社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2003－DBS－130 (18)
2003－ＦＩ－ 71 (18)
2003／5／23

『リアルタイム・データマイニングと相関関係の可視化』

石井 義興　須賀京子　大場 達生
株式会社ビーコン インフォメーション テクノロジー
〒163-1507 東京都新宿区西新宿 1-6-1 新宿エルタワー
{ishii, kksuga, oba}@beacon-it.co.jp

Abstract

企業経営の戦略の要として CRM(Customer Relationship Management)が推進されるとともに、顧客情報の収集手段として、コンタクトセンタや SFA、インターネット（Web サイトや電子メール）を媒体とした方法が広く普及してきている。その結果、企業が利用可能な顧客情報は、その質、量ともに増大を続けており、効率的な顧客情報の処理が急務となってきている。本稿では、増大しつづける顧客情報を効率よく管理し、かつ効果的なマーケティングを実践するための手段として、顧客情報を集合概念で管理し、顧客の属性と購買傾向などの相関関係を可視化する効果的なリアルタイム・データマイニング手法を提案する。

Real Time Data Mining and Visualization for Correlation

Yoshioki ISHII　Kyoko SUGA　Tatsuo OBA
Beacon Information Technology, Inc
Shinjuku L Tower, 7F. 1-6-1, Nishi-shinjuku, Shinjuku-ku, Tokyo 163-1507
{ishii, kksuga, oba}@beacon-it.co.jp

Abstract

While CRM, Customer Relationship Management, has been promoted as a pivot of business strategies, it becomes prevalent extensively that they collect customers' data through contact centers, SFA information and internet such as websites and e-mails. Consequently, the amount of customer data obtained by a company continues to increase in both quality and quantity and thus efficient processing of the customer data is urgently required. This paper proposes an effective real time data mining technique to visualize correlations such as the one between data attributes and purchase tendency in order to practice effective marketing using the increasing customer data managed in the set theory.

はじめに

第一著者は 1978 年以降、情報検索の分野に集合概念を応用することを考え、商品開発を行った。そのシステムは特許、顧客、色々な事実等の情報検索のため現在でも利用されている [1), 2), 3)]。さらに、それを発展させ、顧客情報管理を中心とし、過去の情報も管理すべくテーブルを二次元から時間軸を含む三次元へ拡張したデータモデルを提案し [4)]、商品開発を行った [5), 6)]。

第一著者は情報の有効活用を目的とするデータ・ウェアハウスを管理するためには、データベースとは異なる管理システムが求められることを主張し、色々なタイプの DWMS（Data Warehouse Management System）が存在することを提案した [7)]。そして、その実現型の一つとして開発された TimeCube についての最近の利用法については 2001 年 11 月 ER2001 で発表された [8)]。情報検索分野に集合概念を導入するアプローチは、データ・マイニングの基を成すものと考える。

１．リアルタイム・データマイニング

1.1 リアルタイム・データマイニングの要求

リアルタイム・データマイニングとは、データマイニングを行う担当者が、必要なときに必要なデータを容易に取得し、データ間の相関関係をはじめとした分析を実行し、その結果をもとに意思決定を行うというプロセスを、リアルタイムに実行する手法をいう。

一般のデータマイニングには、利用者が意図した結果を得るまでに次のプロセスが必要である。

① データマインニングを行うためのデータの準備
② サンプリングデータの抽出
③ データマイニングの実行

④　予測モデルの作成
⑤　全体データへの適用
⑥　最終結果の作成

ふつうこれらのプロセスが完了し、結果を得るまでには、データマイニング実行時の変数の重みづけや適用する統計解析手法を変更するなどして数回以上繰り返すことが多い。そのたびにデータモデルの再計算が行われるため、データマイニングの実行には数日間以上を要することが多い。これは、現在の企業が入手できる顧客情報が質・量ともに日増しに膨大になってきているからであると考えられる。質と量の２つの側面から次のように問題を捉えることができる。

- データマイニングなどによる分析を行うとき、対象データが量的に膨大すぎる。
- データマイニングなどによる分析を行うとき、顧客を表すデータが多様すぎ、どれが本当に有意なデータであるか判断が困難である。また各種のデータの相関関係が把握しづらい。

本稿ではこれらの問題を解決する仕組みを『リアルタイム・データマイニング』と定義する。リアルタイム・データマイニングは、time cube データモデルを基盤とした集合概念の応用と相関関係の可視化を実現するものであり、これまでにはなかったまったく新しい手法である。

