1F4-04

エージェント技術を用いた複数データベースからのデータマイニング

Data Mining from Databases with Multiagent

新美 礼彦[*1]
Ayahiko Niimi

[*1]公立はこだて未来大学 システム情報科学部
Future University-Hakodate, School of Systems Information Science

In data mining, not only a single database but also two or more databases might be used. Data mining with some databases is able to show different view which data mining with a single databases shows. In this paper, we propose a method for using multiagent technology in data mining intended for two or more databases on network. In our proposed method, when data mining is done by using two or more databases, the Input/Output of data is described by using XML. The agents in multiagent system are made to access to databases, applications of data mining algorithm, and an arrangement of its result. First of all, we introduce some typical techniques as a technique of data mining to text database. Next, multiagent technology is described. The proposed technique is applied to document databases, and discuss its results.

1. はじめに

本論文では、複数のテキストデータベースを対象としたデータマイニングにおいて、マルチエージェント技術を取り込んだ手法を提案する。提案手法では、複数のデータベースを用いてデータマイニングを行う際、データの入出力を XML を用いて記述する。さらに、エージェントにデータベースへのアクセス、データマイニングアルゴリズムの適用、結果の整理をそれぞれ行わせる。これにより、複数のデータベースを用いる際の記述や処理の統一が行え、データベースアクセスやマイニングアルゴリズムの適用を分散して行うことが可能となる。なお、提案する手法は、テキストマイニングに特化した方法ではないが、本論文では、テキストをもとにしたデータマイニング（テキストマイニング）について取り扱う。まず、テキストデータベースへのデータマイニングの手法として代表的なものを紹介する。次に、使用するマルチエージェント技術について述べる。そして、マルチエージェント技術を取り込んだデータマイニングについて、提案する。提案した手法を文献データベースに適用させ、その結果について考察した。

第 2. 章では、分散データベースを用いたデータマイニングについて述べる。第 3. 章では、本論文で扱うテキストからのデータマイニングに関する手法について述べる。第 4. 章では、提案するマルチエージェントを用いたデータマイニングについて述べる。第 5. 章では、提案手法を文献データベースからのデータマイニングに適用した。その適用のさせかたと結果について考察する。第 6. 章は、まとめと今後の拡張について述べる。

2. 分散データベースによるデータマイニング

データマイニングとは、膨大なデータの中から意味のある知識や役に立つ知識を見つけるという研究分野であり、統計手法や人工知能的手法を用いて、さまざまな種類のデータからの知識抽出を行う。

連絡先: 〒 041-8655 北海道函館市亀田中野町 116-2
公立はこだて未来大学 システム情報科学部
新美 礼彦
TEL:0138–34-6222 FAX:0138–34–6301
E-mail:niimi@fun.ac.jp

データマイニングでは、単独のデータベースのみだけでなく、複数のデータベースを使うこともある。複数のデータベースを使い、データマイニングを行うことにより単独でデータベースを用いるのと違った見方ができる可能性がある。また、単独でデータベースを管理するよりも、分散して管理したほうが管理しやすいという利点もある。さらに、分析する目的/対象に応じて、使用するデータベースの組み合わせを変えることにより、目的に沿ったデータマイニングが行いやすいという特徴もできる。そこで、本論文では、主にテキスト情報を扱う複数のデータベースの組み合わせによるデータマイニングについて検討する。

一般に、複数のデータベースは、同じ属性で記述されているわけではなく、データに対して同じ操作が行えるわけでもない。複数のデータベースを同時に扱う際に、データ操作の統一とデータの記述属性の統一が不可欠である。これを、使用する目的ごとに毎回1から設計し直すのは、データマイニングをするのに手間がかかってしまう。そこで、本論文では、データに対する操作と記述属性の統一をマルチエージェントを用いることにより行う手法を提案する。

複数のデータベースを用いてデータマイニングを行う際、データの入出力を XML を用いて記述する。個別のデータベースへのアクセス法の違いは、XML への変換の際に吸収しておく。これによりデータの入出力をエージェントと切り分けてシステムを構築することが可能となる。また、エージェント間のデータのやり取りは XML のメタ属性により、属性間の変換が可能となる。さらに、エージェントを使うことにより、ネットワーク上で分散して実行することが可能なデータマイニングシステムが構築できる。

