Distributed Web Integration with Multiagent Data Mining

Ayahiko Niimi[1], Hitomi Noji[2], and Osamu Konishi[1]

[1] Department of Media Architecture, Future University-Hakodate
116–2 Kamedanakano-cho, Hakodate 041–8655, Japan
{niimi,okonishi}@fun.ac.jp
[2] Goodwill Engineering
6–10–1 Roppongi Minato-ku, Tokyo 106–6137, Japan

Abstract. We proposed a technique for using multiagent technology in data mining intended for two or more text databases. In this paper, we discuss data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining. First, we proposed data mining technique using multiagent technology. The proposed technique is applied to document databases, and discuss its results. In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file.

1 Introduction

In KES2003 and KES2004, we proposed a technique for using multiagent technology in data mining intended for two or more text databases. [1, 2] We applied our proposed approach to data mining from the document database, and discuss its problems. To apply proposed approach, we constructed only a minimum mounting which runs only UNIX local machine with process communications as agent communication and file system as black board model. It was confirmed to be able to switch the database and the data mining algorithm that used the constructed data mining system. We discussed data mining method based on text (text mining), but our proposed method is not a method of specializing in text mining.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 513–519, 2005.

Section 2 describes proposed data mining approach that uses multiagent techniques, and our proposal approach is applied to data mining from document databases. Chapter 3 describes the Web information integration system with multiagent data mining. Section 4 describes conclusion and enhancing in a future.

2 Multiagent Data Mining with Databases

In KES2003, the multiagent technology is defined as a technology that processed information by cooperatively operating two or more independent programs (agent). [1]

Generally, multiagent technology is discuss with an autonomous control of an individual agent, but in this paper, we do not discuss it mainly.

A communication between agents between one to one, one to multi, multi to multi. In this paper, we use one to one communication by UNIX process communication, one to multi by Black board model.

2.1 Agent Definitions

The definition of agent which is used for data mining in this paper is defined as follows.

Query agent: Query agent receives used the database and the data mining algorithm from a user, and generates other agents. Query agent is generated at each demand of a user.

Mining agent: Mining agent generates DB-access agent, acquires data from DB-access agent, and applies data mining algorithm. Mining agent is generated of each applied mining algorithm.

DB-access agent: DB-access agent acquires data from the database, and sends it to mining agent. DB-access agent is generated of each database and of each mining agent.

Result agent: Result agent observes a movement of mining agents, and obtains result from mining agents. When result agent obtains all results, result agent arrangement/integrates, and shows it to a user.

Black board(BB): Place where results from data mining agent is written.

2.2 Flow of System

A flow of proposed system is defined as follows. (Fig. 1 shows flowchart of proposed system.)

1. A user generates Query agent, with setting the used database and the used data mining algorithm as its parameter.
2. The place of black board(BB) is set with Query agent.
3. Query agent generates Mining agent, and the place of BB is transmitted.
4. Query agent generates Result agent, and the place of BB is transmitted.

5. DB-access agent is generated, and Mining agent is accessed to the database.
6. DB-access agent gets data from the database.
7. Mining agent receives data from DB-access agent, and applies the data mining algorithm.
8. Mining agent writes the result of data mining on BB.
9. Result agent checks BB, and if all results are written, arranges the results and presents to the user.
10. All agents are eliminated.

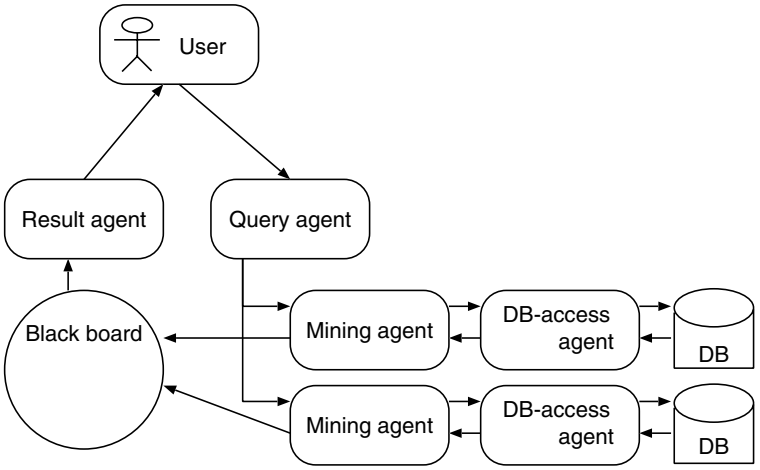

Fig. 1. Flowchart of Proposed System

2.3 Feature of Proposed Method

The proposal method has the following features.