1.2 リアルタイム・データマイニングの定義

本稿ではリアルタイム・マーケティングに必要な要素を次のように定義する。

①　データウェアハウス：必要なデータを事前にアプリケーションにあわせ最適なスキーマ（モデル）で定義しておき、かつデータは過去から現在に至るまで時系列に沿って恒常性が保たれて管理されている。

データマイニングによって一貫性のある結果を導き出すためにはデータの恒常性が求められる。また顧客分析においては顧客属性のようにゆっくり変化する時制要素をもつ定性情報、および購買履歴などの時系列的なトランザクションデータが重要である。従って、これらのデータを効率よく保持・保管するデータウェアハウスが必要である。

②　分析のリアルタイム性（１）：データマイニング担当者が必要なときに、必要なデータをすぐに取り出すことができ、かつ有意な分析ができる。

リアルタイム・データマイニングの利用者が任意の条件でデータを抽出し、さらにその結果を用いて分析を行うことが、オンラインで、かつリアルタイム性をもって実行できなければならない。

③　分析のリアルタイム性（２）：可能な限り分析結果の取得と意思決定までの時間の短縮を実現する。このためデータの管理に集合概念を用いる。

リアルタイム・データマイニングでは可能な限り意思決定までの時間を短縮しなければならない。ハイパフォーマンスな分析環境が要求されるため、データの管理に集合概念を用いる。集合概念の利用・応用については２章で述べる。

④　相関関係の可視化：有意な分析と意思決定の支援のため、集合と集合間の相関関係を可視化することができる。

データマイングで有意な結果を得る、あるいは予測を行うためには、マイニング結果の明細データや集計表などのテキストデータだけでは不充分である。データの視覚化や相関関係の可視化が有効である。相関関係の可視化については３章で述べる。

⑤　仮説検証手法の適用：分析と意思決定の結果が正しかったのかどうかを検証することができる。

データマイニングを経営活動のなかで実践していくにあたっては、予測モデルを作成して結果に基づいて行動を起こすことを１回限りで終わらせることは効果をなさない。繰り返し実行結果を検証し、検証結果を次の予測モデル作成の指標にすることが要求される。また、CRM（Customer Relationship Marketing）やロイヤルティ・マーケティングなどのマーケティング手法においても仮説に基づいた実行とその結果の検証を行いつつ、顧客に継続的にアプローチを行い、顧客満足度や忠誠度、あるいは顧客生涯価値を高めていくことが目的のひとつである。従ってリアルタイム・データマイニングにおいても仮説検証手法が実践できることが必要である[8]。

２．リアルタイム・データマイニングへの集合概念の応用

2-1. リアルタイム・データマイングでのデータウェアハウスの利用

リアルタイム・データマイニングでは、前章で述べたように、まずデータウェアハウスを準備し、様々な仮説に基づき各条件を評価し、これを繰り返す。このとき条件の評価をデータウェアハウスに対する SQL で実行すると、毎回SQLを実行することになり処理効率が悪い。データウェアハウスが正規化された設計であれば、SQL の内容によってはテーブルジョインが発生し高速なレスポンスを実現できない可能性がある。また極力正規化を排除した設計であった場合でも、データの冗長性からSQLの対象レコード数やレコード長が大きくなりがちであるので、同様に高速なレスポンスを実現できない可能性がある。従ってリアルタイム性を実現するには、最適なデータウェアハウス・マネジメント・システム（DWMS）を採用することが重要である。DWMS には３つのタイプが存在する [7]。

- トランザクションデータ・モデル

小売業における POS データに代表される大量のトランザクションデータを取り扱うモデルである。リレーショナル・モデルを基本として処理の高速化にスター型モデル（スキーマ）を適用することが多い。

