Some Aspects of Data Integration in Data Mining Systems

Mirela DANUBIANU, Emanuela – Alina BOLOHAN, Radu TIGANESCUL- AMARITII
Stefan cel Mare University of Suceava
str.Universitatii nr.13, RO-720229 Suceava
mdanub@usv.ro; embolohan@stud.usv.ro; rtiganescul@stud.usv.ro

Abstract — **The aim of this paper is to present some aspects of data preparation in the Knowledge Discovery in Databases process. We have made some experiments regarding data migration from MS Access to Oracle, respectively to DB2 UDB, as a first step in data preparation for data mining algorithms. We have decided to implement a data mining system in order to improve the personalized therapy of speech disorder assisted by Terapers system. Due to the limitation of MS Access, the DBMS used for Terapers, we intend to use for this data mining system a database management system to provide data mining features incorporated. The reason of these experiments is to find the best solution for that.**

Index Terms — **Knowledge Discovery in Database, data mining, data integration, data migration**

I. INTRODUCTION

Data mining occurred in response to technological development produced in recent decades, reflected by the appearance of powerful systems and media storage that enable the acquisition and processing of huge volumes of data. They revealed a tremendous amount of data collected in databases or other files, which can form the basis of surprising information. Since traditional methods of data querying and processing are not efficient for large data it was necessary to develop methods and techniques adapted to this situation.

II. KNOWLEDGE DISCOVERY IN DATABASE AND DATA MINING

Data mining is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. So, data mining is defined as the process of extracting interesting and previously unknown information from data.

The experience of the last years showed that discovering knowledge from huge databases (KDD) involve much more than simply applying a sophisticated data mining algorithm to a predefined dataset.

One of the most important problems in KDD research is the understanding of KDD as a "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*"*[1]

Although there are several differences between process models [1][2][3], the key message is the same: data mining that is applying a sophisticated mining algorithm to a dataset, is just one of several steps in a KDD process. The KDD process is interactive and iterative, involving numerous steps with many decisions being made by the user [2].

According to CRISP–DM, a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 1. The phase sequence is adaptive. That means the next phase in the sequence often depends on the outcomes associated with the preceding phase.

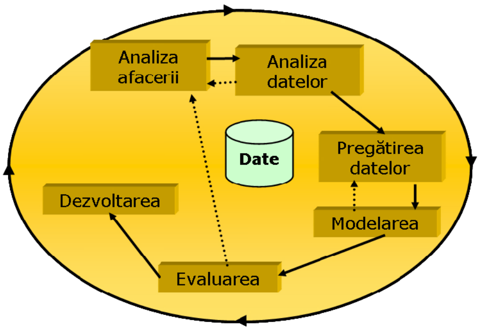


Figure 1. CRISP-DM Model for Data Mining.

- Business understanding - this initial phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a DM problem definition and a preliminary plan designed to achieve the objectives;
- Data understanding- starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information;
- Data preparation- is the phase which covers all activities to construct the final dataset from the initial raw data;
- Modeling-in this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values;
- Evaluation-at this stage the model (or models) obtained are more thoroughly evaluated and the steps executed to construct the model are reviewed to be certain it properly achieves the business objectives;
- Deployment-creation of the model is generally not the end of the project. Even if the purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it [3].

III. DATA PREPARATION FOR DATA MINING

Data from various sources is not the best set for data mining and require different types of transformations to convert them to a usable form in solving problems such as prediction or description. Naturally, these data contain missing values or erroneous, inadequate samples or values with types unsuitable for data mining algorithms.

Many experts agree that one of the critical steps in the KDD process is data preparation and processing, although in literature, this task is often passed background because it is considered to be specific to each application. However, certain parts of the data preparation process or, in some cases even the whole process can be described independently of the application or data mining method used.

Consequently, in real applications of data mining considerable effort are consumed to prepare the data for the application of data mining algorithms. Achieving this goal requires two major tasks:

- organizing data in a standard form
- preparing data to optimize performances

Preparing data for data mining algorithms include:

a) preprocessing - elimination of unnecessary data, checking consistency, detect and remove erroneous data, elimination of extreme values (outliers);
b) data integration - the combination of data from different sources;
c) transformation of variables: through standardization, by crossing the logarithmic scale;
d) separation of the database into three categories of data: data for training, data for validation and data test.

Consequently, data preparation phase aims to cover all activities to construct the dataset that will be fed into the data mining tool from the initial raw data. Data preparation tasks are likely to be performed multiple times and not in any prescribed order. Tasks include table, record and attribute selection and integration from various sources as well as transformation and cleaning of data for data mining tools.

