The Future Direction of New Computing Environment for Exabyte Data in the Business World

Katsutoshi Yada *Kansai University*
Yukinobu Hamuro *Osaka Sangyo University*
Naoki Katoh *Kyoto University*
Kazuhiro Kishiya *Kansai University*
{yada,kishiyak}@kansai-u.ac.jp hamuro@adm.osaka-sandai.ac.jp naoki@archi.kyoto-u.ac.jp

Abstract

With the rapid spread of the Internet and the computerization of trading a huge amount of data on the Internet and of transaction database in enterprises has been accumulated. The purpose of this paper is to explain the significance of the technology to process of exabyte-scale data and presents the business application, CODIRO, which will make it possible to integrate various types of large scale data. CODIRO is a consumer research system which discovers new knowledge by integrating the huge amount of different types of data both on the Internet and within companies. This paper will demonstrate the business implications for exabyte-scale information technology research, by explaining an example of the analysis of the sales effectiveness of television commercials using CODIRO.

1. Introduction

In recent years, various data has been accumulating around the world and has now reached tremendous proportions. Due to the rapid progress of information technology innovations, real economic activity and business have been computerized, and many companies have been developing and managing colossal databases. With the explosive spread of the Internet, communication between consumers in virtual space has come out; a new form of communication within BBS communities has been created, and it is now possible to transmit various types of information with such media as blogs by which communication data, which was never available before, can be accumulated.

This vast and various data is, however, generating new problems. First, just as any one company's business data has expanded into the terabyte size, it has become even more costly and time-consuming to integrate the data of multiple companies, because of the huge size of these data. Second, communication data between consumers on the Internet, such as BBS and blogs, has accumulated astronomically, but third parties have not been able to manage these communication and data format. As the data has not been structured, it has been difficult to integrate and use. Consequently, as far as we know, there have been no concrete examples where the immense data of multiple companies and the unstructured communication data on the Internet have been consolidated and then linked to business.

It is conceivable that, if this kind of terabyte or exabyte-class data constellation can be organically consolidated, then new business chances could be generated. The purpose of this paper is to expose the importance of research into data processing technologies of an exabyte scale by introducing the case example of a business application which organically integrates the immense, various and unstructured data, based on MUSASHI as system architecture to generate new knowledge.

2. Data mining platform: MUSASHI

2.1. MUSASHI

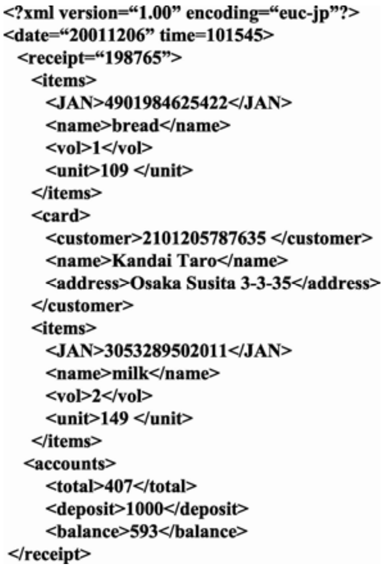
MUSASHI, the Mining Utilities and System Architecture for Scalable processing of Historical data, is a data mining platform [1][2] that efficiently and flexibly processes large-scale data that has been described in XML data. One of its strong points lies in the powerful and flexible ability to preprocess the knowledge discovery process from various amounts of raw data. The development of MUSASHI has been progressed as open source software, and everybody can download it freely from [3].

Proceedings of the The 2005 Symposium on Applications and the Internet Workshops (SAINT-W'05)
0-7695-2263-7/05 $20.00 © 2005 IEEE

MUSASHI has a set of small data processing commands designed for retrieving and processing large datasets efficiently in data extraction, cleaning, reporting and data mining, which specialize in single functions. By mounting such commands as a shell script, it is possible to process enormous amounts of data in various ways. MUSASHI also uses XML as a data structure to integrate multiple databases, by which various types of data can be represented. MUSASHI makes it feasible to carry out the flexible and low-cost integration of the abovementioned structured and vast business data in companies with the unstructured communication data on the Internet.