- アグリゲート・モデル

集計データを扱い多角的な分析を可能とするモデルである。多次元（マルチ・ディメンジョナル）モデルともいう。

- ディメンジョナル・モデル

トランザクションデータではなくディメンジョナル（マスタ、属性）データを取り扱うモデルである。顧客分析に適用することが多い。

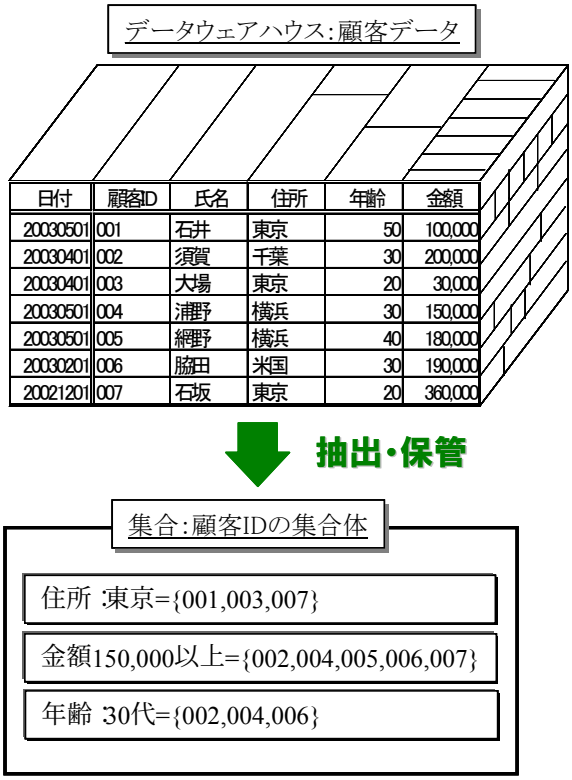
一般のデータマイニングではトランザクションデータ･モデルを利用することが多いが、リアルタイム・データマイニングではデータの抽出・分析・意思決定の時間を最大限に短縮することが目的であるので、SQL に依存する仕組みではこの目的を満たせない。本稿で提案するリアルタイム・データマイニングは主として顧客の属性（ディメンジョナル）データを分析対象とする顧客分析モデルであり、time cube データモデルというディメンジョナル・モデル型の DWMS を採用する。また、処理の高速化を図るために、集合概念を取り入れた仕組みを利用する。

2-2. 集合概念の利用

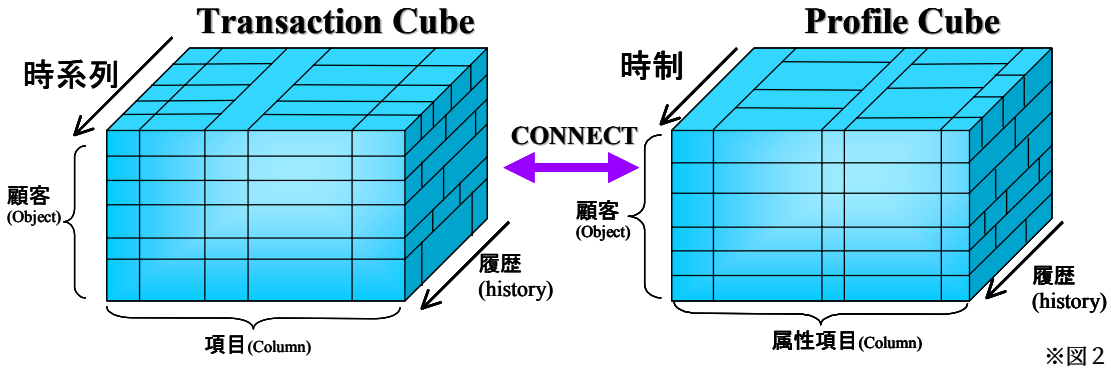
集合概念を利用する方法においては、データウェアハウスから予め様々な条件で分類（セグメント）したデータを集合（サブセット）として、データウェアハウスの主たるテーブルとは別に管理しておく。例えば次のように示すことができる（※図１）。

- データウェアハウスの顧客データに管理されているデータ:顧客 ID、氏名、住所、年齢、昨年度購買金額、等
- 集合の管理
  - 住所：東京={001,003,007}
  - 年齢：30 代={002,004,006}
  - 金額 150,000 以上={002,004,005,006,007}
    ※ 各要素は顧客 ID となる

つまり予めデータの分類を想定して、バッチ処理で事前にその分類に応じたクエリを実行し、結果を集合（主たるテーブルのサブセット）として保管しておく。このとき集合は必要以上の情報をもつ必要はない。このケースにおいては集合で管理すべき情報は顧客 ID のみでよい。集合はある条件で分類された顧客 ID だけを管理することから、冗長性を持たず少ないリソースでデータを管理することができる。よってデータマイニング担当者がリアルタイムに様々な条件で集合を評価するとき、冗長性を持たない省リソースの集合を用



※図１



※図２

いることで、毎回クエリを実行するのに比べて高速に結果を得ることができる。

データマイニングを行う際は繰り返し集合を評価・演算して結果を求める。毎回の評価の際には主に演算の結果のレコード件数（該当する顧客の数）や全体に対する比率（顧客構成比）が、分析（演算）結果要求されることが多い。件数や比率は顧客 ID だけをもつ集合から導出可能であるので、各集合が明細の項目を管理している必要はない。顧客データの明細を見ることは、最終的にこの集合演算の結果が必要なのだという時点（例：実際に DM を印刷・発送するとき顧客氏名や住所が必要）であるので、データウェアハウスの主たるテーブルにはこの時点のみ問合せを行えばよい。

