SIMILARITY BASED TIME-SERIES DATA EXPLORATION

Hiroshi Sugimura and Kazunori Matsumoto
Course of Information and Computer Sciences,
Graduate School of Engineering, Kanagawa Institute of Technology
1030 Shimo-ogino, Atsugi-shi, Kanagawa 243-0292, JAPAN

ABSTRACT

Time-series data has a huge number of applications. We typically start with partially specified ambiguous information, and search for explicit and useful knowledge of fragment patterns. The search task requires an exploratory search a whole database with a long period of trial and error. We then develop an intelligent exploration tool that supports this task. The main contribution of this paper is threefold; first we propose a time-series query language that has an affinity to the standard database query language SQL; second we develop a similarity handling method that effectively realizes a matching of sub-patterns with certain ambiguity; finally we show a combinatory use of the exploration tool and a datamining technique increases the capability of the knowledge discovery.

KEYWORDS

Time series data, exploratory search, user interface, query language, datamining

1. INTRODUCTION

Time series data occurs frequently in business applications and in science [1,2,3,4]. Well-known examples include daily stock prices in the Tokyo stock market, hourly volumes of TV audience-rate rating, and monthly sea surface temperature in the equatorial Pacific. This type of data is typically plotted via a line chart. When analyzing a time series database, such as a collection of historical stock price data or weather information, a frequent goal is to find objects whose series match a given pattern. The search task requires an exploratory search a whole database with a long period of trial and error. We then need an intelligent exploration tool that supports this task. In [5], they propose a method that describes a query to join two patterns. This method specifies the order and the length of two patterns. And entire query is described by simple logical description between a pair of compound queries (AND and OR). We aim to describe entire query more flexibly. For example, a user specifies of the length of the blank between patterns and specifies how often that the pattern is allowed to occur. For this purpose, we combine advantages of the standard SQL and XPath. We also show a method that deals with ambiguity of patterns, and provide a flexible matching among patterns. By using the proposed language with the flexible matching, we semi-automate the explorative search task. In this case, we propose a numerical measure of the usefulness of time-series patterns. We show how the measure is used in the search process.

2. TIME SERIES QUERY LANGUAGE

The standard SQL is widely used as a database query language, which mainly focus on the logical aspects of data. Recently in the community of XML, another query language Xpath becomes an influential position. The both are insufficient in the case of a time-series database. We propose a new language named Temporal Query Language, hereafter *TQL* for short. We show a brief outline of TQL in Table 1 and Table 2. The idea of the path represents logical parts of time-series, and ambiguity is expressed by the wildcards and the quantifiers. A query of TQL is given by a combination of patterns and these operators.

Table 1. Syntax of the PTL of queries

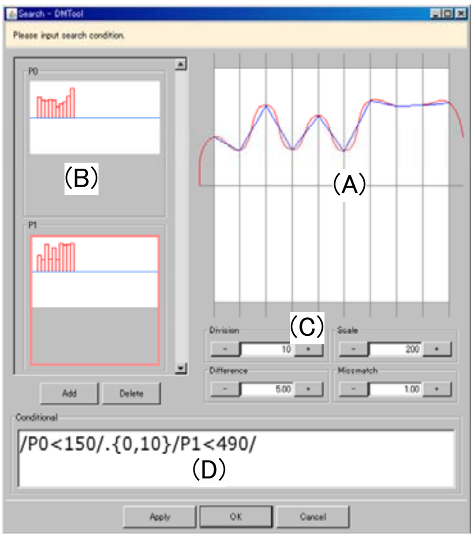
Query ::= Path │ Path Op Query
Path ::= Pattern │ Pattern ″/″ Path
Pattern ::= String │ wildcard
Op ::= ″and″ │ ″or″
Wildcard ::= Refer to Table 2

Table 2. Wildcard and quantification

.	Matches any single data
∗	Zero or more of the preceding element
+	One or more of the preceding element
?	Zero or one of the preceding element
{n,m}	Matches the preceding element at least m and not more than n time

3. IMPLEMENTATION

We show the query dialog in Fig. 1. A user gives a line graph by handwriting into (A). The input graph is stored by the system. (B) displays all stored line graphs. The entire query is represented by TQL into (D), and a search is carried out. For matching between the time series data and a inputted graph, this system uses the Dynamic Time Warping algorithm, DTW for short. DTW is a technique to align two sequences in order to obtain a dissimilarity measure using non-linear temporal alignment. According to several costs and allowance dissimilarity, DTW regard they are equal. These properties for DTW are specified by (C). In Fig. 2 shows the result dialog. (A) displays searched data. (B) is the position (such as index) that (A) displays the current displayed area in entire time series data. (C) represents a total number of searched data.



Please input search condition.

Division 10

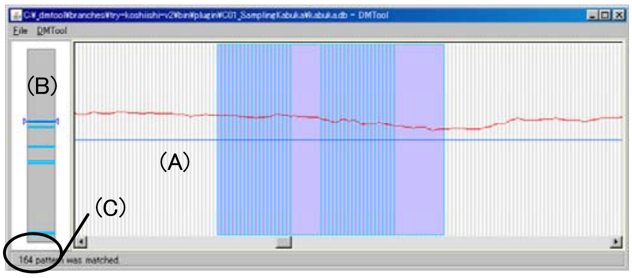
Scale 200

Difference 5.00

Mismatch 1.00

/P0<150/.{0,10}/P1<490/

Figure 1. Query dialog


C:¥_dmtool¥branches¥try-koshiishi-v2¥bin¥plugin¥C01_SamplingKabuka¥kabuka.db - DMTool
File
DMTool
(B)
(A)
(C)

164 pattern was matched

Figure 2. Result dialog

4. CONCLUSION

This paper describes a search tool and query expression for time series data. By using the tool, the knowledge discovery process for time series data becomes easy. We show further information more in detail in the demonstration.

REFERENCES

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

[2] Hiroshi SUGIMURA and Kazunori MATSUMOTO, 2010. *Datamining tool with exploratory search and feature discovery*, Proceedings of the IADIS International Conference Intelligent Systems and Agents 2010 (MCCSIS 2010), pp. 147-150.

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

[4] Shigeaki Sakurai, Youichi Kitahara, et. al. 2008. Discovery of Sequential Patterns Coinciding with Analysts' Interests. Journal of Computers, Vol 3, No 7, pp.1-8.

[5] Haigh, K.Z. and Foslien, W. and Guralnik, V., 2004. *Visual Query Language: Finding patterns in and relationships among time series data*, Proceedings of the seventh Workshop on Mining Scientific and Engineering Datasets.