マルチエージェントについては、4. 章で詳しく述べる。

3. テキストマイニング手法

テキストデータベースに対するデータマイニングをテキストマイニングという。テキストマイニングアルゴリズムはデータマイニングアルゴリズムと同じものが多い。本論文で取り上げるテキストマイニングでは、主にキーワードを用いたデータマイニングアルゴリズムを扱う。

キーワードをデータマイニングで用いるためには、テキストからキーワードを自動抽出しなければならない。キーワード

抽出法として、さまざまなものが提案されている。提案されているキーワード抽出法を大きく分けると、形態素解析を用いるもの、形態素解析を用いないもの、文章の構造をもとに解析するものなどがある。 [市村 01] 以下に、本論文で使用した手法を述べる。

3.1 形態素解析

形態素解析とは、入力文を言語学的に意味をもつ最小単位である形態素に分割し、各形態素の品詞を決定するとともに、活用などの語変形化をしている形態素に対しては原形を割り当てることである。 [松本 99]

日本語では、単語が空白で切られていないため、形態素解析は重要である。英語では、形態素解析は語尾変化 (時制、単数 or 複数)、suffix,prefix などの解析に有効である。

例えば、「発表会を行いたい。」という文で形態素解析を行うと、というように分析される。(表 1 参照)

表 1: 形態素解析の例

出現形	基本形	品詞
発表	発表	名詞-サ変接続
会	会	名詞-接尾-一般
を	を	助詞-格助詞-一般
行い	行う	動詞-自立
たい	たい	助動詞
。	。	記号-句点

表 1 で、左側が文中の形のまま分割したもの、中央がその原形、右側がその品詞である。

形態素解析で分割された単語を要素単語という。要素単語に分けることにより、頻度解析や特定品詞へのフィルタリングが行えるようになる。

3.2 出現頻度による抽出

出現頻度分析では、形態素解析で分割された各要素単語の出現頻度を調べ、出現頻度の高い要素単語をキーワードとして抽出する。出現頻度の高い要素単語をキーワードとして抽出するため、どんな文書からも最適なキーワードを抽出しやすい手法である。しかし、助詞などのキーワードとして適切でない語を抽出する傾向があるため、抽出後のフィルタリングが重要になる。

3.3 連続名詞の抽出

連続名詞の抽出によるキーワードの抽出は、情報検索の世界では名詞概念をキーワードとして抽出する傾向が強いということを利用している。 [那須川 01] 一般的には、形態素解析を用いて名詞を抜粋し、キーワードの抽出をおこなう。「発表会を行いたい。」という表現を形態素解析を行った結果、「発表」、[会」、「を」、「行う」、「たい」の 5 つの要素単語に分割される。「を（助詞）」、「行う（動詞）」、「たい（助動詞）」は、名詞ではないのでキーワードとして抽出せず、この場合「発表」、「会」といった名詞をキーワードとして抽出する。ただし「発表」、「会」といった単位では、頻度は高いが具体性が低いため、「発表会」という、長い単位で語句を抽出することにより語の具体性を上げることができる。

3.4 N-グラム

N-グラム (N-gram) は長い文字列から部分文字列を取り出す方法であり、N には 2 や 3 などの数をとることができる。N-グラムのアルゴリズムでは 1 文字ずつずらしながら、連続する N 文字を取り出し、取り出した文字列の共起頻度を調べ、その集合の中で共起頻度の高い語をキーワードとして抽出するというものである。 [那須川 01] あらかじめ文書に品詞付けを行う必要がなく、任意の数の文字数を設定することができる。しかし、品詞付けを行わないで解析すると、単語の一部分を含んだ文字列をキーワードとして抽出する恐れがる。これを改善するために、本論文では形態素解析を行い、要素単語に分けた後で、その要素単語の連続を調べる手法も検討した。

