

サービス・コンピューティングに向けた データマイニングプラットフォーム MUSASHI の拡張

市川 晃平[†] 矢田 勝俊^{††} 鷺尾 隆^{†††}

[†] 大阪大学情報基盤推進本部 〒 567-0047 大阪府茨木市美穂ヶ丘 5-1

^{††} 関西大学商学部 〒 564-8680 大阪府吹田市山手町 3-3-35

^{†††} 大阪大学産業科学研究所 〒 567-0047 大阪府茨木市美穂ヶ丘 8-1

E-mail: fichikawa@cmc.osaka-u.ac.jp, tyada@kansai-u.ac.jp, washio@ar.sanken.osaka-u.ac.jp

あらまし 本稿の目的は、我々がマーケティング分析のために開発してきたデータマイニングツール、MUSASHI (Mining Utilities and System Architecture for Scalable processing of HIstorical data) をサービス・コンピューティングに向けて改良し、その適用可能性、技術的課題を明らかにすることである。従来のバージョンはユーザビリティが低く、導入には専門的な知識が必要であった。一方、新たに開発した3つのタイプのMUSASHI、つまりGUIを備えたWINDOWS版MUSASHI、ASP指向の分散型MUSASHI、クラウド展開版MUSASHIは、典型的なユーザのニーズに適合したプラットフォームになっている。本発表では企業データでの簡単な処理を参考に、現在の開発状況を紹介する。

キーワード データマイニング, サービス・コンピューティング, MUSASHI

Expansion of Data Mining Platform MUSASHI Towards Service Computing

Kohei ICHIKAWA[†], Katsutoshi YADA^{††}, and Takashi WASHIO^{†††}

[†] Office for Information Infrastructure Design, Osaka University

5-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047, Japan

^{††} Faculty of Commerce, Kansai University

3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680 Japan

^{†††} The Institute of Science and Industrial Research, Osaka University

8-1 Mihogaoka, Ibaraki-shi, Osaka 567-0047, Japan

E-mail: fichikawa@cmc.osaka-u.ac.jp, tyada@kansai-u.ac.jp, washio@ar.sanken.osaka-u.ac.jp

Abstract The objective of this study is to extend a data mining tool, MUSASHI (Mining Utilities and System Architecture for Scalable processing of HIstorical data), which we have been developed for marketing analysis. This paper introduces the further development of MUSASHI towards service computing and clarifies the applicability and technical problems. The original version of MUSASHI requires deep technical knowledge, and it is not suitable for the typical users of marketing analysis. The three types of newly developed MUSASHI, namely WINDOWS version of MUSASHI, ASP oriented MUSASHI, and Cloud based MUSASHI provide a platform that is better suited for the demands of typical users. This paper introduces the status of development of these three MUSASHI with simple examples using real commercial data.

Key words Data mining, Service computing, MUSASHI

1. ま え が き

本稿は、我々がこれまでに開発してきたデータマイニングツール、MUSASHI [1] のサービス・コンピューティングに向けた拡

張について紹介する。具体的には新たに開発を進めている3つのタイプのMUSASHI、つまり、WINDOWS版MUSASHI、ASP指向の分散型MUSASHI、クラウド展開版MUSASHIに関して、その開発状況に関して報告する。

```
<?xml version="1.0" encoding="euc-jp"?>
<xmldata version="1.1">
<header>
<field no="1" name="顧客"></field>
<field no="2" name="店"></field>
<field no="3" name="日付"></field>
<field no="4" name="商品名"></field>
<field no="5" name="数量"></field>
<field no="6" name="金額"></field>
<field no="7" name="売単価"></field>
</header>
<body><![CDATA[
10001 101 20100102 パン 2 400 200
10001 101 20100102 カレー 1 200 200
10001 101 20100103 パン 1 110 110
10001 101 20100103 牛乳 1 120 120
10002 102 20100102 お茶 2 300 150
10002 102 20100102 カップ麺 3 600 200
10002 102 20100103 おにぎり 1 105 105
]]></body>
</xmldata>
```

図 1 XMLtable の例

Fig. 1 An example of XMLtable.