2.2. Data representation in MUSASHI

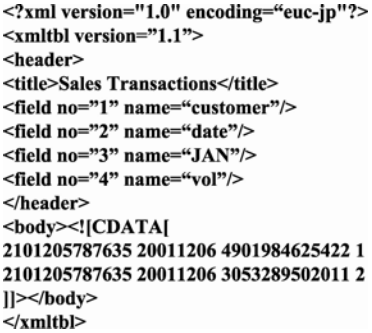
As MUSASHI employs XML as its default data structure, it is possible to efficiently process, not just business data which is structured, but also a variety of data which is generated in operational routines in enterprises and communication processes between consumers. For example, the point-of-sale cash register of recent years has the function to output operations performed by the operator as a log; electronic journal data. MUSASHI can convert that electronic journal data into XML data as in Figure 1, and store them in the form of so-called XML Table, a table data structure, such as in Figure 2, in order to efficiently process large-scale structured data.



```
<?xml version="1.00" encoding="euc-jp"?>
<date="20011206" time=101545>
  <receipt="198765">
    <items>
      <JAN>4901984625422</JAN>
      <name>bread</name>
      <vol>1</vol>
      <unit>109 </unit>
    </items>
    <card>
      <customer>2101205787635 </customer>
      <name>Kandai Taro</name>
      <address>Osaka Susita 3-3-35</address>
    </customer>
    <items>
      <JAN>3053289502011</JAN>
      <name>milk</name>
      <vol>2</vol>
      <unit>149 </unit>
    </items>
   <accounts>
      <total>407</total>
      <deposit>1000</deposit>
      <balance>593</balance>
 </receipt>
```

Figure 1. XML data that converts electronic journal data output from POS cash registers

The XML table is a complete XML document. The root element named <xmltbl> has two elements, <header> and <body>. The table data is described in the body element using a very simple text format with one record on each line, and each field within that record separated by a comma. Names and positional information relating to each of the fields are described in the <field> element; and it is possible to access data via field names. The data title and comments are displayed in their respective <title> and <comment> fields.



```
<?xml version="1.0" encoding="euc-jp"?>
<xmltbl version="1.1">
<header>
<title>Sales Transactions</title>
<field no="1" name="customer"/>
<field no="2" name="date"/>
<field no="3" name="JAN"/>
<field no="4" name="vol"/>
</header>
<body><![CDATA[
2101205787635 20011206 4901984625422 1
2101205787635 20011206 3053289502011 2
]]></body>
</xmltbl>
```

Figure 2. Sales data which has been converted into an XML table

MUSASHI converts the various types of data which exist within companies and on the Internet into XML data. Then, through the provision of commands that efficiently process that data, it forms the basis of a business application that organically consolidates that data.

3. Business chances that are realized through the integration of various types of data

3.1. CODIRO on MUSASHI

CODIRO [4] [5], the consumer research system that we developed on MUSASHI, is a business application that discovers new knowledge through the integration of vast and varied data that exist on the Internet and within companies. With CODIRO, we can gather and process data that has been built up within different companies in cooperation with these companies, as well as communication data which has occurred on the Internet between consumers such as BBS or blogs. We can then acquire new knowledge that could not previously be obtained when the data was isolated.

0-7695-2263-7/05 $20.00 © 2005 IEEE

At present, as is described below, CODIRO can discover useful marketing-related knowledge through the consolidation and analysis of data, which has been accumulated in different companies or on the Internet (See Figure 3). There are four types of data. First are the product-related databases which have accumulated in the manufacturers. For example, this includes data and information related to new products or sales promotion such as in-store promotion in marketing divisions within the manufacturers.

