データウェアハウスとデータマイニングの概要

河野　浩之

1．はじめに

データウェアハウス（Data Warehouse）やデータマイニング（Data Mining：データ発掘）に関連したソフトウェアなどが増えており，その関心が高くなっている．そこで，本稿では，データウェアハウスとデータマイニング関連分野の動向を，その背景となっている研究を含めながら紹介する．

まず，2章では，データウェアハウスとデータマイニングの各々の研究背景を示す．3章では，データウェアハウスとデータマイニングの両者の枠組みを述べる．4章では，データウェアハウス構築に関わる研究動向を，データベースの視点を加えながら述べる．そして，5章をむすびとする．

2．それぞれの研究背景

データウェアハウスとデータマイニングは，複数のコミュニティ（データベースや人工知能，さらには，統計等）に関わる研究であるため，さまざまな理論と技術が重なりあっており，その方向性が把握しづらい．そこで，以下に，それぞれの研究背景を簡単に整理する．

2.1　データウェアハウスの研究背景

データベースが，企業情報システムにおいて重要な役割を果たし続けていることは，1960年代に登場したEDPS（Electronic Data Processing System：電子データ処理システム）やMIS（Management Information System：経営情報システム）などの情報システムの発展を見れば明らかだろう．また，SIS（Strategic Information System：戦略情報システム）を構築する上でも，時々刻々と変化するデータを効率的に扱うデータ処理技術が要求され，例えば，POS（Point of Sales：販売時点管理システム）のような情報処理システムが成功をみている．

しかしながら，同様にデータベースに蓄積されるデータをさまざまな角度から活用することを試みたDSS（Decision Support System：意思決定支援システム）は，必ずしも成功したとは考えがたい．これは，定型的なプロセス管理と非定型的な意思決定支援，基幹系システムと情報系システム，これら両者の相異なる要求を満たすシステム統合をうまく実現できなかったからだと言える．

しかし，データウェアハウスを「サブジェクト指向」「統合化」「時系列」「恒常性」をもつ情報システムとするInmon氏の定義以後，エンドユーザー指向のDSSとしてデータウェアハウス構築が活発化した[6]．また，ちょうどその時期は，情報システム統合化を推し進める製品の市場拡大時期と重なっている．

1．高性能計算機・大容量記憶媒体

計算機性能の向上は，GUI（Graphical User Interface）の操作性を高め，大量の時系列データを高速に処理することを可能にした．また，長期・短期で異なる要約レベルにあるデータ蓄積が，さまざまな記憶媒体の利用により可能となった．

2．ネットワーク製品

TCP/IPを中心とした標準的なネットワーク・プロトコルの普及と，ネットワークの高速化が，各種業務系システムの接続を容易にした．

3．関係データベース（RDB）製品

異なるサブジェクトに対する各種業務系システムが，RDBを用いて構築（または，再構築）された．また，RDB製品が提供する標準的なインターフェースが，システム統合化を促進した．

かわの　ひろゆき
京都大学　大学院情報学研究科システム科学専攻
〒606-8501 京都市左京区吉田本町

しかしながら，恒常的なデータを中心に扱うデータウェアハウスでは，典型的なデータベースで重要な更新操作を必要としない．また，関係データベースを用いたシステム構築が最適とも言えず，従来のデータベース（legacy databases）のデータ設計を見直すことが必要になった．そこで，これらに関係する研究について4章で触れる．

2.2 データマイニングの研究背景

データマイニングは，人工知能分野の動向に強く影響された研究である．1989年にAAAI（American Association for Artificial Intelligence）のワークショップとしてKDD（Knowledge Discovery in Databases：データベースからの知識発見）が開催され，実働するデータベースに蓄えられたデータに対して，主として機械学習で提案された決定木などの学習アルゴリズムを適用する研究が盛んになった[9]．なお，数多くの優れた研究が人工知能の分野ですでに行われていたにも関わらず，KDDワークショップが開催された背景には，トイ・プロブレム（toy problems）に代表される人工知能分野の閉塞感から抜け出すきっかけを，実データを対象としたシステム開発に求めた部分が強い．

なお，データマイニングの用語が定着するにつれて，KDDワークショップは，知識発見とデータ発掘（Knowledge Discovery & Data Mining）に関する国際会議となった．そして，「データから情報へ，さらには知識へ」という目標に向かって，人工知能だけではなく，データベースや統計学の研究領域との交流が試みられている．ただし，統計分野において人工知能の各種手法を研究する必要性があり，さらに，データ解析に伴う本質的な問題解決に向けて挑戦することは貴重な試みであるとはいうものの，「知識」を発見するための壁は厚いというトーンもある（脚注1）．

