Scalable Web Mining Architecture for Backward Induction in Data Warehouse Environment

Dongkwon Joo and Songchun Moon, *Member, IEEE*

***Abstract*— For web mining, the biggest problem is the scarcity of data. To overcome the problem and prepare as many needed data as possible for business intelligent information, we propose backward induction in web mining. Web mining itself is an iterative process where data mining techniques are used back and forth and iteratively. To support backward induction and web mining characteristics, the scalable web mining architecture in data warehouse environment is proposed. The proposed web mining architecture has three kinds of scalabilities. These are the scalabilities of operational database, the scalabilities of data model and the scalabilities of data mining engines. By implementing the scalable web mining architecture having three kinds of scalabilities in data warehouse environment to support backward induction procedures, we can extract the business intelligent information from web mining.**

***Index Terms*—Backward Induction, Data Mining, Data Warehouse, Web, Web Mining, World Wide Web**

I. INTRODUCTION

As numerous business activities including electronic commerce have been done in the world wide web environments, the analysis of customer action behavior on the web has been required to achieve higher business performance. Also business style transformation and technology progress made it necessary to thorougly analyze the web sites in the perspective of web usage mining [1]. Web mining is the use of data mining techniques to automatically discover and extract information from web documents and services.

Although web mining is an application of classical data mining, it is different in that its data source is restricted solely to web log gathered from web server, which is a big drawback because data mining itself focuses on extracting meaningful information from as many data as possible. To overcome this defect and achieve business intelligence, backward induction enabling the identification of as many related data items as possible in early web mining process is required.

In previous web minings which only dealt with only web log, it was impossible to extract intelligent information related to business environment. But, by using backward induction to identify the needed data items in web and business environment, we can get the sophisticated business intelligent information from web mining. Example 1.1 shows why previous web minings failed to extract the intelligent meaningful information.

Example 1.1 (Customer Attraction Failure in Customer Relationship Life Cycle): The core concept of customer attraction in customer relationship life cycle [2] is the selection of new prospective customers in web environment. To do so, the rule information used in identifying the prospective customers must be extracted from existing customer information. For this process, customer profile data, web log data and sales data are needed. However, this has been impossible in previous web-log oriented web minings.

End of Example 1.1 ■

Data mining including web mining is a complex process where data mining techniques are used back and forth and iteratively. To support this characteristics and backward induction, the system architecture where any data items can be incorporated easily in any stages of web mining process without affecting existing architecture is necessary. In this way, the system architecture must have the scalability.

II. WEB MINING ARCHITECTURE

2.1 Backward Induction

Backward induction identifies the needed data items in accordance with the web mining objectives. It is different from previous forward induction that assumes there exists needed data ex ante. In the web and business environment, it is common that the information needs and mining objectives change rapidly. Also, the needed data items frequently change in web mining process. As a result, the data preparation is very important because it is affected by web mining business objectives. What is important is the appropriate direction of reasoning in web mining process.

If there is not related data that has the inherent information, the information cannot be extracted from web mining. This is presented minutely in [3]. For example, let us assume there exists

Dongkwon Joo is with Graduate School of Management, KAIST (Korea Advanced Institute of Science and Technology), Seoul Korea. He is now with Bill Technology Inc., as a researcher. (Telephone: +82-2-958-3333, e-mail: sotr@kgsm.kaist.ac.kr)

Songchun Moon is with Graduate School of Management, KAIST (Korea Advanced Institute of Science and Technology), Seoul (Telephone: +82-2-958-3315, e-mail: scmoon@kgsm.kaist.ac.kr)

IEEE Catalogue No. 01CH37239
0-7803-7101-1/01 / $10.00 © 2001 IEEE.

one data set. However, in this case, it is impossible to get the correlation between given data set and other data sets. From this simple example, we can evaluate the importance of backward induction. The backward induction procedures in web and business environment are shown in table 2.1.

Step	Measures
Business Objectives Identification	- Business Assessment - Mining Area Analysis - Expected Benefits Analysis
Data Preparation	- Data Item Identification - Data Preprocessing
Warehouse Modeling	- Multi-Dimensional Modeling - Loading - Interface Design
Mining Engine Implementation Web Mining	- Mining Technique Identification - Mining Engine Implementation - Evaluation and Modification

Table 2.1 Backward Induction Procedures
This table shows the backward induction procedure in a concrete form. But this procedural list is exhaustive. It is subject to change in accordance with the business or technological environment where the web mining is done.

Step 1: Identification of Business Objectives for Web Mining

What is most important in web mining is the benefit given by data mining. If there is no benefit in mining, web mining must not be done. As a result, as the business objectives are identified better, the benefits become much bigger. In this stage, identification of business objectives for web mining, three jobs must be done. These are *Business Assessment*, *Mining Area Analysis* and finally *Expected Benefits Analysis*.

In *Business Assessment* stage, the issues about what is our primary objective of the business and strategic decision must be explained. In *Mining Area Analysis* stage, the issues about web mining must be specified. In other words, the scope which data mining is applied into must be specified. Finally, in *Expected Benefit Analysis* stage, the benefits from web mining and its cost must be calculated accurately to reflect the situation. In this sense, the fundamental question is whether this web mining is profitable.

Step 2: Data Preparation for Web Mining

In this step of *Data Preparation for Web Mining*, there exist two stages of *Data Item Identification* and *Data Preprocessing*. *Data Item Identification* is for identifying data items needed in the process of web mining, which is done to assure the business intelligence. *Data Preprocessing* deals with processing the data into the suitable form for data mining and analysis.