IV. Data preparation framework for Logo-DM system

Logo-DM Objectives

Logo-DM system aims to improve the quality of logopaedic therapy by applying some data mining techniques on data obtained from TERAPERS system, developed within the Center for Computer Research in the University "Stefan cel Mare" of Suceava [4]. TERAPERS project has proposed to develop a system which is able to assist teachers in their speech therapy of dislalya and to follow how the patients respond to various personalized therapy programs.

It is a complex system that contains a fix part - the intelligent system called LOGOMON, placed on computers located in therapists' offices and a set of mobile devices that guides independent work of children affected by dislalya, follows the way of achieving the tasks and provides activity reports. One component part of LOGOMON is an expert system which aims to assist the speech therapy. Starting March 2008 the system is currently used by the therapists from Regional Speech Therapy Center of Suceava.

In present, because the needs of efficient use of time or due to the economic needs, have become of interest information such as [5]:

- how is the estimated duration of therapy for a particular case,
- what is the predicted final state for a child or what will be its state at the end of various stages of therapy,
- what are the best exercises for each case and how can dose their effort for effectively solve these exercises,
- how is associated the family receptivity - which is an important factor in success of the therapy - with other aspects of family and personal anamnesis.

All this may be the subject of predictions obtained by applying data mining techniques on data collected by using TERAPERS. It is also interesting, as part of the knowledge discovered by data mining algorithms, to be used to enrich the knowledge base of expert system embedded in LOGOMON.

Adapting the therapy programs involves a complex examination of children and recording of relevant data relating to personal and family anamnesis.

Complex examination of how the children articulate the phonemes in various constructions allows a diagnosis and classification in a given category of severity. Anamnesis data collected may provide information relative to various causes that may negatively influence the normal development of the language. It contains historical data and data provided by the cognitive and personality examination.

On provide to the applied personalized therapy programs data such as number of sessions/week, exercises for each phase of therapy and the changes of the original program according to the patient evolution. In addition, the report downloaded from the mobile device collects data on the efforts of child self-employment. These data refer to the exercises done, the number of repetitious for each of these exercises and the results obtained. The tracking of child's progress materializes data which indicate the moment of assessing the child and his status at that time.

All these data are stored in a relational database, build using MS Access DBMS.

MS Access is part of Office Professional package, is cheap and easy to use. It allows development of databases relatively small, with one user. It is very sturdy and can hold control of multi-user transactions. All information in the database is kept in a single file.

After analysis of available technologies, it was concluded that the effective implementation of the system can be made with DBMS which entails multi-user, increased security and, last but not least, to provide facility for analysis and to have implemented data mining algorithms.

We consider two DBMS that meet these conditions: Oracle and DB2 UDB.

In this context a problem which must be solved is the one concerning migrating data from MS Access to the considered DBMS.

Migrating data from MS Access to Oracle

Using Oracle SQL Developer Migration Workbench we have the possibility to migrate data from MS Access to Oracle.

There are four stages in the data migrating process [6][7]:

- Capturing the database source. The first step is to "instantly" capture the database from MS Accesss. This can be done in two ways:

    - Online capture: this process requires the creation of a connection from SQL Developer to the MS Access database. Using JDBC the data about the MS Access database can be accessed and created with Capture Model.
    - Offline capture: this involves using the Exporter to extract the data about the database in a file outside the MS Access database.
- Converting the database capture. Oracle SQL Developer Migration Workbench uses the capturing method to convert the captured objects in Oracle format representing the structure of the destination database. This structure is called Converted Model.
- Generating the Oracle database Oracle SQL Developer Migration Workbench generates DDL declarations for creating a new Oracle database based on the objects of the Converted Model. Running the DDL declarations generates the objects in the Oracle database.
- Migrating data The last step is represented by the migrating data process and it can be made in two ways:
    - Moving the data online (activation): We can create a connection from Oracle SQL Developer to a MS Access and migrate data.
    - Moving data offline (unpacking): We can export data from MS Access using Exporter SQL Developer which will create a series of files that can be executed from a batch file in order to load the data in Oracle.

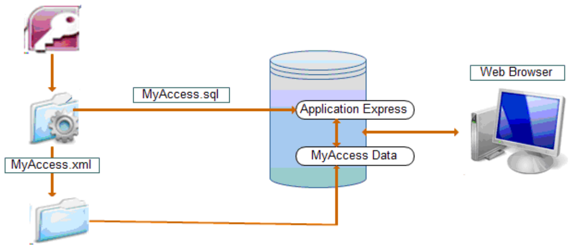


Figure 2. Migrating data from Access in Oracle [ ].

