G-021

ビジネスデータからの知識発見システム：MUSASHI

羽室　行信[1]　加藤　直樹[2]　矢田　勝俊[3]
Yukinobu Hamuro　Naoki Katoh　Katsutoshi Yada

1. はじめに

本稿の目的は、我々がこれまでに企業との共同研究において開発を進めてきた[2][4][5]、ビジネスデータからの知識発見システム（MUSASHI: Mining Utility and System Architecture for Scalable processing of Historical data）[3][10]を紹介することにある。MUSASHI は、大規模な XML データを処理可能な、基幹系、情報系を包含するシステムアーキテクチャである。

MUSASHI は KDD（大規模データベースからの知識発見）[6]プロセスでもっとも労力を要するとされる前処理[7]にその強みがあり、リレーショナルデータベースやデータウェアハウスなど特別なソフトウェアを導入することなしに、大規模データを効率的かつ柔軟に処理できる仕組みを提供する。またいくつかの基本的なデータマイニングツールも既に実装されてはいるが、将来的には、各研究者がそれぞれに開発してきたデータマイニングアルゴリズムを動作させるプラットフォームとして MUSASHI を利用してもらうことを考えている。

2. ビジネスにおける知識発見の問題点

我々は、ビジネス分野での知識発見において、以下に示す問題を指摘することができる。1)商用の知識発見システムの導入には、通常莫大な投資が必要となるが、投資効果がわからないために、システムの導入を躊躇してしまう。2)リレーショナルデータベースを中心に業務システムを構築していると、更新処理に伴って、知識発見で必要とされる履歴データが消失してしまう。3)ビジネスでの知識発見では、時には数千万件から数億件のトランザクションデータを処理する必要がでてくるが、このような莫大なデータを柔軟かつ効率的に処理することは非常に困難である。4)知識発見を進める中で、多様なデータ項目の作成が要求されるが、その要求に応えることのできる柔軟なシステムは少ない。5)例えなんらかの知識発見システムを導入したとしても、ユーザは具体的に何をすればよいかわからない。

3. MUSASHI のコンセプト

上記の問題を解決するために、MUSASHI は以下に示すコンセプトを提案している。

1) MUSASHI は、オープンソースソフトであり、その API も含めて全てを公開している[8]ので、ソフトウェア自体のコストを必要としない。
2) 業務システムで発生するデータは、例え業務上必要がないと判断されても、それら全てを XML[9]で記録する。XML を採用することにより、a)テキストファイルで記述されるため、特別なソフトウェアを利用しなくても、容易に内容を閲覧でき、b)タグをユーザが自由に設定することができ、多様なデータ構造を表現することができ c)XML は標準のデータ記述言語として注目されており、今後多くのアプリケーションソフトが XML に対応することが予測されるため、基本的なソフトウェア技術の多重利用が期待できる。
3) XML のデータ構造は冗長性が高く、データ量が増大してしまう。そのための処理速度の大幅な低下を回避するために、我々は XMLtable と呼ぶデータ構造を採用することにした。XMLtable は、表構造データを伴った完全な XML 文書である。処理速度であるが、1000 万行 600Mbyte のデータに対して、ソーティングで約 11 分、ソーティングを伴わない処理で、1～2 分で処理可能である(Pentium III 800Mhz, メモリ: 512Mbyte, OS: Linux)。
4) MUSASHI は、データ処理のためのプログラムとして、単一の機能に特化した小さなコマンド群を提供する(コマンドの一部を表 1 に示す)。そしてこれらのコマンドを組み合わせ、シェルスクリプトとして実装することによって多様な処理を実現している。この特徴は、特に新しいものではなく、UNIX で伝統的に受け継がれてきた考え方である[1]。複雑なデータ要求に対しても，こうしたコマンドの組み合わせだけで柔軟に対応できるため，開発時間･コストが飛躍的に低減できる。

表 1 MUSASHI が提供するコマンド一覧(抜粋)

コマンド名	機能	コマンド名	機能
xml2xt	xml→xmlTable 変換	xtagg	集計
xt2xml	xmlTable→xml 変換	xtcount	行数計算
xtcut	項目の抜き出し	xtslide	項目を一行ずらす
xtcal	項目間演算	xtpattern	パターン文字列生成
xtsel	行の条件選択	xtchgnum	数値を範囲で変換
xtuniq	重複行の単一化	xtsort	並べ替え
xtjoin	単純結合	xtshare	シェア計算
xtnjoin	自然結合	xtclassify	決定木モデル生成

5) これまでにも多くの優れたデータマイニング技術が開発されてきているが、そららの多くは、ユーザ個々のビジネスの背景を考慮したものでないために、それらの優れた技術をどのようにビジネスに応用するかについて高度な専門知識が要求される。そこで MUSASHI では、これまで我々が蓄積してきたビジネスにおける知識発見の応用例をシェルスクリプトを含めて積極的に公開していくことによって、ビジネスの分野での利用を促進しようと考えている。その一例として、ブランドスイッチ分析[5]について次節にて紹介する。

4. MUSASHI を用いたビジネス応用

メーカーにとって、他社の新製品の動向は自社の既存ブランドの売上に大きく影響を及ぼす可能性があり、できる

[1] 大阪産業大学経営学部
[2] 京都大学大学院工学研究科
[3] 関西大学商学部

かぎり早くに、その影響度合いを把握することが求められる。そこで、ある新商品ブランドの購買シェアの高い顧客グループと低い顧客グループには、その購買行動にどのような違いがあるかを発見することを考える。その時、各顧客グループに特徴的な購買行動を表すと考えられる様々なデータ項目を作りこんでいくことが重要なタスクとなる。そこで以下に、顧客購買履歴データから、顧客別の「ブランドスイッチ回数」の顧客属性を、MUSASHI を用いて作成する様子を簡単に示す。