Step 3: Warehouse Modeling for Web Mining

In this step of *Warehouse Modeling for Web Mining*, the multi-dimensional model to store the data items effectively and meaningful is designed. The accessibility of data items and resulting useful information extraction depends upon how well this data warehouse modeling is done. This step consists of *Multi-dimensional Modeling*, *Loading* and *Interface Design*.

In *Multi-dimensional Modeling* phase, the data model to support storing the data generated in web mining process must be constructed. The data model has to facilitate the analysis and referring to data in data warehouse. In *Loading* Phase, the data in other operational database must be loaded into data warehouse. Finally, in *Interface Design* Phase, the access methods to the data warehouse must be designed.

Step 4: Mining Engine Implementation for Web Mining

Mining engine is an application of its own mining purpose. In this step of *Mining Engine Implementation*, two tasks must be done sequentially. These are *Mining Technique Identification* and *Mining Engine Implementation*. First, we identify what techniques are needed to derive the result we want. An we design the mining engine to support the techniques. These engines can always be developed when we want to apply the new techniques. This scalability is possible because of the data warehouse.

Step 5: Web Mining

In this step of *Web Mining*, the real web mining is done through previously implemented web mining system and the evaluation of the system is done. Also, in accordance with the performance evaluation result, the revision must be done.

2.2 Scalable Architecture in Data Warehouse Environment

The web mining architecture must have scalabilities because of the features in web mining using backward induction. Our proposed web mining architecture in data warehouse environment has three kinds of scalabilities as shown in Fig. 2.1. These are scalabilities of operational database, scalabilities of data model and scalabilities of data mining engines.

First, the scalabilities of operational database provide the abundance of data items in web mining. As we mentioned above, many data items from different areas provide sophisticated business intelligent information. For example, customer profile, web server content information, web structure information, sales information, accounting information, and network information are needed in web mining. Our proposed architecture provides the scalabilities of operational database by adopting data warehouse, which plays an integral role of central data repository. Because of data warehouse, the different data from different sources can be merged, organized and structured in it. The addition of new operational database is very simple without affecting whole web mining system architecture.

Second, the scalabilities of data model provide the flexibility in web mining process. The multi-dimensional model made up of one fact table and many surrounding dimension tables supports the web mining process whose objectives and data items change frequently. If a new perspective of the analysis must be added, it can be easily added in terms of a new dimension table. This is a

advantage of the multi-dimensional model used in data warehouse. Also the multi-dimensional model provides the methods with which we can store the data for web mining in a structured and organized way.

Third, the scalabilities of data mining engines provide the extensibilities of the analysis. To put it another way, many data mining techniques can be applied into the data in accordance with the objectives and characteristics of web mining. By providing single interface to web mining engines, we can acquire the scalabilities of web mining.

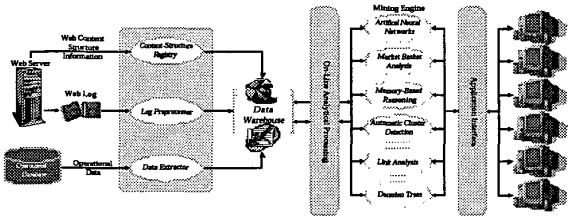

Fig. 2.1 Scalable Web Mining Architecture in Data Warehouse Environment

In this fig 2.1, a condensed web mining architecture is given. This architecture is given for reference. The detailed system architecture can be changed to meet the requirements when the system is implemented.

III. CONCLUSIONS

In this study, we understand the inherent characteristics of web mining and propose the scalable web mining architecture for backward induction in data warehouse environment. Backward induction helps us understand the web mining objectives and environments and prepare as many needed data as possible. This backward induction provides the fundamental way of overcoming the scarcity of data in web mining area. And the scalable web mining architecture in data warehouse environments supports the backward induction and mining process by providing three kinds of scalabilities: the scalabilities of operational database, the scalabilities of data model and the scalabilities of data mining engines.

For the further studies, the study on web mining in extensible Markup Language (XML) environment must be done. Because XML itself has many strengths that simple Hypertext Markup Language (HTML) doesn't have, today's web environment and web trends are evolving into XML environment. In this sense, the study on web mining in XML is very important.

In addition, mobile Internet environment is also very important. Many people browse or surf the net through mobile device such as Personal Digital Assistant (PDA), pocket-sized PC, and Wireless Application Protocol (WAP) browsers. The log collection and data mining will be very important sooner or later. Related with this topic, we made a notable achievement, which is on the location-specific web access patterns in mobile Internet environment [4]. To find out the web access patterns related with geographic locations, we plan to mine with the help of clustering algorithms.

ACKNOWLEDGMENT

The authors thank Namshik Ahn, Minkyo Lee and Namgyu Kim at Database Lab., KAIST (Korea Advanced Institute of Science and Technology) for their helpful notes and discussions.

REFERENCES

[1] J. Srivastava, R. Cooley, M. Deshpande, Pang-Nin Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," *SIGKDD Explorations*, ACM SIGKDD, Jan. 2000

[2] A. Buchner, M. D. Mulvenna, "Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining," *SIGMOD Record*, 27(4): 54-61, 1998

[3] M. Monticino, "Web Analysis: Stripping Away the Hype," *IEEE Computer*, Volume 31, Number 12,1998

[4] Namshik Ahn, Dongkwon Joo and Songchun Moon "Discovering Location-Specific Web Access Patterns in Mobile Internet Environments," Submitted for ACM SIGKDD 2001, San Francisco, CA, USA.