データベースシステム 110－3
（1996. 10. 24）

視覚的データマイニング支援方式の検討

黒川 清　　飯塚裕一　　磯部成二
{kurokawa, iizuka, isobe}@dq.isl.ntt.jp
NTT情報通信研究所
〒238-03 横須賀市武1-2356

データベース中に隠れた情報を有効に活用し、ビジネスに役立てる技術としてデータマイニングが注目されている。データマイニングは、大量のデータから意思決定に有効なパターンやルールを発見することであり、関連分野から様々なアプローチが試みられている。そのような中で、我々は、特にエンドユーザへの結果提示インタフェースに着目し、視覚的なマイニング方式の検討を行っている。本稿では、視覚的データマイニング支援方式の構想とその実現方式について述べる。

A Study on Visual Data Mining Support System

Kiyoshi Kurokawa Yuichi Iizuka Seiji Isobe

NTT Information and Communication Systems Labs.

1-2356 Take Yokosuka-Shi Kanagawa 238-03

Data mining is paid attention as a technique for introducing business success. Data mining is to find effective patterns and rules for supporting decision making processes from a large amount of data. Various approaches are attempted in some research fields. In view of convenient utilization for end-users, we pay attention to result presentation interface on the visual data mining support system. In this paper, the plan and realization of the visual data mining support system are described.

1. はじめに

コンピュータパワーの飛躍的な向上に伴い、企業は膨大なデータをデータベースに蓄積することが可能になってきた。また、インターネットをはじめとしてネットワーク経由で得られるデータも豊富になってきている。一方、企業間の市場競争は激しさを増しており、これらのデータを如何に活用してビジネスチャンスを獲得するかが最大の関心事となっている。これに応えるためのデータベース活用技術として、多次元データ分析(OLAP)やデータマイニングが注目を集めている。

データマイニングは、大量のデータから意思決定に有効なパターンやルールを発見してユーザに提示する技術を指しており、主に機械学習の成果を応用したデータベースのインタフェース技術として研究が進められてきた。これまでは、データベースへの問合せの強化や推論機構の付加、大量データ処理性能の向上といった方向の研究に重点が置かれており、処理結果の提示についてはあまり研究されていなかった。今後、ビジネス分野へのデータマイニング技術の適用を考えると、エンドユーザへの結果提示インタフェースの簡易化が重要な課題になると考えられる。

本稿では、データを視覚化表現し、直観的にパターンやルールを発見する視覚的データマイニング支援方式について述べる。本方式は、ビジネス分野のエンドユーザでも簡易に利用でき、エンドユーザの業務知識や分析スキルを最大限に活用できると考えている。以下、2章では、データマイニング技術の現状と課題について述べ、我々の目的を明確にする。3章では、その目的を達成するためのアプローチを示す。4章では、前章で説明したアプローチのうち、具体的に検討している視覚的データマイニング支援方式について述べる。5章では、視覚的データマイニング支援方式の課題と考察について述べる。最後に、6章でまとめる。

2. データマイニングの現状と我々の目的

データマイニングとは、データベースを金鉱に見立てた採掘テクニックであり、データベースからの知識発見(KDD)プロセスの1ステップであるという見方が一般的である[1]。獲得された知識は、新たなデータに対する予測／分類、現存するデータの解釈／説明、意思決定を容易にするための大規模データの内容要約、などの目的に利用される。KDDの主なプロセスは、ターゲットデータベースの生成、データクリーニング、パターン抽出／検証、頑健化、知識化であり、処理結果をフィードバックし知識を強固なものにする。データマイニングの視点からKDDプロセスを見ると、ターゲットデータベースの作成とデータクリーニングは前処理、パターン抽出抽出／検証は本処理、頑健化と知識化は後処理、と考えることができる。

主に科学技術計算分野に適用されていたデータマイニング手法を一般の業務データベースに適用する場合の問題点は以下のようにまとめられる[2][3]。

1) 結果表現
   ユーザへの結果の提示は、論理式や決定木により表現される
2) データベース更新
   随時更新されるデータベースに対する無矛盾性の保証が必要
3) 記述空間規模
   データ件数や情報量が膨大なものとなる
4) サンプリング
   サンプリングされたデータが妥当なものか確認が難しい

このように、マイニング手法の選択、結果の解釈などは、マイニングに関する知識を持つものにとっても、容易なものではない。

意思決定を行うためにデータマイニングを利用したいと思っているエンドユーザのためには、よりユーザフレンドリで柔軟なデータマイニング技術の確立が必要であり、ユーザの持つ背景知識、業務知識などの利用が課題となる。システムが自動的に知識発見を行うだけでなく、システムとユーザ間のインタラクションによる効率的な知識発見が必要となる。そのような中で、データベース利用分野での知識発見の

