Rule-based Statistical Data Mining Agents for an e-Commerce Application

Yi Qin, Yan-Qing Zhang, K. N. King and R. Sunderraman
Department of Computer Science, Georgia State University, Atlanta, GA 30303, U.S.A.

ABSTRACT

Intelligent data mining techniques have useful e-Business applications. Because an e-Commerce application is related to multiple domains such as statistical analysis, market competition, price comparison, profit improvement and personal preferences, this paper presents a hybrid knowledge-based e-Commerce system fusing intelligent techniques, statistical data mining, and personal information to enhance QoS (Quality of Service) of e-Commerce. A Web-based e-Commerce application software system, eDVD Web Shopping Center, is successfully implemented using Java servlets and an Oracle8i database server. Simulation results have shown that the hybrid intelligent e-Commerce system is able to make smart decisions for different customers.

Keywords: Data Mining, Knowledge-based Systems, Statistics, e-Commerce

1. INTRODUCTION

This shift in computing priorities largely reflects the gradual commercialization of AI (Artificial Intelligence) to make data processing and knowledge discovery in e-business more intelligent and more efficient. The Knowledge-based System (KBS) is a byproduct of research in AI that started being commercialized during the early 1980s. The KBS is a computer-based system that uses extensive domain-specific knowledge (typically human expertise) to solve problems and/or automate (usually partially) decision processes. Knowledge is extracted from information by assigning it meaning and interpretation (i.e., Knowledge Discovery (KD)), and information is obtained from data by finding internal relations and useful patterns (i.e., Data Mining (DM)). The knowledge interpretation or meaning is typically given by human experts to represent the domain-specific knowledge that a KBS attempts to capture. The primary focus of the KBS is on solving problems (as done by human experts) and on the automation of decision processes. In general, both high-level knowledge discovery techniques and low-level data mining methods can be used to enhance QoS of e-Commerce because an e-Commerce company has to make smart financial decisions based on huge amounts of data sets in databases. Therefore, knowledge processing and applied AI can play an important role in e-Commerce [11][14-16]. The KBS can leverage organizational knowledge by providing a real and practical approach to capturing, preserving, and distributing critical commercial knowledge. One of the most successful schemes for knowledge representation in KBSs has proven to be the production system (or rule-based system) [1][4][12][19].

As e-Commerce rapidly develops, human beings rely more and more on computers to accumulate data, process data and use data to do e-Business. Both DM methods and AI techniques can help people do better e-Business. DM methods can be used in a variety of application areas, such as commercial databases (DBs), telecommunication alarm sequences, epidemiological data, etc. The area combines techniques from DBs, statistics, and machine learning. Now a challenging problem is how to process exponentially growing amounts of e-Commerce data by DM tools with limited intelligence. Therefore, there is broad common interest in intelligent DM technology among scientists and engineers from statistics, DM, KD, decision sciences and AI [2][7-9][13][18][20][21].

In this paper, a hybrid knowledge-based e-Commerce system fusing artificial intelligence, statistical data mining and personal information is proposed to enhance QoS (Quality of Service) of E-Commerce. A Web-based e-Commerce application software system, the eDVD Web Shopping Center, is successfully implemented using Java servlets and an Oracle8i database server. The real demonstration of the eDVD Web Shopping Center has shown that the smart E-Commerce system can make better marketing because the system can encourage costumers to buy more products based on their purchases and preferences.

Data Mining and Knowledge Discovery: Theory, Tools, and Technology V, Belur V. Dasarathy, Editor, Proceedings of SPIE Vol. 5098 (2003) © 2003 SPIE · 0277-786X/03/$15.00

2. E-COMMERCE MARTTING SCENARIO

DM is used in an automated approach to exhaustively exploring and bringing to the surface complex relationships in very large datasets, which are largely tabular in nature, having most likely been implemented in relational DB management technology. These techniques can be applied to other data representations, including spatial data domains, text-based domains, and multimedia (image) domains.

In today's business world there is an abundance of available data and a great need to make good use of it. Many businesses would benefit from examining customer habits and trends and making marketing and product decisions based on that analysis. However, the process of manually examining data and making sound decisions based on that data is time consuming and often impractical [7]. DM tools can discover critical information from these huge data sets. Many successful organizations are turning to DM for better decision-making. Using powerful analytical techniques, DM makes it possible to turn raw data into information that can be used to gain a marketplace advantage.

