EMPLOYING EMPIRICAL KNOWLEDGE IN PRACTICE

Yoshiaki Takano[1], Yasushi Tanaka[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 presents a knowledge management system that puts invalidated empirical knowledge into practice. In the most of professional or daily situations, we usually have been acquiring diverse kinds of knowledge. Such one may not have a solid backing theory so that its reliability is hard to estimate, and it thus in many cases ends up up being an informal addition. In the traditional AI technologies, unofficial knowledge is gradually transformed, by a cooperation of domain expert and a knowledge expert, into a set of applicable official knowledge, which is expressed in an adequate AI language. Since this work often needs unacceptable costs, it becomes the bottleneck in the knowledge engineering process. Problems in this process are identified first, and we propose a solution that is based on data mining. Then we also discuss practical issues by using an example in finance domain. We also state the importance of a method that handles ambiguity and reliability of knowledge. The latter part of this paper addresses this issue, and proposes a hypothesis driven method that assists to evaluate the reliability of knowledge and to promote its validation. This method is further improved by a cooperation mechanism with data mining which uses knowledge template.

KEYWORDS

knowledge acquisition, knowledge management, data mining, knowledge reliability

1. INTRODUCTION

We usually have being acquiring knowledge in both professional and daily life. In the most cases, the knowledge has small background theory so that its reliability is low or unknown. For this reason it often stays in informal use. There is a large requirement to convert such the empirical knowledge in an official and validated one. This task can be carried out in the traditional AI technologies [10] by using the expert skills of a knowledge engineer. This manual process is common to the developments of classical expert systems, however it costs too much in general, thus we need to develop a tool that works in semi-automatic manner.

Data mining [7] is a process of discovering and refining useful knowledge from large raw data. The discovered knowledge can be applied to make a prediction, classification, characterization, and so on. A wide variety of techniques are developed [5] to perform data mining depending on the purposes. In the most cases, such a single technique is insufficient and fails to discover useful knowledge from data. To solve this problem, several data mining methods must be cooperatively combined with other technologies. In the case of association rule mining [4], which is a successful subfield of data mining, an explosion of discovered rules becomes a crucial problem. We therefore need a mechanism to focus on a manageable and relatively small portion of the rules. The rule template [9,11] is devised for this purpose, however, it has several unsolved difficulties. We point out here that discovering useful rule templates also becomes a difficult problem. Sometimes it is necessary another exploratory search method for discovering the templates. Then we show the empirical knowledge can be used as a guide in a search for the templates. We further discuss a method that deals with ambiguity and vagueness in knowledge. These ideas are explained in the following.

2. EMPLOYING EMPIRICAL KNOWLEDGE

We show an outline of the system in Figure 1. Empirical knowledge is given to the system through *GUI* and is stored in *EMP KB*. By using knowledge in *EMP KB* and *Mined KB*, *Predictor* makes a prediction about a future situation based on data in *Raw DB* and current real-time data obtained through *I/F*. The prediction is evaluated by a human expert. According to the evaluation result, *Selector* adjusts the reliability grade for each part of knowledge which is used in this process. Flows in data mining are also shown with the arrows in the figure. *DB Miner*, which realizes whole tasks that will be defined in Figure 3, discovers knowledge from *Raw DB*, and stores it in *Mined KB*. In this case, we also use knowledge in *EMP KB*. As we will explain in the following, the top-down data mining requires a guide of the process. *DB Miner* takes a part of knowledge from *EMP KB*, and then uses it as a guide. Note that knowledge in *Mined KB* is also associated with reliability grade, which is adjusted by *Selector*.

