み合わされて購買されるパターンを求めるものがある．Rakesh らは，このバスケット分析手法を提案するとともに，分析アルゴリズムを Apriori アルゴリズムとしてを公開した[1]．バスケット分析手法は，アルゴリズムの効率化や応用分野の拡張等，盛んに研究されて現在に至っている．例えば，購買の時間的な順序を考慮した時系列相関関係分析は，インターネット市場における顧客の購買パターン分析に応用されている．それ以外にも，アイテムとして自然言語の文や，タンパク質といった構造体も扱えるようにしたグラフマイニングなども提案されている．

研究代表者らは，データマイニング技術を実際の産業に応用することを目指して，製造業における故障診断や問題追跡などに，OLAP 手法やマイニング手法を応用することを提案し[2]，実際にガラス瓶製造工程での不具合製品の解析をしている[3]．ところが，製造工程で実際に得られるデータは，一定時間ごとに集計されたデータが大半であり，Apriori アルゴリズムなどの従来手法が前提としている粒度のバスケットが得られないことが多い．なぜならば，バスケットはアイテムの集合であることが前提とされているため，アイテムが複数個ずつ含まれているような多重集合としてのバスケットは考慮されていないためである．従って，入力バスケットが単純集合という前提を外すことによって，製造業や流通業の現場において得られるような集約データからの故障･欠点の相関分析や，家族構成や買い物頻度も考慮した，より詳細な商品購買分析などができるようになる事が期待される．

[1] R.Agrawal, et.al, Mining association rules between sets of items in large databases, Proceedings of ACM SIGMOD 1993 International Conference on Management of data, pp.207-216, 1993

[2] S. Hido, H. Matsuzawa, F. Kitayama, M. Numao: Trace Mining from Distributed Assembly Databases for Causal Analysis, Proceedings of The 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD2009), LNAI, Springer , pp.731-740, 2009.

[3] 沼尾雅之，松尾総一朗：プロセス産業のための履歴テーブルに基づく品質分析法の提案，日本データベース学会論文誌, Vol.10, No.1，pp.79-84，2011.

２．研究の目的

大量の購買データから，複数の商品が一緒に購買されるパターンを見つけるバスケット分析は，データマイニングの主要な方法であり，顧客分析，市場分析等の分野に応用されている．しかし，購買の単位であるバスケットは，商品が高々1 個だけ含まれる集合を前提としているため，複数のバスケットの要素が混ざってしまった集約バスケットや，その結果として，商品が複数個ずつ含まれるような多重集合型バスケットは，そのままの形でバスケット分析に用いることができない．本研究の目的は，このような集約バスケットから，商品間の相関を求める手法を開発することである．

すでに多くの研究がされているバスケット分析であるが，入力バスケットの粒度が一様でないことに問題点を見出し，そこから，多重集合バスケットを入力としたマイニング技術を再構築することは，学術的にも新規性があり，また応用分野開拓においても意味のある事である．本研究は，データマイニングの研究領域に新しい技術課題を示すことにもより，この分野の研究を活発化させることとなる．一方，製造業や流通業などの実用領域において，検査器からの集約データからの故障・欠点の相関分析や，家族構成や買い物頻度も考慮したより詳細な商品購買分析など，これまでできなかった分析や，より高い精度の相関関係分析ができることも期待できる．

申請者らが開発している集約バスケットからの復元法を改良するとともに，バスケットに内在する集約性を考慮したマイクロバスケット分析手法を確立する．

３．研究の方法

研究では，単純バスケットが複数集まったものとして集約バスケットを定義し，その集約バスケットからの相関関係抽出法を開発する．ここで，図１に示すように，集約バスケット(b)から，それを構成する単純バスケット(c)を復元する事を考える．もし，単純バスケットが復元できれば，従来手法を用いて相関関係は抽出できる．これが可能かどうかを検

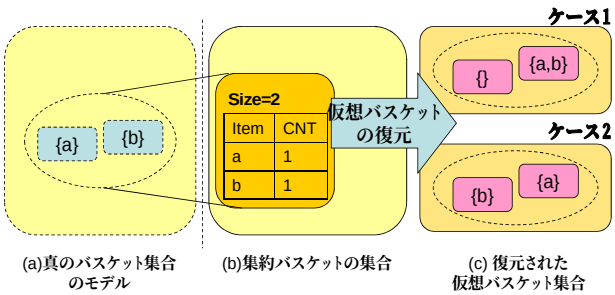

図１．仮想バスケットの復元

証するために，まず，テストデータとして，単純バスケット(a)から始めて，これを集約することで，集約バスケット(b)を人工的に作り，そこから復元を試みる事によって，復元されたバスケット(c)が，真のバスケット(a)にどの程度近いかによって，復元法を評価する事ができる．まず，この枠組みを使って，復元法の理論的解析と実データによる実験・評価を進める．

