IDEA GROUP PUBLISHING
701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com
ITP4544

Knowledge Discovery Process from Sales Data

Katsutoshi Yada
Faculty of Commerce, Kansai University
Yamatecho, Suita, Osaka, 564-8680 Japan,
Email: yada@ipcku.kansai-u.ac.jp
Tel&Fax: +81-6-6368-1121

ABSTRACT

This paper describes the framework of knowledge discovery process in sales data and how the active mining system is applied to the data in the real business world by using the domain knowledge. First the framework of the knowledge discovery process in database is reviewed. It is not clear how users construct actual data mining process and use the domain knowledge in the existing model. We propose two-dimensional matrix of knowledge for sales data analysis to understand knowledge discovery process from purchase history. We distinguish data mining process from creation of business action. We point out that efficient knowledge discovery can be achieved by intensively introducing domain knowledge of experts to the creation of business action.

1. INTRODUCTION

With the propagation of information technology typically represented by Internet, enormous amount of data can be accumulated, and there are now strong interests among researchers and businessmen on the study of data mining (Hamuro, 1998). Despite of the efforts of the analyzers, the knowledge cannot be effectively discovered very often. In Japan, many companies have also been studying the introduction of data mining system, and effective management of business process by data mining is considered to be very important in future.

In cooperative research and study with many firms, we have successfully discovered useful knowledge by data mining (Hamuro, 2001; Ip, 2000; Ip, 2002). In the present article, we try to elucidate the process of knowledge discovery from sales data and to construct a process model for efficient data mining based on these experiences. In the conventional process model, steps of typical data processing are expressed, and it gives no clear explanation as to which kind of knowledge it is converted in the analyzing process or how the domain knowledge should be introduced. By the use of "two-dimensional matrix for type of knowledge", we clearly identify the type of knowledge to be converted in the data mining process and the route of the conversion. We also assert that business action is created from interaction between tacit knowledge and explicit knowledge of the data analyzers and the marketing staffs and that domain knowledge should be efficiently introduced to discovery process.

2. REVIEW OF THE EXISTING STUDIES

Here, we will review the existing studies on knowledge discovery process. The problems in the framework of conventional knowledge discovery process are pointed out, and we will clearly define the primary aim of the present article.

2.1 Framework of Knowledge Discovery Process

Matheus et al. (1993) explained a model of the entire system and its elements along the knowledge discovery process. As major domains, they cited acquisition of data, processing, extraction of pattern, expression of knowledge, and evaluation.

In a narrow sense, data mining is a process to extract patterns between data. In this case, important elements are expression of knowledge, criteria for evaluation, and development of algorithm. The existing study on knowledge discovery process (Valdes-Perez, 1999) puts emphasis on the interaction between the analyzer and the system. The analyzer utilizes knowledge base currently existing in and out of the system such as the analysis in the past or opinions of experts and extracts useful and beneficial rules. In the knowledge discovery in reality, it is important how human action can intervene into the knowledge discovery process (Langrey, 1998), and it is asserted that we must be definitely conscious about the utilization of the introduction of domain knowledge.

Here, we will review the study of Fayyad et al. (1996) on the knowledge discovery process (Fig. 1). According to their study, the starting point of the knowledge discovery is to define the ultimate purpose and to understand the application area and related domain knowledge. Then the data sets necessary for the discovery are accurately defined, and various attribute groups are prepared (preparation of target data). Normally, such data include a plenty of noises, abnormal values, and defective values, and it is a pre-process of these data under a certain rule. Important attributes are estimated and selected from the advance analysis, and the adjusted data set is prepared. By collating these data with the purpose of analysis, the most adequate data mining algorithm is selected. Or, is it a regression model? These problems must be clearly recognized and due consideration must be given on the characteristics of the data set used, and the most appropriate method is selected. And a pattern really interesting is extracted.

