KNOWLEDGE ACQUISITION AND MAINTENANCE BASED ON DATA MINING WITH PREFERENCES

Yasushi Tanaka[1], Yoshiaki Takano[1], Hiroshi Sugimura[2], and Kazunori Matsumoto[1, 2]
[1]*Department of Information and Computer Sciences, Kanagawa Institute of Technology*
[2]*Course of Information and Computer Sciences, Graduate School of Kanagawa Institute of Technology*
[1, 2]*1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN*

ABSTRACT

This paper is concerned with an information systems that manages knowledge acquisition and maintenance. The importance of knowledge based intelligent system has been well recongnized in most of the industrial areas. It is also well known that the knowledge acquisition process becomes a bottleneck in bulding intelligent systems. Thus the problme is one of providing useful mechanism to acquire effective knowlege in a automatic or semi-automatic manner. Data mining is an expecting technology for this purpose, however, the current methods have several difficulties to be cleared. First, the value of knowledge is relative to a situation, which includes the purpose, the user's expectation, and so on. We assume these information relating to the situation is formalized as preference relations. We thus provide a method that controls data mining process using the given preference relations. Second, the knowledge obtained by data mining may includes some types of fault, or a set of the knowledge as a whole may become inconsistent. We propose a validation mechanism of the extracted knowledge.

KEYWORDS

knowledge acquisition, data mining, time-series data

1. INTRODUCTION

In accordance with the varieties of data and knowledge, various types of information systems are used in industrial areas. In particular, time-series data is the most commonly appeared in so many practical areas such as fluctuations of stock prices, of exchange rates, scientific data, sensing histories in industrial plants, and so on. Although there exists advanced mathematical theories [6] dealing with these data, experimental and heuristic knowledge about them are still important. We therefore have strong requirements for an information system that manages experimental and heuristic knowledge. For this reason, developments of that type of information system are now actively going on. Technologies in artificial intelligence are applicable in such the development, in particular, data mining [6,9] is one of the most expected practice, which aims at an extraction or discovery of useful and unknown knowledge from a large amount of data.

The system proposed in this paper applies data mining, and we here point out two major problems in using the current methods in practice. First, the value of knowledge is relative to a situation, which includes the purpose of knowledge, the user's expectation, and his/her applicable experiences. We thus provide a method that controls data mining process depending on current situation. Second, the knowledge obtained by data mining may includes some types of fault, or a set of the knowledge as a whole may become inconsistent. A validation mechanism of the knowledge is proposed here to solve this problem. This short paper also show an outline of the entire system and a methodology according to which we make the best use of the system.

2. PREFERENCE RELATIONS

In the most cases, we naturally have an explicit preference relation on objects which is being under considerations. The meaning of a preference relation is determined depending on the intension of a user. Take a meeting plan making tool for example; we give the information about member's individual schedules

and the reservation lists of meeting rooms. The tool outputs a list of possible plans in which every members can attend and a room is available. Since each member has own preference on days and on rooms, the plan is required to satisfy every preferences as much as possible. Similar situations frequently occur in the case of data mining. We have a preference relation on pieces of knowledge, and expect a rule which has higher preference is discovered earlier. There is a small study however on a data mining strategy under a preference relation. In this chapter we will discuss more in detail on preferences, and begin with some definitions.

We assume a universe of discourse U is separated into disjoint n categories, the universe is covered with the categories, i.e. $U = C_1 \cup C_2 \cup \cdots C_n$. A preference relation on C_i is any partial order relation $\prec_i$ that is defined over this category, we simply say a preference on C_i, and we omit the subscript. This preference has of course its intended meaning, however, it is not affect the formal discussion here. We naturally extend the relation over U, which is denoted as $\prec$ as the follows. Let $X, Y \in U$ be subsets of the universe, then $X \prec Y$, read as Y is preferable to X, holds if both of the next two conditions holds;

(1) for any $x_i \in X$,there exists an element of $y_j \in Y$ that belong to the same category of X and $x_i \prec y_j$, and

(2) for any $y_j \in Y$, there exists an element of $x_i \in X$ that belong to the same category of Y and $x_i \prec y_j$.

