

バスケット解析のためのブラウザーの試作

5 E-4

木村 雅彦 清水 周一 沼尾 雅之

日本アイ・ビー・エム株式会社 東京基礎研究所

1. はじめに

近年、企業内に蓄積されている膨大な量のデータベースを活用するための手法として、データ・マイニングが脚光を浴びつつある。データ・マイニングは、一言で言えば「データベースなどの大量の生データから隠れた情報を発掘する (mining)」、いわゆる知識発見型の分析手法である。

筆者らは、汎用クラスタリング・エンジン TEWAT[1] などを使って、データ・マイニングの流通や教育などの分野への応用、分野特有のノウハウの蓄積、およびそれらの成果を実装したプロトタイプ・システムの開発に取り組んでいる。本稿では、その中でも、TEWAT によるクラスタリングの結果を分かりやすく表示し、対話的に結果の検証やより詳しい分析をするためのブラウザーについて報告する。

2. TEWAT を使ったバスケット解析

バスケット解析とは、POS 端末などから日々入力され蓄積される売上データを分析することであり、その中には、併売パターン（複数の商品が同時に買われること）や順序パターン（同じ客が複数の商品を順に買うこと）などの購買パターンの発見や、購買パターンと顧客の属性との因果関係の分析が含まれる。

解析に使用した TEWAT は、約3万までの2値属性を持つ約100万件までのデータを、属性のビット・パターンの類似性から、 $O(n)$ のオーダーで準最適なクラスターに分割する。

これをバスケット解析に応用するには、買われた全ての商品をコードごとに属性として用意し、1つ1つのバスケット・データを商品ごとのある・なしのビット・パターンに変換する。これを TEWAT にかけると、ビット・パターン（＝併売パターン）の似通ったバスケットからなるクラスターの集合が得られる。そこで、クラスターごとに商品構成や併売パターンを調べると、データ全体の併売構造を知る手がかりになる。

3. ブラウザー

「はじめに」で述べた目的を実現するため、ブラウザーには主な画面が3つ用意されている。

まず、TEWAT の出力を表示するために、クラスタ・ブラウザー：

商品×クラスターのマトリックスに、各クラスター内の商品の数を疑似カラーで表示した画面

バスケット・ブラウザー：

商品×バスケットのマトリックスに、指定したクラスター内の各バスケットの持つ商品に*マークをプロットした画面

という階層関係にある2つの画面がある。

クラスタ・ブラウザーは、商品分布の大まかな傾向を知り、より詳しい解析のためにクラスターの絞り込みをするのに使用する。一方、バスケット・ブラウザーは、クラスター内での詳しい分析を1つ1つのバスケットを参照しながら行なうのに適している。例えば、複数の商品をマウスで指定し、その商品を全て (AND) またはいずれか (OR) を含むバスケットをハイライト表示し、併売された商品を併売数でソートして表示すること（多商品併売解析）が可能である。これにより、TEWAT が作成したクラスターが本当に有用な併売パターンを持っているのかを検証できる。

最後に、併売ブラウザー：

商品×商品のマトリックスに、2商品間の併売数を疑似カラーで表示した画面

は、TEWAT とは関係なく、2商品間の併売数を、全てのバスケットについて数え上げた結果を表示する。TEWAT を使うとクラスターに分割されて分かりにくい商品の併売関係を調べるのに有効である。

また、以上の画面では、

- 商品 ID、商品数、併売数などによるソート
- パン・ウインドウによる概略表示及び移動先の指定
- 別ウインドウでの商品等の詳細情報の表示

などの機能が備わっており、解析を支援したり、操作性を向上させたりするのに役立っている。

4. 適用例

あるホームセンターの POS データ 104,503 件（商品数 2897）に適用し、665 のクラスターが得られた例を紹介する。

図 1 は、クラスター・ブラウザーの一部を拡大したものである。図中、横方向の帯は、売れ筋の商品がクラスターによらず広範囲で買われていることを示している。また、1つのクラスターに、近い商品コードの商品が集中している場合は、縦の線として現れる。

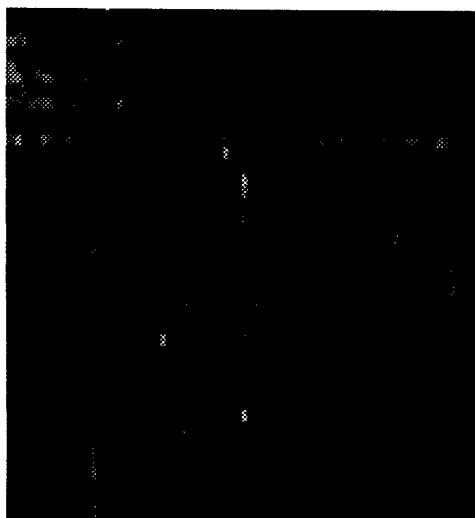


図 1: クラスター・ブラウザー（一部）

図 2 は、バスケット・ブラウザーで多商品併売解析をしているところ（一部）を表している。左端に、商品名の一部、() 内に併売数、クラスター内での商品数が順に見える。上端の縦の数字は、バスケット ID を表しており、そのバスケットに含まれる商品のところに、*のマークがプロットされているのが分かる。

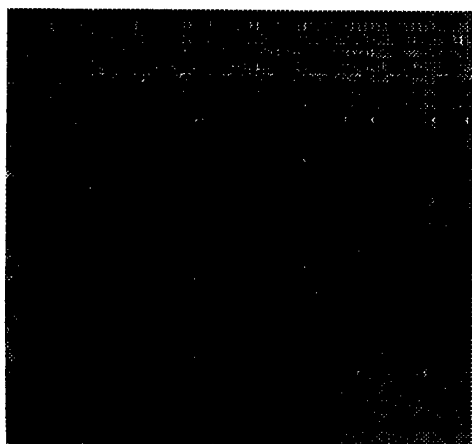


図 2: バスケット・ブラウザー（一部）

図 3 は、併売ブラウザーのパン・ウインドウである。

図からは、1) 上部に L 字型の筋が数本見えること、2) 対角線付近に明るいかたまりが並んでいること、3) それ以外の所に、所々ボツボツと明るい点が見えること、などが分かる。1) は売れ筋の商品がどの商品ともよく併売されること、2) は同じく近い商品コードの商品同士がよく併売されること、3) は商品コードの遠い商品同士の組合せを表している。1), 2) が大半であることを考えると、3) は予想外の注目すべきケースと言える。



図 3: 併売ブラウザー（パン・ウインドウ）

5. おわりに

汎用クラスタリング・エンジン TEWAT を使用したバスケット解析のためのブラウザーを試作した。3種類の画面を組み合わせることにより、商品分布の大まかな傾向を探ったり、1つ1つのバスケットを参照しながら多商品併売解析を行ったり、意外性のある商品の組合せを発見したり、といったことが快適に行なえることがわかった。

試作をして気がついたのは、データ・マイニングにおける人間の認識能力の有用性である。例えば、図 3 を見れば、人間は用意に全体の傾向をつかみ、予想外の商品の組合せを発見することができる。このように、ユーザー・インターフェースを工夫することによって人間の高い認識能力を知識発見に利用することは、今後もデータ・マイニングにおける有効な手段の一つとなると思われる。

参考文献

- [1] Chantal Bédécarrax, et al.: "A New Methodology for Systematic Exploitation of Technology Databases", *Information Processing and Management*, Vol.30, No.3, pp.407-418, 1994