MUSASHI は限られた単機能の処理, 例えば, 項目の抜き出し, 集計計算, 平均を計算するといった処理を行うコマンド群を提供し, これらの単機能のコマンドをパイプやリダイレクションによって組み合わせることで, 多様な処理を実現可能としている. これは UNIX のコマンド群が伝統的に行ってきた手法と共通し, 単純でかつ柔軟性に優れた効率的な情報処理システムを構築する上で非常に有効に機能する. また, 単機能に限定された個々の MUSASHI のコマンド群は高速にデータを処理することを可能としている.

一方で, 実際にユーザ環境において MUSASHI を導入するにあたっては, UNIX ベースの MUSASHI の思想は逆に障壁になる. MUSASHI を導入するためには UNIX 互換プラットフォームが必須であり, かつ柔軟なデータ分析を行うには UNIX 上で動作する何らかのシェルスクリプトに関して精通している必要がある. 我々は MUSASHI が有効な分野として, マーケティング分野におけるデータマイニングを対象に現在まで研究開発を行ってきた. しかし, 一般的な規模の小売店においては, UNIX 互換プラットフォームや技術者を維持しておらず, また外注依頼するほどのコストを投資出来る余裕はない. したがって, 現在でも MUSASHI を含め, データマイニング手法を用いたツールの利用が進んでいないのが実情である. 本研究ではこのような現状を鑑み, MUSASHI をよりユーザビリティの高いデータマイニングプラットフォームとして展開するために, データマイニングをサービスとして提供するサービス・コンピューティングに向けた展開に関して模索している.

本稿では, 具体的に現在取り組んでいる展開方法として, GUI を備えた WINDOWS 版 MUSASHI, ASP 指向の分散型 MUSASHI, クラウド展開版 MUSASHI に関して述べる. 従来, 多くのアプリケーションが一般的なユーザに広く受け入れられる過程として, スタンドアロンの PC 上で GUI アプリケーションとして動作する段階, Web ベースのサーバ・クライアント型サービスとして動作する段階, そして現在ではインターネットと大規模な計算資源を基盤とするクラウドコンピュー

表 1 MUSASHI コマンド群の例

Table 1 Examples of MUSASHI command tools.

コマンド名	機能	コマンド名	機能
xtagg	集計	xtcal	項目間の計算
xtcat	ファイルの併合	xtcut	項目の抜き出し
xtjoin	単純結合	xtsel	行の条件選択
xtsort	並べ替え	xtuniq	重複行の単一化

```
xtcut -f 店, 日付, 金額, 数量 -i input.xml |
xtagg -k 店, 日付 -f 金額, 数量 -c sum |
xt2csv -F -o result.csv
```

図 2 売上推移分析のスクリプト例

Fig. 2 An example script to analyze changes in sales.

ティングにおいて動作する段階を踏んできた. 本研究では, この各段階における MUSASHI の展開方法を検討している.

本稿の構成は以下の通りである. まず, 現状の MUSASHI の特徴に関して簡単に紹介する. 次に, WINDOWS 版, ASP 指向版, クラウド展開版の MUSASHI に関して, それぞれの展開手法に関して報告し, 最後に今後の課題に関してまとめる.

2. MUSASHI の特徴

MUSASHI は表形式のデータを XML で表現した XMLtable 形式データの処理に特化している. XMLtable 形式データは表形式データを単純にスペース区切りで表現したデータ部分と, カラム属性などのメタデータ部分からなる XML ドキュメントであり, データマイニングの対象となる大規模データを柔軟かつ効率良く表現できる形式である (図 1). MUSASHI はこのような XMLtable 形式データを処理する単一機能に特化したコマンド群を提供する. 表 1 は MUSASHI が提供するコマンド群の一部である. 現在, こうしたコマンドは 76 個あり, パイプやリダイレクションを用いてこれらのコマンドを組み合わせることによって, 複雑なデータ処理も実現可能である.

