A System for Retrieving Time-Series Data Based on Linguistic Expression

Naoya Otsuka
Graduate School of Informatics
Kansai University
Osaka, Japan
k589149@kansai-u.ac.jp

Daiki Hasui
Graduate School of Informatics
Kansai University
Osaka, Japan
k234107@kansai-u.ac.jp

Mitsunori Matsushita
Faculty of Informatics
Kansai University
Osaka, Japan
mat@res.kutc.kansai-u.ac.jp

***Abstract*—This paper proposes a system for retrieving time-series data based on a linguistic query given by a user. Our proposed system uses a line chart as a query. The system generates a linguistic query by verbalizing the chart first, and then it retrieves similar charts by using the obtained linguistic query.**

***Keywords*-exploratory data analysis; time-series data; information retrieval**

I. Introduction

In recent years, we have been able to easily obtain large volumes of time-series data from the Internet. These data are useful in finding trends in marketing, in identifying significant patterns prior to fluctuations in the stock exchange market, and in predicting natural disasters.

We find such trends and patterns from the data by combining a computer's ability of handling huge data and a human's ability of finding subtle differences in patterns. In such collaboration, a statistical chart is often used to visualize the data. However, users find it difficult to search a line chart for a graph based on their requirements.

For example, users might want to search for time-series data based on their interests and requirements. Although they have a graph, such as a convex graph, they might not be able to verbalize their thoughts. In such cases, users cannot clarify their request for information. By repeatedly paraphrasing a request, the users' request gradually becomes clearer [1].

When users want to access the time-series data, they submit queries, such as "Which stocks showed characteristic changes around August of last year?" and "Which items showed slow changes compared to other items?" By comparing a variety of time-series data, the users would find data that meets the specified conditions. However, users are required to understand the data set portion where data that share the target characteristics are located. Users must explore a segment of a graph according to their thinking and interest.

The goal of our study is to provide flexible ways to access time-series data by implementing a system that retrieves data that matches a user-specified trend and then narrows the data to a user-specified time period. We extract the required time-series data by matching the user's request against a set of linguistic expressions generated in advance [2]. This approach requires the following three steps: (1) pre-generating linguistic expressions based on the time-series data, (2) interpreting a query represented by a linguistic expression, and (3) matching the query in (2) against the predefined linguistic expressions in (1). In this paper, we focus on (1) and (3) and propose a system for retrieving time-series data using linguistics expressions based on a line chart that illustrates the trends in the data.

In our proposed system, the user selects a time period or a segment of the time-series data from a prepared example graph, and then the system generates a linguistic expression that describes the characteristics of the graph transition within that period. Based on the generated linguistic expressions, users can retrieve a similar graph. In this proposed system, if a user's information request is not clarified at first, they can select a period within the time-series data by describing the shape of the graph or the trend that they want.

II. Related Work

Various methods that use natural language processing have been used to understand a set of numerical information, such as time-series data.

Ahmad et al. and Kobayashi et al. have performed studies on graph analysis [3], [4]. Ahmad et al. proposed a method to generate text based on the characteristics of a graph by extracting an inflection point and the swing cycle of the graph through wavelet analysis [3]. Kobayashi et al. proposed a method to verbalize the relation between multiple time-series data by comparing multiple graphs encoded using SAX (Symbolic Aggregate approXimation) [4].

Kukich and Kobayashi et al. have studied the generation of text from time-series data [5], [6]. Kukich proposed a method to express the overall condition of time-series data by generating and integrating a set of messages from a prepared domain knowledge base [5]. Kobayashi et al. proposed a method to generate a document that describes the pattern of a graph based on the viewpoint given by a user [7], [6].

978-1-4799-2528-5/13 $31.00 © 2013 IEEE
DOI 10.1109/.55

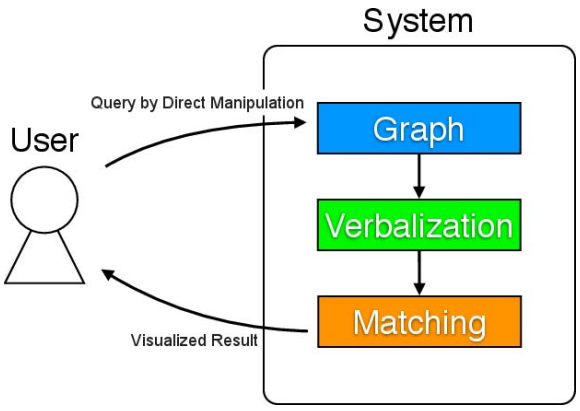

Figure 1. Overview of our proposed system

III. Our Approach

Our proposed system includes a method to generate linguistic expressions that describe a graph and another method to match user-generated expressions with the predefined expressions. Figure 1 shows the overview of the our proposed system.