3.5 相関ルール抽出

1 文中に現れる文字や単語の相関から、キーワードを抽出することが考えられる。その相関をルールとして抽出しキーワード (群) とすることが相関ルールによるキーワード抽出である。N-グラムを用いたアルゴリズムと同様に、形態素解析を行わなくてもキーワードを抽出することが可能である。相関ルールを高速に抽出する手法として、apriori アルゴリズムがある。[Agrawal 94] これも、N-グラムと同様に、単語の一部分のみを抽出する可能性を減らすため、本論分では形態素解析を行った後の、要素単語間の相関ルールからキーワードを作成した。

3.6 文章構造の解析によるキーワード抽出

文章構造を用いてキーワードを抽出することも考えられる。ニュースなどでは、話題になる文を先頭のほうにおく傾向が強い。また、HTML や LaTeX ではタイトルや、章見出しなどにタグをつけて記述することから、これらを情報からキーワードを抽出することもできる。

4. マルチエージェントデータマイニング

本論文では、マルチエージェント技術を複数の独立したプログラム（エージェント）を協調動作させることにより、情報を処理していく技術と捕らえた。一般的にマルチエージェント技術では個々のエージェントの自律的な制御に注目されるが、本論文ではそれに関しては考慮しないものとする。

エージェント間通信では、1 対 1 のもの、1 対多のもの多対多のものがある。本論文では、1 体 1 の通信として、UNIX のプロセス間通信、1 対多の通信として Black board モデルを使用した。

4.1 使用するエージェントとその定義

本論文では、以下で定義するエージェントを用いる。

Query agent: ユーザから使用するデータベースやデータマイニングアルゴリズムなどを受け取り、他のエージェントを生成する。Query agent はユーザからの要求ごとに生成する。

Mining agent: DB-access agent を生成し、DB-access agent を通じてデータを取得し、データマイニングアルゴリズムを適用する。Mining agent は、適用するマイニングアルゴリズムごとに生成する。

DB-access agent: データベースからデータを取得し、Mining agent に送信する。DB-access agent は Mining agent ごと、データベースごとに生成する。

Result agent: Mining agent の動きを監視し、Mining agent からの結果が集まると、それを整理/統合し、ユーザに提示する。

Black board(BB): データマイニングエージェントからの結果が書き込まれる場所

4.2 システムの流れ

提案するシステムの流れは、以下の通りである。また、各エージェント間の情報のやり取りを図 1 に示す。

1. ユーザは Query agent を生成し、使用するデータベース、データマイニングアルゴリズムなどを設定する。
2. Query agent が Black board(BB) の場所を設定する。
3. Query agent が Mining agent を生成し、BB の場所を送信する。
4. Query agent が Result agent を生成し、BB の場所を送信する。
5. Mining agent は DB-access agent を生成し、データベースにアクセスする。
6. DB-access agent はデータベースからデータを取得する。
7. Mining agent は DB-access agent からデータを受け取り、データマイニングアルゴリズムを適用する。
8. Mining agent はデータマイニングの結果を BB に記入する。
9. Result agent は BB をチェックして、結果が全て書き込まれたら、その結果を整理してユーザに提示する。
10. 全てのエージェントを消滅させる。

