データベースシステム 110－2
（1996．10．24）

データマイニングサーバKnodiasにおける属性情報利用方式の検討

†和田 信義、†田中 秀俊、†白石 将、†石川 洋、†石井 篤、
†安田 智、†小幡 康、‡杉崎 元、‡三石 彰純、†和田 雄次

†三菱電機（株）情報技術総合研究所
‡三菱電機東部コンピュータシステム（株）

データマイニングサーバの研究について報告する。現在開発中のシステムKnodiasは、適用分野における業務知識を相関ルール発見の際のバイアスとして用いることを特長の一つとする。業務知識は多様であるが、今回はルール形式のものと、オントロジと呼ばれる用語概念辞書とを用いた知識発見について述べる。ルール形式のものは、例えば有益な知識を優先的に抽出するフィルタとして用いることができる。オントロジは、そこに記述された用語階層と属性情報を用いて、外部にて発見された知識を同化させたり、適切な切り口から知識を作ることにも利用でき、また頻度情報を付すことによって、知識発見の効率向上に用いることもできる。

An Approach for Utilizing Attribute Information in Data Mining Server, Knodias.

†Nobuyoshi Wada, †Hidetoshi Tanaka, †Masashi Shiraishi, †Hiroshi Ishikawa, †Atsushi Ishii
†Satoshi Yasuda, †Yasushi Obata, ‡Gen Sugizaki, ‡Akizumi Mitsuishi, †Yuji Wada

†Information Technology R&D Center, Mitsubishi Electric Corporation
‡Mitsubishi Electric Computer System (Tokyo) Corporation

One of the features of Knodias, our data mining server under development, is the functionality to utilize the knowledge of the application domain as bias in mining association rules. The ways of applying two types of knowledge, i.e., rules and ontology, are presented in this paper. Both refer to the attributes of items in the domain. Rules enable to filter out association rules of less use. Ontology allows externally mined rules to assimilate into the target domain, with hierarchical and attribute information, as well as gives the appropriate level of grain size for each association rule. If the count of emergence for each item is added as an attribute information to it, the discovery process can be more efficient.

1. はじめに

データマイニングにおいて、対象となるデータベースに格納される項目に関する属性情報を用いることで、 (1) 前処理、後処理の手段が豊富になる。
(2) 相関ルール生成を高速化できる可能性がある。
といったメリットがある。このアプローチを、ここでは、属性指向のマイニングと呼ぶことにする。本報告では、属性指向のマイニングの効果について検討した結果を報告する。

2. 研究の動機

・理想のデータマイニングとは因果関係まで求めるもの

現在のデータマイニングでは、表面的に成立する規則性を機械的に抽出し、その事実が示唆する因果関係は、ユーザが推定して付すこととなる。いわば、規則性を結論とする物語を創造するのであり、この物語があってはじめて規則性に説得力が加わる。理想上のデータマイニングとは、こうした因果関係の説明まで出力するものであると考える。一方、データベースに格納されているデータ単体では、この物語（あるいは因果関係）を生成するに必要な情報を含んではいない。背景知識と呼ぶべき付帯情報を外部から与える必要がある。その最も原始的な形が、データとして格納されている項目の内容を属性値の列挙の形で表現したものである。これが、各領域で共有されている概念の構造を反映したものであるとき、これをオントロジ[6]と呼ぶことができると考える。

・属性を用いて発見の品質を高めたい

相関ルール生成の代表的アルゴリズム Apriori[3]では、データ中の項目名だけを見てマイニングするのが基本だが、結果を説明するのはむしろその属性値間の共通性であったり、そこに記された機能概念間に成立する依存関係の場合もあると考えられる。また、マイニングの結果得られた知識から不要なものを捨てる際、自明、あるいは偶然の一致などの判断をするが、そこでは、物語性がない、もしくは予測可能なので陳腐、といった評価がなされていると考えることができる。こうした理由から、マイニング結果を説明するのに属性情報を利用すること、さらには属性情報を含めてマイニングすること、とを検討する必要があると考えた。

・個人の水準の背景知識も統合したい

さて、一方で、オントロジの属性項目が共通なものについては、属性値を表形式で整理しうるので、データ本体が関係表で表現されているならそれと join することで、属性を含んだ形のマイニングへの単純な拡張が可能である。しかしここで重要なのは、データ本体は一般には既存で共有されているものであり、オントロジはユーザが必要に応じて作成する点である。属性情報を利用したマイニングを行うためには、論理的空間的に遠隔での、共有データと局所的属性情報を統合利用する機構が必要である。