2-3. time cube モデル

第一著者が提唱した time cube モデルは、次に挙げる特徴をもつデータモデルである（図２）。

- 従来の RDBMS の２次元のテーブルに時間の概念を加えた３次元の論理モデル（Cube）でデータを管理する。Cube には Transaction Cube と Profile Cube の異なる２つのキューブが定義可能である。
- オブジェクト（管理対象）とカラム（属性項目）、およびオブジェクトのヒストリ（管理対象の履歴）から構成される３次元モデルである。
- トランザクションデータを時系列に管理する Transaction Cube モデルをもつ。
- マスタデータを時制で管理する Profile Cube モデルをもつ。
- Transaction Cube と Profile Cube は時系列に沿って自動的に結合可能である（利用者が時間概念を明示的に指示して結合を行う必要がない）。
- time cube モデルでは問合せ結果をオブジェクト ID の集合として保管でき、その後いろいろな集合演算による検索が可能である。

株式会社ビーコン IT は、time cube モデルに基づいた DWMS として TimeCube という製品を開発している。TimeCube は時系列的にデータを管理し、かつ集合概念を採用していることから、リアルタイム･データマイニングを実現するシステムとして機能する。

TimeCube による顧客データウェアハウスの構築においては、顧客マスタ情報（定性データ）を Profile Cube に、購買履歴情報（定量データ）を Transaction Cube に、それぞれ時系列に蓄積・管理する設計が一般的である。TimeCube は DWMS であることから、データの蓄積はトランザクション処理による更新ではなく、一括ロード方式による更新･追加を行っていく。つまりオブジェクトの履歴が追加されていくことになる。このとき TimeCube はオブジェクト、およびヒストリにユニークな番号（ObjectID、HistoryID）を割り当てる。TimeCube で管理される集合は、このユニークな番号の集合である。

2.4 時系列データと集合管理

TimeCube は時系列でデータを管理するため、集合も時系列的に管理することが可能である。時系列的な集合の管理は次の３つに分類することができる。

- TimeCube のデータウェアハウス内に管理されている過去から現在までのデータに対しクエリを実行した結果の集合

例１：「99 年に東京に住んでいた顧客」
例２：「今年１～３月に５万円以上購買した顧客」

- TimeCube のデータウェアハウス内に管理されているが現在は条件に一致しないものを含む集合

例３：99 年当時に「現在横浜に住んでいる顧客」という条件で作成した集合。このなかには、現在（03 年時点）横浜に住んでいない顧

客も含まれている可能性がある。

- TimeCube データウェアハウス内に現在管理されていないトランザクションデータで作成された集合。

例４：90 年に「本年１年で 10 万円以上購買した顧客」という条件で作成した集合。現在 90 年当時のトランザクションデータが、TimeCube データウェアハウス内に存在していなくとも、この集合と顧客マスタが存在していれば管理可能である。

これらの例が示すとおり、TimeCube では主たるデータを Cube で管理するのとは別に、集合という概念が存在し、これらはセット DB に管理される。例１～２が示すように任意の時点での条件で集合を作成することができるとともに、例３が示すように過去に作成したものを保管しておくことで再度データウェアハウスに問合せを行わなくとも過去時点での集合を再利用することができる。さらに例４が示すように過去のトランザクションデータを削除した場合においてもマスタデータだけ一貫性を保って保管しておくことで、削除した顧客データについても過去の状態の集合を再利用することができる。

リアルタイム・データマイニングには、過去から現在に至るまでの膨大なデータの効率的な管理と高速なレスポンスが要求されるが、データの実体を Cube としてもつ一方、集合概念を用い、実体とは切り離したセット DB で管理することで柔軟性のあるデータ管理が実現できる。

３．集合概念の活用と可視化

リアルタイム･データマイニングを行うときに time cube モデルによるデータウェアハウスを利用することで集合概念による効率的なデータ管理と高速なレスポンスでの分析が可能となる。しかしながらデータマイニングの担当者にとっては単にデータ管理やアクセス効率のみを追及しても企業の利益向上や ROI 最大化にはつながらない。企業担当者にとってのデータマイニングに対する最大の要求は、利益の源になる顧客を発掘することである。そのためには仮説をたてて検証を繰り返すことで予測の精度を向上したり、購買してくれそうな顧客はどのような属性・要因をもっているのかについての相関性を把握することが必要である。顧客ごと、あるいは特定の顧客集団（集合）ごとに予測率や相関度を算出して分析することも必要であるが、複雑な条件になればなるほど数表やリストでは把握しにくくなる。このような問題を解決するためには、相関性を視覚化、可視化するのがよい。視覚化、可視化によりデータマイニング担当者はより直感的に意思決定を行うことができるようになる。