DM uses a plethora of algorithmic tools such as statistics, regression models, neural networks, fuzzy sets, evolutionary methods, rough sets, and clustering [2][7][20][21].

The area of DM is inherently associated with DBs, in this sense DM significantly augments DBs by making them more user-friendly and thus helping people to feel more comfortable dealing with the vast amounts of data and making use of them. One could easily store records of important transactions, retrieve them without any difficulty and make some decisions augmented by the facts found in the DBs [2].

Over the past 20 years, computers have been used to capture detailed transaction information in a variety of corporate enterprises. Retail sales, telecommunications, banking, and credit card operations are examples of transaction-intensive industries. These transactional systems are designed to capture detailed information about every aspect of business. Only five years ago, DB vendors were struggling to provide systems that could deliver several hundred transactions per minute. Now, for instance, Wal-Mart alone generates around 20 million transactions a day. The knowledge being stored at a continuously growing pace becomes less and less comprehensible while it is buried in gigabytes of records. No human can use such data in an efficient way and be able to understand basic trends and thus make rational decisions. The information, even stored, becomes less and less useful as we are faced with difficulties of retrieving it and making it available in an easily comprehensible format at higher levels of summarization. For instance, if a result of a retrieval session from the DB is a collection of thousands of records then the data is as useless.

Advances in data gathering, storage, and distribution technologies have far outpaced computational advances in techniques for analyzing and understanding data. There is an urgent need for a new generation of tools and techniques for automated DM and Knowledge Discovery in Databases (KDD). KDD is a broad area that integrates methods from several fields including statistics, DBs, AI, machine learning, pattern recognition, machine discovery, uncertainty modeling, data visualization, high performance computing, optimization, management information systems and KBSs.

KDD refers to a multi-step process that can be highly interactive and iterative. It includes data selection, data sampling, preprocessing and transformation for subsequent steps. DM algorithms are then used to discover patterns, clusters and models from data. These patterns and hypotheses are then rendered in operational forms that are easy for people to visualize and understand. DM is a step in the overall KDD process.

3. HYBRID KNOWLEDGE-BASED E-COMMERCE SYSTEM

Due to more and more customer-oriented business trends, DM techniques have been widely studied. Figure 1 shows major components in the knowledge-based e-commerce system. In Figure 1, the system accepts validated userID and password, and stores facts in DB, if applicable. A global DB and a set of production rules are vital components in rule-based systems. The DB acts as a context buffer which records the conditions evaluated by the rules and the information on which the rules act. In this design, the system invokes a rule engine to navigate production rules, which also sit in the global DB, by analyzing and triggering appropriate rule(s) augmented by the facts found in DB. The rule engine returns (sequential) action(s) to a user interface after concluding which rule(s) is (are) fired. The user interface can interact with the DB without invoking the rule engine.

One special aspect of user interface design is the challenge of displaying clearly tables, charts, maps, and diagrams. These special items are grouped under the term information-visualization and emphasize an aspect of the user interface development that requires separate, special attention.

Rules can be used to express: (1) Heuristic knowledge: rules are particularly useful for representing surface knowledge about relations between different input and output variables, for example, the rule "**IF** persons A and B have the same nationality **THEN** their native languages are the same" expresses a heuristic relating the native languages of persons to their nationalities.

