ANALYSIS METHODS OF WORKFLOW EXECUTION FOR DATA BASED ON ROUGH SET THEORY IN DATA MINING

Anoop Shrivastava: Aditya College of Technology & Science,Satna(M.P.)

Abstract:

Exploring the trivial workflow data needs high performance data processing technology. In this research work we put forward analysis method of workflow execution data based on data mining. The main idea of it is to retrieve the workflow data to a data warehouse and adopt OLAP technology and data mining method to support customers to select different measures and view the corresponding data in different dimensions and different abstract levels, which is important for them to make decision. This research work presents the use of a relatively new method, the Rough Set (RS) theory for knowledge acquisition in time sequence condition monitoring. An additional attraction of the RS theory is that it allows automated generation of knowledge models, offering clear explanations to the inferences performed in diagnosis.

I. Introduction

Effective enterprise management and decision cannot leave corrective evaluation and analysis for enterprise operation performance. Because operation process has integrated these factors of human, resource, application, and operation and reflected directly management capacity and state of enterprise, so, a majority of evaluation system are all developed surrounding enterprise operation process. Workflow management system provides valuable data source for teal operation evaluation as information system of operation process definition, execution and management. These data record all process and execution locus in system, including human, resource, application and operation logic, and so on. These data is true and real-time renewed relative to simulation data. At the same time, information organization mode making operation process as core of workflow system provides ideal structure for integrating another information source and evaluating operation performance analysis from different point of view. So, workflow execution data is important component part of enterprise operation evaluation system.

II. Analysis Method of Workflow Execution Data

A. Data Warehouse Technology

A data warehouse is a specialized type of database. More specifically, a data warehouse is a "repository of information gathered from multiple sources, stored under a unified schema, at a single site". A data warehouse should be separately maintained from the organization's operational database since the functional and performance requirements of online analytical. Data warehousing as a process of organizing the storage of large, multivariate data sets in a way that facilitates the retrieval of information for analytic purposes.

In present day, more and more enterprises realize to sufficiently adopt only and mine present data in order to realize the best enterprise benefits. Most of enterprise is not lack the data other than redundancy of data and disaccord. Traditional application system of database faces the design of operation work. It reduces working intensity of idiographic manipulator, but the leader of enterprise is not possesses relevant system. Enterprise needs new technology to compensate insufficiency of former data system and need to integrate data gathered widely to data warehouse in order to pick up useful information and help them make a timely and corrective judge in operation management and development. Basic system structure of data warehouse is shown as figure 1.

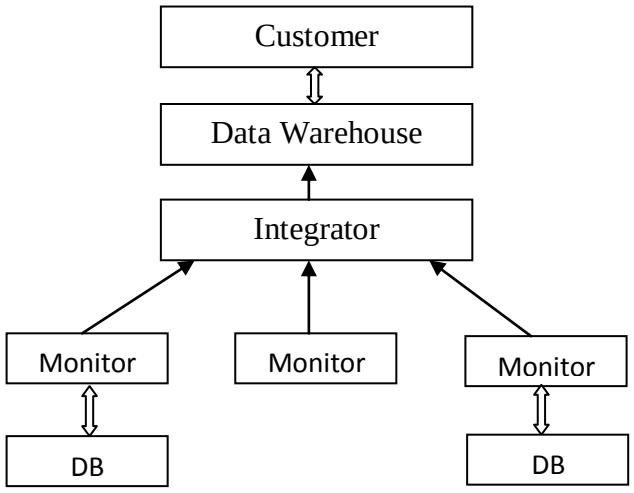

Figure 1: Basic System Structure of Data Warehouse

The aim of data warehouse is to build a systemic data storage environment and separate plentiful data needed by analysis and decision-making from traditional operation condition. It makes transfer dispersive and disaccord operation data to integrated and uniform information. The member of enterprise in different branch can find new views and question new analysis and idea through applying inner data and information in this single condition, and accordingly develop institutional decision system to gain more management benefit. In the recent decades, with the development and application of database technology, people try to reprocess data in database and form a synthetic and face-to-

analysis condition so that it supports best decision analysis and form data warehouse technology. Data warehouse system includes warehouse technology and on-line analytical processing (OLAP) technology and data mining technology (DM).