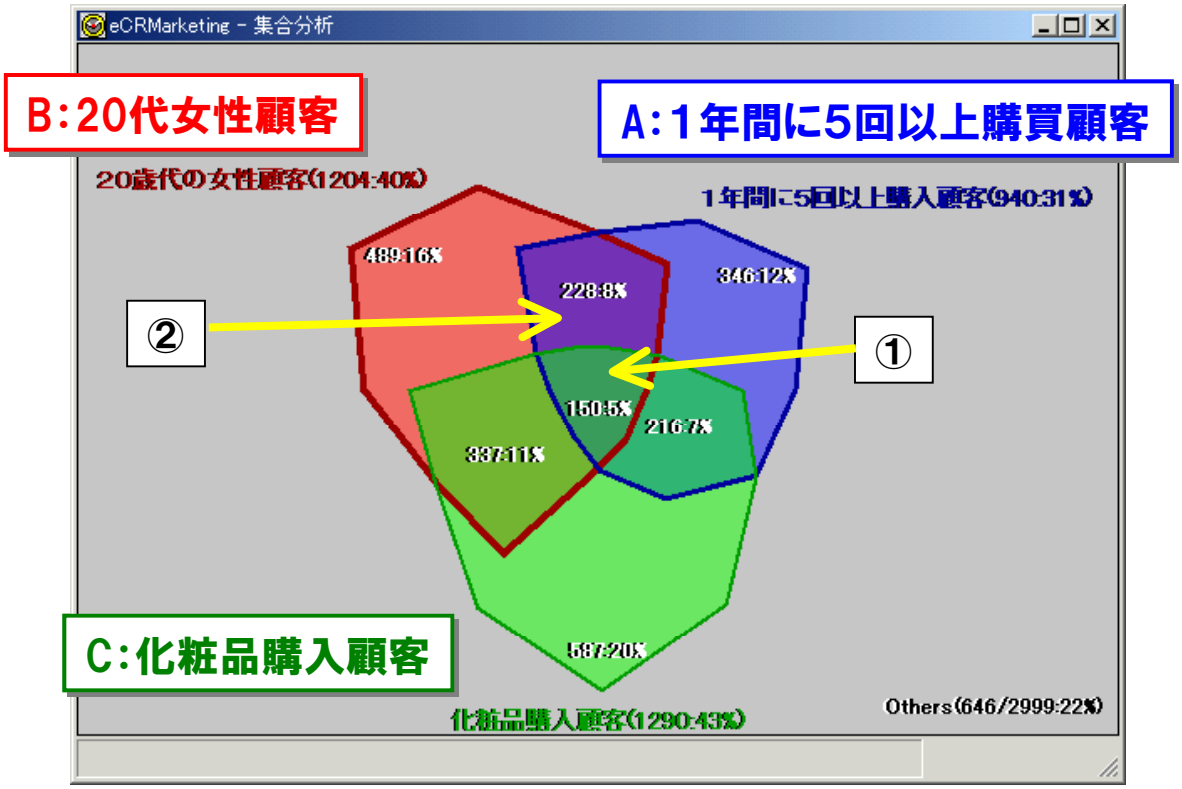
TimeCube はデータウェアハウスに集合概念を応用した機能をもつが、あわせて集合概念による分析の可視化機能も提供する（製品名：Targeting Palette）。可視化機能には全く新しいインターフェイスが４種類あり、それぞれを KaleiDiagram、Cosmos、Float、Mosaic と名付けた。これらはすべて特許申請が行われ、一部は既に特許取得が行われている。

- KaleiDiagram（可視化機能その１）

ベン図の表現方法を採用したものである。KaleiDiagram ではベン図で表現される各集合に含まれる要素数とその面積が正確な比率で表現される特徴を備えている。多角形で表現される技術は特許を取得している。