(1) 集約バスケットの定義

本研究は，アイテムが複数個ずつ含まれる多重集合バスケットから，アイテム間の相関関係を抽出するマイニング手法を，その定義から応用に至るまで総合的に確立することである．まず，集約バスケットを定義する前に，従来の単純バスケットとそこからの相関関係分析を Apriori アルゴリズムに基づいて説明する．商品の集合としてのアイテム集合を

$$I=\{a_1, .., a_n\},$$

とし，バスケットはその部分集合

$$t_i \subseteq I$$

と定義する．そしてバスケットが集まったものをトランザクション集合

$$T=\{t_1,..,t_N\}$$

とする．次に複数商品間の相関関係を表す２つの指標を定義する．あるアイテムの部分集合 **A** が，いくつのバスケットに含まれていたかを，全体のバスケット数で除したものを支持度

$$Supp(A)=|\{t \in T | A \subseteq t\}|/|T|$$

とし，これが高いアイテム集合を頻出パターンと呼ぶ．また，あるアイテム集合 **X** が買われた時に，同時にアイテム集合 **Y** も買われる度合いを確信度

$$Conf(A)=Supp(X \cup Y)/Supp(X)$$

で与え，これら２つの閾値を与えてそれを両方満たすルールを見つけ出すものが Apriori アルゴリズムである．

次に，集約バスケットは，アイテムカウント集合として，

$$at_i=\{<id_i, cnt_i> \mid id_i \in I\}$$

と定義する．これは，アイテム id とその個数 cnt のペアを要素とする集合で表現するものであり，集約バスケットに現れないアイテムも<id, 0>という形で表現される．最後に集約バスケットが集まった集約トランザクションは

$$AT=\{at_1,..,at_N\}$$

となり，これが本提案での入力データとなる．

(2) 集約バスケット分析アルゴリズムの開発と評価

アルゴリズムは仮想バスケットの復元と，そこからの相関関係抽出に分けられる．前者については，集約バスケットに含まれているアイテムを単純バスケットに分配するアルゴリズムの開発が必要になる．これは，集約バスケットの統計的な性質を守りながら，単純バスケットを生成するようなトランザクションデータ生成プログラムを開発することと等価と考えられる．ところが現在提案されているデータ生成法は，頻度分布など統計量に人工的な偏りがあり，現実のバスケットを反映していない．したがって，まずこれら既存のデータ生成アルゴリズムを精査し，頻出パターンの分布が，現実データからのものと同等になるようなアルゴリズムを開発する．さらに，生成データについての統計的解析や，アルゴリズム自体の効率の計算量的評価も行う．

(3) マイクロバスケット分析アルゴリズムの開発と評価

通常のバスケットは，より細かい粒度のバスケット（これを**バスケットプリミティブ**と呼ぶ）の集約という前提で，ここから，アイテム間のより精度の高い相関関係を抽出する方法を，新たにマイクロバスケット分析と定義し，（２）と同様にバスケットからバスケットプリミティブを復元する問題として定式化する．この際，集約度は未知数であるので，集約度を変化させながら，真のモデルに収束させていく分析方法を確立する．また，この手法の応用として，懐中電灯１個と乾電池２個といったアイテムの個数間の関係の抽出方法も開発する．

(4) 製造および流通の実データによる評価および検証

製造業においては，検査装置などの制約により，欠点を個品単位ではなく一定時間ごとの個数として検査していることが多い．したがって，得られる品質データは必然的に集約バスケットとなり，さらに集約度も大きいため，今まではバスケット分析の対象にはならなかった．また，流通業，小売業におけるバスケット分析についても，粒度の多様性に起因するノイズによって，有用な相関の発見にいたらない場合も多かった．そこで，品質管理データから欠点種ごとの相関関係抽出，および，購買データからのマイクロバスケット分析を行うことにより，数百万個のバスケットからなる実トランザクションによって提案手法を評価し，既存のバスケット分析に対する優位性を検証する．

４．研究成果

(1) 集約バスケットの理論的解析

図１のように，集約バスケットは真のバスケットモデルが集約されているという仮定で，真のバスケットモデルの統計的特徴が，どのように変換されるか，また，復元された仮想バスケットの何割が真のバスケットモデルの特徴を保持するかについての理論的解析を行った．たとえば，アイテム a,b が１つずつ含まれた集約度２の集約バスケットからは，２つの仮想バスケットが構成されるが，組み合わせは２通りあり，どちらかが，真のバスケットモデルと一致するはずである．一

