STUDIA INFORMATICA 2004
Volume 25 Number 2 (58)

Marek SIKORA
Silesian University of Technology, Institute of Computer Sciences

DATA CLEANING AND TRANSFORMATION – THE FIRST STAGE OF DATA MINING PROCESS

Summary. A system enabling the data originating from various sources to be integrated and managed, has been described in this paper. The system aims at preparing the data for a data mining process. It has the open architecture allowing its function to be extended by the next users. The software presented in the paper has been developed to integrate and manage the data getting from the systems used in the mining industry.

Keywords: databases, data management, data mining

CZYSZCZENIE I PRZEKSZTAŁCANIE DANYCH – PIERWSZY ETAP PROCESU EKSPLORACJI DANYCH

Streszczenie. W artykule opisano system umożliwiający integrację i manipulację danymi pochodzącymi z różnego rodzaju źródeł. Celem systemu jest przygotowanie danych do procesu data mining. System posiada otwartą architekturę umożliwiającą rozszerzanie jego funkcji przez kolejnych użytkowników. Opisane oprogramowanie powstało z myślą o integracji i zarządzaniu danymi pochodzącymi z systemów wykorzystywanych w przemyśle wydobywczym.

Słowa kluczowe: bazy danych, zarządzanie danymi, data mining

1. Introduction

The data to be analysed by means of the data mining methods are stored in the diverse databases. At the beginning of the data mining process, the data should be transformed into a form to be accepted by a system of data analysis. This task is executed mainly by means of the open-access import and export mechanisms being implemented usually to the systems of data management and analysis. However the data, in particular being acquired by the industry systems [1][3], is frequently stored in "non-standard" data format for which the commonly used data exchange procedures do not exist.

The first stage of the data mining process is associated also with all activities aiming at the data conversion (e.g. computation of certain aggregated values, to add or delete the columns) and elimination the errors or deleting the wrong values. These tasks are to some extent just carried out by the data analysis systems or should be programmed by the users.

The paper presents a system joining both options. The described system may be treated as a platform between various data formats (including non-standard ones) and a tool allowing the various useful operations to be performed at the first stage of the data mining process.

The developed system has the open architecture. The extension of its functionality is made by implementation of so-called "plug-ins", which may be made independently by different users and attached to the existing application.

Although the system has been developed to manage the information being stored in the monitoring systems of natural hazards and process engineering in hard coal mines, it can be also used to carry out any tasks concerning the data cleaning and transformation.

2. Existing solutions

In this chapter there are presented briefly the systems and their capabilities to manage the data in order to execute the first stage of the data mining. In this paper we take into consideration neither the solutions dedicated to the specific database management systems nor the creation of data warehouses related to a specific problem of the analysis [5]. The ready programs for data cleaning and transformation (including aggregation) are the subject matter of our interest.

The typical representatives of the software in which we are interested, may be the following programs: WinPure, DataManager, Monarch, SPSS Data Editor and others. The first three of the mentioned programs are the applications performing only the data management. The last one is a part of a powerful data analysis system being made available to users by the SPSS company. Majority of the big data analysis systems includes the data management modules. The following systems can be classified as powerful ones: Statistica, Alice`dSoft, PolyAnalyst, Insightful Miner, SAS Analyst, SAS Enterprise Miner, Salford-Systems.

The capabilities of the above-mentioned programs are similar (we have tested their evaluation versions). However the essential differences concern the graphical interface, the purchase price and the operation speed.

Below there are given all capabilities, which the tested programs provide:

- Data import from the extern sources: test format, defined source ODBC
- Grouping of data: grouping of lines of a few identical tables, grouping of tables into one table (sum of columns)
- Performance of SQL queries related to just imported tables (the applications use usually a query wizard)
- Changes of tables' framework: names and types of fields, indexes
- Changes of values according to a criterion defined by a user (e.g. to change a given text into another text, to change a numerical value into another numerical value or a text, to change no value into a value or a text
- Data filtering considering the advanced filtering criteria formulated by a user
- Data editing
- Computation of the simple descriptive statistics and values of correlation coefficients
- Adding a column including values, which depend on the values in the other columns – a user defines a function according to which a new column should be completed.