図 2 のスクリプトは図 1 のデータに対し, 店舗ごとの売上金額, 売上数量の日次推移を分析するスクリプトである. 1 行目のコマンドで分析に必要なカラムのデータだけを抽出し, 2 行目のコマンドで店舗, 日別の売上金額及び売上数量の合計を算出している. 3 行目は得られた結果を CSV 形式に変換して, result.csv という名前で保存している. MUSASHI のアーキテクチャは, このようにコマンド群を組み合わせることによって多様な分析を可能としている. このような特徴を持つ MUSASHI は, UNIX 互換プラットフォームを備え, シェルスクリプトに精通した技術者にとっては, 非常に強力なツールとなり得る. しかし, その一方で, そのような環境や技術を持ち合わせていないユーザにとっては活用が進んでいないのが現状である.

3. サービス・コンピューティングに向けた取り組みと開発の現状

現在, 我々は MUSASHI をより典型的なユーザのニーズに合わせた形態で提供する方法に関して模索している. 以下に説

```

title="@title:タイトル:text:売上推移_日別@"
comment="@comment:処理:text:日別に店の売上金額・数量を集計@"

function child {
  xtcut -f 店, 日付, 金額, 数量 -i"$1" |
  xtagg -k 店, 日付 -f 金額, 数量 -c sum -o"xx$2.txt"
}
@files:child:files@
xtcat -i"xx*.xt" |
xt2csv -F -o"@file:ファイル:file@"
rm xx*

```

図 3 売上推移分析のテンプレートスクリプト例

Fig. 3 An example template script to analyze changes in sales.

明する 3 つのタイプの MUSASHI はその取組の現状である。

3.1 GUI を備えた WINDOWS 版 MUSASHI

典型的なマーケティング分野でのユーザにとって、MUSASHI の導入を検討する際の最大の問題は、UNIX 互換プラットフォームが必須であり、かつ CUI ベースのインタフェースしか持ち合わせていないことである。我々は現在、この問題の解決に向けて、WINDOWS ベースのスタンドアロン PC 上で GUI を備えて動作する MUSASHI の開発を進めている。具体的には以下の 3 つのアプローチを取っている。つまり、1) MUSASHI コマンド群の WINDOWS 環境への移植、2) GUI から MUSASHI スクリプトを実行可能とする GUI ツールの開発、3) 典型的なマーケティング分析スクリプトの提供に取り組んでいる。

MUSASHI コマンド群の WINDOWS 環境への移植により、MUSASHI の実行にあたって UNIX 互換プラットフォームは必須ではなくなる。これにより、ユーザは MUSASHI を利用するために、新たに UNIX 互換プラットフォームを導入する必要はなく、現在、既に業務で利用している PC を利用して MUSASHI を使用することが可能となる。本研究では WINDOWS 環境への移植には、MinGW (Minimalist GNU for Windows) [2] を利用した。MinGW は gcc や make などの GNU 開発環境 (GNU ツールチェーン) の WINDOWS 移植版であり、UNIX ベースアプリケーションの WINDOWS への移植を支援する。

我々は上記の CUI ベースの移植に加えて、スクリプトを GUI を用いて実行可能とするための GUI ツール、MUSASHI-R を開発している。スクリプトを実行するには、入力ファイルと必要に応じてパラメータを指定する必要がある。CUI であればコマンドラインを通じてパラメータリストを受け取ることが可能なため単純であるが、GUI を介する場合、スクリプトの種類と入力すべきパラメータに応じて動的に GUI パーツ (テキストボックス、ファイル選択ボックスなど) を構成する必要がある。MUSASHI-R はスクリプトのテンプレートを受け取り、テンプレート内の可変変数部分に応じて動的に GUI を構成する手法を取っている。図 3 に図 2 の売上推移分析と同等のテンプレートスクリプトを示す。テンプレート内の「@xxx:yyy:zzz@」という記述部分が可変変数部分であり、この記述内容に応じて、MUSASHI-R は GUI を動的に生成する。図 4 にこのスクリプト実行時の MUSASHI-R のスクリーンショットを示す。

我々はこの MUSASHI-R を基盤に、この MUSASHI-R 上で動作するテンプレートスクリプトとして、マーケティング



図 4 MUSASHI-R のスクリーンショット

Fig. 4 A screenshot of MUSASHI-R.