このように，データマイニングは，限定された領域の研究を指し示すというよりも，統計，データベース，パターン認識，学習，視覚化（visualization）など数多くの研究と関連しながら，知識発見という困難な問題に対してシステム構成手法を与える視点を失わずに挑戦するフレームワークと言える．

3．データウェアハウスとデータマイニングの関係

データから知識発見を行うプロセスは，以下のようにまとめられている（図1）[3]．そして，実用性の高い知識を求めるために，各プロセスのバランス良い実行，特に収集データに対する前処理と，導出されたルールに対する後処理が重要とされる．つまり，質の高いデータウェアハウスを構成することが，データマイニングの成否を決める重要な鍵となってくる．

知識発見プロセス

(a) データが生成される領域の特性の理解に努め，すでに知られている性質などに考慮してデータ収集を行い，データベース/データウェアハウスを構築する．
(b) データに対する前処理として，選択操作/サンプリング（selection/sampling）によってデータクリーニング（data cleaning）を行う．
(c) 実用的な時間で処理が可能な範囲になるようにデータの次元低減などの前処理／データ変換（preprocessing/data transformation）を行う．
(d) データマイニング・アルゴリズムを実行する．
(e) 生成されたルールに対する後処理として，求められた記述の解釈／評価（interpretation/evaluation）を行うとともに，検証（verification）を行う．
(f) 最終的に得られたルールに対する評価を用いて，知識とする．

なお，データマイニングでは市場データのような実データを対象とした知識発見過程を扱い，データ要約やデータ分析を通じて求められた知識は，データウェアハウスを用いた意思決定支援に役立つと期待される．

さらに，求められた規則性(regularity)，制約(constraint)，ルール(rule) などの「知識」に対して，理解可能性(understandability)を高める技術として，視覚化技術の援用が重視されており，さまざまなGUIの研究も活発に行われている．

また，文献[3]では，知識発見において最も重要な

(脚注1) アメリカ統計学会（American Statistical Association）に引き続いて開催されたKDD-97における，Peter J. Huber博士（University of Bayreuth）による“Keynote Address”など．なお，KDD-98はデータベースの国際会議であるVLDB（Very Large Data Base）と連続開催された．また，“http://www.kdnuggets.com/”によれば，今後は，ACM SIGKDD（Special Interest Group on Knowledge Discovery ＼& Data Mining）の会議として開催される予定である．