The primary concept of data warehousing is that the data stored for business analysis can most effectively be accessed by separating it from the data in the operational systems. Many of the reasons for this, separation have evolved over the years. In the past, legacy systems archived data onto tapes as it became inactive and many analysis reports ran from these tapes or mirror data sources to minimize the performance impact on the operational systems.

These reasons to separate the operational data from analysis data have not significantly changed with the evolution of the data warehousing systems, except that now they are considered more formally during the data warehouse building process. Advances in technology and changes in the nature of business have made many of the business analysis processes much more complex and sophisticated. In addition to producing standard reports, today's data warehousing systems support very sophisticated online analysis including multi-dimensional analysis.

Data warehousing systems are most successful when data can be combined from more than one operational system. When the data needs to be brought together from more than one source application, it is natural that this integration be done at a place independent of the source applications. Before the evolution of structured data warehouses, analysts in many instances would combine data extracted from more than one operational system into a single spreadsheet or a database. The data warehouse may very effectively combine data from multiple source applications such as sales, marketing, finance, and production. Many large data warehouse architectures allow for the source applications to be integrated into the data warehouse incrementally.

B. OLAP

OLAP is an acronym for On Line Analytic Processing. It is an approach to quickly provide the answer to analytical queries that are dimensional in nature. It is a part of the broader category business intelligence, which also includes Extract Transform Load (ETL), relational reporting and data mining. The term On-Line Analytic Processing - OLAP (or Fast Analysis of Shared Multidimensional Information - FASMI) refers to technology that allows users of multidimensional databases to generate on-line descriptive or comparative summaries ("views") of data and other analytic queries.

OLAP supports queries and data analysis on aggregated databases built in data warehouses. It is a system for collecting, managing, processing and presenting multidimensional data for analysis and management purposes (Figure 2).

Another way, OLAP is a query technology for data warehouse. It applies specially to analyze fast large numbers of data. OLAP is a concept of user interface rather than a data storage technology. OLAP can make various analysis action chipping slice, chipping piece, rotating for data of data warehouse and make user observe data of data warehouse from multi-angle and multi-side, consequently, know deeply information and connotation in data. OLAP includes multidimensional OLAP (MOLAP), relational OLAP (ROLAP) and hybrid OLAP (HOLAP). Data in MOLAP is stored in mode of multi-dimensional that its query efficiency is good, but its loading time is long. Data in ROLAP is stored still in relational database that its query efficiency is relatively low, but its loading time is short. HOLAP is eclectic method.

OLAP systems have been traditionally categorized using the following taxonomy:

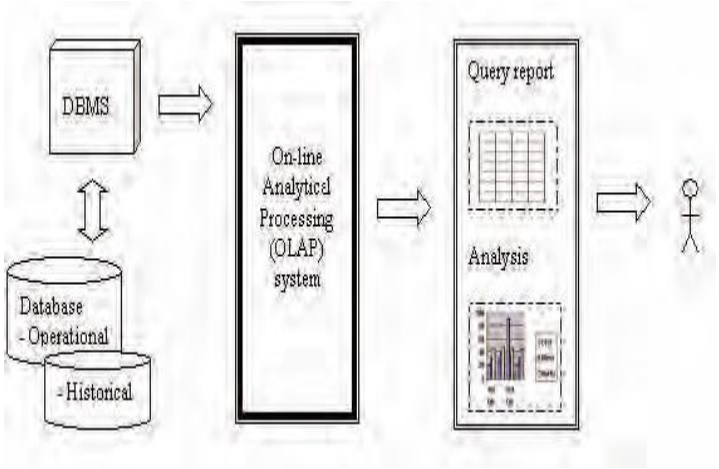

Figure 2: OLAP System

MOLAP (Multidimensional OLAP)

In MOLAP, both the source data and the aggregations are stores in a multidimensional format. MOLAP is the fastest option for data retrieval, but requires the most disk space. Disk space is less of a concern these days with lowering storage and processing cost.

MOLAP is the 'classic' form of OLAP and is sometimes referred to as just OLAP. MOLAP uses database structures that are generally optimal for attribute such as time period, location, product or account code. The way that each dimension will be aggregated is defined in advance by one or more hierarchies.

