Detecting and Imaging Irregularities in Time-series Data

Qian Zhang
Rolls-Royce@NTU Corporate Lab
Nanyang Technological University
Singapore 637460
zhangqian@ntu.edu.sg

Feng Lin
School of Computer Science and Engineering
Nanyang Technological University
Singapore 639798
asflin@ntu.edu.sg

Hock Soon Seah
School of Computer Science and Engineering
Nanyang Technological University
Singapore 639798
ashsseah@ntu.edu.sg

***Abstract*—Imaging and visual analytics are of great importance for problems that need closely coupled human and machine analysis. In this paper, we propose an interactive system to show irregularities in a time-series dataset. The key technique is a bar-chart-like irregularity plot that gives user a quick insight of the entire time series dataset, with detected status such as normal, missing value, extreme value and possible outlier marked in different colors. The timestamp alignment plot that shows time-related changes and trending information can be used to evaluate patterns and validate automatic detection result in irregularity plot. Technical descriptions of the detection methods and results are presented. Data analysts can benefit from the dataset overview provided by the system before proceeding further data cleansing operations.**

Keywords—imaging and visual analytics; multimedia system; timestamp alignment; irregularity detection; time-series.

I. INTRODUCTION

In the information age, the capacity to collect and store new data grows rapidly, which leads to new challenges in the analysis process. Finding patterns and knowledge hidden in the data becomes a major problem for many analysts, decision makers, engineers and researchers. In "Illuminating the Path" [1], Thomas and Cook define visual analytics as the science of analytical reasoning facilitated by interactive visual interfaces. This emerging field handles the heterogeneous and massive volumes of data by integrating human judgement in the analysis process by means of visual representations and interactions.

However, the real world datasets are usually dirty, with corrupt or inaccurate records as data gathering methods are often loosely controlled. Analyzing data that have not been carefully inspected can result in misleading findings. The phrase "garbage in, garbage out" is particularly applicable to the data analysis process. Data cleansing or pre-processing is thus needed to prepare the dataset before modeling. Much effort has been devoted to this topic and various approaches are proposed for detecting outliers in temporal datasets [2], but to generate an overview of a dataset still requires certain level of programming knowledge to perform simple data manipulations.

In this paper, we propose an interactive system for imaging a raw time-series dataset and providing insights to users. The dataset is first screened to detect possible incomplete or inaccurate records, and the system presents the detected status of data segments as colored blocks. Four possible statuses of the time-series data are identified, i.e. normal, missing value, extreme value and possible outlier. Each detected status is marked with a different color. Detailed information is shown through user interactions. The user can easily discover the start time and end time of the blocks by using tooltips. To provide additional information to the users, the timestamp alignment plot presents the changes over time. It helps users to determine whether the detected statuses are appropriate based on the users' expertise and domain knowledge.

II. IMAGING AND GRAPHICAL REPRESENTATION

Fig. 1 shows the user interface of our system. The user interface (UI) consists of three parts: a table, an irregularity plot with detected statuses, and a timestamp alignment plot presenting the data changes over time. We will introduce the detailed functionalities and interactions of these graphical representations in the following subsections.

A. Irregularity Detection Plot

The motivation behind our proposal of the irregularity detection plot is the grunt daily work of analysts and researchers to visualize and clean a dirty dataset. This plot with a detection algorithm intends to free them from the repetitive coding and provide a quick overview of the time series data.

As shown in Fig. 1, the irregularity plot takes the form of a bar chart, or more precisely a Gantt chart [3], with an x axis of time and a y axis of series names. This representation provides an overview of the dataset quality by presenting the detected statuses of data blocks. The categories of status are also denoted in the legend. In this screenshot, we show one flight instance of an aircraft from the Flight Data Recordings Fuel Efficiency Dataset [4]. A multivariate time series is chosen for demonstration here, i.e., ALT_Mean (mean of samplings of pressure altitude) and FF (fuel flow movement through delivery pipes).

Due to the limited visual space in the plot, detailed information of a detected data segment cannot be displayed. An interactive tooltip is used to address this issue. When a user hovers the cursor over a time block, its status category and detailed starting and ending time will be shown in the tooltip (Fig. 2). This provides additional information to the users for further manipulations such as data cleansing and analysis.