After data migration from MS Access a series of modifications were required on data types. Such a data type, subject of conversion, is Date and Timestamp.

We must ensure that the system has the time mask and data format like the one below:

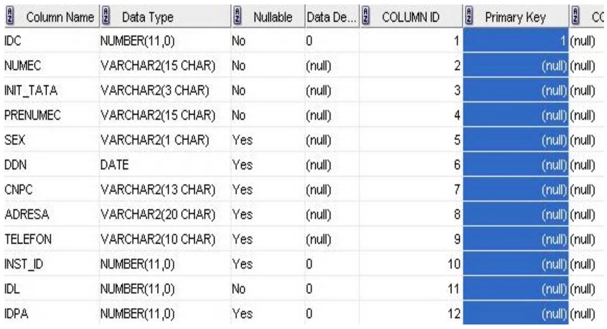
```
yyyy-mm-dd HH24:mi:ss.ff3
OR
mm/dd/yyyy HH24:mi:ss
```

If the time mask and data do not coincide then we must add in SQL Developer the proper Date and Timestamp format from MS Access in order for the data migration to end successful.

The migrating process from MS Access to an Oracle database is simple to achieve but some objects and syntaxes aren't automatically migrated. Thus a manual intervention is required. Analyzing the captured model, identifying the number, type and object complexity we can calculate the necessary time for a required manual task.

Once the data is imported data constraints are kept.

A small sample of data types resulted in Oracle, from Logomon data is presented in Fig. 3.



Column Name	Data Type	Nullable	Data De...	COLUMN ID	Primary Key	CC
IDC	NUMBER(11,0)	No	0	1	1	(null)
NUMEC	VARCHAR2(15 CHAR)	No	(null)	2	(null)	(null)
INIT_TATA	VARCHAR2(3 CHAR)	No	(null)	3	(null)	(null)
PRENUMEC	VARCHAR2(15 CHAR)	No	(null)	4	(null)	(null)
SEX	VARCHAR2(1 CHAR)	Yes	(null)	5	(null)	(null)
DDN	DATE	Yes	(null)	6	(null)	(null)
CNPC	VARCHAR2(13 CHAR)	Yes	(null)	7	(null)	(null)
ADRESA	VARCHAR2(20 CHAR)	Yes	(null)	8	(null)	(null)
TELEFON	VARCHAR2(10 CHAR)	Yes	(null)	9	(null)	(null)
INST_ID	NUMBER(11,0)	Yes	0	10	(null)	(null)
IDL	NUMBER(11,0)	No	0	11	(null)	(null)
IDPA	NUMBER(11,0)	Yes	0	12	(null)	(null)

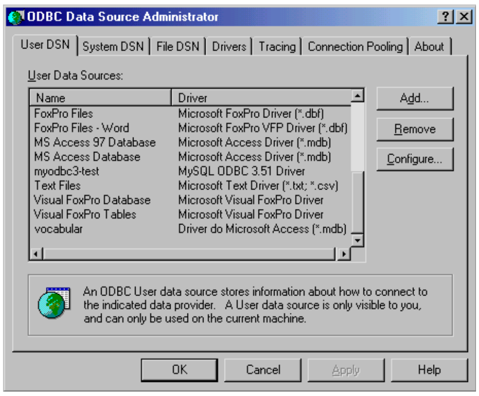
Figure 3. Sample of converted data types in Oracle.

Migrating data from MS Access in DB2 UDB

DB2 didn't allow to directly import a database from Microsoft Access. Thus we chose to import the tables through ODBC-Open Database Connectivity [8].

ODBC provides an application programming interface (API) common for accessing databases stored on different Relational Database Management System (RDBMS). Through ODBC an application program can unitarily access different RDBMS platforms with the help of a common code program. This way the user program will implement a code for accessing the database independent from the RDBMS platform. The only thing the program needs is the ODBC driver.

The installation and configuration of the ODBC driver is made with the **ODBC Data Sources** component from (Fig. 4).