ROLAP (Relational OLAP)

All data, including the aggregations are stored within the source relational database. This will be a concern for larger data warehousing implementations which have higher usage needs. ROLAP is the slowest for data retrieval. Whether an aggregation exists or not, a ROLAP database must access the data warehouse itself. ROLAP is best suited for smaller data warehousing implementations.

ROLAP works directly with relational database. The base data and the dimension tables are stored as relational tables and new tables are created to hold the aggregated information depends on a specialized schema design.

HOLAP (Hybrid OLAP)

HOLAP is a combination of both the ROLAP and MOLAP methodologies. HOLAP allows storing part of the data in the MOLAP store and another part of the data in ROLAP store. Also we can say that, HOLAP databases store the aggregations that exist within a multidimensional structure, leaving the cell-level data itself in a relational form. Where the data is pre aggregated, HOLAP offers the performance of MOLAP, where the data must be fetched from the tables. HOLAP is as slow as ROLAP.

There is no clear agreement across the industry as to what constitutes "Hybrid OLAP", except that a database will divide data between relational and specialized storage. The degree of control that cube designer has over this partitioning varies from product to product.

C. Data Mining

Data mining, *the extraction of hidden predictive information from large databases*, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve.

Data mining techniques are the result of a long process of research and product development. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

Data mining is a process distilled connotative, unknown beforehand, potential useful information and knowledge from plentiful, incomplete, noisy, fuzzy and stochastic database. It processes data and makes maximum value and transfer information in data to useful knowledge. The desirable features of a data mining system are: the capability of interrogating and learning from any existing database; the ability to learn and evolve continuously using newly acquired data; and the ability to remove redundant data from the system.

- If enterprise market decision refers:
- It needs to start synchronously abundant database list and need to gather the data many list according to determinate rule and form data content supported decision problem.
- Data of enterprise is conformed distributed in many systems.
- Recorder in database list is needed to filter data according to the problem of decision supporting.
- Data storage cannot modify continually.

D. Data Mining Vs OLAP

Is OLAP data mining? As we have seen, OLAP is enabled by a change to the data definition of a relational database in such a way that it allows for the pre-computation of certain query results. OLAP itself is a way to look at these pre-aggregated query results in real time. However, OLAP itself is still simply a way to evaluate queries which is different from building models of the data as in data mining. Therefore, from a technical point of view we cannot consider OLAP to be data mining. Where data mining tools model data and return actionable rules, OLAP allows users to compare and contrast measures along business dimensions in real time.

It is interesting to note, that recently a tight integration of data mining and OLAP has occurred. For example, Microsoft SQL Server 2000 not only allows OLAP tools to access the data cubes but also enables its data mining tools to mine data cubes

III. Analysis Structure Based on Data Mining

Workflow data provide source of valuable information for us to implement operation and analysis and evaluation of enterprise operation, while information organization mode taking process as core in workflow data provide favorable logic structure to integrate other information source and constitute more perfect operation analysis and evaluation and control. Workflow data is not limited by log list of workflow management system because data amount of workflow log list is less. Thus, in present day data source used by research about workflow execution data analysis involve entire database of workflow management system, including example warehouse and model warehouse. Whatever workflow model warehouse and example warehouse are all data warehouse of relation type which emphasis on renewal of data. These data are distracted in many list and not recombined so that it necessarily affect efficiency of analysis. Therefore, we put forward establish workflow execution data to analyze data warehouse based on data mining because data warehouse technology is the most developmental and popular technology supporting storage and analysis of great capacity data. Its online analysis management technology and data mining can hold out the multi-angular and multi granularity analysis for data warehouse data and its analysis speed meet customer's real interactive needs. The framework is shown as figure 3.