ためのデータベース探索として、視覚化技術の適用が試みられている[4][5]。従来、データマイニングにおける視覚化の利用は、値分散、特異点、項目間の簡単な関係などの内容を把握するために、2D-3D散布図やヒストグラムで表現されるものが多かった。このように、知識発見のために視覚化が有効に利用されているとは言い難い。

我々は、システムとユーザ間で能力を相互補完することにより、データベースからの知識発見と的確な意思決定支援ができると考えている。そこで、システムからユーザへの情報提示の方法として視覚化に着目し、それを適用した、ユーザ介在型のデータマイニング方式について検討を行っている。視覚的な表現により情報分析の着目点や切り口を際立たせた情報表現を可能にし、人間が行うべき頭脳的分析・判断を迅速かつ総合的なものとすることができる。次章では、従来の決定木による知識発見と視覚的なデータ表現による知識発見について比較考察する。

3. 視覚的データマイニングの予備検証

データマイニング結果をユーザに提示する表現形式と解釈の観点から、従来の決定木による方法[6]と視覚化表現による方法の比較を行う。表3.1に使用したデータを示す。このデータには4種の属性と2つのクラスがあり、各々の事例は"天候(outlook)"に従ってグループ分けしてある。これらの事例を属性値をもとに"開催(Play),中止(Don't Play)"のクラスに分割する方法について考える。

表3.1 サンプルデータ

outlook	temp. (F)	humidity (%)	windy	class
sunny	75	70	true	Play
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
overcast	83	78	false	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
overcast	64	65	true	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
rain	75	80	false	Play
rain	71	80	true	Don't Play

3.1 従来の決定木による方法

表3.1の事例は単一のクラスに属さないので、それらの分割を試みる。まず、"天候"を選んだ場合、その出力値は、"晴(sunny),曇(overcast),雨(rain)" の3通りある。"曇"のグループは全て"開催"のクラスからなるが、"晴"と"雨"のグループでは複数のクラスが混在する。そこで、"晴"のグループを"湿度(humidity)"の"75%"で分割し、また"雨"のグループを"強風(windy)"の"真(true),偽(false)"で分割すれば、その結果得られる各グループはすべて単一のクラスからなるようにできる。