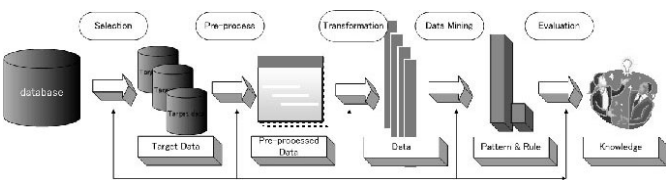
The pattern discovered from the above steps is offered to the person concerned in charge of evaluation. Up to the time when useful meaning will be found, the procedure is turned back to the preceding step, and try-and-error attempt is repeated.

2.2 Problems in the Existing Studies

The framework of conventional knowledge discovery process does not give useful suggestion almost at all for the data mining of the business field (Yada, 2002). Basically, the above model is a general model relating to knowledge discovery, and it is difficult to give sufficient suggestion to a specific problem such as purchase data analysis for the customer.

The following two points are the most important: First, what is really converted in the knowledge discovery process is the part of data with the meaning called "information" or "knowledge". In the process model as given above, it is the part where only data processing is handled.

Figure 1: The existing framework of knowledge discovery process



The knowledge discovery processing is a process where a wide variety of information and knowledge are integrated and are converted to new knowledge (Cowan and Foray, 1997; Tell, 1997). Nevertheless, it is not elucidated what kind of knowledge it is or how it is converted. In the course of knowledge discovery process, we must indicate the route of concrete knowledge conversion and efficiently carry out the discovery process.

The second problem can be summarized as follows: In all of the existing models, it is suggested that the introduction of domain knowledge is indispensable for the discovery of useful knowledge, but none of these models clearly indicate how the domain knowledge should be introduced in reality. Typically, there are many models, which advocate the introduction of domain knowledge to all processes. However, this is practically impossible. It is difficult to obtain suggestion as to at which stage the introduction of domain knowledge leads to more efficient knowledge discovery. For the purpose of efficiently discovering useful knowledge, we must present accurate definite strategy on how the introduction of domain knowledge is to be utilized in the knowledge discovery process.

In the present article, a framework is presented, by which the type of knowledge to be discovered can be classified, and the positioning of analysis to be carried out in the knowledge discovery process is defined. By the use of this framework, it is possible to understand how the discovery process is advanced and also to offer the directivity (principle) of the analysis. Then, a strategy is presented as to how the data of domain knowledge can be used in the knowledge discovery process for the purpose of efficiently utilizing valuable domain knowledge inside and outside of the company.

3. KNOWLEDGE DISCOVERY PROCESS FROM SALES DATA

The purpose of this paper is to develop a framework for giving guidelines for knowledge discovery process useful in the analysis of purchase history of the customers such as POS data with ID. In order to increase the effectiveness of the knowledge discovery process, discussion will be made on how the domain knowledge is introduced.

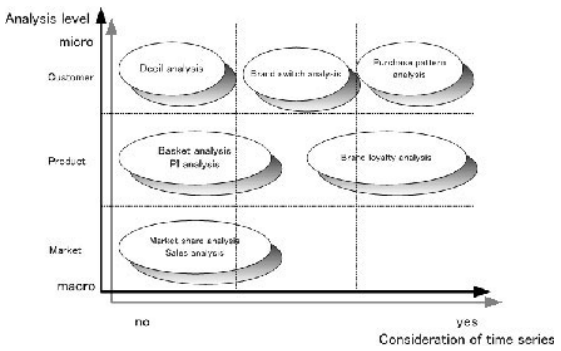
3.1 Two-dimensional Matrix of Knowledge Type Extracted from Purchase History

To evaluate the knowledge discovery process, we focused the attention, not on data processing, but on the meaning of data to be processed and converted, i.e. on information and knowledge.

The data to be processed in the knowledge discovery process from purchase history and the type of knowledge to be discovered can be easily understood if these are expressed in two-dimensional matrix as shown in Fig. 2. The first dimension is the dimension of analysis level. The analysis level ranges from macro-level such as the entire market to micro-level at the position of customers. In actual analysis, the data can be classified according to these levels.