致する割合は，仮想バスケットの復元法に依存するが，これを等確率とした場合，仮想バスケットから得られた 2 つのアイテムの相関の支持度 $Supp^V(\{a,b\})$は，真のモデルでの支持度 $Supp^T(\cdot)$を用いて以下のように表せることがわかっている（雑誌論文④）．

$$Supp^V(\{a,b\}) = (1/k)\ (Supp^T(\{a,b\}) + (k-1)Supp^T(\{a\})\ Supp^T(\{b\}))$$

ただし，k は集約度

ここから，集約度が大きくなるにつれて，真のモデルでの相関が再現される割合が減り，独立な事象と認識されていくことがわかる．

(2) 集約バスケット分析アルゴリズムの開発と評価

まず，単純バスケットのトランザクションデータ生成方法についての既存研究をサーベイし，統計的な偏りの少ない生成法を提案した（学会発表⑥）．

次に，仮想バスケット復元法として，確率的分配方式と個別分配方式を提案し，比較・評価した．前者は，集約バスケットに現れるアイテムの出現確率を計算し，これに基づいて，単純バスケットに振り分ける方式である．一方，後者は，集約バスケットを分割して単純バスケットを作る方法である．前者では．各アイテムが集約バスケット内に出現する確率に基づき分配されるため，一見，集約前のバスケットをうまく復元できるように思えるが，例えば，集約バスケット A に 1 個だけ含まれていた出現確率 1/2 のアイテムを，2 つのバスケット A1,A2 に分配する場合，A1 か A2 に分配される以外に，A1 と A2 のどちらにも分配されない場合や，どちらにも分配されてしまう場合が発生する．そのため，全ての集約バスケットに含まれていた各アイテムの個数の総和が，全ての仮想バスケット中に含まれる個数の総和と異なるという問題が発生する．これは，仮想バスケットを復元した後に，アイテムの出現頻度を数え上げる際に大きな誤差をもたらす事になり，後者の方が優れている事が示された．

さらに，複数の復元アルゴリズムを定量的に比較するために，**復元率**を新たに定義した．これは，真のバスケットモデルにおける上位頻出パターンのいくつが，復元された仮想バスケットから得られた上位頻出パターンに含まれるかで評価する．実際にトランザクションデータを生成して復元率を計算した結果，データセットや，集約率によって復元率が異なることが確認され，理論的解析結果と一致することが示された（図２：横軸が頻出パターン上位の数で，縦軸が復元率，A,B は異なるデータセット，また，AX,BX の X は集約サイズを表す）．

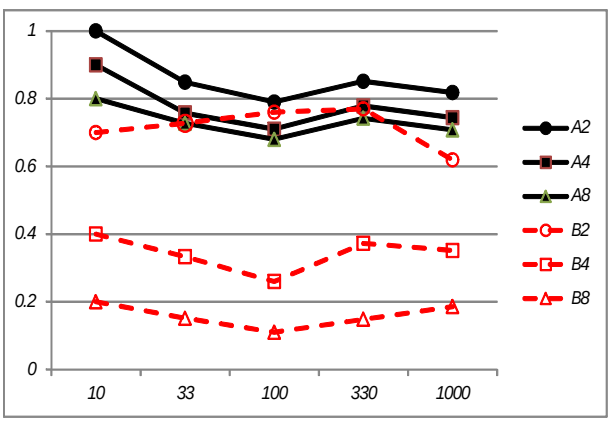

図 2．復元率

(3) マイクロバスケット分析アルゴリズムの開発と評価

(2)と同様にアルゴリズムは仮想バスケットの復元と，そこからの相関関係抽出に分けられるが，集約度が未知数になるため，アルゴリズム的には探索法や繰り返しによる収束法が必要になる．ここでは繰り返し近似法をベースとしたバスケット復元アルゴリズムを開発した。まず，集約バスケットの複数のアイテム重複度を比較する事によって，集約度の範囲を絞り込む事ができる事が示された．これを利用して，比較的少ない回数で集約度を推定して，分割候補を生成できる事を示した．この方法は，全体の集約度が低い場合にはうまくいくが，集約度が高い場合は誤差が大きくなり，これが，さらに集約度の推定を妨げるという悪影響を与えて，集約度が収束しないこともあった．これについては，今後の更なる研究が必要である．

(4) 製造および流通の実データによる評価および検証

製造業における欠品データから，実際に欠品種類の相関関係を分析した．従来は，欠点種類ごとに異なる検査装置によって検査されているために，欠点はすべて独立なものとして処理されている．しかし，欠点間の相関パターンを調べることにより，冗長な検査装置を発見したり，あるいは，欠点間の因果関係を突き止めることができる．