User Data Sources:

Name	Driver
FoxPro Files	Microsoft FoxPro Driver (*.dbf)
FoxPro Files - Word	Microsoft FoxPro VFP Driver (*.dbf)
MS Access 97 Database	Microsoft Access Driver (*.mdb)
MS Access Database	Microsoft Access Driver (*.mdb)
myodbc3-test	MySQL ODBC 3.51 Driver
Text Files	Microsoft Text Driver (*.txt; *.csv)
Visual FoxPro Database	Microsoft Visual FoxPro Driver
Visual FoxPro Tables	Microsoft Visual FoxPro Driver
vocabular	Driver do Microsoft Access (*.mdb)

An ODBC User data source stores information about how to connect to the indicated data provider. A User data source is only visible to you, and can only be used on the current machine.

Figure 4. ODBC Data Sources component.

ODBC is an SQL based language for accessing data. When an application program desires to access data from a database it sends an SQL sequence to the ODBC Driver Manager, which will correctly load the ODBC driver for accessing data. This driver converts the SQL sequence from the program user format to the RDBMS format. RDBMS finds the data in the database and gives them to the application program via the driver and ODBC Driver Manager.

ODBC provides the user programmer a cursor library which allows access to a set of columns extracted from the database without maintaining a permanent access to it. As any API, ODBC API application is offered complete support for database operations. But, as Windows API programming, it involves a large amount of code and it is difficult to program the application [9].

The data stored in tables was correctly transferred in DB2. Data types were converted as it follows:

TABLE I. TRANSFORMATION SUFFERED

MS Access	DB2
Number	Integer
Number	Smallint
Text	Varchar
Memo	Long Varchar

If the data type in Microsoft Access is Number type and consists of three digits or less in DB2 will be converted to Smallint.

V. EXPERIMENTEL RESULTS

We have shown some of the problems we encountered in the process of migrating data from one DBMS to another. These were related mainly to convert data from one data type to another, but there might be some other problems.

In order to evaluate the migration process from MS Access, we have used the data from table “Fise”, which contains personal data for the children with speech impairments. Its structure contains 103 fields and their data type varies from *Text* or *Number* to *Yes/N* and *Date*. In this table we have recorded data about 50 children.

In the Table 2 we present a short comparison between the two migration processes.

TABLE 2. COMPARISON BETWEEN DATA MIGRATION IN ORACLE AND DB2 UDB

Criteria	Oracle	DB2 UDB
Number of correct transferred records	47	49
Number of fields whose data type to be transformed	51	103

VI. CONCLUSIONS

Migrating data from Microsoft Access in Oracle or in DB2 UDB is an important step in preparing the data to be integrated in a Data Mining System. We have tested two variants for migrating data from MS Access to Oracle respectively to DB2 UDB. Each of the two processes had strengths and weaknesses. In Oracle, there are data types similar those from Access, which does not require conversion, but there are cases where manual intervention is necessary for a correct transfer. This leads to lower performance.

For DB2 is necessary a conversion of all types of data from Access, but it is made automatic.

Next we will make an assessment for the other stages of data preparation for data mining algorithms to find which of the two systems offer the best solution.

REFERENCES

[1] U. Fayyad, G. Piatetsky-Shapiro, and P Smyth, (1996) "The KDD process for extracting useful knowledge from volumes of data". Communications of the ACM, 39(11):27-34, November.

[2] R.J. Brachman, and T. Anand. (1996) "The process of knowledge discovery in databases: A human centered approach". In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, Advances in Knowledge Discovery and Data Mining, chapter 2, pages 37-57. AAAI/MIT Press.

[3] R. Wirth, and J. Hipp, (2000) "CRISP-DM: Towards a standard process model for data mining". In Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, pages 29-39, Manchester, UK.

[4] M. Danubianu , S.G. Pentiuc, O Schipor, M Nestor, I Ungurean "Distributed Intelligent System for Personalized Therapy of Speech Disorders", Proceedings of ICCGI08, 2008, Atena.

[5] M Danubianu, S.G. Pentiuc, T. Socaciu, "Towards the Optimized Personalized Therapy of Speech Disorders by Data Mining Techniques", The Fourth International Multi Conference on Computing in the Global Information Technology ICCGI 2009, Vol: CD, 23-29 August, Cannes - La Bocca, France, 2009.

[6] http://www.oracle.com/technology/obe/hol08/sqldev_migration/msaccess/migrate_microsoft_access_otn.htm#t6.

[7] http://download.oracle.com/docs/cd/B25329_01/doc/appdev.102/b25108/toc.htm.

[8] http://www.ibm.com/developerworks/data/library/techarticle/dm-0605bhogal/index.html.

[9] http://users.utcluj.ro/~valean/II/bd.pdf.