The analysis of the extreme macro-level is the analysis relating to the entire market such as market share, transition of sale, etc. In the more detailed level, basket analysis, brand switch analysis (Berry and Linoff, 1997), etc. are included as the analysis of commodity level. Further, in the analysis of customer level, detailed analysis is conducted such as transition analysis (decil analysis) on the amount of sales proceeds for each customer and purchase pattern analysis (Woolf, 1993; Hawkins, 1999).

Figure 2: Two-dimensional matrix of knowledge type



The second dimension relates to how to handle the time series data in the contents of analysis. In typical POS data with ID, lines of receipts of the customer are accumulated in time series, and purchase behavior of the customer can be identified for a considerably long period. The second viewpoint is how far the changes of situation over the course of time should be considered by the use of these time series data in the analysis.

In a typical POS analysis, the accumulated time series data are converged together and sales proceeds (e.g. sales proceeds for each commodity) are calculated. In such type of analysis, time series data are not effectively utilized almost at all. On the other hand, as the analysis for effectively utilizing the detailed time series data, there are analyses with more emphasis on the changes over time: the analysis of the changes before and after business action, i.e. verification of sales promotion effects such as coupon sale, or an analysis of purchase pattern for long period (Hamuro, 2001).

3.2 Knowledge Discovery Process and Creation of Business Action

1) Knowledge Discovery Process

In any of the cases, the analysis of purchase history is started from left lower portion of the matrix. That is, to acquire basic knowledge of the entire market, it is carried out from basic analysis such as analysis of brand share, total sales amount, etc. In these analyses, the analyzer gives no consideration on the changes over time. The time series data are put together and the tendency in the market is identified. Then, it is turned to the analysis with the time taken into account such as expansion of the entire market, transition of sale of each individual commodity, etc.

Next, as the analyses of commodity level, more detailed analysis of PI (purchase index) value per number of commodities sold or basket analysis is conducted. The former is the calculation of sales proceeds per 1,000 customers who visited the store, and this is also used for the comparison of selling powers of commodities between the stores with different scale. Basket analysis is an analysis to define the characteristics of the commodities purchased at the same time. Many of these analyses are analyses for each commodity in a certain fixed period, and much consideration is not given on time base in most cases. Also, brand loyalty analysis is an analysis to elucidate loyalty and commitment of a consumer to commodity for long period. The change of situations due to time is taken into account in this analysis.

The analysis on the extreme micro-level and complicated analysis based on sales data is an analysis with attention focused on behavior of the customers. For instance, decil analysis is an analysis for customer management. In this analysis, the customers are divided to 10 groups depending on sales amount for each customer in a certain fixed period, and the results are utilized for identification and management of excellent customers for the store. In addition, there are brand switch analysis or purchase pattern analysis for the customers in long period using time series data.

As described above, typical knowledge discovery process is started from the simplest basic analysis on left lower portion of the matrix, and it is advanced to the analysis of customer level by taking more detailed and complicated time base into consideration. Naturally, it is needless to say that the more it is advanced toward time series analysis of the customer level on right upper portion, the higher technical ability and analyzing ability are required. However, it depends on the purpose of project, budget, duration of the project and strategy of the firm as to up to which step the analyses should be performed in the process. Useful knowledge for execution of business action is acquired from a part or all of these processes. Useful knowledge is not necessarily discovered from

the detailed analysis. However, in a typical analytical process, the most detailed analysis is performed, and the results of all analyses are evaluated one after another, and the usefulness of the knowledge is evaluated from the viewpoint of the execution of the business action. If the knowledge is not acquired from these processes, the viewpoint must be changed, and it should be started again from new basic analysis.

2) Knowledge Discovery and Creation of Business Action

Next, we will discuss the knowledge in business and the knowledge creating process. In Western Epistemology, knowledge is considered as "a justified true belief", and attention is focused on explicit knowledge (Nonaka and Takeuchi, 1995). The knowledge handled in existing and our researches relates only to the explicit knowledge. Is it really possible to explain knowledge discovery only from the explicit knowledge?