３種類の集合を、ベン図を応用した視覚的な画面で重ねることができ、同時に８つの顧客セグメントを表現することができる。次ページの図３の例では A,B,C の集合の AND、OR、XOR を含め８種類の顧客セグメントが表現されている。従来の SQL や OLAP ツールでは、ふつういくつかの条件を AND 条件で指定してクエリを実行すると、①のみが抽出される。②は過去１年以内で５回以上購入していて（A）、かつ 20 代の女性（B）だが、化粧品を購入したことがないセグメントになる。このような複雑な条件や「購入していない」といった購買データのトランザクションの存在しない条件で検索や分析を行う場合には、集合演算とその可視化が大きな効果をもたらす。

KaleiDiagram を始めとする各可視化機能は、前述のとおり各集合を表現する部分の面積が正確に顧客数（集合の要素数）に比例して表現されるため、同時に表現される８つの集合の面積を見比べることで（要素数と全要素数に対する割合も表示することができる）、どの条件の集合がもっとも多いのか、あるいは少ないのかなどを把握し、集合間の相関性を把握することも可能である。

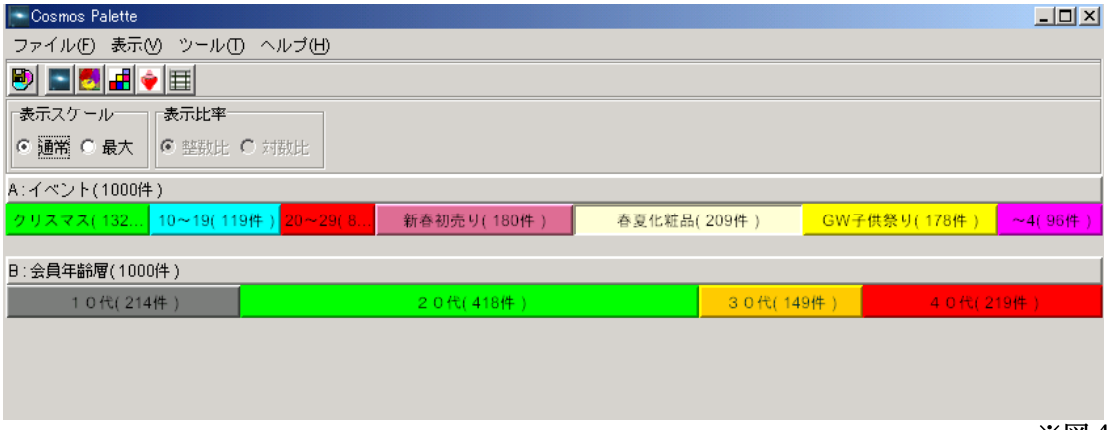


※図３

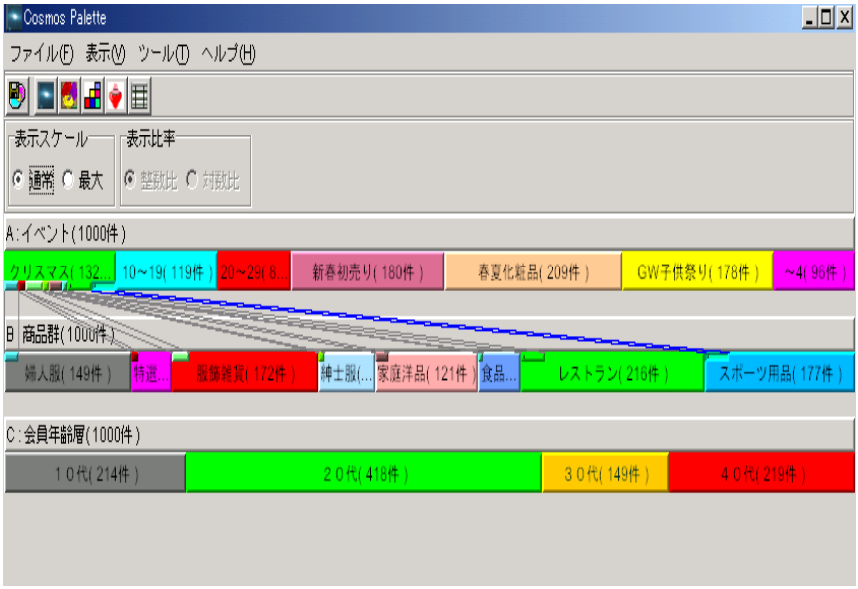
他の DWMS で同様の結果を得ようとすると、KaleiDiagram と比較するならば各条件の SQL クエリを８回実行しなければならないが、KaleiDiagram では３つの集合を操作するだけである。しかも集合演算時には、セット DB へのみアクセスするだけで、実際のデータベースにアクセスをしないため、高速に分析ができる。したがってこのような集合概念を応用した分析手法こそがリアルタイム・データマイニングである。