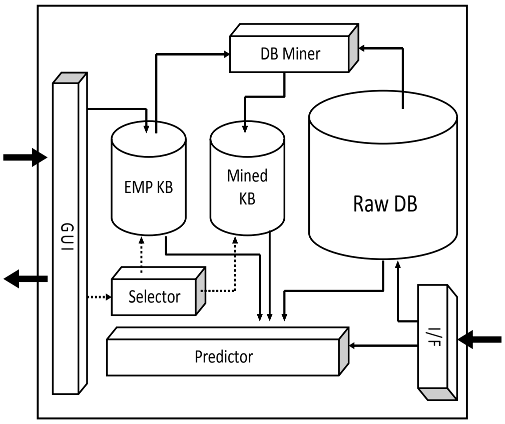

Figure 1. Outline of System

3. EXPERIMENT IN FINANCIAL DOMAIN

An experimental system including essential parts of the proposed approach is implemented, and has been evaluating in the finance area especially for analyzing and predicting stock prices. Diverse kinds of tools and systems have put to practical use, some of them are based on complex theories having advanced mathematical background. On the other hand, there are many tools depending upon traditional knowledge which is acquired socially or personally in experience. Even in introductory articles [1,6], we could obtain nearly several hundreds pieces of knowledge. Patterns of stock charts, such as named Double top, Double bottom, and so on, are small knowledge that say chart behaviors. Elliott's wave theory, Granville's theory and Dow's theory are also well known rules of thumb. All of these are not engineering theories, and thus we could not expect clear descriptions on their applicability conditions, probabilistic reliability, expected returns, and so on. Take Elliott's wave theory for example, which is often explained by using the abstract chart patterns and English as we shown in Figure 2. The patterns appearing in this theory is so abstracted that we cannot expect occurrences perfectly the same ones in real situations. An applicability of this theory depends on how we deal with a similarity among patterns, we therefore this process inevitably includes ambiguity.

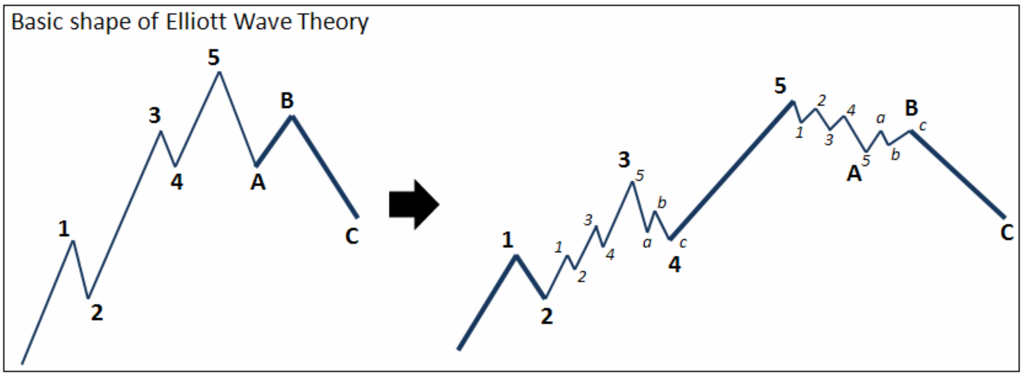

Figure 2. Example of Typical Representation of Knowledge

3.1 Knowledge Representation and Inference with Ambiguity

The knowledge representation with **IF THEN** rules [10] is a base of our approach. The effectiveness of this type of knowledge is well understood and needless to say here. As we point out above, empirical knowledge in finance area requires a method to manipulate ambiguity. The ambiguity becomes twofold: one is relating to the reliability of a rule, and the other is relating to the membership vagueness both in the premise and conclusion parts of a rule. We briefly describe how to deal with these issues.

First, we assume a raw data is defined as a time-series data that consists of numerical stock values. For a given time-series data, any sub-sequence of it can become a pattern. Patterns are regarded as equal ones if they have enough similarities and have within ignorable differences. Note that the comparison should be carried out even the patterns having different lengths. A ten-year pattern may become similar to that of three-week in some cases, for example. The dynamic time warping [8], DTW for short, is a method that computes, under the predefined set of parameters, the similarity measure $s(t_1, t_2)$ for a given pair of time-series data t_1 and t_2, whose length can be different. More similar pair has a smaller value, and it becomes zero if they are exactly the same ones. For a given threshold value σ, we regard t_1 and t_2 are equal if $s(t_1, t_2) < \sigma$. The value of σ is called a maximum allowance, which is closely related to the level of ambiguity in knowledge.

For a rule **IF** *X* **THEN** *Y* with a given maximum allowance, where *X* and *Y* are patterns of time-series data, the premise *X* matches any similar patterns *X'*, and the conclusion *Y* can vary with similar *Y'*. In other words, DTW with the maximum allowance manages the membership ambiguity of patterns. We extend this idea to handle more complex rules, where *X* or *Y* is not a mere pattern but is a sequence of patterns.

3.2 Hypothesis Driven Knowledge Validation

There is no simple solution to deal with another ambiguity relating to the reliability of a rule. Most of the existing AI systems [10] assign a real value between 0 and 1, that is intended to express the degree of a reliability. The rule is logically necessary if its value becomes 1, is conversely a contradiction if the value is 0. This numerical approach has an inherent difficulty in the value assignment task. We adopt entirely different approach based on a unique investigation over the practical applications in finance. We provide a mechanism that assists to infer the reliability of a rule instead of assigning a value to each rule.

A hypothesis *H* is a pattern that is under consideration. Since *H* can be any pattern, it may be a conclusion of some rule, or may be a prediction of future. The problem is one of assisting to judge a degree with which *H* is correct or not. The final judgment of this is a responsibility to a user of *H*. We thus collect and show a set of information that supports *H*, and at the same we also focus on the information that negatively support *H*.