978-1-5386-2615-3/18/$31.00 ©2018 IEEE

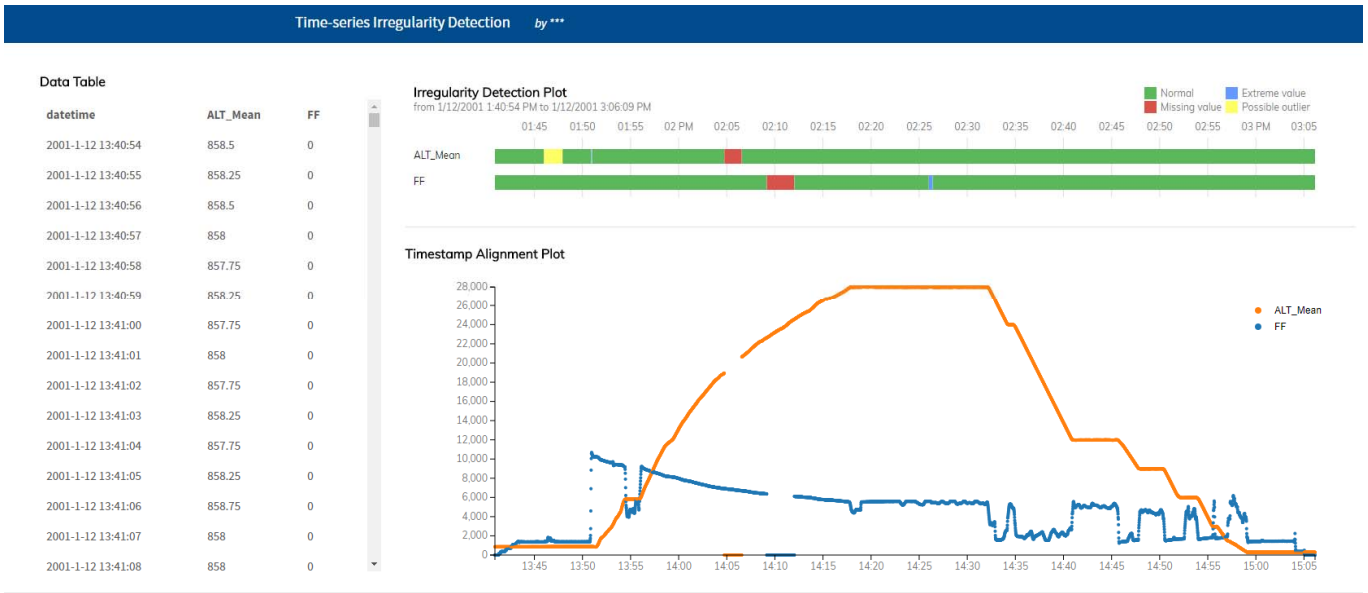

Figure 1 User interface of our imaging system. It consists of three parts, a data table (left), an irregularity detection plot (top right), and a timestamp alignment plot (bottom right). These graphical representations provide users a quick insight of the raw time-series dataset, which can help in further data cleansing process.

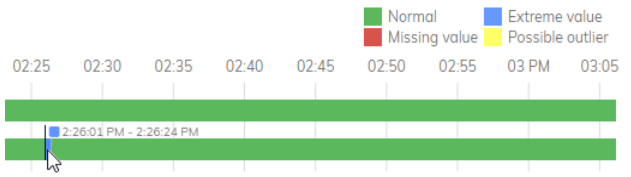

Figure 2 Interactivity of the irregularity detection plot. When a user hovers the cursor over a data segment, a tooltip will show the start and ending time of the time block.

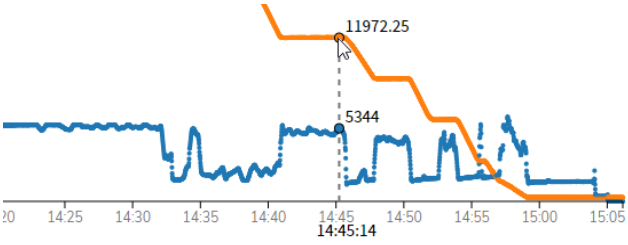

Figure 3 Interactivity of the timestamp alignment plot. A user may hover the cursor over a data point to show the complete record at the same timestamp and detailed information.

B. Timestamp Alignment Plot

Time series are frequently plotted via line charts. However, the timestamps of each data point are not visible in this representation form. In order to clearly show the timestamp of the records as well as the sampling rate, we propose the timestamp alignment plot.