In addition to these capabilities, the above-mentioned programmes provide a user with the functions non-related directly to preparing the data to be analysed, but rather the functions are related to the initial database intended to create the various types of diagrams, registers and reports.

3. The “Rubin” system

The “Rubin” – integration and data management system has been developed within the framework of the research project being carried on in our Institute. It may be regarded as a member of the family of programs having been mentioned in the previous chapter. This program differs from the typical representatives of the mentioned family. The difference consists in the implemented mechanisms of data downloading from the non-standard databases (including the databases of the SCADA industrial visualization systems and the compound binary file databases) [4]. Furthermore the program has the open architecture and therefore it is available to complete the next functions by various users.

The system aims mainly to integrate and to prepare for analysis the information stored in the databases of the mining process monitoring systems installed in hard coal mines. One of the main features of the mining process monitoring systems is, that they maintain to a large degree the non-standard format databases (binary files – not infrequently coupled each other – [1][3]). One of the tasks of the “Rubin” system is the integration of data originating from the mentioned systems.

3.1. Architecture of the system

The “Rubin” system is a module system operating on the principle of program “plug-ins” (modules), which are coupled to the primary application (kernel of the system). Each module makes a discrete program operating however on the basis of the data tables being common to the kernel. The communication with the kernel consists in sending by the module of a definite type of request (usually it is the order to vary on a data table) and return of the result. Depending on the result of the request, the kernel initiates an additional action (e.g. to write down a new data table being result of the operation made by the module). The kernel operates in a way on the principle of a server, listening whether the modules send their requests. In addition to the listening of requests being sent by the modules, the most frequently occurring operations have been programmed in the kernel. These operations are common for all modules of the system.

The communication between the kernel and modules is made by means of the component *mxPlugin*, which may be used by the programming environment Delphi and Borland C++.

The kernel of the system includes also a file server for management the relational database. The file server allows the SQL queries to be sent to the data just included in the system.

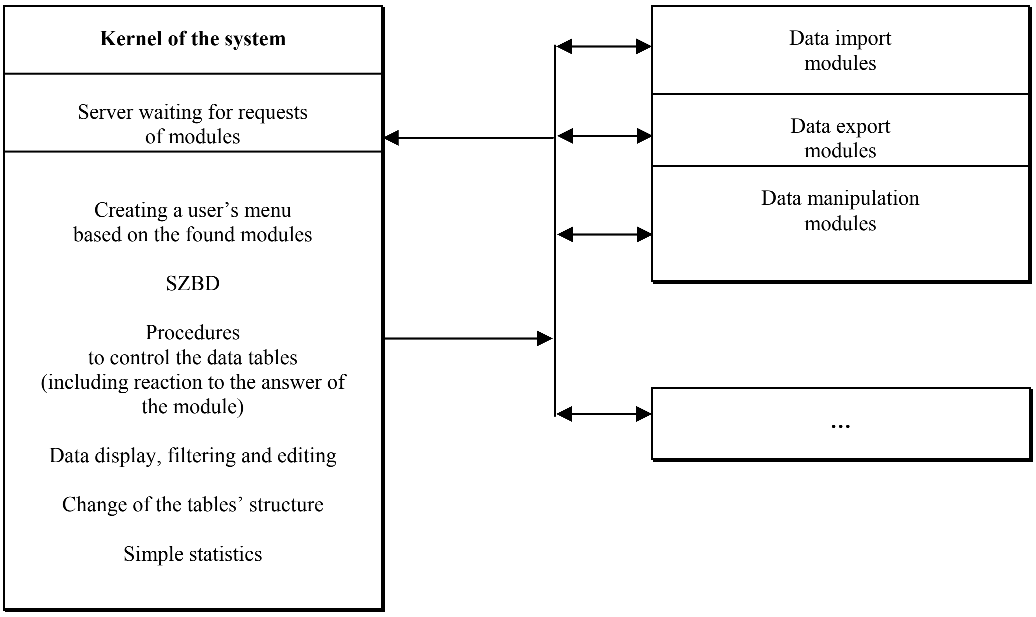

Kernel of the system
Server waiting for requests of modules
Creating a user’s menu based on the found modules
SZBD
Procedures to control the data tables (including reaction to the answer of the module)
Data display, filtering and editing
Change of the tables’ structure
Simple statistics
Data import modules
Data export modules
Data manipulation modules
...