最終的に得られた決定木を図3.1に示す。この決定木から、"開催,中止"の判断基準が知識として得られる。

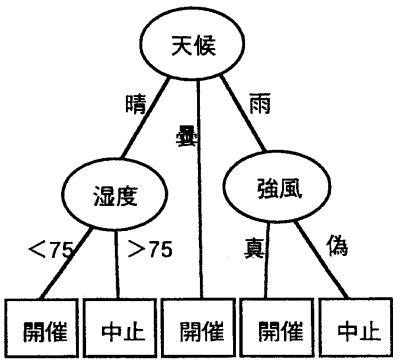

図3.1 決定木

3.2 視覚化表現による方法

次に、図形の配置、形状などに情報を付与することで多次元情報の相関を表現するような視覚化により分析した結果について述べる。

この方法では各々の事例を一つの図形で表わしている。まず、"気温"や"湿度"など、値が連続的に変化するようなデータで配置を、カテゴラ

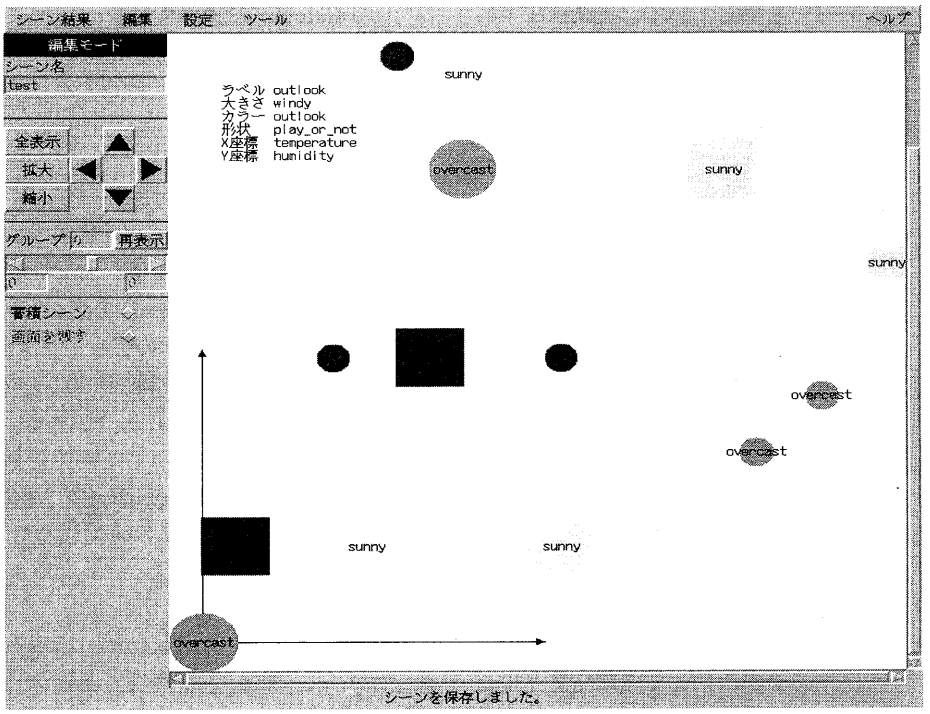

図3.2　視覚的データマイニングの結果例（その１）

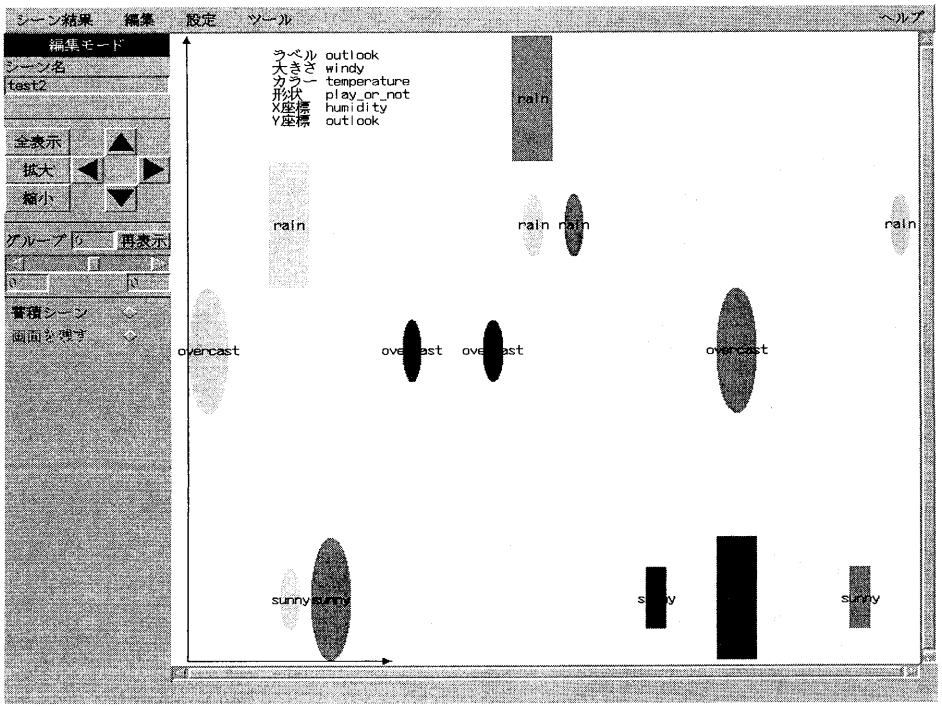

図3.3　視覚的データマイニングの結果例（その２）

イズできる質的データで色や大きさを表す方法で検討した。図3.2は、"天候"を図形のラベルとカラーで表し、"気温"と"湿度"の値をもとに配置させた多次元散布図である。この例では、図形の大きさにより"強風"の"真(大),偽(小)"を表現し、形状により"開催,中止"を表現している。楕円は"開催"、矩型は"中止"を示し、形状を見ることにより開催可否の判断ができる。図3.2から、"曇"の場合は他の条件によらず開催であり、"晴"の場合はY軸のある値を境として"開催,中止"が分割され、さらに"雨"の場合は大きさによって分割される。"天候"の各値に注目して結果を解釈することで、"開催,中止"の判断基準を知識として得ることができる。しかし、楕円、矩形の各図形が分散して配置しているため、全体を一覧して判断基準を読み取ることは難しい。この例のように次元数が少ない場合には比較的容易に解釈できるが、より多次元な情報について事例の分割を行う際には、最適な表現を如何にして得るかが重要な課題となる。