We believe that it is difficult to explain all of the processes to discover new knowledge using only the explicit knowledge. In the knowledge, there is also tacit knowledge (Polanyi, 1995), which cannot be expressed by words and is difficult to transfer to the other person. New knowledge is born not only from the integration of the explicit knowledge but also from interaction between tacit knowledge and explicit knowledge (Nonaka, 1994; Yada, 1998). In organizational knowledge creating process, Nonaka (1995) defined the process of "socialization", by which two or more people have tacit knowledge in common, and the process of "externalization", by which an individual converts the tacit knowledge shared in common to explicit knowledge and demonstrated that new innovation is developed from the interaction process between explicit knowledge and tacit knowledge in individuals and organizations. In the conventional process model, due consideration is not given to such interaction of knowledge.

We consider that the knowledge expressed by two-dimensional matrix as given above is entirely different from the new business action. The knowledge expressed by 2-dimensional matrix is produced by the existing analytical method or it is born through integration of these knowledge. However, business action is not automatically born from these knowledge. An expert of marketing, who interprets the meaning from these knowledge and who has sufficient tacit knowledge on the market, gives birth to the business action as a new idea. Therefore, we propose a model shown in Fig. 3 as the knowledge discovery process including the new business action.

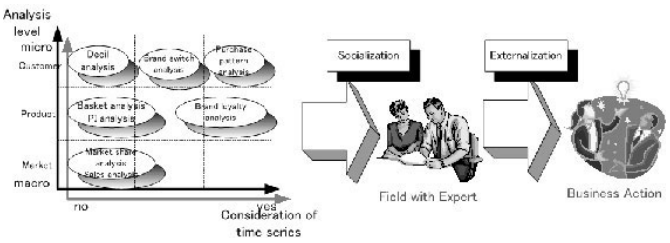
We do not believe that useful patterns and rules obtained from the data are automatically converted to business action. First, in a field where a data analyzer and an expert (on marketing) commonly share the context of analysis results, tacit knowledge acquired in the data mining process is shared in common. This is what Nonaka (1995) called a process of socialization. Next, a person in charge of marketing integrates this with the accumulated experiences, and it is transferred to the process of externalization where new business action is developed. Business action is not automatically given by the data mining system. The patterns and the rules obtained are fused with the existing knowledge, i.e. it is born from interaction of explicit knowledge and tacit knowledge.

3) Introduction of Domain Knowledge

As pointed out in many studies, the introduction of domain knowledge is indispensable for the discovery of the knowledge useful for business. Many of the existing studies recommend that the domain knowledge should be introduced in all steps of the knowledge discovery process, but this is not very efficient and not very practical. Therefore, for the purpose of efficiently introducing domain knowledge of the person in charge of marketing in the knowledge discovery process, it appears to be essential to concentrate the efforts from the scene of internalization to the step of externalization for the creation of business action. We have been making it possible to achieve efficient knowledge discovery by limiting the introduction of domain knowledge to the interpretation of rules, to common sharing of the context of analysis, and to creation of business action.

In order to efficiently introduce domain knowledge, the expert must use the expression easily understandable – in other words, the rules, which can create "a clue" for the new business. In some cases, even when the accuracy of model may be low, suggestion valuable for the marketer may be offered. Osawa (2001) expressed this phenomenon by a concept of "chance discovery" and performed a study from multilateral approaches. In order to introduce domain knowledge more efficiently, a study of "chance discovery" based not only on the extraction of rules but also on the development of business action has important meaning. In future, the application of this type of study to the field of business may offer big business chance.