Second are databases in which the enormous amount of sales history data of each customer has been accumulated within retailers. Usually a chain store has a membership of at least tens of thousands per a store and has accumulated sales history data of them for several years. About 0.5-5 terabytes of sales data has been accumulated annually.

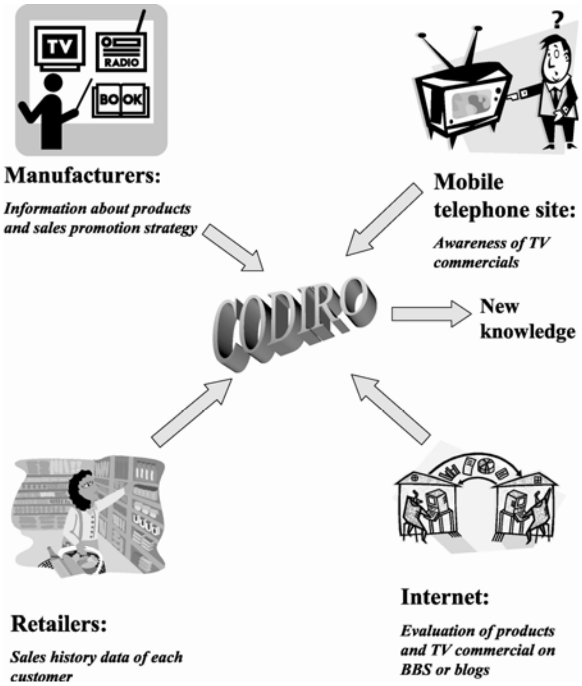

TV
RADIO
BOOK
Manufacturers:
Information about products and sales promotion strategy
?
Mobile telephone site:
Awareness of TV commercials
CODIRO
New knowledge
Retailers:
Sales history data of each customer
Internet:
Evaluation of products and TV commercial on BBS or blogs

Figure 3. CODIRO deals with vast and varied data in database of enterprises and on Internet

Third are both data which is related to surveys that gauge the awareness rates of television commercials, and which can be collected anytime from site members with the cooperation of mobile telephone websites. It has a wide membership throughout Japan that includes many different age groups and their extensive information that can be broken down according to age, sex and residence area, etc. It is possible for us to collect accurate data concerning specific consumer group reactions to a given TV commercial and goods in a short time.

The last is text data on Internet such as BBS which has cumulated opinions of consumers with regard to products and television commercials. On Internet such as BBS or blogs, a huge amount of communication data concerning the evaluation and the reputation about various goods has accumulated. It is valuable for marketing stuff in manufactures to plan and implement sales promotion strategy.

In order to analyze and then discover the relationships between this type of structured and unstructured data within companies and on Internet, CODIRO has built into it, system- and data-mining technology to preprocess large-scale data, as well as text-mining technology that can extract knowledge from text data. These types of data include: questionnaire-type data which can be collected from mobile telephone websites; Internet communication data; and the vast amounts data related to customer purchase histories. CODIRO will make contributions [4] to the field of marketing research as follows:

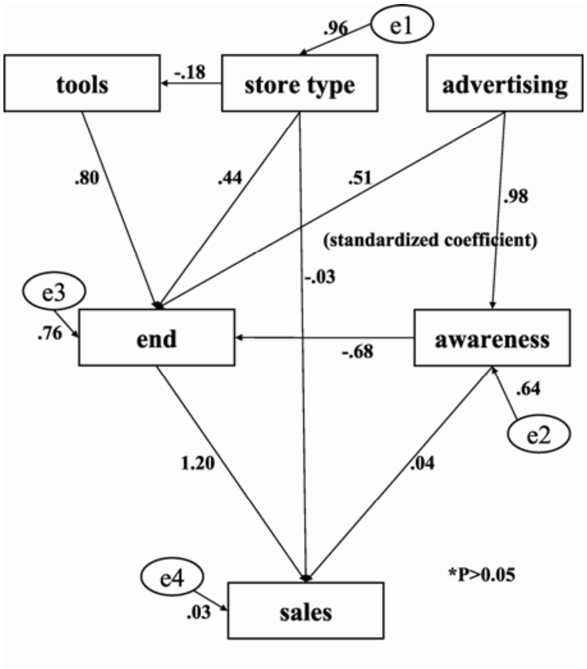
- The use of mobile telephones for awareness level research.
- It is possible to include detailed variables related to in-store merchandising methods and customer purchase history with data about TV commercials by using data mining technique.