We show an example in Figure 1. As we see in the example, the preference relation is an extension of standard IS-A hierarchy, we therefore can regard it as an ontology on the universe. We further extend the relation over knowledge which is expressed in association rule. An association rule is written in **IF** *X* **THEN** *Y*, where *X* and *Y* are disjoint subsets of the universe. For simplicity, we express this rule as $X \rightarrow Y$. Discovering association rules is an important research issue in data mining. For two rules, $A = X \rightarrow Y$ and $B = V \rightarrow W$, we say *B* is preferable to *A*, denoted as $A \prec B$, if $(X \cup Y) \prec (V \cup W)$ holds. The purpose of a data mining of association rules with preference, hereafter DMAP for short, is to discover more preferable association rules first. We explain a system that carries out DMAP in the next chapter. We conclude this chapter with the proposition, which becomes the mathematical base of DMAP. This assures correctness and powerfulness of DMAP.

Proposition. Let U be a universe of discourse. There is an algorithm that satisfies the followings;

(1) for $X, Y \in U$, is Y outputted first if $X \prec Y$, and

(2) any subset $V \subseteq U$ is always outputted at sometime.

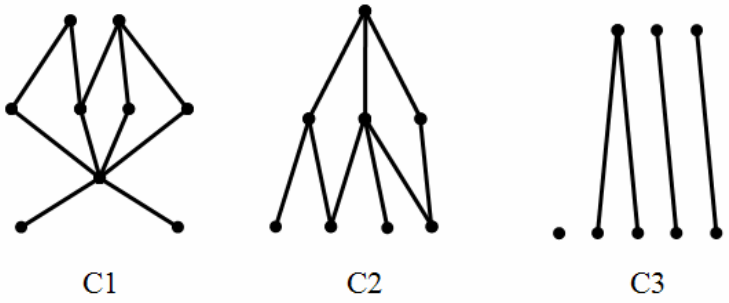

Figure 1. Example of Preference Relation

Ontology as Preference Relation

As we mention in the above, a preference relation can be regarded as a special case of ontology that defines IS-A hierarchies on the universe. For a given IS-A hierarchy, let X be a concept that is more general than the concept Y, this situation is expressed as $Y \prec X$. Under this interpretation, the data mining algorithm of

DMAP outputs the association rules first which consists of more general concepts. Remark here we can easily modify the algorithm running reverse order, which outputs more specific concepts first manner. Other studies [3] propose mining methods in similar settings. A pioneer work [11] provides an efficient algorithm that runs with IS-A hierarchies, however, the purpose is different from DMAP in several points of view. In particular the main aim of it is increasing discovery performance by using general concepts. Then it does not output is a stratified manner, i.e. association rules belonging to different conceptual levels are found at the same time.

3. DMAP SYSTEM

A system named DMAP has an outline shown in Figure 2. Typical flows in the data mining phase are shown by solid arrows, and the dotted lines show flows of storing data. *SOURCE DB* is a database which stores raw data. Preference relations are maintained in *STD PREF* and *PRS PREF*. In general, preference relations become large and hard to manipulate so that a predefined set of standard relations are applicable to many cases. *STD PREF* is a subsystem including database that maintains the standard preference relation. On the other hand, *PRS PREF* deals with personal preferences which belong to each user, thus this maintains a separate database for each user. A user may specify a part of preferences and may use the standard ones for the rest part. *PREF MNG* has a function to mix up different preferences. In this task, we carry out validations of the given definitions and exclude errors. *GEN* step wisely generates a set of concepts with higher preference first manner. *ASOC* Miner is a data mining engine that realize DMAP.