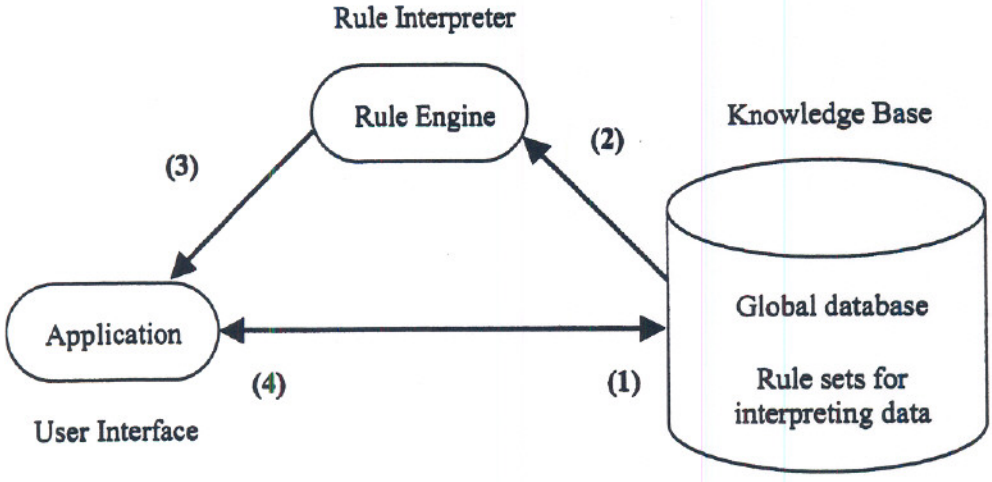

Figure 1 Structure of a knowledge-based e-commerce system

Due to the nature of rules applied in system design, heuristic knowledge is not adopted. (2) Domain models: rules can be used to represent known relations between different components / objects in the domain. For example, in system design, the following rule is given:

IF frequencyM(x) & totalH(x) | frequencyH(x) & (totalM(x) | totalH(x))

THEN excellentCust(x).

This rule expresses a simple fixed relation between shopping frequency, total purchasing amount and customer categories. A set of such rules can cumulatively define the model of a particular domain. (3) Action sequences –rules can also be used to represent actual action sequences, like the following rule presented in our system design.

IF junior(x) & PG13(x, y) | R(x, y) **THEN** juniorWarning

The rule engine (or the central control system) serves as a rule interpreter and sequencer. It scans the production rules for those active or applicable, i.e. **IF** condition is **TRUE**. This step generates a list of active rules (which might be a null list). If more than one rule is active, then deactivate those rules that would duplicate characteristics already on the WM. This step prevents redundancy. It fires the lowest numbered active production rule. If there are no applicable rules, exit the loop. The best guess will be the top item on the WM list. It turns the **IF** part of all production rules to **FALSE** and go to the control statement. This provides an iterative structure.

4. eDVD SHOPPING CENTER

The Main features adapted in this system design and implementation have been introduced. Now it is ready to experience how a knowledge-based eDVD shopping center, an e-business world, works to achieve better QoS.

Fig. 2 shows the customer received 10% discount. Lets' pretend to be rule engine again to figure out how 10% discount is granted.

Rule engine inferences rules by forward chaining, navigating each "postCond":

(a) For Rule 3 "postCond", only one token frequency(x)<=5, see frequency(x)<=5 → search up "data" for frequency(ua1234567)<=5, search succeeds, set frequency(x)<=5 TRUE → frequencyL(x) (in "action") is TRUE.

(b) For Rule 4 "postCond", only one token amtAccumulate(x)<=500, see amtAccumulate(x)<=500 → search up "data" for amtAccumulate(ua1234567)<=500, search succeeds, set amtAccumulate(x)<=500 TRUE → totalL(x) (in "action") is TRUE.

(c) For Rule 5 "postCond", token by token frequencyL(x), totalL(x), totalM(x), |, frequencyM(x), totalL(x), &, |, see frequencyL(x) → search up "data" for frequencyL(ua1234567), search fails, then

(c.1) earch up "action" for frequencyL(x), search succeeds, execute each token (one token in this case) in its "postCond" part by recursion, until exhaust all rules. Here frequencyL(x) is TRUE because "postCond"

frequency(x)<=5 is TRUE, see totalL(x) → search up "data" for totalL(ua1234567), search fails, then repeat (c.1) for totalL(x) → totalL(x) is TRUE, see totalM(x) → search up "data" for totalM(ua1234567), search fails, then repeat (c.1) for totalM(x) → totalM(x) is FALSE, see operator token |, calculate last two tokens (see Appendices) → "totalL(x) totalM(x) |" is TRUE, see operator token &, combine previous calculation with frequencyL(x) (see Appendices) → frequencyL(x) totalL(x) totalM(x) | & is TRUE, Rule 5 is fired → fairCust(x) is TRUE,