In this plot, the sampled parameter at each timestamp is shown as a dot. Aligned with the irregularity detection plot, this time-series plot displays the data changes over time, which can also assist analysts to critically analyze the detected status of data with their own domain knowledge. Integrating human judgement by means of visual representations and interaction techniques in the analysis process sits at the core of visual analytics.

As shown in Fig. 3, a user may hover the cursor over a data point to display the complete record at this timestamp as well as the detailed information, i.e. the exact value and time.

III. Irregularity Detection

We address the problem of detecting irregularities in the raw time-series data. The term "irregular" depends on the context in which the "regular" or "valid" are defined. It is not realistic to expect an explicit definition or norm for determining the irregularities of different kinds of time-series datasets. We pose the problem of detecting irregularities in the dataset with a purpose of data cleansing and visual analytics.

The system partitions the time series into different blocks and classifies each block into one of the four categories: missing value, extreme value, possible outlier and normal. These categories are proposed according to the data quality criteria [5]. We will discuss three of the categories in the following subsections. The data segments that are not in any of the three categories are regarded as normal.

A. Missing Value

Completeness is the degree to which all required measures are known. It is almost impossible to fix incompleteness accurately with a data cleansing methodology as we cannot infer facts that were not captured when the data in question was initially recorded. The knowledge of the absence of certain measure in a time period is still useful for the modeling step in data analysis.

We want to detect the missing values in the dataset. By "missing" we simply mean null or "not present for whatever reason", whether the entries are left empty or filled by some specific values or numbers. Typical special strings to denote missing data are "NaN", "False", "NULL" or "###", which are often used in the numerical measure columns. Thus, we can detect the data type and filter out the null values.

However, this step is not applicable to the datasets whose missing values are represented by an extreme number beyond the normal data range, such as 999999. Calculating the distributions of the time series will help us to identify those missing values. According to the Chebyshev's inequality, for a wide class of probability distributions, no more than a certain fraction of values can be more than a certain distance from the mean. It means that no more than $1/k^2$ of the distribution's values can be more than k standard deviations away from the mean. Here, we use k=3. Repetitive data values that are more than 3 standard deviations from the mean are regarded as suspicious, which could be a special representation of missing values.

Considering the above two conditions of missing value, the detection can be performed by following steps:

1) We detect the major data type of the measure, and mark the other data entries as missing values. For example, in the Fuel Efficiency Dataset, data samplings of all the measures are numerical, and the strings in the data entry will be marked as missing.

2) After filtering out the not-a-number values in the series, we calculate the distributions of the data. If some of the data values are more than 3 sample standard deviations from the sample mean and the occurrence of these data are higher than expected occurrence, we mark them as missing values.

We illustrate this missing value detection process in Fig. 4. After filtering the empty strings in the dataset, we plot the distribution of FF measure. As shown in the figure, a few data values are more than 3 standard deviations of the mean. The values are 999999. These values are then marked as missing, and will be ignored for other detections like extreme value and possible outliers.

B. Extreme Value

The validity of data is the degree to which the measures conform to defined rules or constraints. Data constraints fall into several categories. Typically, there are range constraints, which mean that numbers should fall within a certain range. In other words, they have minimum and/or maximum permissible values.

We detect the extreme values of the time series. As it requires external source of information about the dataset, we leave it for human users to determine whether these minimum and maximum values are valid. In the irregularity detection plot, the extreme values are marked in blue.

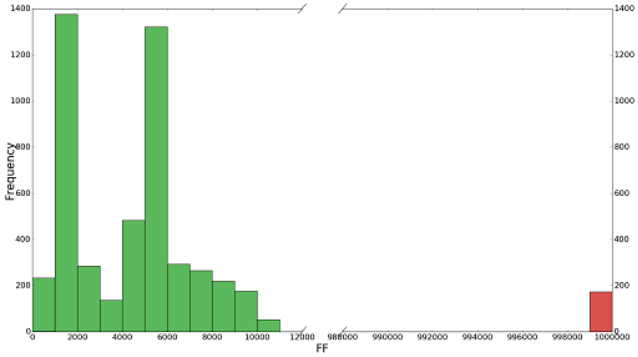

Figure 4 Data distribution of the FF measure from the Fuel Efficiency dataset. Some datasets use a specific number beyond the normal data range to denote missing values. Here, we detect the values of the red rectangle and mark them as missing values in time-series FF.