In our proposed system, the time-series data are described using linguistic expressions that are stored in a database. When a user selects a segment of an example graph, the system describes the trends and characteristics of that segment using linguistic expressions. By matching the linguistics expressions from the user and from the database, the system retrieves the graphs that meet the user's requirements. By iterating such a process, users can access and explore the time-series data.

A. Generating A Linguistic Expression from Time-Series Data

To give a linguistic expression to a graph, we focused on three characteristics: (1) the fluctuation in the graph, (2) the degree of the change, and (3) the general form of the graph. We used three expressions, which are "rising," "declining," and "stable," to represent the fluctuation in the graph.

We determine the expression of the fluctuation using Equation 1. Let $x_t \in X$ be an element at time $t \in T$ in time-series data X, and let t_{start} and t_{end} be the start point and the end point, respectively, of a period selected by the user. If the absolute value of the *fluctuation* is less than one-tenth of $\max(X) - \min(X)$, the fluctuation is determined to be "stable." If the value of the *fluctuation* is positive or negative, the fluctuation is determined to be "rising," or "declining," respectively.

$$fluctuation = x_{t_{end}} - x_{t_{start}} \tag{1}$$

Next, we derive the degree of change using Equation 2. The degree of change denotes the slope between the start point and the end point of the user-specified range. If the

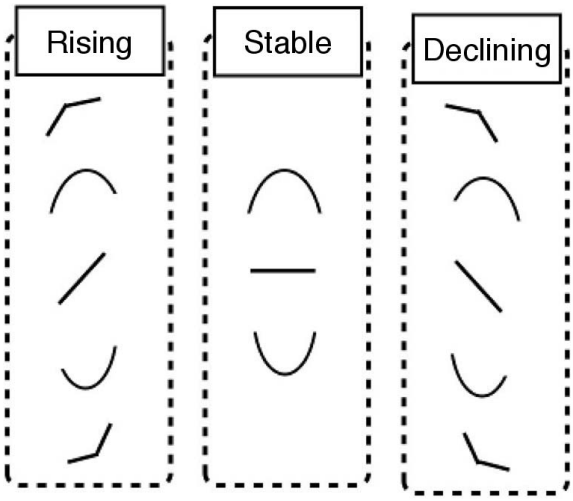

Figure 2. General forms of graph

degree is greater than 2.0, the degree of change is described as "greatly." If the value of *degree* is less than 1.0, the degree of change is described as "minimally." Otherwise, the degree of change is described as "smoothly."

$$degree = \frac{x_{t_{end}} - x_{t_{start}}}{t_{end} - t_{start}} \tag{2}$$

Finally, the general form of the graph is derived based on the positional relations among start point, middle point, and end point. We classified the general forms of the graph into five basic types: "convex upward," "convex downward," "the first half is heavy and the second half is gradual," "the first half is gradual and the second half is heavy," and "linear." We describe the graph pattern or the fluctuation of the graph as "rising," "declining," or "stable." Consequently, the general form of the graph can be classified into one of the 13 patterns shown in Figure 2.

If the fluctuations of the first half and the second half of the graph are different, it is described as "convex upward" or "convex downward." For example, if the first half is "rising" and the second half is "declining," it is described as "convex upward." If the graph is "convex upward," then it is described as "The first half is rising and the second half is declining." If the graph is "convex downward," then it is described as "The first half is declining and the second half is rising." The system generates the same linguistic expression whether the general form of the graph is rising or declining, but the system handles the forms differently. If the fluctuation of the graph is "stable," it is described as "It changes in an upward convex but finally becomes stable." or "It changes in a downward convex but finally becomes stable."

If the fluctuation in half of the graph is the same or only slightly different according to Equation 1, the graph

is described as “The first half is heavy and the second half is gradual.” or “The first half is gradual and the second half is heavy.” In addition, if the graph is “rising” and “the first half is heavy and the second half is gradual,” it is described as “The first half is significantly rising, but the second half is not rising much.” Similarly, if the graph is “rising” and “the first half is gradual and the second half is heavy,” it is described as “The first half is not rising much but the second half is significantly rising.”

If it does not meet any of the above conditions, it is described as a “liner.” For example, if the graph is “rising” and the fluctuation value is high or low, it is described as “It is greatly rising globally.” or “It is minimally rising globally.,” respectively. If the fluctuation is “stable,” it is described as “It is globally stable.”