実際に使ったデータは，ガラス瓶の製造プロセスにおける欠品データである．これは一定時間における欠点の種類と，その数を示しており，これは集約バスケットに対応させる事ができる．集約度に対応する製品数も入っているために，集約度既知の集約度バスケット分析手法を採用する事ができた．約 15000 行の欠品データがあり，集約度は約 18 であった．集約度が大きいために，復元率は低かったが，欠品間の相関ルールを得る事ができた．

5．主な発表論文等

（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計　4　件）

① J. ZHOU, X. LI, H. CHEN, R.CHEN, and M. NUMAO: "3D Objects Tracking by MapReduce GPGPU-Enhanced Particle Filter", IEICE TRANSACTIONS on Information and Systems, Vol.E98-D, No.5, pp.1035-1044, 2015.（査読有り）
② 中野隆介, 沼尾雅之：無線 LAN アクセスポイントへの検索要求を用いた屋内混雑度推定手法, 日本データベース学会論文誌, Vol.12, No.1, pp.121-126, 2013.（査読有り）
③ 高橋 麻美 , 根路銘 崇 , 沼尾 雅之:利用者単位の消費電力量測定手法と家庭における節電指標の提案, 情報処理学会論文誌,53(7),pp.1711-1720, 2012.（査読有り）
④ 松澤裕史, 沼尾雅之：集約バスケットからの相関関係マイニング, 電子情報通信学会論文誌 D,Vol. J95-D No.2, pp.170-182,2012.（査読有り）

〔学会発表〕（計 9 件）
① 篠原 透, 沼尾 雅之：シーケンシャルパターンマイニング拡張による特徴的なコード進行の抽出手法, 第 7 回データ工学と情報マネジメントに関するフォーラム (DEIM2015), F1-2, 2015 年３月２日, 磐梯熱海ホテル湯の華（福島）.
② 安藤勇, 沼尾雅之：RFID を用いた電波強度による異常行動認識システムの提案, 第 13 回情報科学技術フォーラム（FIT2014), K-044, 2014 年 9 月 5 日, 筑波大学（つくば市）.
③ 横堀哲也, 沼尾雅之：プローブ要求を利用したスマートフォンユーザー向け屋内位置推定手法, 第 13 回情報科学技術フォーラム（FIT2014), 査読付き論文, RO-004, 2014 年 9 月 3 日, 筑波大学（つくば市）.
④ 篠原 透, 沼尾雅之：シーケンシャルパターンマイニング拡張による特徴的なコード進行の抽出手法情報処理学会, 音楽情報科学研究会 104-3, 2014 年 8 月 25 日, 京都大学（京都市）
⑤ ZHOU Jieyun, NUMAO Masayuki, LI Xiaofeng, CHEN Haitao：3D Objects Tracking by GPGPU-Enhanced Particle Filter Algorithms, 第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM2014), D9-2, 2014 年 3 月 5 日, ウェスティン淡路（淡路市）.
⑥ 松石 浩輔, 沼尾 雅之：相関分析のためのラティス構造に基づくデータセット生成器, 第 6 回データ工学と情報マネジメントに関するフォーラム(DEIM2014), C6-5, 2014 年 3 月４日, ウェスティン淡路（淡路市）.
⑦ 清水 智可良, 沼尾 雅之:RSS を用いたセンサーデータマッシュアップのためのウェブアーキテクチャ, マルチメディア, 分散, 協調とモバイル(DICOMO2013)シンポジウム, 4H-1,2013 年 7 月 11 日, ホテル大平原（北海道）.
⑧ 中野 隆介, 沼尾 雅之 ：無線 LAN アクセスポイントへの検索要求を利用した屋内混雑度推定, 第 5 回データ工学と情報マネジメントに関するフォーラム (DEIM2013), E6-5, 2013 年 3 月 4 日, ホテル湯の華（福島）.
⑨ 服部陽彦, 沼尾雅之：電源環境の変化に強い家電認識手法の提案, 情報処理学会ユビキタスコンピューティングシステム研究会, 2012 年 11 月 1 日, お茶の水大学（東京）.

〔その他〕
ホームページ等
www.nm.cs.uec.ac.jp

６．研究組織
(1)研究代表者
沼尾　雅之 （NUMAO Masayuki）
電気通信大学・情報理工学研究科・教授
研究者番号：９０５０８８２１

(2)研究分担者
丸山　宏 （MARUYAMA Hiroshi）
統計数理研究所・モデリング研究系・教授
研究者番号：９０６０９７２８

(3)連携研究者
なし