C. Possible Outlier

Outlier or anomaly detection is an important step in the data cleansing process. In statistics, an outlier is an observation that is distant from other observations, which may be due to variability in the measurement or a measurement error. Statistical methods are often used. By analyzing the data using values of mean, standard deviation, and range, a domain expert may find values that are unexpected indicating possible invalid observations. Thus, here we also use statistical method to detect possible outliers, and leave it for the human users to determine whether they are plausible or not.

We use the moving average and standard deviations to detect possible outliers. Moving average is a calculation to analyze data points by creating a series of averages of different subsets of the full data set. It is also commonly used with time-series data to smooth out short-term fluctuations and highlight longer-term cycles. However, when it is used to estimate the underlying trend in a time series, the moving average is susceptible to rare events such as rapid shocks or other anomalies. It is optimal when the fluctuations about the trend are normally distributed [6], without very large deviations from the mean, which is not the case in fuel flow data. This makes it less suitable for outlier detection. Thus, we combine the methods of moving average and moving standard deviations for detecting outliers.

We take an equal number of data on either side of a central value, and calculate the unweighted mean and standard deviation of these data within the window. This ensures that variations in the mean and standard deviation are aligned with the variations in the data rather than being shifted in time.

As shown in Fig. 5, we use this algorithm to detect possible outliers. The time series of measured FF from the Fuel Efficiency Data is drawn in the figure. In this paper, we use a

window size of 300 data points. We set the limit of normal values of the data as within 3 window standard deviations from the window mean. Thus, the data segment highlighted by the red rectangle is regarded as a possible outlier.

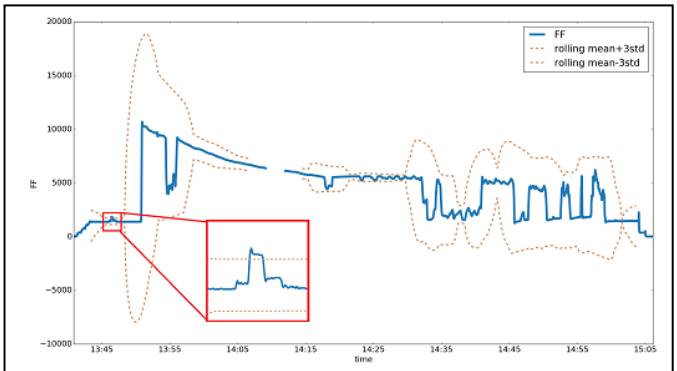

Figure 5 Outlier detection using moving average and standard deviation. We can see that a data segment is more than 3 window standard deviations from the window mean (highlighted in red rectangles). This segment is regarded as a possible outlier in the time-series data.

IV. Conclusion

We propose an imaging and graphical representation system for irregularity detection of the time-series data. The new visual representations, along with the automatic detection algorithm, provide a quick overview of the entire dataset for further data manipulations. This multimedia system with generated images, text and user interactions will benefit data analysts and researchers and free them from repetitive coding needed to generate charts. Future work on this system may be to integrate more interactions which allow the user to manipulate the data right away when they see the detection results.

Acknowledgments

This work was conducted within the Rolls-Royce@NTU Corporate Lab with support from the National Research Foundation (NRF) Singapore under the Corp Lab@University Scheme.

References

[1] Cook, Kristin A., and James J. Thomas. "Illuminating the path: The research and development agenda for visual analytics." (2005).

[2] Gupta, Manish, et al. "Outlier detection for temporal data: A survey." IEEE Transactions on Knowledge and Data Engineering 26.9 (2014): 2250-2267.

[3] Wilson, James M. "Gantt charts: A centenary appreciation." European Journal of Operational Research 149.2 (2003): 430-437.

[4] CrowdANALYTIX, “What makes airplanes fuel efficient?”, https://crowdanalytix.com/contests/what-makes-airplanes-fuel-efficient-

[5] Müller, Heiko, and Johann-Christph Freytag. “Problems, methods, and challenges in comprehensive data cleansing.” Professoren des Inst. Für Informatik, 2005.

[6] Arce, Gonzalo R. “Nonlinear signal processing: a statistical approach.” John Wiley & Sons, 2005.