Automatic data mining for telemetry database of computer systems

C.-H. Wu a, C.-H. Yang a,⁎, S.-C. Lo b, N. Vichare c, E. Rhem c, M. Pecht d,e

a Graduate Institute of Mechanical and Electrical Engineering, National Taipei University of Technology, Taipei, Taiwan
b Graduate Institute of Information and Logistics Management, National Taipei University of Technology, Taipei, Taiwan
c Dell Client Systems Reliability Engineering, Dell Computer Corp., Austin, TX, USA
d Center for Advanced Life Cycle Engineering (CALCE), University of Maryland, College Park, MD 20742, USA
e Prognostics and Health Management Center, City University of Hong Kong, China

A R T I C L E   I N F O

Article history:
Received 3 April 2010
Accepted 13 September 2010
Available online 12 October 2010

A B S T R A C T

This study focuses on the development of automation platform for performing data mining on a telemetry database of computer systems. It is common for computer systems to encounter failures in an unexpected manner. It is therefore valuable to have prognostics capability for computer systems to minimize the effects of unexpected system failure. Data acquisition schemes employing telemetry techniques are considered the most effective method for collection of in-service information for computer systems. Analysis of an enormous telemetry database of high complexity must be completed before useful knowledge can be extracted. In this research, an automatic data mining platform is reported for the extraction of useful knowledge from the telemetry database. This paper describes the structure and basic theories underlying the data mining of the telemetry database. Also, an automatic computer program capable of performing database management, filtering, data analysis, and reporting is described. Some useful data generated by the platform are reported for the telemetry database.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Computer systems commonly encounter various failures due to electrical component or system issues under different service conditions. Users sometimes lose data or experience un-availability at critical moments. More seriously, critical business users may suffer tremendous financial loss from unpredicted failure. From the manufacturer's point of view, the reliability of a computer system is directly related to its specifications and leads to the customers’ loyalty to the product. Good understanding and control of computer systems’ reliability is fundamental and crucial to computer system manufacturers. In general, environmental factors and the quality of the computers’ parts contribute to the failure of computer systems. For manufacturers, in-service environmental factors cannot be avoided. However, the quality of individual components in each system can be evaluated and controlled. In computer manufacturing, it is therefore valuable to have prognostics capability for computer systems for the minimization of unexpected failures.

To collect in-service information from computer systems for health monitoring, data collection methods employing remote telemetry are regarded as very effective. Often, the telemetry task generates a very large database that has complex data types as well. With the data collected from a variety of sensors in field applications, the telemetry database will inevitably contain faulty data information. Therefore, analysis of the telemetry database is another important task following the telemetry work. Fig. 1 describes the typical mythology [1] regarding the acquisition and analysis of the telemetry database. This methodology uses the measured data and traditional statistics to develop a physics-based model to predict the tendency of parameters in the future. Also, it must be noted that factors leading to the failure of computer systems are multi-dimensional and need multi-variate analysis.

Kumar et al. [2] presented an approach for selecting important factors for health management of a computer server. Since failure is a multi-dimensional problem, their study uses principal component analysis (PCA) approach to reduce the dimension of observed parameters to several available parameters. In another study, Kumar and Pecht [3] discussed and characterized the performance of a computer system in various environmental and usage conditions. They discuss a method for baselining a computer system by considering different factors and demonstrate empirical equations to compare the temperature, humidity, and usage of different components. Begin new paragraph, Gu et al. [4] developed a method called sum of the loading cycle range (SLCR) to build relationships between field usage data and lab test data. From these relationships they can make the test data from the test process more meaningful. Vichare et al. [5] monitored the temperature variation during different power cycles, usage history, and CPU computing. They utilized traditional statistics to analyze the telemetry data and found the relationships that are applied in damage estimation and remaining life predictions. Begin new paragraph, C.-H. Yang presented an approach for selecting important factors for health management of a computer server. Since failure is a multi-dimensional problem, their study uses principal component analysis (PCA) approach to reduce the dimension of observed parameters to several available parameters. In another study, Kumar and Pecht [3] discussed and characterized the performance of a computer system in various environmental and usage conditions. They discuss a method for baselining a computer system by considering different factors and demonstrate empirical equations to compare the temperature, humidity, and usage of different components. Begin new paragraph, Gu et al. [4] developed a method called sum of the loading cycle range (SLCR) to build relationships between field usage data and lab test data. From these relationships they can make the test data from the test process more meaningful. Vichare et al. [5] monitored the temperature variation during different power cycles, usage history, and CPU computing. They utilized traditional statistics to analyze the telemetry data and found the relationships that are applied in damage estimation and remaining life predictions.

⁎ Corresponding author. Tel.: +886 2 27712171.
E-mail address: cyang@ntut.edu.tw (C.-H. Yang).