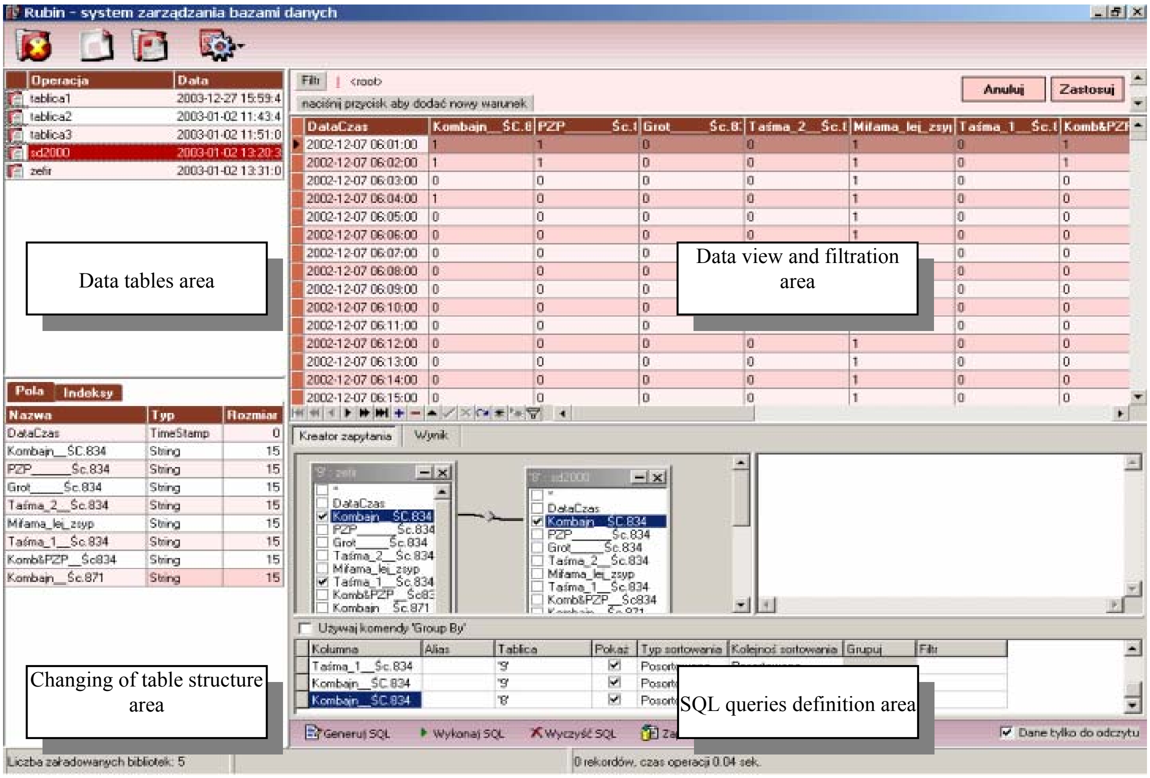
Fig. 1. Architecture of the “Rubin” system
Rys. 1. Architektura systemu “Rubin”

In the figure 1 the architecture of the “Rubin” system has been presented. All attached modules are located in a discrete directory. The kernel of the system creates a user’s menu on the base of the content of this directory. A user can activate the modules by means of the user's menu.

3.2. Functions of the system

The data are imported to the system by using a few modules. The access to any database managed by SZBD is available. A user may generate a SQL query (manually or by means of a wizard) enabling the interesting data to be got.

The data included in the file bases and SCADA bases may be downloaded to the "Rubin" system by means of the specialized modules. A user gets in this case a list of objects being defined in the base (e.g. sensors). He can select from this base the interesting for him items and then he defines a time period within which the data should be downloaded. The present version of the system enables the data to be got from the file bases of the following systems: Zefir, SMP, SMOK, longwall-complex operation and mine drainage monitoring systems. After getting the data, a new data table occurs in the system. A user can perform the next operations with this data table.