図３の例において仮説として、②の集合にダイレクトメールを送付し顧客に来店を促す。ふつう、このダイレクトメール送付とその後の顧客購買行動の結果を検証したいとき、顧客データベースにダイレクトメール送付済みフラグ等の新たな情報を付加する、あるいは送付者のデータを新しいテーブルとして作成し、後日購買トランザクション等と突き合わせをするのが一般的である。

これに対し、TimeCube では②を新しい集合としてセット DB に名前を付けリアルタイムに保存しておく。検証は、後で購買をした顧客の集合を作成し、それぞれを重ね合わせるだけで（KaleiDiagram による集合演算）、容易に検証ができる。TimeCube を用いない検証方法では、事前にダイレクトメール送付フラグの追加や新しいテーブルを作成することが必要であるが、TimeCube では集合演算の結果、任意の集合を別の集合としてその場でリアルタイムに保存できる。また結果検証に利用す



※図４



※図５

る集合は誰でも使用できるので（集合の共有機能)、利用者にとって自由度が高い画期的な手法である

● Cosmos（可視化機能その２）

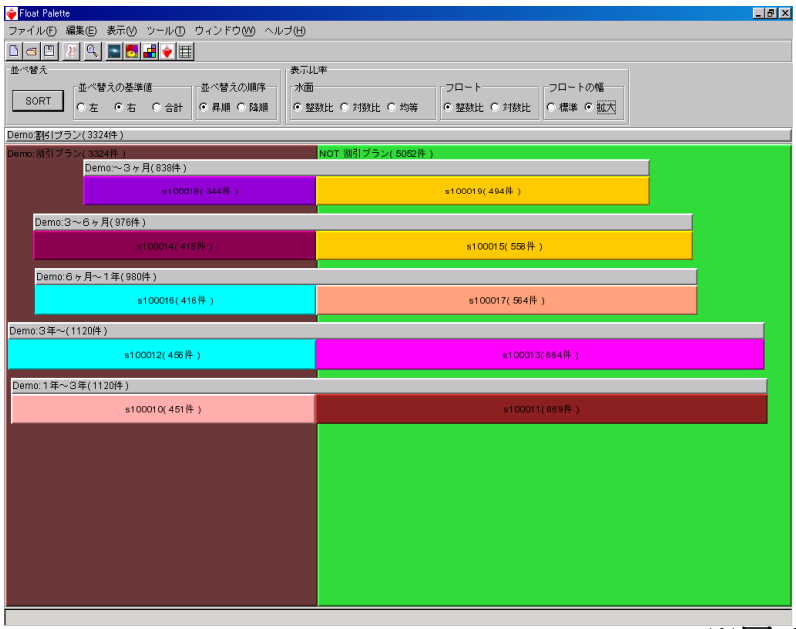
任意の属性の集合からなる和集合を一本の棒グラフで表現する（図４)。一本の棒グラフの面積は、それを構成する集合の要素数に正確に比例して分布表現（面積）される。複数の属性の和集合を一度に表示することができるので、ある属性に対する分布を俯瞰的に分析する場合や、多種類の切り口から全体の構成比率を把握する場合に有効な表現方法である。図５では集合間の相互の相関関係を多対多の関係で表現している。この例では、昨年のクリスマスイベントに来店した顧客の購買商品群と年代層の相関関係となる。

Cosmos では、和集合の下層の集合も階層構造を保持することが可能で、その階層数には制限がなく集合のドリルダウンにより詳細な集合について深堀りすることも可能である。そのほかにも各集合間を結び付けて、各集合間での面積比を補足的に表示することも可能である（図５)。

● Float（可視化機能その３）

ウィンドウ内をまず任意の集合の真偽で２つに分割して表現する。このときそれぞれの集合が表現される面積は各集合の要素数に比例する。次に任意の属性からなる和集合を、各集合ごとに最初に表現した集合の真偽と論理演算を行った結果で棒グラフを用いて表示する。左右に分かれる各集合の論理演算の結果（真と偽の部分）もその要素数に応じて面積が比例する（図６)。