Available online 12 October 2010
paragraph, in the area of multi-variante analysis of computer systems, there are many studies focused on the isolation and identification of faulty parameters in a computer system. In traditional statistics analysis, principal components analysis is used to reduce the dimensions of parameters. Kumar et al. [6,7] used Mahalanobis distance (MD) to isolate the faulty parameters from the observed data and determined the health condition of a computer system. They showed that MD is a good health index for measuring multi-variante monitored parameters. Many studies have damage models by analysis of the collected data and have developed prognostics models. Gu and Pecht [8] developed a physical of failure (PoF) damaged model based on telemetry data that provides information about the product’s life cycle in different loading conditions. Gu improved this approach to have higher reliability than the traditional reliability methods. According to fundamental theories, such as hypotheses testing, Lopez [9] developed a Continu- ous System Telemetry Harness (CSTH) method which integrates the Sequential Probability Ratio Test (SPRT) and Multivariate State Estimation Technique (MSET) algorithms.

Although there are many studies focused on data processing for electronic products, there are few studies discussing the analysis structure or the automatic platform for the analysis of a telemetry database. This paper describes the structure and basic theories underlying the data mining of a telemetry database. Also, an automatic computer program capable of performing database management, data analysis, and reporting is described. Some useful highlights generated by the platform are reported for the telemetry database.

2. Computer data mining platform

With the large and complex nature of the telemetry database, extensive analysis of the database is demanded before useful knowledge can be extracted from the database. It is therefore an important task to develop a data analysis protocol with accompanying enhanced analytical tools. To accomplish this task, an automatic data mining platform is developed to extract useful knowledge from a large and complex telemetry database.

Fig. 2 shows the structure of this automatic data mining platform. The kernel of this platform is a computer program designed to manage the automatic data mining process. The backbone functions of the platform include pre-processing and filtering of the database, analysis computation, and generation of data mining reports. The database task is responsible for automatic database functions such as querying, sorting, and temporary storage. The statistics task interfaces with external third-party statistical analysis functions for the purpose of conducting robust statistical analysis. The main functions for the developed automatic platform include database management, pre-processing of the database for analysis, data analysis, and analysis reporting.

3. Database and pre-processing

The database management functions are responsible for retrieving and querying data from the telemetry database stored in the MS-SQL format. The telemetry data is collected from a large number of dispersed users. The users own different computer systems with a large variety of hardware configurations, such as CPU, hard disks, and memory, which are called hardware parameters. Therefore, the telemetry database has a large size and very complex nature. Each user can have multiple records, and each record contains hundreds of different data with various data types. Also, the data records have uneven length, uneven spacing between records, complex types of data (e.g., real numbers, integers, Boolean numbers, and text) and missing or out-of-range values. Fig. 3 shows part of the database. As shown in Fig. 3, the database is stacked up with multiple records uploaded from different users at multiple times. Each record has very complex data types, including text description integers, Boolean, real numbers, etc. Therefore, the development of a data pre-processing module is a serious and challenging issue. In fact, the pre-processing task becomes the most critical part of the entire data mining platform for the telemetry database.

The platform is capable of performing question-orientated analysis (QOA). In QOA a targeted question is assumed by the user. According to the QOA assumption, the platform guides the user to prepare the data from the telemetry database and perform the desired analysis. Fig. 4 shows the structure of the data filtering task in the data mining platform. By means of a multi-staged filtering function, the data management task is responsible for the pre-processing task. The smart filter will prepare valid data for analysis. With the targeted analysis question, the program prepares a suitable subset of the database for subsequent analysis. To filter out the discussed topic is an important step before the analysis. The developed filter function allows the platform to answer certain types of questions for specific situations. The filter can be used, for example, to test various hypotheses asked in the form of a question, such as what is the difference in correlation between two groups of users? How does the computer system model affect the correlation between a select numbers of parameters of interest? What are the dominant parameters affecting a response parameter? How different are the dominant parameters between computer system models?
With the filter function, the datasets can be used for further analysis. An integral part of the analysis lies in the processing of the filtered datasets. After a filter is applied to the database, four categories of algorithms can be used for analysis: (1) descriptive statistics; (2) groupings, clusters, and similarities; (3) factors influencing the response parameter; and (4) correlation changes. Each of these is reserved for a specific type of analysis. Descriptive statistics can be provided for specific filter configurations. The statistics provided here aim to summarize and present various aspects and characteristics of the data under filter configurations. Regression analysis of factors influencing the response parameter is mostly focused on general linear regression, the Loess method, and logistic regression to accommodate binary and categorical response variables. Groupings, clusters, and similarities analysis investigates clusters and groups in multi-variate data and tries to determine the similarity between groups, even in different filter configurations. The search for groupings and clusters will be approached primarily through non-parametric search techniques such as K-means algorithms. Hypothesis tests are used to test the similarities between groups. Correlation matrices are used to measure the linear association between variables for specified filter configurations. The analysis also determines the correlation between parameters by considering their frequency signatures. The primary technique for this analysis relies on coherence analysis [10]. With coherence analysis the user can correlate the data generated at different periods of time with a particular response variable of interest under particular filter configurations.