さて、こうした性質を有する属性指向のマイニングを実現する方法を、現在研究開発中のデータマイニングサーバ Knodias[5]を前提に考察した。Knodias はデータマイニングを目的とする一連のシステムの総称である。属性指向マイニングの議論をする便宜上、まず Knodias の全体のコンセプトを次節で説明し、その後、本題の属性指向のマイニングについて報告することにする。

※本論文で用いる用語と概念の定義など

本論文で『属性』とは、データが指し示すオブジェクトを表現する属性であり、例えば『バター』というデータでは、製造者や、製品カテゴリ、などがあげられる。また、オントロジとデータ辞書の違いは、オントロジが各領域で共有される意味辞書であるのに対し、データ辞書は普通は意味管理まで含まず、またデータベースごとに定義されるのものという点である。データ辞書の内容に加え、意味情報を補う形でオントロジを用いる。今回オントロジを持ち出した背景には、ＣＡＬＳで提唱されているように、データを共有化する必要が社会的に認識されてきていることにある。データ共有のためには意味的な水準でも合意事項が必要で、その実現方法としてオントロジを用いることがある、という事情がある。分散された異種データベースを統合利用するニーズはデータマイニング以外にも共通に存在するのであり、そのソリューションのために、オントロジを共通に用いよう、というのが適用の動機である。

3. データマイニングサーバ Knodias

・データベースサーバに置かれる

近年、計算機資源は分散化の道をたどって来た一方で、企業の基幹データベースを始め、重要なデータは依然として集中的に管理される場合が多い状況といえる。データウェアハウスも基本的にはこの形態のデータベースを中心に置いており、この集中された大規模データを扱うデータマイニングシステムを、そうした共通のデータベースサーバ上に置くのは自然といえる。Knodias はこうした中央集中的データマイニングを指向したサーバとして研究を進めているシステムである。図1に示すように、大福帳データベースと多次元データベースとを主な構成要素とするデータウェアハウス環境を土台として、その上にデータマイニング機能を構築する、というのが基本思想である。

・プラグインのプラットフォーム

また、次々に提案される新しいマイニングアルゴリズムを迅速に取り込んでゆくために、基本部をアルゴリズム独立の形で作成し、各アルゴリズムをプラグインシステムの形で実装できる、いわばＡＩにおける「黒板」に近い、オープンな性質のシステムをめざしている。

・イントラネット対応

一方、共有資源に対して多様な計算機環境からユーザがアクセスする利用方法を考えた場合、ユーザの使用環境の相違から利用方法に著しい制約を受けないようにするために、いわゆるイントラネット対応とし、もはや大多数のユーザが利用するウェッブブラウザからアクセスする形式とし、クライアントを Java[1]で記述することによってポータビリティを確保する、という方式を採用することは、むしろ自然といえる。

・オントロジを分散管理

データマートは、データウェアハウスのなかで必要な一部をユーザ環境に降ろしてきたものといえるが、個々のユーザがローカルに管理するデータがそれとは別に存在する場合があり、マイニングの際にはこのローカルデータ群と共通資源のデータとを統合して実行する場面もあると考えられる。本来整合性の保証されないデータベースを統合利用するには、意味的に等価でしかし表現の異なるデータを、同一のものとして明示的に対応づけ、同じ表現ながら全く異なるデータを、異なるものとして扱うようにやはり明示的に、調停（メディエート）を行う必要がある。Knodias では、図2に示すように、共有資

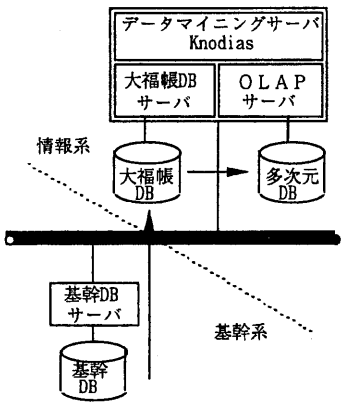

図1　Knodiasの基本構成

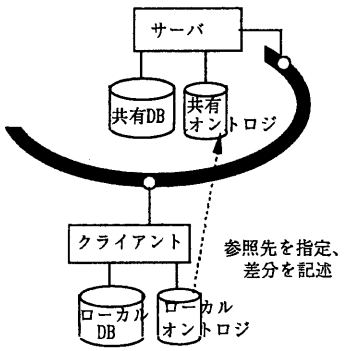

図2　分散するオントロジの統合利用

[1] Java はサンマイクロシステムの登録商標である。

源のデータはもとより、各分散資源に対して、概念辞書（オントロジ）を付すことで、この調停を容易化することを狙う一方、本報告で以下に述べるように、属性指向のマイニングを実現することをめざしている。