次に、配置を"天候" と"湿度"の値をもとに決定し、カラーの変化により"気温"の高低を表現したものを図3.3に示す。その他の属性（形状、大きさ）は図3.2と同じである。一瞥して図3.3の右下および左上の部分に分布する図形は、矩型であり"中止"であることが分かる。詳細に見ていくと下側に配置された図形は"晴"を表しているが、X軸のある値を境に"開催,中止"が分割される。また、上側に配置された図形は"雨"を表し、X軸の値や色、すなわち"湿度"や"気温"では分割できないが、大きさに注目すると大きい図形が"中止"を表す矩型であり、"強風"によって分割が可能であることが分かる。"雨"で"湿度"が高い事例において中止になる可能性があることも考えられるが（図3.3の右上に矩型の図形が配置される可能性）、"天候"や"湿度"、"強風"の値は互いに独立であるとは考え難い。この例では、湿度が高くしかも強風でない事例は存在しないことが分かる。

以上のように、決定木と同様な知識を視覚的に表現することができるため、特に図3.3における"天候"や"湿度"などの注目点を捉えることができれば直感的に理解可能な表現、つまり"開催,中止"のクラス分けが一瞥して分かり、視覚化による知識発見が有効であると考えられる。視覚化による表現と決定木による結果との対応関係は、形状が"開催,中止"を表すクラス分けに、配置や大きさなどが決定木の各頂点に相当する。

4. 視覚的データマイニング支援方式の検討

2章で述べた視覚的データマイニング支援方式の実現に際し、我々が研究開発した、データベース情報ビジュアル化環境の適用を検討する。

4.1 INFOVISER

データベース情報ビジュアル化環境(INFOVISER)[7],[8],[9]は、文字数値の情報集合をノード型やライン型などのオブジェクトモデルとして表示する、ノードラインビューモデル[10]に基づいて図形表現する。INFOVISERにおける視覚化の中心的な機能は、実体抽出機能、情報変換機能、結果表示機能であり、各々の機能についてGUIを通したユーザ定義機能をもつ。

1) 実体抽出機能
   図形表示の対象データをデータベースやファイルから取り出し、組合せることにより、オブジェクトモデルとして取り込む
2) 情報変換機能
   実体抽出機能により取り込まれたオブジェクトモデルを、画面オブジェクト群に変換・生成する
3) 結果表示機能
   生成した画面オブジェクトをユーザ定義にもとづいて画面に表示するとともに、画面オブジェクトの直接操作を可能とするインタフェースを提供する
4) ユーザ定義機能
   GUI画面を通じ上記3機能の処理定義を行うものであり、ユーザの視点変更を容易にする

4.2 視覚的データマイニングの機能モデル

我々が検討している、ユーザとのインタラクションを考慮した、視覚的データマイニングの

機能モデルを図4.1に示す。

抽出：分析対象となるデータ群を取り出す

変換：文字数値情報から図形情報へのメディア変換を行う

表示：結果を表示する

仮説生成：抽出や変換を行う際の仮説を生成する

分析：データ集合を分析し特徴を抽出する

探索：仮説にもとづいて必要となるデータの抽出を支援する

予測：判断するために有効と考えられる視覚化表現を支援する

INFOVISERでは図4.1の抽出、変換が上述の1),2) に相当し、ユーザが背景知識をもとに生成した仮説に基づいてGUIを通して定義する。したがって、ユーザが満足するビジュアル化表現が得られるか否かはユーザ自らによる定義によって左右される。3章で述べたように、注目点が明確に認識されなければ大量データに埋もれた有意な情報を知識として発見することは難しいので、対象データの抽出や図形変換などを行う際に指標となる仮説生成および探索や分析、予測などの機能が重要となる。このような仮説生成から、システムによる情報提示までを機能として付加することにより、注目点が定かではない非定型な分析が有効にできると考えられる。

4.3 仮説生成

図形変換のための仮説生成に注目し、特徴に基づく変換定義の生成（仮説生成から変換までの機能）について検討する。INFOVISERに対して、ユーザが任意に利用できる機能として、分析機能、仮説生成支援機能を備えた、分析·定義支援ユニットを付加した視覚的データマイニング支援方式を提案する。図4.2に構成図を示す。