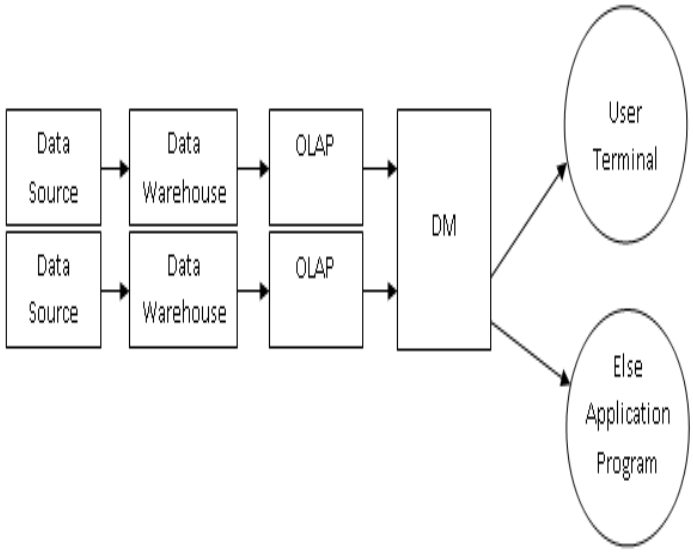

Figure 3: Analysis structure of workflow execution

IV. Data Mining Arithmetic

Workflow execution data provides valuable information for process analysis and evaluation of enterprise operation. At the same time, information organization mode using process as the core of workflow data provides favorable logic structure for integrating another information source and building perfect operation analysis, evaluation, control. But, because multi-dimension and multi-granularity workflow execution data is multifarious and not aggregate, so, we put forward effective data mining method to make data analysis that can analyze various data in different dimension and detail level. Based on OLAP data mining can provide mining in different data-set and detail level. It can improve greatly agility and ability of data mining.

A. Time-sequence Data Mining Method Based on RS Theory

Time-sequence means a series of observation value gained according to time sequence. The "time" indicates data arranged according to not only time sequence, but also space sequence. With system operating in workflow system, workflow execution information and data accumulated are more and more. We only mine the newest data of workflow log to make increment update, rather than query whole example database in order to ensure the efficiency of data update.

RS theory is a mathematical method that analyzes and manages data for possessing imprecise and fuzzy and uncertainty. It doesn't need initial or additional information of any related data. Rough Set is a mathematical tool that extracts information automatically from data by using matrix algebra and Set Theory to investigate correlations amongst data. When the RS method is applied to data mining and knowledge acquisition, it starts with an information system S which contains a pair (U, A), where U is a non-empty, finite set of objects, called the universe or search space, and A is a non-empty finite set of attributes.

Rough set theory introduced by Zdzislw Pawlak in the early 1980's is a mathematical tools to deal with vagueness and uncertainty. Rough set theory defines an indiscernibility relation that partitions the universe of example into elementary sets. A concept is rough when it contains at least one elementary set that contains both positive and negative example.

For example, in time-sequence data-set of process control workflow system, if we collect data of product fault-rate according to month, then, every value indicates change-rate up to before-month. "-1" indicates percentage from 0 to 5 declined. "0" indicates percentage from 0 to 5 gone up. "-1" indicates percentage from 5 to 10 gone up. So, time sequence information list shown as Table I

TABLE I - TIME SEQUENCE INFORMATION LIST

Month	Product output	Fault-ratio	Decision
1	0	0	1
2	-1	0	-1
3	-1	-1	-1
4	-1	0	-1
5	0	0	1
6	0	0	-1

The Time sequence information list is executed as follows using some arithmetic steps.

Defining time-sequence information system (TIS) is

$$St= (U, A \cup \{d, t\}, <) \quad (1)$$

In which U is object-set. A is attribution. d is decision attribution ($d \notin A$) . t is sequence attribution ($t \notin A$) . $\prec$ is a sequence relation of sequence attribution t .

Then, we transfer TIS to information system (IS) in RS. Assuming sequence performance possess n difference integer from 1 to n, we assume |U| = n in arithmetic.

Defining input: TIS

$$St = (U, A \cup \{d, t\}, <)$$

tracking scope is Δ .

Output: positive-rule decision list

$$T= (U^,A^U\{d^\}). \quad (2)$$

Arithmetic step:

(1) For whole $a \in A$, we assign only index between 1 and A , A = { A a ,a ,...,a 1 2 } .

(2) $\Phi \rightarrow \hat{A}$, $\Phi \rightarrow B$, for i from i =1 to Δ, $A \rightarrow B$.