Rubin - system zarządzania bazami danych
Filtr
<root>
naciśnij przycisk aby dodać nowy warunek
Anuluj
Zastosuj
Pola
Indeksy
Kreator zapytania
Wynik
Generuj SQL
Wykonaj SQL
Wyczyść SQL
Za
Data tables area
Data view and filtration area
Changing of table structure area
SQL queries definition area

Operacja	Data
tablica1	2003-12-27 15:59:4
tablica2	2003-01-02 11:43:4
tablica3	2003-01-02 11:51:0
sd2000	2003-01-02 13:20:3
zefir	2003-01-02 13:31:0

DataCzas	Kombajn__ŚC.8	PZP______Śc.8	Grot______Śc.8	Taśma_2__Śc.8	Miłama_lej_zsy	Taśma_1__Śc.8	Komb&PZP
2002-12-07 06:01:00	1	1	0	0	1	0	1
2002-12-07 06:02:00	1	1	0	0	1	0	1
2002-12-07 06:03:00	0	0	0	0	1	0	0
2002-12-07 06:04:00	1	0	0	0	1	0	0
2002-12-07 06:05:00	0	0	0	0	1	0	0
2002-12-07 06:06:00	0	0	0	0	1	0	0
2002-12-07 06:07:00	0	0	0			0	0
2002-12-07 06:08:00	0	0	0			0	0
2002-12-07 06:09:00	0	0	0			0	0
2002-12-07 06:10:00	0	0	0			0	0
2002-12-07 06:11:00	0	0	0			0	0
2002-12-07 06:12:00	0	0	0	0	1	0	0
2002-12-07 06:13:00	0	0	0	0	1	0	0
2002-12-07 06:14:00	0	0	0	0	1	0	0
2002-12-07 06:15:00	0	0	0	0	1	0	0

Nazwa	Typ	Rozmiar
DataCzas	TimeStamp	0
Kombajn__ŚC.834	String	15
PZP______Śc.834	String	15
Grot______Śc.834	String	15
Taśma_2__Śc.834	String	15
Miłama_lej_zsyp	String	15
Taśma_1__Śc.834	String	15
Komb&PZP__Śc834	String	15
Kombajn__Śc.871	String	15

'9' : zefir
□ *
□ DataCzas
☑ Kombajn__ŚC.834
□ PZP______Śc.834
□ Grot______Śc.834
□ Taśma_2__Śc.834
□ Miłama_lej_zsyp
☑ Taśma_1__Śc.834
□ Komb&PZP__Śc83
□ Kombajn__Śc.871

'8' : sd2000
□ *
□ DataCzas
☑ Kombajn__ŚC.834
□ PZP______Śc.834
□ Grot______Śc.834
□ Taśma_2__Śc.834
□ Miłama_lej_zsyp
□ Taśma_1__Śc.834
□ Komb&PZP__Śc834

□ Używaj komendy 'Group By'

Kolumna	Alias	Tablica	Pokaż	Typ sortowania	Kolejnoś sortowania	Grupuj	Filtr
Taśma_1__Śc.834		'9'	☑	Posort	Posortowane		
Kombajn__ŚC.834		'9'	☑	Posort			
Kombajn__ŚC.834		'8'	☑	Posort			

☑ Dane tylko do odczytu

Liczba załadowanych bibliotek: 5

0 rekordów, czas operacji 0.04 sek.

Fig. 2. Main window of the "Rubin" system
Rys. 2. Okno główne systemu „Rubin"

The main window workspace has been divided into four basic fields. In the first one a user can see a list of accessible tables. In the second one the data of any table may be browsed. In this field there are also available the options of filtering and grouping of data according to given by a user criteria (the data being filtered can be put down in a new table). It is also possible to delete a selected group of columns. In the context menu, the option *Statistics* is available. The option informs on the basic values of descriptive statistics and also linear correlation coefficients to another selected column. Depending on the types of columns, the Pearson's or Goodman-Kruskal's [2] linear correlation coefficients are used. The third field of the window workspace enables a user to change a structure of the selected column (e.g. exchange/conversion of a type and/or name of columns; to set or delete indexes of selected fields permitting the operation of computation to be significantly accelerated). And finally in the forth field of the workspace the SQL queries may be created. This field enables the data from a few tables to be grouped together. A user can optionally key-in a query or create it by a wizard.