a) 分析機能

分析対象データの種別（質的／量的）や目的（相関の抽出、クラスタリングなど）により分析手法を選択し、結果として特徴を抽出する

b) 仮説生成機能

分析機能により抽出された特徴を有効に提示するために、分析対象データと図形の対応付けを行う

分析機能における特徴把握の方法としては、以下を考えている。

1) 値分析

統計処理などを用いて、データ集合から特異

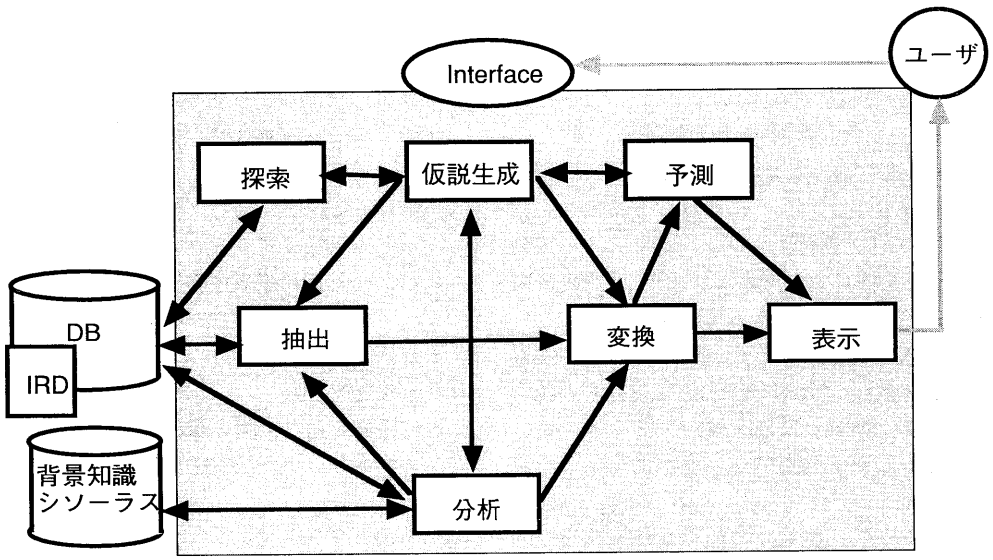

図4.1 視覚的データマイニング支援の機能モデル

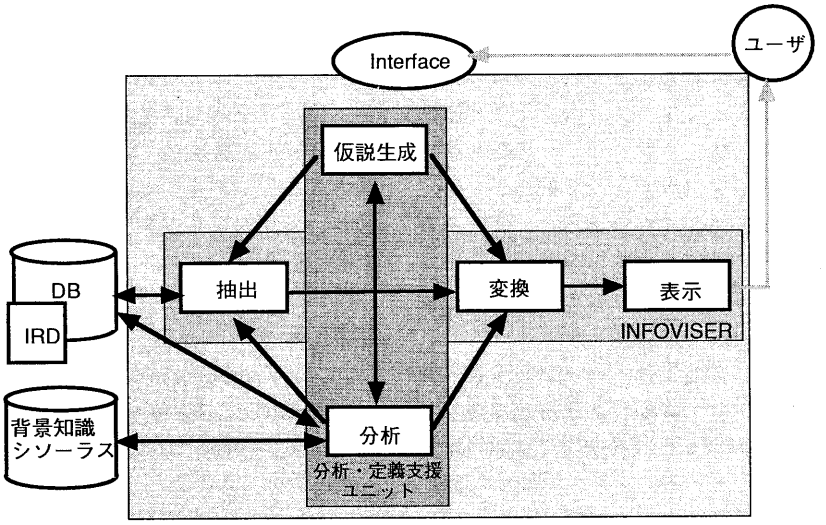

図4.2　INFOVISERを利用した視覚的データマイニング支援方式の機能構成図

点や相関関係などを導き出す

2) 意味分析

IRD(Information Resource Dictionaly)などを用いてデータの表現・意味内容を表すデータ（メタデータ）を参照し、背景知識やシソーラスを用い解釈を加えることにより特徴を得る

表示結果の検証は人間が行い、必要があれば、意思決定者の意図に沿った図形表現を得るために定義のカスタマイズをおこなう。以上の機能を実現し、システムから分析対象データの特徴をもとにした図形表示を提案することで、ユーザのもつ背景知識や業務知識などを有効に利用した、より迅速な分析が可能となると考えられる。

5. 今後の課題と考察

これまで述べてきたことから、視覚的データマイニング支援方式と統計や機械学習といった従来のデータマイニングの長短を比較すると以下のようにまとめられる。従来の方式には、

・対象データに制約がある
・事前に情報を必要とする
・ｱﾙｺﾞﾘｽﾞﾑが複雑になると処理が遅くなる
・簡単に仮説を導きだせない
・結果の解釈に専門知識を必要とする