B. Matching Method

In order to match the user-defined section of a graph with the stored graphs, the proposed system determines the degree of coincidence by using the three characteristics; i.e., (1) the fluctuation in the graph, (2) the degree of the change, and (3) the general form of the graph. First, the system compares the degree of the fluctuation within the graphs to determine if they are both “rising”, “declining”, or “stable”. If they are not the same, the degree of coincidence is determined to be 0%.

Second, if the fluctuations match, the degree of the changes and the general form of the graph account for 60% and 40% , respectively, of the degree of coincidence. When users compare graphs, they emphasize the general form of the graph more than the degree of the changes; therefore, our proposed system assigns a greater weight to the general form of the graph. If both characteristics are completely consistent, then the degree of coincidence is 100%.

If half of the graph is heavy and the other half is gradual, the fluctuations of the graphs are significantly different, whether they are “rising” or “declining”; therefore, the degree of coincidence also considers the fluctuation of the graph. The degree of coincidence is higher between graphs that are described as “rising,” “the first half is heavy and the second half is gradual,” and “convex upward.” Similarly, the degree of coincidence also higher between graphs that are described as “declining,” “the first half is heavy and the second half is gradual,” and “convex downward.” In contrast, the degree of coincidence is lower between graphs that are described as “declining,” “the first half is heavy and the second half is gradual,” and “convex upward.” Consequently, if the general forms of the graph (described as “convex upward,” “convex downward,” “the first half is gradual and the second half is heavy,” “the first half is heavy and the second half is gradual,” or “liner”) are the same and the fluctuations of the graph are different, the degree of coincidence is lower. In addition, the degree of coincidence is lower between “the first half is heavy and the second half is gradual” and “the first half is gradual and the second half is heavy.” Tables I and II show the matching rates based on the degree of changes and the general form of the graph, respectively.

Table I
MATCHING TABLE BASED ON THE DEGREE OF CHANGES

	low	**middle**	**high**
low	**40**	20	10
middle	20	**40**	20
high	10	20	**40**

IV. IMPLEMENTATION

We implemented a prototype system based on our approach.

A. Target Data

We obtained meteorological time-series data from Statistics Japan[1] as the target analysis data. We selected data collected between March 2009 and February 2012 and used it to build a test data set.

B. Design Guideline

In our proposed system, a query given by users is not a linguistic expression but periods and trends in a graph. First, the system generates a linguistic expression for the periods and trends of the time-series data within the range specified by the user. The generated linguistic expression is used as the query applied to the user-specified range, and the system presents matching graphs from the database. In our prototype system, users can select periods and trends of a graph using one of two methods: (1) selecting a graph from a variety of example graphs and their respective periods, or (2) selecting trends and periods separately. Consequently, users can retrieve similar graphs by selecting a specific period of a graph or by selecting periods and trends of a graph.

Second, the system generates a linguistic expression based on an example graph that users selected. In the prototype system, we described a section of graph selected by users using two factors: fluctuations within a graph and the general form of a graph. The fluctuations denote the graph trend, such as “rising,” “declining,” and “stable.” The general form denotes whether a trend is the same pattern, whether the slope of a graph inverts, and whether the degree of increase or decrease changes.

Finally, we developed a method for matching user-selected graphs against target graphs stored on the database. We implemented a function to present the results to users in descending order, according to the matching rate between graphs based on the fluctuations and the general form.

[1]http://www.stat.go.jp/

Table II
MATCHING TABLE BASED ON THE GENERAL FORM OF GRAPH

	convex upward	heavy to gradual (rising)	gradual to heavy (declining)	liner	heavy to gradual (declining)	gradual to heavy (rising)	convex downward
convex upward	**60**	50	50	30	20	20	20
heavy to gradual (rising)	50	**60**	50	30	20	20	20
gradual to heavy (declining)	50	50	**60**	30	20	20	20
liner	30	30	30	**60**	30	30	30
heavy to gradual (declining)	20	20	20	30	**60**	50	50
gradual to heavy (rising)	20	20	20	30	50	**60**	50
convex downward	20	20	20	30	50	50	**60**

Table III
COLOR CORRESPONDING TO THE FLUCTUATION OF GRAPH AND THE DEGREE OF CHANGES

		fluctuation		
		low	**middle**	**high**
degree	**stable**	light yellow	yellow	dark yello
	rising	light red	red	dark red
	declining	light blue	blue	dark blue

C. How to Manipulate

Figures 3 and 4 show the interface of our prototype system. The example graph and the retrieved results are shown the right side and the left side of window, respectively (see left sides of Figures 3 and 4).