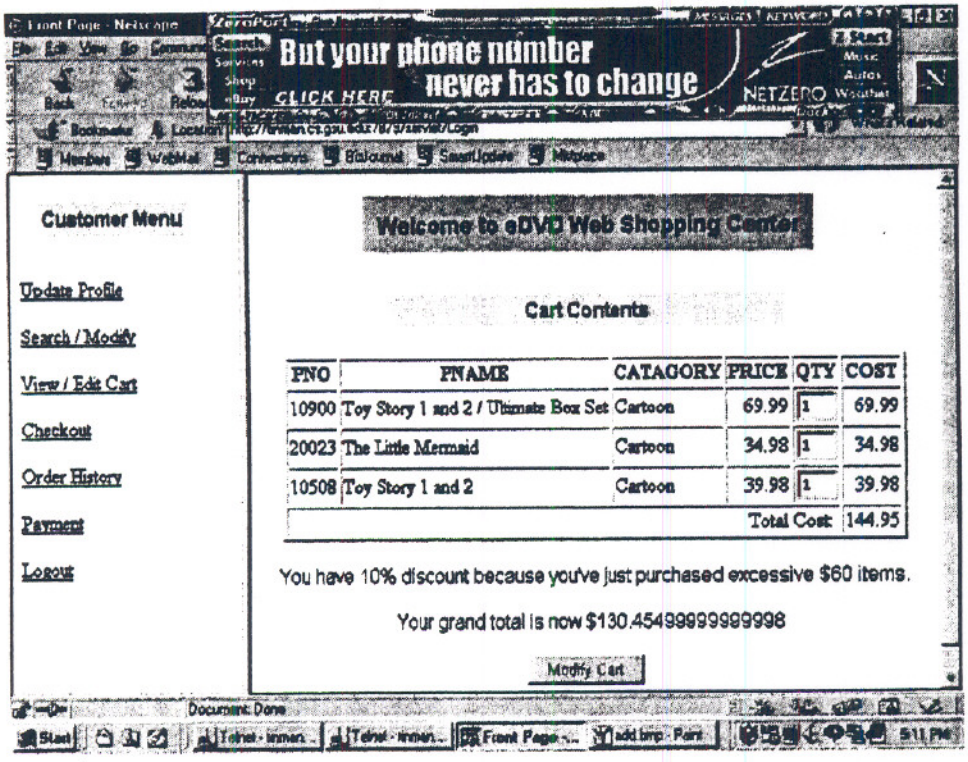

Figure 2 the eDVD Web shopping Center

(d) For Rule 6 "postCond", token by token goodCust(x), fairCust(x), total60(x), &, |, see goodCust(x) → search up "data" for goodCust(ua1234567), search fails, then

(d.1) search up "action" for goodCust(x), search succeeds, execute each token (one token in this case) in its "postCond" part by recursion, until exhaust all rules. Here goodCust(x) is FALSE, see fairCust(x) → search up "data" for fairCust(ua1234567), search fails, then, repeat (d.1) for fairCust(x) → fairCust(x) is TRUE, see total60(x) → search up "data" for total60(ua1234567), search succeeds, set total60(x) TRUE, see operator token &, calculate last two tokens (see Appendices) → fairCurst(x) total60(x) & is TRUE, see operator token |, combine previous calculation with goodCust(x) (see Appendices) → goodCust(x) fairCust(x) total60(x) & | is TRUE, then Rule 5 is fired, goodFairDiscount is applied.

5. TECHNICAL ANALYSIS

DM is the data analysis component of KDD. According to its exponents, KDD encompasses all steps from the collection and management of data through data analysis. It is a broad area that integrates methods from several fields including machine learning, statistics, pattern recognition, AI, and DB systems, for the analysis of large volumes of data. In recent years the power of machine learning and statistical techniques to discover interesting patterns in raw data has manifested itself in many applications. As these techniques have matured in sophistication and power, industry has become interested in them and become directly involved in their promotion and use, particularly in various conferences on KDD.

Statistics is defined as the science of collecting, analyzing and presenting data. If this admittedly broad definition is accepted, then KDD is statistics and DM is statistical analysis. KDD has a spin that comes from DB methodology and from computing with large data sets, while statistics has an emphasis that comes from mathematical statistics, from computing with small data sets, and from practical statistical analysis with small data sets.

Large data sets often contain, for the purposes for which the data will be used, a relatively small number of independent items of information. Thus a large volume of data can be reduced to a much smaller summary form, which can enormously aid the subsequent analysis task. This is important when considering whether the huge extent of a data set may allow the use of analysis methods, which for small or medium size samples, would have made very poor use of the data. Frequency (number of purchase a customer made) and total (total amount a customer accumulated) within a time frame, 90 days in this application, to characterize the small data sets. These two small data set elements are segmented into three different levels respectively, as low, medium, and high, in terms of density.