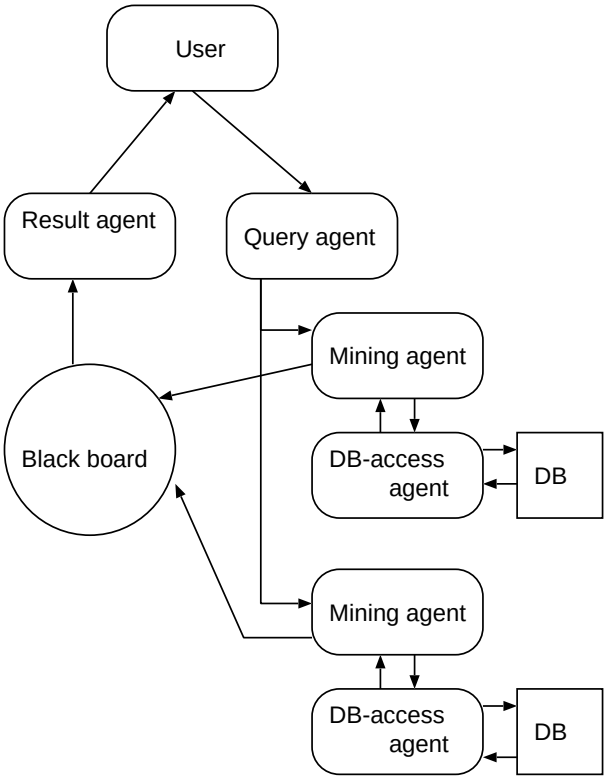

図 1: システムにおけるエージェント間の情報の流れ

データベースやマイニングの目的によっては、結果をみてから処理を変更したい場合も考えられる。この場合、上記のアルゴリズムを複数会行うことにより対応する。また、結果を整理してユーザに提示するとき、同義語などの重複を防ぐために、シソーラスを使う場合が考えられる。この場合は、Result agent からシソーラスの機能を持ったエージェントを起動し、処理結果を送信してもらう。この方法では、マイニングの結果のみに対して、シソーラスを参照すればよいので、効率がいい。しかし、実験では実装の容易さとシソーラスの大きさを考慮して、シソーラスを Mining agent の１つとして実装した。つまり、シソーラスエージェントはシソーラスの登録リストを返すというマイニングを行うエージェントとして捕らえた。この結果を Result agent が他の Mining agent の結果と区別して処理を行うことになる。

4.3 提案手法の特徴

提案する方法は、以下の特徴をもっている。

まず、Minig agent としてシソーラスエージェントを組み込み、シソーラスデータベースにアクセスできるようにすることにより、マイニングの結果をより意味のある形にまとめることができる。

次に、Query agent が複数の Mining agent を生成することにより、複数のデータマイニングアルゴリズムを並列に実行することが可能となる。また、データベースにアクセスする DB-access agent とデータを処理する Mining agent を分離することにより、データベースへのアクセスとデータの処理を分けてシステムを構築することが可能となる。

また、結果を整理するエージェントをおくことにより、それぞれのデータマイニングアルゴリズムの処理とそれの整理/統合を分けて考えることが可能となる。また、ユーザの目的に応じた整理/統合をシステムに組み込むことも容易となる。

このシステムにより、システム利用者は DB Agent と Mining Agent を再利用して、Query Agent と Result Agent のみを作り直すことにより、目的に応じたシステムを構築することが容易となる。

今回の実装では、UNIX 上でのプロセス間通信とファイルによる Black board モデルを使ったが、これを TCP/IP 上での通信に拡張することは簡単に行える。これにより、インターネット上に分散しているデータベースへの適用へ簡単に拡張することが出来る。本手法の問題点は UNIX 上のプロセス間通信を使っていることではなく、Black board モデルを使っていることである。使用するデータベース数とデータマイニングアルゴリズムが増えると、Black board への書き込みが問題になり、一番遅いエージェントの動作に全体の動作が引きづられてしまう。そのため、データベースへのアクセスとデータマイニングアルゴリズムの処理は並列化できても、結局、Black board でのチェック時に処理が停留してしまう。Black board への書き込みチェックに最大待ち時間を設定する、ユーザに逐次的に結果を見れるなどの対処をする必要がある。

5. 実験環境の構築

提案した手法を検証するため、データマイニングアルゴリズムをマルチエージェントに組み込んだ実験環境を構築した。構築した実験環境は、以下の通りである。

実験環境は UNIX システム上に構築した。エージェントは、環境中の各プログラムと定義した。これにより各エージェント (プログラム) は独立して動作する。

プログラム間の通信 (エージェント間通信) は、起動時のオプションと標準出力の取り込みと、Black board を利用する。

Black board はファイルを利用し、UNIX 上で Query agent 起動時の Process ID からユニークなファイル名を生成してそれを用いた。