3.3 Cooperation with Data Mining

We in this paper mainly use a method of association rule mining [10], which is extendedly applicable to the case of time-series data [4,11]. An outline of whole data mining process [12,13], which is realized *DB Miner* defined in Figure 1, is shown in Figure 2. The process is not a single task but includes several different tasks. The process can be carried out in either a top-down or a bottom-up approach. In the top-down approach the user defines concepts or strategies according which a direction of the tasks are accomplished. In [3,10], a given set of rule templates become the concepts or rules that control the process. On the other hand, in the bottom-up approach, knowledge discovery runs with relatively uncontrolled manner in the first step, and then an initial concept or rule is identified by the user, who has enough experience in the domain. In the both approaches, these tasks may be repeated until enough knowledge is discovered. Both approaches of course have merits and demerits.

For an empirical knowledge expressed in **IF** *X* **THEN** *Y*, we use this rule with the following operations as a top-down guide of data mining. The first category of the operations is relating a selection of focus points. Since both of the premise *X* and the conclusion *Y* are obtained from experience, they can be regarded as hints of important points to be noteworthy. Sub-patterns of *X* and *Y* are also notable for the same reason. Once we designate a piece of information taken from *X* or *Y*, it becomes the template of a target association rule that is to be discovered. The search in the mining is restricted within the scope including the given information.

Another category includes adjustment of matching against patterns. As we already explained in the above, a given maximum allowance defines the degree of allowable similarities in the matching. By changing the allowance values, we either narrow or magnify the matching, and then the grain of extracted association rules.

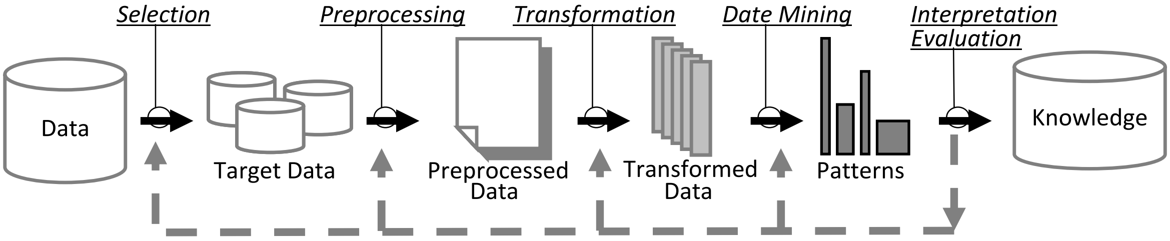

Figure 3. Outline of Data Mining Process

4. CONCLUSION

We propose a system that employs empirical knowledge in practical use. The system is partly implemented in finance domain. In most cases, empirical knowledge lacks clear background theory so that it is difficult to guarantee its reliability. Thus many of it stays only in private use. We first identify the difficulties in dealing with such the empirical ones. We point out that two types of ambiguity are necessary to realize the purpose. Making an assistance of knowledge validation, we devise a hypothesis driven approach, in which we identify and demonstrate a set of supports and a set of anti supports of hypothetical knowledge. Using these sets the user infers the reliability of it. Another important proposal of this paper is a top-down guide of data mining process. To avoid an extraction of explosive number of association rules, existing knowledge is used as a guide of focusing the extraction. The two categories of operators applied on the knowledge narrow or magnify the scope of exploration in the process. Experimental results will be shown in a full paper.

REFERENCES

http://stockcharts.com/

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.

Jean-Marc Adamo, 2001. *Data Mining for Association Rules and Sequential Patterns: Sequential and Parallel Algorithms. Springer*, New York ,USA.

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

J. M. W. Tadion, 1996. *Deciphering the Market: Principles of Chart Reading and Trading Stocks, Commodities, and Currencies*. John Wiley & Sons Ltd, San Francisco, USA.

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

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

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

Nils J. Nilsson, 1998. *ARTIFICIAL INTELLIGENCE: A New Synthesis*. Morgan Kaufmann Publishers, San Francisco, USA.

Ruey S. Tsay, 2005. *Analysis Of Financial Time Series*. Wiley-Interscience, Malden MA 02148, USA.

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, 1996. From Data Mining to Knowledge Discovery: An Overview. *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING*. Vol. 1, No. 1, pp. 1-34

Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, Ramasamy Uthurusamy, 1996. *ADVANCES IN KNOWLEDGE DISCOVERY AND DATA MINING*. Mit Pr, Cambridge, MA 02142-1493, USA.