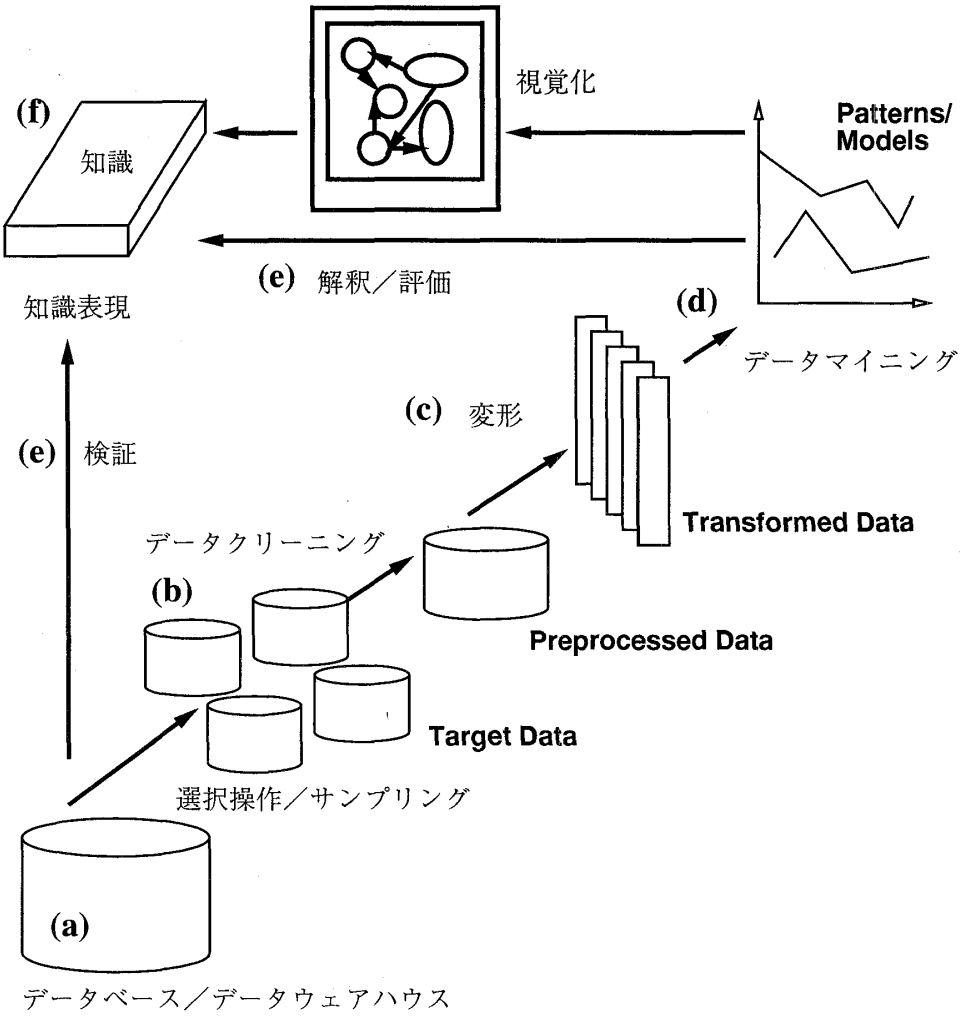

図1　知識発見プロセス

問題である知識の質の評価基準として「確実度(certainty)，妥当性(validity)，新規性(novelty)，潜在的有効性(potentially useful)，理解可能性(understandability)，興味深さ(interestingness)」からなる評価関数の構成を示し，定められた閾値を満たす知識が重要であるとしている．しかしながら，多くの研究は手順(a)～(d)を扱っているため，発見科学（Science Discovery）の研究などの展開によって，今後，知識の本質に迫ることが必要である．

4．研究動向

オペレーションズリサーチの視点から，データウェアハウスやデータマイニングをとらえる上で，例えば，Paul Gray 教授による“The New DSS：Data Warehousing, OLAP, MDD, and KDD”[4]が参考になる．そこで，講演[4]で述べられた内容にも注意を払いながら研究動向をまとめる．

4.1　データウェアハウスの研究

2章で述べたデータウェアハウス構築に関わる問題が，データベース領域の研究で全く扱われていなかったわけではない．当然ながら，データベースを用いて作成するレポートには統計処理が必要であり，高度な要約（summary）処理を要求する研究もなされた．しかし，解析的処理の必要性を認めていたにも関わらず，システム性能を極端に劣化させるおそれのある操作を，データベースの基盤部分に実装する場合には，慎重な検討が行われた．

まず，トランザクション処理性能を保証するには，単一業務を指向したシンプルなシステム設計が要求された．さらに，記憶領域に関わるコストを最小化するために，正規形に代表されるスキーマ設計の枠組みが不可欠であったと言える．そして，このような厳密なスキーマ設計は，解析処理には不利であったが，多くの業務系システムで重要なデータ挿入・更新・削除におけるトランザクション処理の一貫性を保証する上では非常に有効であった．

しかし，計算機性能の向上と記憶媒体の大容量化が，

データ要約を伴う解析処理を現実的なものとした．例えば，ジョイン操作によって非常に大きなテーブルを生成し，GroupBy操作を用いて合計・平均などの集約演算（aggregation）を行う処理が該当する．（これは，例えば，売上データを，属性$\{a_1, a_2, \ldots, a_n\}$の組合せや日付けの範囲$(d_1, d_2)$をさまざまに変えてデータ解析を行う，ドリルダウン（drill-down）やロールアップ（roll-up）と呼ばれる操作で必要とされる．）

そこで，解析的処理を実時間で行うOLAP（On-Line Analytic Processing）では，多次元データベース（MDD：Multidimensional Databases）が用いられる．なお，OLAPの実現には，関係データベースを用いたROLAPと，多次元データベースを用いたMOLAPの枠組みを比較することも多いが，両者それぞれにデータ処理上の利点がある．また，この種の解析的処理を効率よく実現するために，既存の設計手法と異なるスキーマ構成手法（star schemaや，snowflake schemaと呼ばれる）が提案されている．