Users can change the example graph from the menu on the top left side of the window. By dragging and dropping on the example graph, users can select the range of the graph to search. The range can also be adjusted by dragging the vertical lines that indicate the start or end point of the range, or by using the triangular handles on the lower end point of the lines. The system generates a linguistic expression based on characteristics of the range and displays the text at the bottom of the window (Figure 3). Multiple ranges can be selected by dragging and dropping a range other than those already selected. To make it easier to select overlapping ranges, the handles that denote the start and end points of the ranges can be moved vertically.

The background color between the points is determined by the fluctuation and the degree of changes within the selected range of the graph, as shown in Table III.

The button at the bottom left of the window toggles between two search modes: (1) based on selecting a range from the example graph (Figure 3) and (2) based on selecting a trend and a range (Figure 4). The difference between these modes is whether the user selects a trend by using the example graph or not.

V. CONCLUSION

To provide flexible ways to access time-series data, we considered the approach of extracting specified time-series data by matching users' information requests. In this paper, we proposed a method to generate linguistic expressions from user-selected periods in a line chart. Additionally, we proposed a system to retrieve time-series data based on the linguistic expression. The user provides a graph shape as a query. In order to assign a linguistic expression to the graph shape, we focused on three characteristics: (1) fluctuations in the graph, (2) degrees of changes, and (3) general form of the graph.

To build the prototype system, we used the same methods for generating linguistic expressions and for matching them based on various factors. However, those methods need to be elaborated.

In the future, we will verify the usefulness of our proposed system. Additionally, we will continue to improve the system for to increase the number of ways to access the time-series data.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 22300048.

REFERENCES

[1] F. Hartwig and B. Dearing, *Exploratory data analysis*. SAGE Publications, 1979.

[2] R. Sueyoshi, M. Matsushita, and N. Shirozu, "Generating linguistic expression of charts based on comparison of multiple time-series data," in *Proc. 1st Interactive Information Access and Visual Mining*, 2012, pp. 14–19, in Japanese.

[3] S. Ahmad, P. C. F. de Oliveira, and K. Ahmad, "Summarization of multimodal information," in *Proc. 4th International conference on Language Resources and Evaluation*, 2004, pp. 1049–1052.

[4] M. Kobayashi and I. Kobayashi, "An approach to linguistic summarization based on comparison among multiple time-series data," in *Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on*, 2012, pp. 1100–1103.

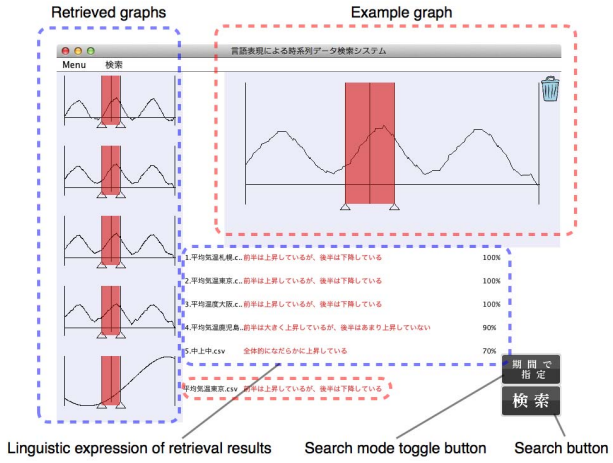

Figure 3. Search mode based on selecting a range from the example graph in the prototype system

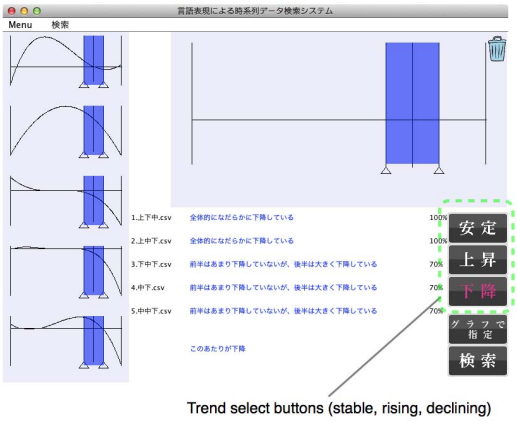

Figure 4. Search mode based on selecting a trend and range in the prototype system

[5] K. Kukich, "Design of a knowledge-based report generator," in *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1983, pp. 145–150.

[6] C. Watanabe and I. Kobayashi, "Intelligent information presentation corresponding to user request based on collaboration between text and 2d charts," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 12, no. 1, pp. 10–15, 2008.

[7] I. Kobayashi, "A study on text generation from non-verbal information on 2d charts," in *Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing*, ser. CICLing '01. Springer-Verlag, 2001, pp. 226–238.