3.2. A case of analyzing the effects of advertising

In this section, as a case example of discovering knowledge using CODIRO, we will explain the analysis of sales effectiveness of television commercials for retort pouch foods produced by the food manufacturer, Company A. In this case we analyzed the association among (1) data held by Company A which is related to television commercials and in-store promotions, (2) data collected from mobile telephone websites which measure the awareness of the television commercials, (3) freely expressed communication data which is related to the commercials and/or product, and (4) purchase history data of members, obtained with the collaboration of the supermarket.

In this case, new knowledge was exposed that was unattainable from the conventional data or with existing research methods. For example, it was apparent that, three weeks after the television commercial for a new product went to air, at around the time when consumer awareness level stopped rising, in-store promotions had the highest sales effectiveness.

Furthermore, in order to understand the complex relationship between the various types of data included in CODIRO, such as data on advertising, in-store promotions, consumer product awareness, sales, etc, covariance structure analysis was carried out, and the results are illustrated in Figure 4. From the model obtained, it could not be said that the consumer awareness and the amount of television commercials that went to air had a direct effect on sales performance, however it was discovered that they indirectly increased sales effectiveness through in-store promotions such as end-sales (position of sales space within the store).



	RMR	GFI	AGFI	PGFI	AIC	BCC	BIC	CAIC
Model	10.771	0.962	0.842	0.229	37.325	43.379	66.232	82.232

Figure 4. Results of covariance structure analysis

4. Conclusion

This paper has introduced CODIRO, the consumer research system that discovers new knowledge by integrating a variety of types of data from within multiple companies and communication data from consumers on Internet. There are still several problems remaining for future research. Now, CODIRO has been built on the premise of a maximum of several hundred GB of data, including sales data from a number of stores and communication data from several thousand consumers. In the future, we must expand the magnitude of data to an even larger scale, and we must implement various new technologies to do so. In addition CODIRO has not been integrated with text mining technique completely [6].

If we can develop and implement the technology that can process data that amount to the order of exabytes and extract useful knowledge from structured and unstructured data, then we believe we can discover knowledge that had previously been unattainable, and we can offer to companies the opportunity to realize new business chances.

References

[1] Y. Hamuro, N. Katoh and K. Yada, "MUSASHI: Flexible and Efficient Data Preprocessing Tool for KDD based on XML," *DCAP2002 Workshop held in conjunction with ICDM2002*, pp.38-49, 2002, pp.38-49.

[2] Y. Hamuro, N. Katoh, and K. Yada, "Data Mining oriented System for Business Applications," *Proceedings of First International Conference DS'98*, LNAI 1532, Springer-Verlag, 1998, pp.441-442.

[3] http://musashien.sourceforge.net/

[4] K. Yada, K. Kishiya and H. Osawa, "The Structure of Scenario Communication: A case study of consumer TV commercial awareness research," *Proc. of the First European Workshop on Chance Discovery (EWCD 2004), in conjunction with 16th European Conference on Artificial Intelligence (ECAI2004)*, 2004, pp.132-140.

[5] K. Yada, K. Kishiya, H. Osawa, C. Miyazaki and A. Miyawaki, "Consumer Research Systems by Using Mobile Device: CODIRO," *IPSJ SIG Technical Reports*, 2004-ICS-136, 2004, pp.115-122. (in Japanese)

[6] Y. Ohsawa, "Chance Discovery for Making Decisions in Complex Real World," *New Generation Computing*, Vol.20 No.2, 2002, pp.143-163.