なお，データキューブ（data cube）は，データウェアハウスにおけるデータ処理に対する理論を与えている．文献[5]では，問合せでひんぱんに利用されるジョイン演算を前処理してデータベースに実体化したビュー（materialized view）として格納する場合のコスト評価が行われている．そして，このような問合せ処理コストと記憶領域コストのトレードオフを考えたデータベース設計を現実のものとする上で，計算機速度の向上と記憶媒体の低価格化は非常に大きい位置を占める．

4.2 実用化を目指すデータマイニング

機械学習を含む人工知能の分野において，学習アルゴリズムが数多く提案され，計算量の面からも研究されているが，多くのアルゴリズムは，理想的（実験的）なデータに対する学習を考えており，スケーラビリティ等の問題も多い．

しかし，データマイニングでは，現実のデータから知識発見を行うアプリケーション・システム開発に有効な技術を強く求めている．そのため，「要領を得ないデータ(inconclusive data)，ノイズを含むデータ(noisy data)，疎なデータ(sparse data)」などを，効率よく処理することがデータマイニングでは重要である．

また，理論や技術面だけではなく，データウェアハウスのユーザーである企業と，その企業が関わるマーケットへのインパクトの分析も重要である．

例えば，小売業のPOSデータや，通信販売やクレジット産業の顧客の購買記録などの市場データに対する有効な手法の研究が様々な角度から行われており，相関ルールはその代表的なものである．

相関ルールは，どのようなアイテム（商品）を組み合わせて購入する場合が多いかということを示すもので，商品配置やカタログ構成などを決定する上で利用できるルールである（バスケット分析）．また，このような購買傾向のパターンを求める相関ルールは，ダイレクトメール送付を行うデータベースマーケティング（database marketing，mailshot response）においても有効である．

しかし，この種の基本的かつ典型的な分析も，非常に大規模なデータベース全体に対して，効率よくアイテムの組合せを求めることは難しい．そこで，頻繁に出現するアイテムの組合せ（ラージアイテム集合）を効率よく求めるアルゴリズムの代表例が，apriorである[1]．

ところで，この種の技術が有効である分野は，例えば，航空会社，銀行，クレジットカード会社，販売，電話，保険などであると分析されている[4]．したがって，これらのアルゴリズムを実装したKDDM（Knowledge Discovery and Data Mining tools）関連ソフトウェア市場を形成する上で非常に重要な領域である．そこで，KDD-97,98では「KDDMツールに関する大会（KDD-CUP）」が開催され，MineSet(SGI)，Enterprise Miner(SAS)などが良い成績をおさめている．その他，各種アルゴリズムの比較評価が，[10]においても試みられている．

以上，データマイニングでは，複数領域の研究を視野にいれながら，市場データや医療データなどの実データを対象とした知識発見システムを開発する研究が活発に行われてきた．また，これらの処理を通して，有用なルールを捨てない頑健なアルゴリズムが明らかにされてきた．なお，データマイニングで扱われる主要なアルゴリズムは，「相関(associations)，時系列(sequences)，分類(classifications)，クラスタリング(clusters)，予測(forecasting)」に大きく分類されている．

4.3 各種データベースとの関係

ところで，前節の最後に示した分類で同種のアルゴリズムであっても，文書データ，画像データ，地理デ

ータなどの対象となるデータによって，その応用形態は大きく異なる[2]．そこで，データベース技術の高度化[7]にも注意を払いながら，幾つかの典型的なデータベースを示し，それらに関連するデータマイニング研究について述べる．なお，詳細な文献は[8]などを参照してほしい．

●関係データベース

関係データベースでは，属性間の関数従属性が非常に重要な役割をもつ．そこで，データから従属性を効率よく求める研究がされている．その他，不正発見（fraud detection）や相関ルールを求めるアルゴリズムも実装されている．

●オブジェクト指向データベース

オブジェクト指向データベースでは，クラス階層が重要な役割を果たす．そこで，COBWEBのような概念クラスタリング（conceptual clustering）の研究を踏まえながら，自動的あるいは半自動的に概念木（classification tree）を生成する研究が行われている．

●文書データベース

文書データベースは，Webシステムなどを用いて情報共有を行う場合などに重要な位置を占める．そこで，相関ルールによる単語共起の分析や，文書クラスタリング，さらに，文書集合提示や要約記述などの文書マイニング（text mining）の研究が行われている．