表２顧客購買履歴データ

(a)顧客 No	(b)日付	(c)ブランド	(d)次ブランド
01001	20011006	020A	020B
01001	20011110	020B	020B
01001	20011201	020B	020A
01001	20011206	020A	*
01002	20011015	020A	020B
01002	20011018	020B	020A
01002	20011027	020A	020B
01002	20011106	020B	020A
01002	20011125	020A	*

注）”*”は null 値を表す。

表２の項目(a)～(c)は、顧客の購買履歴データを表しているとする。データ中の二人の顧客（01001 と 01002）は、それぞれ「020A→020B→020B→020A」、「020A→020B→020A→020B→020A」の順にブランドを購入していることがわかる。ここで，ブランドスイッチとは，ある一人の顧客の隣接する２回の購入について，あるブランドから他のブランドに商品を買い換えたことと定義すると、顧客 01001 は２回、顧客 01002 は４回、スイッチしたことになる。MUSASHI を用いて顧客別のスイッチ回数を求めるスクリプトは図１のようになる。

```
xtslide –k 顧客 No –s 日付 –f ブランド:次ブランド |
xtsel –c ‘$ブランド –ne $次ブランド’ |
xtcut –f 顧客 No |
xtcount –k 顧客 No –a スイッチ回数
```

図１ブランドスイッチ回数を求めるスクリプト

一行目の xtslide コマンドは、顧客 No を単位にして、日付順に並べ、「ブランド」項目の値を一行上にずらしている（スライドさせる）。その結果、表２－(d)の「次ブランド」項目が追加される。そして、ブランドスイッチの定義から、「ブランド」項目と「次ブランド」項目が異なる行が、スイッチを表しているので、その条件に適合する行を、つぎの xtsel コマンドにて選択している。そして、選択された行数が、各顧客のスイッチ回数を表すことになるので、つぎの xtcut と xtcount コマンドにて、顧客別の行数をカウントすることによって、ブランドスイッチ回数を求めている。最終的に作成されたデータは表３－(a)に示されている。

このように、MUSASHI が提供するコマンドを組み合わせることによって、簡単に新しいデータ項目を作成できることがわかるであろう。

表３顧客別ブランドスイッチ回数

顧客 No	(a)スイッチ回数	(b)購入パターン	(c)平均間隔日数
01001	2	ABBA	23.67
01002	4	ABABA	10.25

ここでは一つの顧客属性の作成を例に MUSASHI の利用方法を見てきたが、その他にも、文字列で表したブランド購入パターン（表３－(b)）や、平均購入間隔日数（表３－(c)）など多様な顧客属性を作成していく。そして、それらの顧客属性を用いて、分類モデルを生成し（xtclassify にて）、有用なルールの抽出を試みる、という流れになる。

我々は、ここで紹介したブランドスイッチに関する知識発見以外にも、「ロイヤル顧客の早期発見[5]」、「顧客単位の最適な価格設定法[5]」、「ブランド関連強度分析[4]」など、MUSASHI をベースにした多様な応用事例を開発しており、すぐにでも実際のビジネスに適用可能である。

5. むすび

本稿では、MUSASHI の特徴を、そのコンセプトおよび応用事例を中心にして解説してきた。MUSASHI プロジェクトはまだ始まって１年ほどであり、ユーザーインターフェースの整備や、基幹系業務システムへの適用手法など課題も多い。しかし我々は、「基本となるインフラ技術は皆で共有し、その上での応用について企業が独自に保有する」という理念のもと、MUSASHI にまつわる全ての基礎技術はオープンソースとして公開し[8]、企業における効率的かつ効果的な情報システムの一助を担うことのできるコミュニティの構築を目指している。研究者がそれぞれに開発してきたデータマイニングアルゴリズムを MUSASHI 上の共通プラットフォームで実装していただき、日本発の知識発見システムを構築していきたいと考えている。

謝辞

本研究の一部は、平成１４年度関西大学重点領域研究助成金、および平成１５年度科学研究費補助金（基盤研究（Ｃ）(1)）によって行った。

参考文献

[1] Gancarz, M., *The Unix Philosophy*, Butterworth-Heinemann,1996.
[2] Hamuro, Y., Katoh, N., Matsuda, Y., Yada, K., "Mining Pharmacy Data Helps to Make Profits”, *Data Mining and Knowledge Discovery*, Vol. 2 Issue 4, pp.391-398, 1998.
[3] Hamuro, Y., Katoh, N., Yada, K., "MUSASHI: Flexible and Efficient Data Preprocessing Tool for KDD on XML”, *Proc. of first international workshop on data cleaning and preprocessing*, pp.38-49, 2002.
[4] Katoh, N., Hamuro, Y., Yada, K., "Discovering Purchase Association among Brands from Purchase History", *Proc. of SSGRR 2002w*, 2002.
[5] 加藤直樹, 羽室行信, 矢田勝俊「マーケティングとデータマイニング」システム／制御／情報, vol.46, No.4, pp.190-196,2002.
[6] Piatetsky-Shapiro G, (Editor) *Knowledge Discovery in Databases*, AAAI Press,1991.
[7] Pyle, D., *Data Preparation for Data Mining*, Morgan Kaufmann, 1999.
[8] URL http://musashi.adm.osaka-sandai.ac.jp/
[9] URL http://www.w3.org/XML/
[10] 矢田勝俊, 羽室行信, 加藤直樹, 鷲尾隆, 元田浩, "データマイニングシステム：MUSASHI," 信学技報, Vol.102, No.710, pp.59-64, 2003.