The above-mentioned functions are carried out by a kernel of the system and the included modules perform the remaining operations. The modules enable a user to perform all functions of data transformation being mentioned in the chapter two of this paper. Besides the described features, the following functions being implemented in the modules are available:

- Time intervals. The data from various systems are collected with different frequencies. The "time intervals" module allows the tables to be reformulated in such a way that the tables would include the data occurring at regular intervals. A user indicates a field (it should be of type *DateTime*), according to which a time-laps will be calculated and defines an initial time and interval. After a lapse of initial time, a new record is added to output table. A user defines for the remaining fields a mode of calculation of a value within the output table (it may be: a first value, a maximum within the interval, a minimum, an average value, a sum, a numerator). This proceeding allows the values occurring in the next intervals to be aggregated. The obtained table describes the data integrated from various systems, in which the successive measurements relate approximately to identical time.
- Exchange of values. The module aims to provide a user with information on values occurring in a given field and to allow the previous values to be exchanged for the new ones according to criteria being defined by a user (from a simple exchange e.g. "SC9B" for "Longwall 9B" to an advanced exchange e.g. *IF (pressure<=12.53 and temperature<120) or state=-1then state="no data available"*, where *pressure, temperature* and *state* are the names of fields in the table to be modified). The advanced exchange of values enables also a user to add a new column, which will include values depending on the values within just existing columns.

- Detection of outliers. The outliers can be defined by means of the *Statistics* function providing information on a value of the inter-quartile range for numerical data, a list of all different values of a given field and a percentage of values frequencies. The errors being found that way, may be deleted by means of the modules: “exchange of value” or “deleting records”. Besides these possibilities, the “Rubin” system offers a module, browsing the numerical values of a given field and finding the records, which include values differing more (or less) from the average value of the next records in a percentage threshold given by a user. After marking the records, a user determines himself to exchange the errors being found for the average value of the next records (in the event of a group of values being identified as incorrect, they are exchanged for the values calculated on the basis of two point form of the equation of a straight line passing through extreme values regarded as correct).
- Division a table into parts. This module enables a user to select from a table a certain subset of records according to one of the following criteria of selection: random, percentage, every n-record selection, according to condition defined by a user. Two tables are the result of the operation. The first one includes the records meeting the assigned conditions and the second one the records that do not meet the conditions.

The system has been tested based on tables consisting of a few thousands of objects (a largest table had a size of 1,390,000 records). A most time-consuming operation was the operation of data integration by means of the module “time intervals”; the aggregation for the largest table lasted a dozen or so minutes. The result that has been gained, should be acceptable, considering a fact that the system does not use any specialized database server, which would be able to optimise an operation time of some functions.

3.3. Example of operation

As an exemplary model of data processing, one may quote the integration of data downloaded from the monitoring systems of rock bump hazards in mine workings. A few types of these monitoring systems have been put in use in mines (e.g. ARES Ocena, ARAMIS, Hestia). All of them store the data in the relational databases; therefore the data downloading presents no special difficulty. The main information to which the entries in the tables are related, is a name or number of a specific mine working; the names however may be different within various systems, so they should be standardized by means of the module “exchange of value”. After that operation, the data concerning rock bump being registered, their energy and seismoacoustic activity in a selected group of mine workings and during selected period, may be integrated (SQL query or integration module of tables by means of a wizard), while within the output table, only the fields being interesting for a user can be indicated (e.g. energy of the event, type of bump, hour of its occurrence).

Depending on the purpose of data getting, it is possible then to aggregate the values of bump energy as well as seismoacoustic activity and energy during a given interval (hour, shift or day interval). And next by means of the module of “advanced exchange of value”, there are pasted in the output table the new fields within which the bump energy increase as well as the energy and acoustic activity during the following intervals are calculated (however the increases can be communicated by either numerical or linguistic values like: high increase, increase, constant, decrease, big decrease).

The last step to proceed is to connect the output data table with a table including information on shift progress of a mine working (module of table join – data tables close to each other).

The presented outline of action has allowed us to get from various systems a synthetic information on level of rock bump hazard in mine workings. The data set is being currently analysed from a possibility perspective of bump emergency prediction (one hour or one shift forward).