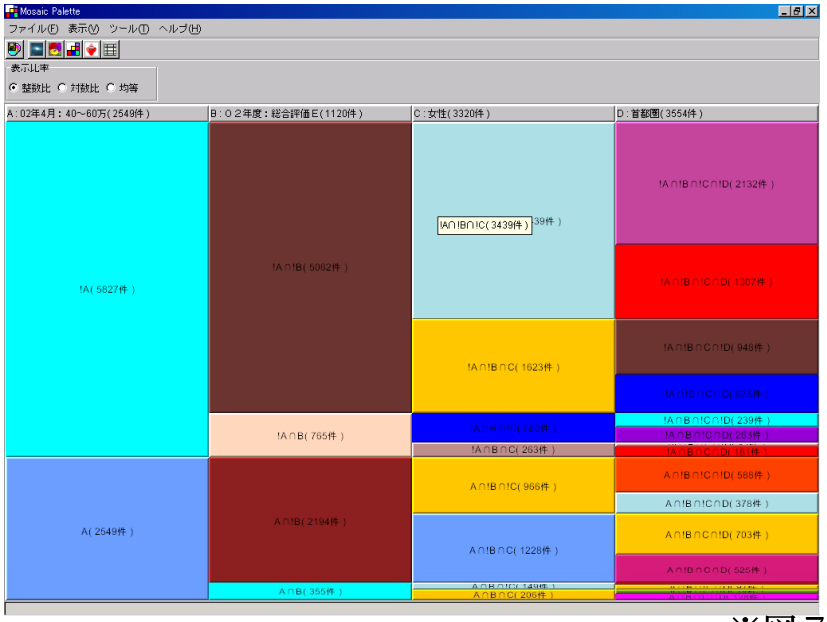
また、Float は各集合の要素数（左側、右側、全体）で棒グラフ部分を並べ替えることも可能である。



※図６

● Mosaic（可視化機能その４）

集合概念を応用した決定木による表現方法である。左から順に任意の集合を評価し、その真偽ごとに次の集合が論理演算された結果が表現される。ここでも各集合が表現される面積は、各集合の要素数に比例して表現される



※図７

（図７）。

これら４つの可視化手法は単に集合を表現するだけではなく、利用者は任意の集合をクリックすることで該当する集合の明細データを画面上にリストアップしたり、CSV 形式のファイルとして保存して利用することができる。

また、各画面間で集合をドラッグアンドドロップして評価することもできる。たとえば、Cosmos 画面で選択した任意の集合（演算結果）と Float 画面で選択した任意の集合（演算結果）を KaleiDiagram 画面にそれぞれドラッグアンドドロップし、新たに集合演算することもできる。

４．むすび

本稿では、意思決定までに時間を要する、あるいは情報（データ）間の相関性の把握が利用者にとって必ずしも容易ではない、といった従来のデータマイニングに内在する課題を解決するものとして、リアルタイム・データマイニング手法を提案した。

ここで述べた time cube モデルに基づく集合概念の採用と４つの表現による集合の可視化の具体的な方法は、従来のリレーショナルモデルと SQL による実現されているデータマイニングの課題である利用の難しさを除去するとともに、非リアルタイム性という欠点を克服した。

株式会社ビーコン IT が開発している TimeCube、および Targeting Palette は、time cube モデルを製品設計・開発の中心に据え、検索や分析の容易性、高速化を実現するために集合概念を DWMS に取り入れている。さらに相関関係の可視化を実現するために本稿であげて４つの手法を実現している。

DWMS として TimeCube を、また相関関係可視化の手法として Targeting Palette を利用することにより、１章で定義したリアルタイム・データマイニングを実現することができる。当該製品は既に商品化され、国内 60 社以上で利用されており、輸出も試みている。

参考文献

1) 石井義興 “SOIR (Set Oriented Information Retrieval) ランゲージ” 第 20 回プログラミング・シンポジウム 1979 年 1 月
2) 石井義興 “会話型情報検索言語 - SOIR -“ ソフトウェア流通 No. 4, 1980 年 7 月
3) 横田一正、石井義興 “会話型情報検索言語 SOAR” bit 1984 年 1 月号
4) 石井義興 “Three-Dimensional DBMS” 情報処理学会 ﾃﾞｰﾀﾍﾞｰｽ研究会, 1989
5) N. Mohan “DWMS : Data Warehouse Management System” in Proceedings of the 22nd VLDB, 1996
6) Y. Ishii, T. Ishizaka, N. Mohan, J. Feng “TimeCube : Efficient storage, Access and Analysis of Temporal (Historical) Data” ER'98 Workshop on Spatio-Temporal Data Management, Springer LNCS 1552 P.474-483, Nov. 1998
7) 石井義興“データ・ウェアハウス” 日本経営科学研究所, 1995
8) T. Oba “Competency of Set Analysis in CRM Closed Loop Marketing” ER2001 Springer LNCS 2224 P.604-606, Nov. 2001