From this experiment, apparently, we can reduce large data sets to manageable size by carefully choosing proper elements out of it without loosing any information, and manipulate the selected elements for sequential statistical DM method in KDD. Furthermore, the smart eDVD shopping center can attract valuable customers by promoting certain products, predict marketing trends.

6. CONCLUSIONS

An eDVD Web Shopping Center is successfully implemented. Major technical merits are (1) a relational DB technique presents its efficiency for large data sets in knowledge-based system implementation, (2) DM plays a very important role in large scale KDD for discovering hidden patterns and relationships in data, with an emphasis on large observational DB, and (3) statistical DM lets us explore data using proven statistics and innovative graphics to discover hidden patterns, with which users can create models for discovering relationships between multivariate data that go beyond the classical models in both power and predictive ability.

REFERENCES

1. A. Barr and E. A. Feigenbaum, *The Handbook of Artificial Intelligence, Vol. 1*, Pitman Books, London, 1981.
2. Krzysztof Cios, Witold Pedrycz, and Roman Swiniarski, *Data Miming methods for Knowledge Discovery*, Kluwer Academic Publishers, 1998.
3. Soumitra Dutta, *Knowledge Processing & Applied Artificial Intelligence*, Butterworth-Heinemann Ltd., 1993.
4. Edward A. Feigenbaum, Avron Barr, and Paul R. Cohen (eds), *The handbook of Artificial Intelligence, Vol. 1-3*j, HeurisTech Press/William Kaufmann, Inc., Stanford, CA (1981-82).
5. Morris W. Firebaugh, *Artificial Intelligence A Knowledge-Based Approach*, PWS-KENT Publishing Company, 1988.
6. P. Harmon. and D. King, *Expert Systems*, pp. 26, John Wiley, 1985.
7. K. L. Hearn and Y. Zhang, "Fuzzy, crisp, and human logic in e-commerce marketing data mining," Proc. of SPIE'2001 Conf. of Data Mining and Knowledge Discovery: Theory, Tools, and Technology, Vol.4384, pp.67-74, April 16-17, 2001.
8. Huan Liu, Hiroshi Motoda, *Feature Extraction, Construction and Selection: A Data Mining Perspective*, Kluwer Academic Publishers, 1998.
9. George F. Luger, *Computation and Intelligence Collected Readings*, AAAI Press / The MIT Press, 1995.
10. Allen Newell and Herbert A. Simon, *Human Problem Solving*, Pretice-Hall, Englewood Cliffs, NJ, 1972.
11. I. Nonaka, *The Knowledge-Creating Company*, Harvard Business Review, 1991.
12. D. W. Patterson, *Introduction to Artificial Intelligence and Expert Systems*, Prentice-Hall, 1990.
13. Emil Post, *Formal Reductions of the General Combinatorial Decision Problem*, American Journal of Mathematics 65, 1943.
14. Wendy B. Rauch-Hindin, *Artificial Intelligence in Business, Science, and Industry Volume I: Fundamentals*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1985.
15. Wendy B. Rauch-Hindin, *Artificial Intelligence in Business, Science, and Industry Volume II: Application*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1985.
16. Wendy B. Rauch-Hindin, *A Guide to Commercial Artificial Intelligence*, Prentice-Hall, Inc., Englewood Cliffs, New Jersey 07632, 1988.

17. *SIGKDD explorations January 2000, Volume 1, Issue 2.*
18. Xue Z. Wang, *Data Mining and Knowledge Discovery for Process Monitoring and Control*, Springer-Verlag London Limited, 1999
19. D. A. Waterman, *A Guide to Expert Systems*, Addison-Wesley, 1986.
20. Christopher Westphal, Teresa Blaxton, *Data Mining Solutions Methods and Tools for Solving Real-World Problems*, John Wiley & Sons, Inc., 1998.
21. Y.-Q. Zhang, M. D. Fraser, R. A. Gagliano and A. Kandel, "Granular Neural Networks for Numerical-Linguistic Data Fusion and Knowledge Discovery," Special Issue on Neural Networks for Data Mining and Knowledge Discovery, IEEE Transactions on Neural Networks, Vol. 11, No. 3, pp.658-667, May, 2000.