4. Highlights of platform performance

Highlights of the performance of the automatic data mining platform developed for the telemetry database are presented below. The results are presented in four categories of algorithms including: (1) descriptive statistics, (2) groupings, clusters and similarities, (3) factors influencing the response parameter, and (4) correlation changes. The platform performance is described as the follows.

4.1. Descriptive statistics

The central tendency and variability characteristics are shown in the following case. This example involves two computer components and six variables in the telemetry database. The targeted question is to analyze the similarities of these two devices based on the six variables. The variables are named A1–A6. Analysis results from the platform are shown in Fig. 5. The central tendency and variability characteristics for the six variables of the two devices are automatically computed and reported by the platform. Visual representations using boxplots are useful for describing the similarities between the six parameters of the two devices. However, there is some uncertainty and inaccuracy associated with the boxplot visual representation. Other statistical methods such as hypothesis testing can play a complementary role in the boxplot description.

The platform uses the Shapiro–Wilk normality test to test the normality of the targeted datasets. The t-test is used to determine the average characteristic similarity between observations. The Wilcoxon Rank Sum test is used to compare central tendency similarities. The Kolmogorov Smirnoff test is used to test the similarities in terms of probability distributions. The three hypothesis test methods are good for judging samples that are similar or not in different phases.

4.2. Groupings, clusters and similarities

For the grouping purpose, principal component analysis (PCA) was used to reduce the dimensions of the variables. The results of the PCA analysis are shown in Table 1 for the eigen-analysis results, Table 2 for the PCA loadings and Fig. 6 for score and loading plot. It is shown that all of the observed parameters, A1–A5, had the same characteristics in the PC1 dimension. However, the characteristics were obviously different in the PC2 dimension. A2 and A4 had the same characteristics in the PC2 dimension. The other parameters were opposite. For the PCA results in Tables 1 and 2, the five variables from A1–A5 were reduced to two major variables of PC1 and PC2.

K-means analysis was then used to analyze the grouping behavior of the PCA variables. In Fig. 7, it can be seen that the groupings for PC1 vs. PC2 and PC1 vs. PC3 are clearly separated for device 1 and device 2. The measured information is more different in PC1 vs. PC2 than in PC1 vs. PC3. The values of PC1 for device 2 are higher than those for device 1. It is also shown that there are no differences in PC2 and PC3 for these two devices.
4.3. Factors influencing the response parameter

In the regression analysis, the bi-plot includes the scatter and regression curve plots of A2 vs. A4, as shown in Fig. 8, for the two devices. The slope coefficients for the two devices were also compared. The fact that the slope of device 2 is larger than that of device 1 reveals that the A4 parameter contributes more...
towards A2 for device 1. And for these two plots, device 1 shows higher variability characteristics than device 2. Stepwise regression is also included for the purpose of eliminating variables that are not meaningful and enhancing the meaningful variables in a regression model. Logistic regression is a statistics classified method. It is used to predict or classify the probability of events by fitting observed data to a logistic curve. As shown in Fig. 9, the logistic regression makes use of a predictor and is used to identify the two devices based on the B1 value of 4976 corresponding to the p value of 0.5 in the logistic regression.

4.4. Correlation changes

The correlation analysis in this platform provides the information that a filter configuration affects the correlation across users, groups, times, and systems. In spatial correlation analysis, the correlation coefficients of two parameters are computed under different filter configurations. Further, the relationship of parameters changes per device and for different periods can also be computed in the frequency domain. In the correlation analysis, the coherence analysis is used to investigate the frequency behavior of the correlations changing over time. Fig. 10 shows the results of the coherence analysis for different filter configurations over a time period of 8 days.

5. Conclusions

Reliability issues in computer systems are being increasingly recognized for more than just their effects on cost and efficiency. Although telemetry techniques are effective in acquiring the in-service data of computer systems, the enormous telemetry database with high complexity still calls for the development of effective analysis protocols and application tools. In this research, an automatic data mining platform extracts useful knowledge from
the telemetry database of a computer system. The data mining protocol for the analysis of a computer telemetry database is developed with integrated functions including (a) access to the large-scale telemetry database, (b) user-defined data filtering and preprocessing, and (c) performing data mining functions. The platform has now been extensively evaluated for its functions and robustness for field applications. Also, this platform has high flexibility for applications to other fields needing analysis of complex databases.

Acknowledgement

This work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU8/CRF/09).

References