Figure 3: Knowledge discovery and creation of business action



4. CONCLUSION

In the past, there has been only a framework as general theory for the model of knowledge discovery process, and it has been difficult to give important suggestion to the knowledge discovery in reality. In this paper, we presented a framework with focus on the change of knowledge type in the knowledge discovery process in the data mining from purchase history data. The knowledge discovery process in reality can be understood as the conversion of knowledge on 2-dimensional matrix expressed by analysis level and time base. It is important to distinguish the knowledge obtained from analysis from business action. We pointed out that it is effective to evaluate the patterns and the rules thus obtained and to introduce domain knowledge in the step to create business action.

However, we did not discuss fully on the analysis of the process to create the business action to be actually constructed. In particular, the externalized business action must be integrated well with the existing explicit knowledge. In the present study, we did not deal with the process of combination to full extent. This remains to be the subject of our study in future.

REFERENCES

Berry, M.J.A, and Linoff, G., Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, 1997.

Cowan, R. and Foray, D., "The Economics of Codification and the Diffusion of Knowledge," Industrial and Corporate Change, Vol. 6 No. 3, pp.595-622, 1997.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. "From Data Mining to Knowledge Discovery in Databases," AI Magazine 17, pp.1-34, 1996.

Hamuro, Y., Katoh, N., Matsuda, N., and Yada, K., "Mining Pharmacy Data Helps to Make Profits," Data Mining and Knowledge Discovery, Vol. 2 Issue 4, pp.391-398, December 1998.

Hamuro, Y., Katoh, N. and Yada, K., "Discovering association strength among brand loyalties from purchase history," Proc. 2001 IEEE International Symp. on Industrial Electronics, pp.114-117, Pusan June 2001.

Hawkins, G. E.. Building the Customer Specific Retail Enterprise, Breezy Heights Publishing, 1999.

Ip, E., Yada, K., Hamuro, Y., and Katoh, N., "A Data Mining System for Managing Customer Relationship," Proceedings of the 2000 Americas Conference on Information Systems, pp.101-105, August 2000.

Ip, E., Johnson, J., Yada, K., Hamuro, Y., Katoh, N. and Cheung, S., "A Neural Network Application to Identify High-Value Customer for a Large Retail Store in Japan," Neural Networks in Business: Techniques and Applications, Idea Group Publishing, 2002, pp.55-69.

Langley, P. "The Computer-aided Discovery of Scientific Knowledge, Proc. 1st International Conference on Discovery Science (Lecture Notes in Artificial Intelligence 1532), Springer-Verlag, pp.25-39, 1998.

Matheus, C. J., Chan, P. K., and Piatetsky-Shapiro, G. "Systems for Knowledge Discovery in Databases," IEEE Transaction on Knowledge and Data Engineering, Vol.5, pp.903-913, 1993.

Nonaka, I., "Dynamic Theory of Organizational Knowledge Creation," Organization Science, Vol. 5 No. 1, pp. 14-37, 1994.

Nonaka, I. and Takeuchi, H., The Knowledge-Creating Company, Oxford University Press, 1995.

Osawa, Y., "The Scope of Chance Discovery," New Frontiers in Artificial Intelligence, LNAI 2253, pp.413, 2001.

Polanyi, M., Personal Knowledge –Towards a Post-Critical Philosophy, Routledge and Kegan Paul, 1962.

Tell, F., Knowledge and Justification –Exploring the Knowledge Based Firm, Linkoping University, 1997.

Valdes-Perez, R. E. "Principles of Human Computer Collaboration for Knowledge Discovery," Artificial Intelligence 107, pp.335-346, 1999.

Woolf, B. P., Customer Specific Marketing, Teal Books, 1993.

Yada, K., Katoh, N., Hamuro, Y., and Matsuda, Y., "Customer Profiling Makes Profits: How did a Japanese firm achieve competitive advantage through the knowledge creation?" Proceedings of The Practical Application of Knowledge Management 98, The Practical Application Company, pp.57-66, March 1998.

Yada, K., "The Future Direction of Active Mining in the Business World," Frontiers in Artificial Intelligence and Applications, Vol.79, IOS Press, pp.239-245, 2002.