The result of data mining can be made for more meaning result by building in the thesaurus agent as Mining agent, and making it can access the thesaurus database.

Query agent generates two or more Mining agent, it becomes possible to execute data mining algorithms in parallel. Moreover, it becomes possible that constructing the system and the access to the database and the processing of data are divided by separating DB-access agent accessed the database with Mining agent that processes data.

It becomes possible that the processing of each data mining algorithm and its arrangement/integration are separately thought by setting the agent which arranges the result. Moreover, it becomes easy to build arrangement/integration according to user's purpose into the system.

The system user comes to be able to construct the system corresponding to the purpose by recycling DB Agent and Mining Agent, and do tuning of Query agent and Result agent.

In this paper, the black board model with the file was handled with the interprocess communication on UNIX, but it can be easily enhanced to the communication on TCP/IP. Then, it is possible to enhance proposed approach to application to database that has been distributed on Internet. The problem of proposed approach is not using interprocess communication on UNIX but using black board model. Writing in the black board becomes a problem when the number of databases and data mining algorithm used increase, then the entire operation is influenced from the operation of the slowest agent. Therefore, the access to database and the processing of the data mining algorithm can be run parallel, but processing stops when checking results in the blackboard. It is necessary to consider that the maximum time is set to the black board writing check, and the system can show the result after each agent process.

3 Construction of Experimental Environment

We constructed an experimental environment which has multiagents with data mining algorithms to verify our proposed approach.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

The constructed experimental environment was following.

We proposed following operation in the system that constructs with multiagent (Fig. 2). In this system, company information can be obtained by inputting URL of the company that wants to examine it is in the Web site that the user specified. The system has two main part of system.

One system is the system that retrieves company information by user's input and extracts information, and another one is a system that integrates company informations.

The system works by the flow from the following 1 to 5. The location of each operation is as shown in Fig. 2.

1. Read URL that the user input, and the Web page is preserved.
2. Information is extracted on the preserved Web page, and it preserves it in the XML file.
3. Two or more extracted XML files are integrated into one XML file.
4. The index is calculated from the integrated XML file, and added to the integrated XML file.
5. The result from XML file is displayed by Web a browser.

Essential information, the financial situation, employment information, and the index of the company are displayed as a result of this system. It is thought

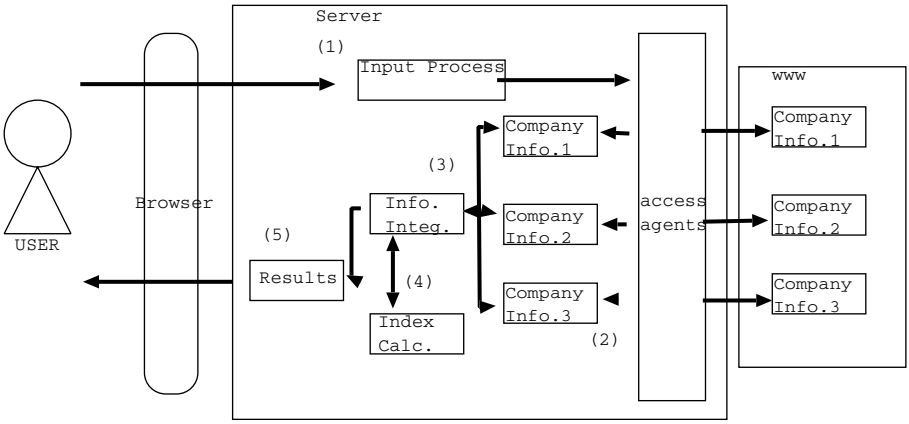

Fig. 2. Web Integration System

that more detailed information can be obtained by using not only basic information but also the index that can judge the company in the third person about the company.

3.1 Information Retrieve and Extract

In this system, at first, input information from Web browser, and the site on Web is preserved, and the system retrieves, and extracts necessary information from the preserved HTML file. In this operation, the text part of the corporate information is extracted from the tag of HTML by using the class of the pattern match of Java. Information on the extracted each item is put in the tag of specified XML. The XML file of each referred site is made by this operation.