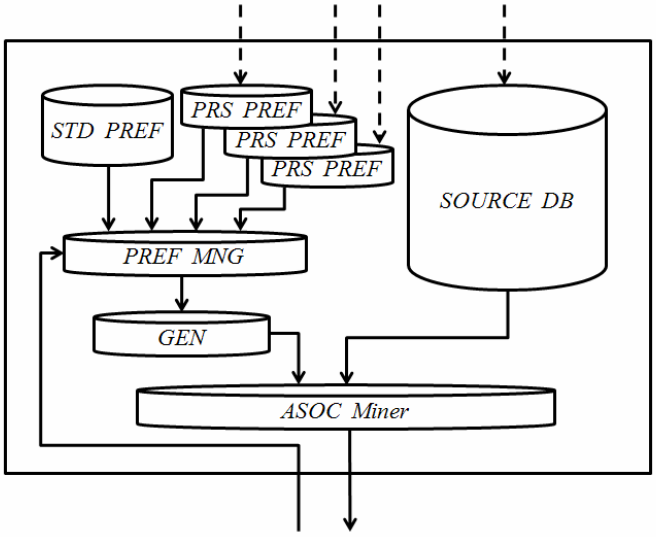

Figure 2. Outline of DMAP

4. EFFECTIVENESS OF DMAP

Data mining is defined as a process [8] that extracts applicable, useful and unknown knowledge. In the standard approach these effectiveness is measured based on numerical values calculated with evaluation formulas. In the case of association rules, the fundamental evaluation is carried out by using the confidence and support [5,7]. Many studies devise other measures [1,2,12], however their effectives depends on properties of databases and aims of mining. Instead of using numerical calculations, the approach in [4,7,10] propose a use of templates. By using the predefined templates, we discard a set of rules which does not satisfy the templates. This method is useful to focus only on the interested part of knowledge, and also useful to increase the efficiency of a mining procedure. The problem is one of defining proper templates that is

applicable to the current situation. This is general becomes a difficult problem, which does not yet have any proper solution. Instead of defining templates, our approach runs with preference relations, which are easy to define, buy a stepwise manner.

Preference relation can be regarded as a type of ontology, which is applicable to increase the ability of data mining. As many studies [8] point out, granularity of knowledge description affects the performance of data mining. Unnecessarily detailed description, that is too small granularity, brings a diversification, and then knowledge extraction often fails. We rough the description by climbing up the hierarchy of ontology and rewriting the description with the upper levels.

5. CONCLUSIONS

This paper presents a concept of information system that can acquire effective knowledge using an approach of data mining. We identify problems in the current technologies of that. In particular, we point out an unfocused use of data mining often cause an explosion of useless knowledge. As a solution for this, we propose a method that depends on preference relations on fragments of information. The intensions, aims, or other kinds of user's preference can be encoded into these relations. Then, data mining process is focused only on the important subset of information. As a result, we can obtain useful and effective knowledge with increased efficiency.

REFERENCES

Carlos J. Alonso Gonzalez, Juan J. Rodriguez Diez, 2004. Boosting Interval-Based Literals: Variable Length and Early Classification. *Data Mining in Time Series Databases*, vol.57, No.7, pp.149-171.

Eamonn Keogh, Selina Chu, David Hart, Michael Pazzani, 2004. Segmenting Time Series: A Survey and Novel Approach. *Data Mining in Time Series Databases*, vol.57, No.1, pp.1-21.

George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, 2008. *Time Series Analysis: Forecasting and Control.* John Wiley & Sons Inc, San Francisco, USA.

Guozhu Dong and Jian Pei, 2007. *Sequence Data Mining*. Springer, New York, USA.

Ian H. Witten, Eibe Frank, 2005. *Data Mining: Practical Machine Learning Tools And Techniques*. Morgan Kaufmann Publishers, San Francisco, USA.

James Douglas Hamilton, 1994. *Time Series Analysis*. Princeton University Press, New Jersey, USA.

Jiawei Han, Micheline Kamber 2001. *Data Mining:Concepts and Techniques*. Morgan Kaufmann Publishers,San Francisco, USA.

M. A. Bramer, 2007. *Principles of Data Mining*. Springer, New York ,USA.

M. G. Elfeky, W. G. Aref, A. K. Elmagarmid, 2005. Periodicity Detection in Time Series Databases. *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, No.7, pp.875-887.

Mark Last, Abraham Kandel, Horst Bunke, 2004. *Data Mining In Time Series Databases*. World Scientific Pub Co Inc, New Jersey, USA.

Michael W. Berry, Murray Browne, 2006. *Lecture Notes in Data Mining.* World Scientific Pub Co Inc, New Jersey, USA.

Xiaoyi Jiang, Horst Bunke, Janos Csirik, 2004. Median Strings: A Review. *Data Mining in Time Series Databases*, vol.57, No.8, pp.173-192.