4. Conclusions

In this paper there has been described a system enabling the indispensable transformation of data to be done in order to prepare them for analysis. The system performs data import and export to various formats with particular consideration for non-standard formats of data repositories being used by monitoring systems of hazards and process engineering in hard coal mines. The system aims mainly to provide a user with synthetic, integrated and transformed information on mining activity from different points of view. The developed application aimed at implementation of the most frequently used operations preparing the data for analysis or reporting. Our application has been patterned after the available world-wide solutions.

Apart from the mining and industrial applications, the system may be used as a universal tool realizing the first stage of the data exploration.

Because of its open architecture, a simple extension of functionality of the system is available. The authors of the system make its documentation available to interested users, allowing them to implement the next modules.

The research being presented in this paper were funded by the State Committee for Scientific Research within the framework of the grant No 5 T12A 001 23.

REFERENCES

1. Dec B., Gajoch A.: System dyspozytorski ZEFIR – Struktura programu. Mechanizacja i Automatyzacja Górnictwa. Nr 4/5 (354). Centrum EMAG, Katowice 2000.
2. Koronacki J., Mielniczuk J.: Statystyka dla studentów kierunków technicznych i przyrodniczych. WNT, Warszawa 2001.
3. Krzystanek Z., Bojko B., Dylong A.: Rozwój systemu SMP Kontroli Zagrożeń Metanowych i Pożarowych. Mechanizacja i Automatyzacja Górnictwa, nr 9/10 (3358). Centrum EMAG, Katowice 2000.
4. Sikora M.: Integracja i zarządzanie danymi pochodzącymi z systemów monitorowania zagrożeń i procesów technologicznych w kopalniach węgla kamiennego. Mechanizacja i Automatyzacja Górnictwa, nr 2 (397). Centrum EMAG, Katowice 2004.
5. Szulim R., Moczulski W.: AI-based control of dynamic processes using knowledge discovery in industrial databases. Proceedings if AI-METH 2002 Gliwice, Poland, November 13-15, 2002.

Recenzent: Dr inż. Grzegorz Drwal

Wpłynęło do Redakcji 23 lutego 2004 r.

Omówienie

W artykule opisano system umożliwiający integrację i manipulację danymi pochodzącymi z różnego rodzaju źródeł. Celem systemu jest przygotowanie danych do procesu data mining.

W rozdziale drugim artykułu krótko omówiono istniejące rozwiązania programowe umożliwiające przygotowanie danych do procesu eksploracji. Wymieniono funkcje, jakie najczęściej można realizować za pomocą tego typu oprogramowania.

W rozdziale trzecim przedstawiono architekturę i funkcje zrealizowanego systemu zarządzania danymi „Rubin”. System ten można zaliczyć do rodziny programów przygotowujących dane do procesu eksploracji. Poza funkcjami typowymi dla tej rodziny programów system rubin realizuje zadania agregacji danych oraz prostej detekcji błędnych wartości pomiarowych.

Opisane oprogramowanie powstało z myślą o integracji i zarządzaniu danymi pochodzącymi z systemów wykorzystywanych w przemyśle wydobywczym. Dlatego też „Rubin” pełni również rolę platformy integrującej dane pochodzące z różnych systemów monitorowania bezpieczeństwa i procesów technologicznych prowadzonych w kopalniach węgla kamiennego. Systemy te nierzadko wykorzystują binarne repozytoria danych, zatem możliwość integracji danych pochodzących z tych systemów jest ważną cechą systemu „Rubin".

Opisany system posiada otwartą architekturę (rys. 1) umożliwiającą łatwe rozszerzanie jego funkcji przez kolejnych użytkowników. Kolejne funkcje systemu można zrealizować za pomocą dołączanych do jądra programu modułów (plug –ins).

System działa pod kontrolą systemu operacyjnego Windows, a wszystkie funkcje (łącznie z zadawaniem zaawansowanych zapytań SQL) realizowane są za pomocą kreatorów (rys. 2) pozwalających na korzystanie z systemu nie zaawansowanym użytkownikom.

Rozdział czwarty zawiera opis przykładowej realizacji zadania przygotowania danych pochodzących z różnych systemów monitorowania zagrożeń tąpaniami w wyrobisku górniczym.

Adres:

Marek SIKORA: Silesian University of Technology, Institute of Computer Sciences,
Akademicka 16, 44-101 Gliwice, Poland, msikora@homer.iinf.polsl.gliwice.pl