●地理データベース

空間データベースは，地理情報システム（GIS：Geographic Information System）などの基盤である．なお，空間データマイニング（spatial data mining）では，クラスタリングアルゴリズムが重要な役割を果たし，CLARANS（Clustering Large Applications Based upon Randomized Search）などが提案されている．さらに，データベース特有のデータ構造（R-treeやPR-Quad tree）を考慮したクラスタリングの高速化も研究されている．しかし，計算幾何学（computational geometry）分野の研究の進展が不可欠である．

上述したように，文字や数値以外を扱うマルチメディア領域へとデータベース応用が広がるにつれて，さまざまなデータモデルやデータ構造（半構造データ等）に応じたデータマイニングが必要とされている．

さらに，システム性能面を考えた場合，データベースを構成するハードウェア面からのアプローチも不可欠であり，並列データマイニング（PDM，parallel data mining）は，処理効率を向上させる上で必要となる．

なお，データマイニングを行う問合せを，データウェアハウスに普及させるには，既存のシステムとの整合性を考えながら適切にインターフェースとなる問合せ言語を拡張することが必要である．例えば，標準的なSQLを用いて相関ルールを求める記述を与え，問合せ実行時に最適化を行う手法は，システム構成上，非常に受け入れられやすいアプローチと言える．

5．おわりに

以上，実データをより高度に活用するシステムとしてのデータウェアハウスと，知識発見システムを指向するデータマイニングの概要について述べた．なお，データウェアハウスは，データマイニングにおける研究成果を広く実践するフィールドとなっている．詳しくは，各種業務システムへと適用した例を解説した，本特集に含まれる論文を参照してほしい．

また，最後になったが，質の高い知識をデータマイニングで得るには，より質の高いデータが蓄積されるデータウェアハウス構築を促進するシステムが不可欠である．よって，質の高いデータが蓄積されるデータウェアハウスを構築する上で，ERP（Enterprise Resource Planning）[11]，CTI（Computer Telephony Integration），SCM（Supply Chain Management），SFA（Sales Force Automation）などの各種システムの活用に関わるストラテジーを，オペレーションズ・リサーチの研究を通じて明確に与えることが期待される．

謝　辞

日頃御指導頂く南山大学経営学部情報管理学科 長谷川利治教授 に深謝致します．本稿の一部は，文部省科学研究費(10780259,08244103)の研究成果を含む．

参考文献

[1] Agrawal, R. and R. Srikant：“Fast Algorithms for Mining Association Rules” Proc. of the 20th International Conference on Very Large Data Bases, Santiago, Chile, pp.487-489, 1994.

[2] Chen, M.-S., Han, J. and Yu, P. S.: "Data Mining: An Overview from a Database Perspective", IEEE Trans. on Knowledge and Data Engineering, Vol.8, No.6, pp.866-883, 1996.
[3] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R.: "Advances in Knowledge Discovery and Data Mining", AAAI/MIT Press, (1996).
[4] Gray, P.,: "The New DSS: Data Warehousing, OLAP, MDD, and KDD", HICSS-31, 1998.
[5] Harinarayan, V., Rajaraman, A. and Ullman, J. D.: "Implementing Data Cubes Efficiently", Proc. of the 1996 ACM SIGMOD, pp.205-216, 1996.
[6] Inmon, W.H. and Hackathorn, R.D., "Using the Data Warehouse", John Wiley & Sons, (1994). (藤本康秀監訳, "よくわかるデータウェアハウス活用法", インターナショナル・トムソン・パブリッシング・ジャパン, (1996).)
[7] Kambayashi, Y., Makinouchi, A., Uemura, S., Tanaka, K. and Masunaga, Y.(eds.): "Advanced Database Systems for Integration of Media and User Environments'98", Advanced database research and development series, Vol.9, World Scientific, (1998).
[8] 河野浩之: "知識発見とデータマイニング" 日本ファジィ学会誌, Vol.9, No.6, pp.851-860, 1997.
[9] Michalski, R.S., Bratko, I. and Kubat, M. (eds.),: "Machine Learning and Data Mining, Methods and Applications", John Wiley & Sons, (1998).
[10] 津本周作他: "共通データに基づく知識発見手法の比較と評価", 第12回人工知能学会全国大会論文集, pp.72-86, 1998.
[11] 安田一彦: "生産スケジューリング・ソフトウェア: 統合業務システムERPと製造実施支援システムMESとの架け橋", 生産スケジューリング・シンポジウム'97講演論文集, pp.7-12, 1997.