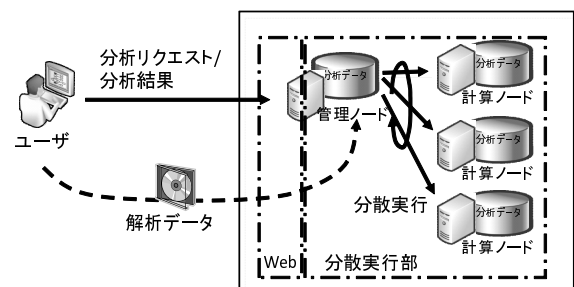


図 5 ASP 指向型 MUSASHI

Fig. 5 ASP oriented MUSASHI.

における典型的な分析に関するスクリプトを提供している。MUSASHI を利用したマーケティング分析用のスクリプトは、それ自身がマーケティングにおけるデータマイニングのノウハウを表わしており、データマイニングに取り組みたいと考える小売店にとっては、非常に有用なツールである。我々は現在までに、36 種類の分析用スクリプトを提供している。MUSASHI-R は既にプロトタイプが完成しており、実際の小売業者への展開を進める段階にある。

3.2 ASP 指向の分散型 MUSASHI

ASP 指向の分散型 MUSASHI はサーバを複数台用いて分散し、大規模なデータマイニングを Web インタフェース経由で容易に実行可能とするものである。MUSASHI はスタンドアロン環境においても数百万件から数千万件規模のデータは扱えるが、それより大規模になると使用に適さない。例えば、100 店舗程から構成される中規模のチェーン店において、数年分のデータの推移分析を行う際には取り扱うデータが数百 GB 程度となり、スタンドアロンで扱える規模を超える。

ASP 指向型 MUSASHI は、コマンドの実行に関しては基本的に WINDOWS 版 MUSASHI の MUSASHI-R の仕組みを踏襲する。ASP 指向型 MUSASHI は、登録されているテンプレートスクリプトに応じて、Web インタフェースを動的に作成し、Web ブラウザ経由でデータマイニング分析要求を受け付ける。

図 5 に ASP 指向型 MUSASHI の構成の概要図を示す。デー

タの大規模化に伴い、分析処理上、最も問題になるのは IO の増大である。したがって、この問題はどんなに高性能なサーバを導入しても 1 台の計算機では解決できる問題ではない。IO の増大を解決するためには物理的に IO のチャンネルを増やす必要がある。つまり、複数台の計算機に処理を分散する必要がある。我々が現在取り組んでいる ASP 指向型 MUSASHI では、ユーザは予め分析対象データを Web インタフェース経由でアップロードするか、もしくは CD、DVD 等の記録メディアを介して計算ノード群にアップロードする。管理サーバと計算ノードは伝統的な Beowulf 型クラスターの構成を取り、Web インタフェースを経由したユーザからの分析要求は管理サーバによって解釈され、入力ファイル別に計算ノードに分散されて実行される。ASP 指向型 MUSASHI は現在、提案・開発段階である。

3.3 クラウド展開版 MUSASHI

クラウドコンピューティング（以下、クラウド）の特徴は導入時の初期費用削減と、計算要求が拡大した際にそのスケールに合わせて規模を容易に拡大できる柔軟性にある。クラウド展開版 MUSASHI は、ASP 指向型 MUSASHI の導入をコスト面で容易にする目的と、さらなる大規模化に向けた取り組みの 2 つの目的がある。

ASP 指向型 MUSASHI を展開する際には、サーバ群を設置するにあたり、ある程度の初期費用が発生する。また、導入前にどの程度の計算要求があるか見積もることは難しく、導入すべきサーバ性能の選定は困難である。さらに一度導入に至れば、管理運用コストが一定して発生することになる。コストの見積が正確にできない状況で、これらの IT 設備に投資をする決断を行うのは非常に困難であり、導入の際の障壁になりうる。