という問題がある。一方、視覚的データマイニングでは、

・多次元データでは軸の組合せが多くなり画面数が膨大になる
・大量データでは画面に表示される図形が多くなり識別が困難になる
・境界領域の値への予測が不明確になる

という問題がある。以上から、下記の３点を今後の課題として検討する必要がある。

5.1 ハイブリッド化

上記で述べた両者の欠点は、改善の必要はあるものの本質的な特徴であることから、改善には限界が予想される。従って、この問題を解決するために、データマイニングプロセスの各段階で、両者をハイブリッド的に有効利用する方式の確立が重要といえる。例えば、視覚的データマイニングの欠点である、判断に必要な情報が無いことに対しては、統計演算結果を結果画面の背景に重畳表示する等はこれに相当する。

5.2 多次元／大量の画面／図形

INFOVISERのような多次元データの次元を図

形の性質に対応させる方式では、次元の数が多くなると、一図形に対応できないために多くの画面に分割することになる。また、次元と図形の性質（配置、形状、カラーなど）の組合せによって多数の画面を生成できる可能性がある。この問題に対しては、4章で述べた、データの統計的特徴やメタデータの特性を利用することにより、組合せ方法を自動生成する方式が有効と思われる。

また、大量データの場合、科学分野の連続的数値データのボリュームビジュアル化は別として、INFOVISERのような図形表現では数百オブジェクトを超えると通常のディスプレイ1画面では認識が困難になる。この問題に対しては、データ分析の視点に応じた画面表示方法の工夫によって対応する必要がある。具体的には、一部を詳細に識別性良く見るためにはフィシュアイ等の部分拡大表示方法が、大量図形の表示には重複排除の図形配置アルゴリズムや近接図形の集約化等の方法が考えられる。

5.3 意味・パターンマッチング

視覚的データマイニングは、データさえあれば誰でも気軽に使用できるという特徴があるが、ユーザには分析対象データに関係する背景知識をフルに活用して、注意深く図形パターンを読み取ることが求められる。従って、有効なパターンが発見できるか否かは、ユーザの特性に依存する部分もある。この問題に対しては、図形の配置や形状のパターンが意味を持つ時、意味とパターンが強い関係を持つことがあるという仮説が成り立てば、これを利用して、図形表現とその表現が意味している特徴のガイダンスを表示することが可能になる。しかし、この課題は多くの経験を積み重ねることが必要であり、前記の仮説の成立を確認し、意味的なパターン分析アルゴリズムの実装に至るまで、長期に渡る課題として取り組む必要がある。

6. おわりに

本稿では、ユーザフレンドリな結果表示インタフェースを持つ、視覚的なデータマイニングの支援方式を提案し、INFOVISERを利用した実現方式について述べた。今後は、課題の解決を図り、視覚的なデータマイニング支援方式の確立を目指す。

参考文献

[1] U. Fayyad, E. Simoudis: "Knowledge Discoverty in Databases", Tutorial Notes, 14th International Joint Conference on Artificial Intelligence(IJCAI-95), 1995.

[2] J. Han: "Data Mining Techniques", Tutorial Notes, ACM-SIGMOD'96, 1996.

[3] 河野, 西尾, J. Han: "データベースからの知識獲得技術", 人工知能学会誌, Vol. 10, No. 1, pp.38-44, 1995.

[4] D. A. Keim: "Databases and Visualization" Tutorial Notes, ACM-SIGMOD'96, 1996.

[5] S. G. Eick, D. E. Fyock: "Visualizing Corporate Data", AT&T Technical Journal, Vol.75, No.1, pp. 74-86, 1996.

[6] J. R. Quinlan(訳:古川): "AIによるデータ解析", トッパン, pp.17-32, 1995.

[7] 石垣、磯部: "データベースのビジュアル化ツール",経営システムVol.5 No.3･4, 1995.

[8] K.Kurokawa, S. Isobe, H. Shiohara: "Information Visualization Environment for Character-based Database Systems", Proceedings of The First International Conference on Visual Information Systems, pp.38-47, 1996.

[9] 磯部、黒川、塩原: "DB情報ビジュアル化技術", NTT R&D, Vol. 45, No. 1, 1996.

[10] 黒川、磯部、塩原、鬼塚: "情報可視化のためのﾃﾞｰﾀﾋﾞｼﾞｭｱﾙ化ﾓﾃﾞﾙ", 情報処理学会研究報告 96-HI-65, Vol.96, No.21, pp.51-56, 1996.