使用したデータベースは、文献データベースと、それに関係するシソーラスデータベースである。このうち、文献データベースは、非線形分野に関する研究会の予稿集から作成した。[新美 03] 今回の実験のため、それを 2 つに分けたて実験で使用した。また、シソーラスデータベースとして、この文献データベースからテキストを抜き出し、形態素解析を行った上で、相関の高い高頻度語を抽出し、さらに専門家によりある程度チェックされたものを作成した。文献データベースは、1 つにおよそ 1200 の非線形問題に関する論文が登録されており、シソーラスデータベースにはおよそ 270 語が登録されている。

異なる形式のデータベースへのアクセスが可能なことを確認するため、文献データベースは RDB 形式、シソーラスデータベースはテキストファイル形式を用いた。

使用したデータマイニングアルゴリズムは、頻度分析、n-gram, 相関ルール分析、シソーラス分析である。ここでのシソーラス分析とは、シソーラスデータベースからシソーラス情報を取得することである。また、形態素解析には ChaSen を用いた。 [松本 99]

構築したデータマイニングシステムを使い、使用するデータベースやデータマイニングアルゴリズムの切り替えが行えることを確認した。得られた結果は、マルチエージェントを用いないで構築したシステムでの結果と同じものであった。構築した環境では、単独でのデータマイニングとあまり差がないため、マルチエージェントで構築した利点が少ないが、提案した枠組みを使えば、ネットワーク上で分散した環境も構築可能である。実験により提案した枠組みが少なくともローカルなマシン上で並列に動作することが確認できた。Black board モデルによるシステムの遅れは、特に体感できなかった。今回のシステムでは、使用するデータマイニングアルゴリズムもデータベースも多くないため、Black board への書き込みの待ち時間が問題にならなかったものと考えられる。それよりも、各データマイニングアルゴリズムの処理の方に時間がかかっていた。

6. おわりに

本論文では、マルチエージェント技術を用いて、分散データベースからのデータマイニング手法を提案した。

提案した手法を実際に文献データベースからのデータマイニングに適用し、その利点と問題点を検討した。

今回は提案した手法を適用するため、エージェント間通信としてプロセス間通信を、Black board モデルとしてファイルシステムを用いて、単独の UNIX システム上で動く最小限の実装しかしていない。このデータマイニングシステムにより、データベースとデータマイニングアルゴリズムを切り替えながらデータマイニングを行えることを確認した。

構築したシステムは単独のマシン上で行ったが、データベースへのアクセスは現在の実装でもネットワークを隔てたデータベースにアクセスできる。

今後は、データマイニング処理を行うエージェントをネットワークを隔てたコンピュータ上で実行できるようにし、データマイニング処理の負荷分散を考慮できるように拡張する予定である。

参考文献

[市村 01] 市村 由美、長谷川 隆明、渡部 勇、佐藤 光弘: テキストマイニング - 事例紹介, 人工知能学会誌 Vol.16 No.2,pp.192–200 (2001).

[松本 99] 松本 裕治、北内 啓、山下 達雄、平野 善隆、松田 寛、浅原 正幸:日本語形態素解析システム『茶筌』version 2.0 使用説明書 第二版 (1999).

[那須川 01] 那須川 哲哉、河野 浩之、有村 博樹:テキストマイニング基盤技術, 人工知能学会誌 Vol.16,No.2,pp.201–211 (2001).

[Nagao 94] Nagao, M., Mori, S.: A New Method of N-gram Statistics for Large Number of n and Automatic Extraction of Words and Phrases from Large Text Data of Japanese, In Proceedings of the 15th International Conference on Computational Linguistics pp.611–615 (1994).

[Agrawal 94] Agrawal, R., Srikant, R.: Fast Algorithms for Mining Association Rules, the 20th International Conference on Very Large Databases, Santiago, Chile, September 1994:32pages (1994).

[永田 01] 永田 昌明、平 博順: テキスト分類 - 学習理論の「見本市」, 情報処理 Vol.42 No.1,pp.32–37 (2001).

[新美 03] 新美 礼彦:カオス文献情報からのデータマイニングによる研究動向調査, 信学技法, AI20002-57, pp.59–64 (2003).