3.2 Information Integration

In the information integration, necessary information is extracted from the Web site by using the Java program, and each Web site is brought together in one XML file. In this research, it thought information was extracted from the Web site of various forms, and XML that was able to correspond to a lot of file formats was used. The XML file that extracts information on each tag in XML by using the Java program when information is integrated, extracts information from each Web site, and makes it individually is brought together in one XML file as information on one company.

3.3 Index Calculation

This system evaluates the company that uses the index as a material judged from a position the third person to know more detailed information about the

corporate information. The index for the valuation of business enterprise is calculated by using extracted information, and it adds it to the XML file of the corporate information.

3.4 Show Results

The XML file that matches the index calculation result of making from the corporate information and such an index calculation is converted into the HTML form, and the corporate information is displayed. At this time, information such as the content of the work of the enterprise, the salaries, and branch offices is displayed in the form of the text besides the calculated index.

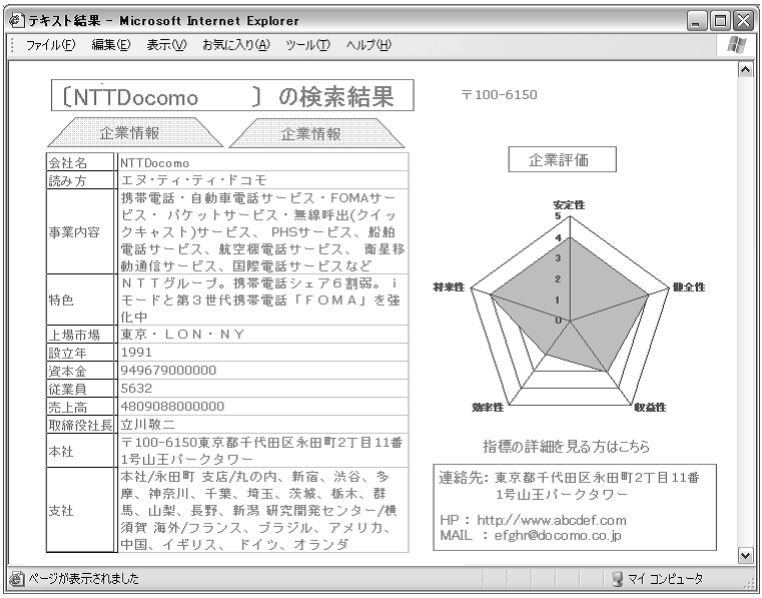

Fig. 3. Show Results

Information that cannot be used for a corporate name and making to the index in year of establishment etc. is displayed in the form of the text as shown in Fig. 3. Moreover, the index is shown in figure and the table like Star-chart for instance.

3.5 Experimental Results

In this paper, We confirm the operation of the information integration, and the index calculation in the system were done and verify the operation proof as the system that offered the valuation of business enterprise of the information extraction,

Three kinds of experiments of (1) information extraction (2) information integration (3) index calculation and the result display were done in confirming the operation in each part. The program operated in each experiment without trouble as the result.

4 Conclusion

In KES2003 and KES2004, we proposed a technique for using multiagent technology in data mining intended for two or more text databases.

In this paper, proposed data mining using multiagent was applied to information integration system on Web, and the effectiveness was verified. In the proposed method, the part of the database access agent was changed to the Web access agent. Also, mining agent was changed to the information extraction agent from the HTML file. The information integration on the Web page can be thought just like the information integration from the database. Similarly, it can be thought that the information extraction operation is one of the text mining algorithms.

We constructed distributed Web integration with multiagent data mining for company information integration, and We verified its system. Three kinds of experiments of (1) information extraction (2) information integration (3) index calculation and the result display were done in confirming the operation in each part. The program operated in each experiment without trouble as the result.

There is XBRL for sharing the corporate information. [3] We want to examine integration with such data format in the future.

References

1. Niimi, A., Konishi, O.: Data Mining for Distributed Databases with Multiagents. KES'2003, Proceedings, PartII, Lecture Notes in Artificial Intelligence 2774, Springer:pp.1412–1418 (2003)
2. Niimi, A., Konishi, O.: Extension of Multiagent Data Mining for Distributed Databases. KES'2004, Proceedings, PartIII, Lecture Notes in Artificial Intelligence 3215, Springer:pp.780–787 (2004)
3. XBRL Japan, http://www.xbrl-jp.org/ (In Japanese)