近年、研究開発が進んできたクラウドは IaaS (Infrastructure as a Service) と呼ばれ、仮想計算機技術を応用し、計算機資源の貸出を時間単位で行うことを可能としている。これは従来の年単位や月単位の計算機資源の貸出と比べ、コスト面において大きなメリットがある。ユーザは計算機資源を利用した際に利用した時間や通信量に応じた料金を支払うのみでよく、初期費用・運用コストともに大きく節約できる。クラウド展開型 MUSASHI はクラウドを基盤として ASP 指向型 MUSASHI を展開することで、そのコストを最小化でき、実際の企業への展開を促進できると考えられる。

クラウドのもう一つの特徴は膨大な量の計算機資源を使用できる点にある。ASP 指向型 MUSASHI は 3.2 節で示すようにデータの分散処理の観点から、それぞれの計算ノードにデータのローカルコピーを持っている。しかし、扱うデータ量が膨大になるとローカルコピーで保持できる限界を超える。例えば、店舗数が数千を超える大規模なチェーン店で、数年分のデータ分析を行う際には、扱うデータが数 TB～数 PB に達する。そのため、クラウド技術の分野では、データを複数の計算機間で分散して保持し、データ分析も分散して行う手法が盛んに研究開発されている。Google が提唱している MapReduce [3] がその代表的な技術である。本研究では MapReduce のオープンソース実装である Hadoop を用いた MUSASHI の分散データ管理・分析手法の確立を目指している。図 6 は Hadoop を用いて、あ

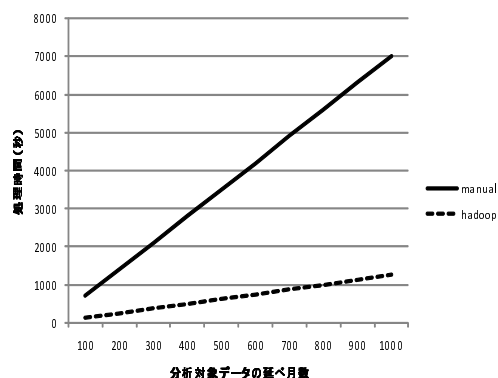


図 6 Hadoop による MUSASHI の処理時間
Fig. 6 Processing time of MUSASHI in Hadoop.

る商品と一緒に購入されやすい商品进行分析する MUSASHI のスクリプトを実行した際の処理時間の例である。使用した計算機は Intel Xeon E5440 (4 コア搭載) をデュアルで搭載しているサーバ 4 台で、図 6 は 1 台のマシン上で手で MUSASHI のスクリプトを実行した際との比較を行っている。延べ 1000ヶ月のデータは 80GB 程度にも達するが、Hadoop を用いて分散した場合は非常に高速に処理が終えていることが分かる。クラウド展開版 MUSASHI も現在、ASP 指向型の MUSASHI の発展及び大規模化の位置づけにあり、開発段階にある。

4. まとめと今後の課題

本稿では、MUSASHI をデータマイニングの基盤プラットフォームとして位置づけ、データマイニングをサービスとして提供するサービス・コンピューティングに向けた展開に関して紹介した。現在、データマイニングに対する投資は研究投資資金を有する一部の企業のみが可能である。本稿で目指す MUSASHI を基盤としたデータマイニングプラットフォームは中規模の小売店においても導入可能なサービスの提供を目指して研究開発を続けている。WINDOWS 版 MUSASHI である MUSASHI-R に関しては既に関心のある企業に働きかけてプロトタイプの使用を進めていく段階にある。また、ASP、クラウド版に関しても今後同様に実際の企業に導入を進めていく予定である。今後はこれら企業からの情報をフィードバックし、基盤となるデータマイニング技術のインフラは我々の研究コミュニティにおいて共有し、その応用・ノウハウに関しては企業に開放するオープンプラットフォームの構築を目指していく。

謝 辞

本研究は科研費特定領域研究 (21013032) の支援を受けて行われた。

文 献

- [1] 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”, 信学技法, **102**, 710, pp. 59–64.
- [2] “MinGW - Minimalist GNU for Windows”, <http://www.mingw.org/>.
- [3] J. Dean and S. Ghemawat: “MapReduce: simplified data processing on large clusters”, Commun. ACM, **51**, 1, pp. 107–113 (2008).