・システムスケーラビリティ

さて、応用ドメインによって、データの量が異なるから、データウェアハウスがスケーラブルである必要があるのと同様に、データマイニングも各応用で必要十分な性能を有するように、小規模データ用パソコンソフトから、大規模のものではプラント時系列データ用分析システムまで、連続に拡張できるスケーラブルな性質が望まれる。Knodiasでは、データと処理の規模に応じた多様なハードウェア構成に対応し、必要な性能を確保できる、システムとしてのスケーラビリティをめざしている。スタンドアロンのパソコン用小規模パッケージから、大規模なサーバシステムまで、機能的に一貫性を有するソリューションの体系を実現したいと考えている。

4. 属性指向のマイニング

ここでは、データマイニングサーバKnodiasのコンセプトを前提に、属性指向のマイニングの実現方法と効果について述べる。属性情報は、ここではオントロジとして記述されていることを前提とし、特に、属性情報を用いた相関ルール生成に課題を絞って検討を試みる。

オントロジは、対象となる専門分野で共有される概念に関する辞書である。自分の扱おうとする領域の用語を整理して、データエントリでの整合性を維持するだけでなく、外部のデータベースとの概念体系の差異によるスキーマ等が合致しない状況において、対応関係を管理もしくは生成し、自分の領域に同化して扱うことを可能とする。用語間の関係、すなわち同義語や、階層の上下関係など、用語辞書としても記述される内容をも包括する。実装の際、オントロジはフレームのような形式で記述されるのが通例である。

マイニング処理本体だけでなく、その前処理と後処理の重要性はよく指摘されるが、ここでも説明の便宜上、図3に示すように、マイニング全体をこうした3段階に分けて、属性指向のマイニングの効果について検討を加える。

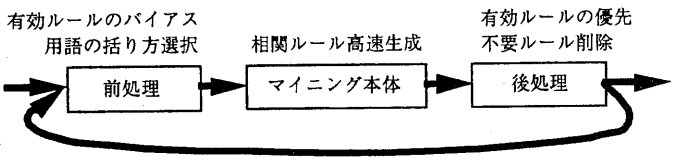

図3　マイニングの各段階での属性情報の利用

4.1 前処理における属性利用の効果

単一のドメインであっても、用語階層は必ずしも一つに定まるものではない。例えば、ＰＯＳで記録される製品のデータは最も詳細のレベルのＩＤと考えられるが、製品名の階層として、その製造会社を上位と考えて括った方が都合がいい場合もあり、一方、製品分類でのより広い括りを上位と考えた方がいい場合もある。理想的には、あらゆる括りに対して相関ルールを生成してみて、有効と思われるものを選択していったほうがいいのだが、これは計算量が膨大となることが容易に推察できる。その妥協点として、前処理の段階でユーザが適切な括を指定し、必要に応じて詳細度を下げて相関ルールを生成する、という方法がとりうる。今回の方式では、大福帳DBが作成される段階で、オントロジとデータ項

目が比較され、出現回数が加算されて頻度情報としてオントロジに対応づけて格納される。新規のトランザクションではこの頻度情報が保守される一方、未知語に対してはその出現を記録し、後でオペレータが既存のオントロジに関係づけ、属性情報を記述することで、オントロジを成長可能のものとすることができる。相関ルール生成の際には、最小頻度として support 値なるものを指定することがある。事前に頻度情報が集計されている場合、使用されている用語とその階層をインデンテーション付のリストで表示し、個々の出現頻度値を付し、support 値を超えているものについては色を変えるなどして、どの用語が相関ルールにかかるかを明示的に示すことができる。さらに、個々の用語で support 値に満たない場合でも、その上の階層で括れば support 値を超えることがあるので、システムではデフォルトとして、上位の方向に探して、そのしきい値を超える最初の用語を採用し、相関ルール生成に用いることができる。ユーザは適宜、所望の水準の用語を採用すべく、使用する用語としてより上位のものを選択することができる。このようにして前処理で使用する用語を指定して相関ルールを生成し、その結果を見てまたこの水準に戻って指定を変えて実行、といった繰り返しがここで想定する利用方法である。

さて、特定のルールに対してバイアスをかける場合、最も簡単な方法は、このルールの右辺と左辺に登場する語に着目して相関ルールを生成する方法である。この時、正確に一致する語だけを選んだのでは、ただひとつの相関ルールの統計量を求めていることになり、それほどの効果はない。そこで、オントロジを用い、同一もしくは類似の属性項目値を有する用語を関係語とみなし、その関係語で出現頻度がしきい値を超える語のみに対し、相関ルールを生成する。

このように、前処理の段階で属性情報を用いることにより、相関ルール生成の方法を制御することが可能で、これによってより柔軟な知識発見が可能となる。

4.2 後処理における属性利用の効果

相関ルールの生成の際には、不要なルールが生成される一方、価値のあるルールが表示の優先順位の下の方になって、ユーザの目から埋もれる可能性がある。これを避けるために、相関ルール生成の後処理として、特定のルールについてはユーザの目に触れる前に削除し、逆に重要と判断される種類のルールでは、表示の優先順位を上げる処理を行う必要がある。