(3) Connecting i to existing index, then

B = {ai1,ai2,……..,ai|A|}, $\hat{A} \cup B \rightarrow \hat{A}$

(4) $\Phi \rightarrow \hat{U}$

(5) For every x Є U , if t(x) > Δ , then Û U y→Û, in which y possesses following property:

d(x)→d(y) and for each avw Є Â, avw (y) =aw (z) among z ∈ U ,aw ∈ A, t(z)= t(x)−v.

So, information system list according to above arithmetic shown as Table II:

TABLE-II INFORMATION SYSTEM LIST

Current month produce output change-ratio	Current month fault-ratio	Last month produce output change-ratio	Last month fault-ratio	Decision
-1	0	0	0	-1
-1	-1	-1	0	-1
-1	0	-1	-1	-1
0	0	-1	0	1
0	0	0	0	1
0	0	0	0	0

We know the relativity of produce data by above list. Its change mode is falling and swing and fixedness and rising.

V. Conclusion

This research work put forward data mining analysis method of workflow execution data based on RS theory aiming at data warehouse of workflow system. Data mining analysis technology analyzing and mining data of workflow log can effectively evaluate operation property of information system and duly find fault and potential problem existing in information system.

VI. Acknowledgment

I would collectively like to thank Dr. J. S. Parihar (Principal of Aditya Engineering College of Technology and Science), our college chairman Mr. Ajay Katare who has been helpful in providing insight and suggestions from a professional standpoint and for proofreading out paper and checking for general coherence. I would also like to thank our friend Miss Ruchi and Miss Shruti, for her efforts in helping improving my paper and reading this paper despite the fact that Rough Set theory evaluates the execution of data. I would also like to thank my family for their general support.

VI. References

[1] M-S Chen, Jiawei Han, and Philip S. Yu, "Data Mining: An Overview from a Database Perspective", IEEE Transactions on Knowledge and Data Engineering, 8(6): 866-883, 1996.

[2] Feyyad UM, "Data Ming and Knowledge Discovery: Making Sense out of data", IEEE Expert (Intelligent Systems), 1996, 11(5): 20-25

[3] Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. ACM SIGMOD Record, 26, 65-74.

[4] Fu Yongjiao, Data Mining: Tasks, "Techniques and Applications", IEEE Potentials, 1997, 16(4): 18-20

[5] Surajit Chaudhuri, "Data Mining and Database System: Where is the intersection?". Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, vol. 21 No, 1 March 1998

[6] Jiawei H, Micheline K, "Data Ming: Concepts and Techniques", Simon Fraser University, Morgan Kaufmann Publishers, 2000

[7] Witten, I. H., & Frank, E. (2000). Data mining. New York: Morgan-Kaufmann.

[8] Judbe. W, Odgers, B, R., Shepherdson, J. M., "Agent-enhanced workflow", BT-Technology, 2000,

[9] Liu, D.T., and Xu, X.W., "A review of Web-based produced data management systems", Computers in Industry, 2001, 44: 251-262

[10] Seidman, C. (2001). Data Mining with Microsoft SQL Server 2000 Technical Reference. Microsoft Press.

[11] Huang, G.H, and Mak, K.L, "Web-integrated manufacturing: recent developments and emerging issues", INT.J Computer integrated manufacturing, 2002, 14(1): 3-13

VII. Biography

Anoop Shrivastava received the B.C.A. degree from the University of Makhanlal Chaturvedi Rashtriya Patrikarita Vishvavidhalaya,Bhopal,Madhya Pradesh, in 2003, the M.Sc. degree in Information Technology from the Makhanlal Chaturvedi Rashtriya Patrikarita Vishvavidhalaya, Bhopal, Madhya Pradesh, in 2005, and the M.Tech.degree in Neural Network and Information Security from the Karnataka Open State University, Mysore, Karnataka 2011, respectively. Currently, He is an associate Professor of Computer Science/Information Technology at Aditya College of Technology & Science affiliated with Rajiv Gandhi Technical University, Bhopal. He is teaching Data Base Management System, Dataware House and Mining, Neural Network, Information Security & Cryptography, Theory of Computation, ERP & CRM. Professor Anoop Shrivastava may be reached at:
anoop.shrivastava@adityacollege.in
shrivastava_ak@hotmail.com