後者、すなわち特定ルールの優先順位を上げるには、生成された相関ルールに対し、バイアスとなるルールに含まれる語を有するルールを選択し、表示優先度を上げることで実現できる。バイアスとなるルールが外部から導入したものの場合、オントロジを参照して、語の置換を行ってから対応するルールを探す、という手順を踏むことで、ルールを同化して利用することができる。

一方、不要なルールが多量に生成されるのはデータマイニングでは大きな問題と認識されているが、ドメインの知識があって初めてルールが自明とか不要と判断できるのであり、フィルタを自動的にかける万能の方法が存在するわけではない。そこで、今回検討した方式では、ユーザが削除指定したルールを記憶しておいて、次回以降表示しないという、簡単な方法を検討した。理想的には、属性情報を用いて類似ルールの削除を自動的に行いたいが、用いる語の水準を変えてマイニングする利用方法を考えたとき、水準を上げる方向でも下げる方向でも自動的に削除すべきかが明らかではない。ユーザに毎回聞く、という方法もあるが、むしろ正確に一致するもののみ自動削除、という簡単な方法のほうが、マイニングの挙動が予測しやすい、見通しのよいシステムとなると考えられる。

4.3 マイニング時の高速化の可能性

大規模データに対する相関ルール生成では、ディスクをスキャンする回数を減少することで、高速化

できる可能性がある。例えばAprioriの例では、Agrawalが指摘[3]する通り、アイテムセットの生成を2回分まとめて実行してメモリに保持できれば、サポートを数え上げる回数を減らせる。枝刈りをせずに処理を進めるために、候補増加によるメモリ爆発が問題となるが、この量を処理可能な範囲に維持できる問題であれば、ボトルネックとなりがちなディスクアクセスを数分の一という程度まで高速化は可能となる。オントロジに頻度情報を付す今回検討した方法では、まず最初の数え上げ段階はない。使用する語の総量を頻度情報をもとに絞り込み、2項組を生成する。この時の候補の数は、使用する語の数をNとすればN(N-1)/2であり、この段階で3項組を生成すればN(N-1)(N-2)/6である。保持すべき候補の数はこの和であり、これがメモリ制約を満たすものか、Nから計算して検査すればよい。今回は、あらかじめユーザが候補数上限数を与える方式を検討した。4項組以降も同様。ただし、3項組以降から候補を生成するとき、枝刈りがありうるので、事前に計算できる候補数はその最大値である。

こうした方法で実際に高速化できるのは、特殊な場合と考えられる。すなわち、ユーザが与えるsupport制約を満たす項目数Nが少なく、データのスキャンがボトルネックとなるほどデータベースが大きいドメインを対象とする場合である。

5. 関連する研究

用語階層上で適切な水準ものを自動的に選択する方式[1][2][7]について、研究例が報告されている。相関ルール生成の高速化[3]については、並列化による研究[4]が知られている。

6. 結論と展望

データマイニングサーバ Knodias で実現しようとしている属性指向の相関ルール生成機能について述べた。これを実装して、知識発見の柔軟性と高速化が実現できることを示し、この方法論の有効性を検証することが今後の課題である。

7. 謝辞

今回の研究の機会を与えていただきました、三菱電機(株)情報技術総合研究所情報処理部門 岩瀬部門長、アーキテクチャ部風間成介部長、システム部坂下善彦部長、知識処理チーム中島克人チームリーダに感謝いたします。

8. 参考文献

[1] Srikant, R., Agrawal,R. : Mining Generalized Association Rules, Proc. VLDB'95.

[2] Han,J., Fu,Y. : Dynamic Generation and Refinement of Concept Hierarchies for Knowledge Discovery in Databases., Proc. VLDB'95.

[3] Agrawal,R., et al. : Fast Algorithms for Mining Association Rules., Proc. VLDB'94.

[4] 新谷, 喜連川 : データマイニングにおける相関関係抽出の並列処理方式の実装とその評価, 並列処理シンポジウム JSPP'96, pp.97-104, 1996.

[5] 石井,他 : 業務ノウハウを活用するデータマイニングサーバKnodias., 第53回情処全大3R-5, 1996.

[6] 西田 : ソフトウェアエージェント, 人工知能学会誌, Vol.10, No.5, pp.704-711, 1995.

[7] Fukuda,T., et al. : Data Mining Using Two-Dimensional Optimized Association Rules : Scheme, Algorithms, and Visualization., Proc. SIGMOD'96, 1996.