Expert system for data migration between different database management systems

BOGDAN WALEK, CYRIL KLIMEŠ
Department of Informatics and Computers
University of Ostrava
30. dubna 22, 701 03 Ostrava
CZECH REPUBLIC
bogdan.walek@osu.cz, cyril.klimes@osu.cz http://www.osu.cz

Abstract: This paper deals with the design of the fuzzy expert system for processing data migration between different relational database management systems (RDBMS). At the beginning we identify current state of data migration between different RDBMS. Then we propose new approach which suggests creating of expert system as a tool for migration of database tables and their data between different RDBMS. The expert system contains a knowledge base which is composed of IF-THEN rules and based on the input data suggests appropriate data types of columns of database tables. The proposed tool, which contains an expert system, also includes the possibility of optimizing the data types in the target RDBMS database tables based on processed data of the source RDBMS database tables. Paper also proposes methodological guidelines for successful migration of database tables and their data between different RDBMS and suggests the possibility of tool extension when the source or target RDMBS is changed. The proposed expert system is shown on data migration of selected database of the source RDBMS to the target RDBMS.

Key-Words: expert system, fuzzy, data migration, database, relational database, data type, relational database management system.

1 Introduction

Actually, the concept of data migration is mentioned very often. Data migration is the process of transferring data between various computer or informations systems, formats and storage types [1]. We can apply this concept to all areas where we work with data, such as information systems, file systems, storage types, databases, etc. There are many reasons for data migration, the most common include the following:

1. Upgrading to a newer software or hardware
2. Change of company policy
3. Investment to IT services
4. Saving measures
5. Integration of multiple data sources into one system

In the area of relational databases [2] we can provide data migration of databases between various types of relational database management systems, in which databases are implemented (further in text referred as RDBMS). Data migration can be also provided to another version of the same RDBMS, if we need to upgrade this RDBMS. During the process of data migration is important to provide correct migration of the logical structure of database and data migration of data stored in database tables with the appropriate data types. We also need to remember that various types of RDBMS have different properties and characteristics. All RDBMS are based on the relational model [2], [3], but the specific properties and parts of databases can be implemented differently. Differences may be in naming and support of data types, SQL commands for creating and editing database tables or the specific features of the RDBMS, which may not be supported in another RDBMS.

2 Problem Formulation

For clarification of terms we present definition of general process for data migration in the area of database, which will be used further in text.

The general process of data migration, which is used for database migration and data between different types of RDBMS, is called ETL (Extract, transform and load) and consists of these steps:

1. Extracting data from the source database.
2. Transforming data into usable form for migration to the target database.
3. Migration of data to the target database.

ISBN: 978-1-61804-118-0

The ETL process uses the terms source and target databases that are refine there:
Source database – database of source RDBMS, from which data are migrated.
Target database – database of target RDBMS, into which data are migrated.
As we mentioned in the introduction, one of the main differences between different types of RDBMS is the naming and definition of data types, which are supported in specific RDBMS. For illustration, here is an example of selected RDBMS and their data types, which are correspond to standard data type CHARACTER LARGE OBJECT of SQL3 standard [4] for storing long texts:

1. MySQL - MEDIUMTEXT, TEXT, LONGTEXT, BLOB, LONGBLOB
2. Oracle - LONG, BLOB, CLOB, NCLOB
3. PostgreSQL - TEXT
4. MSSQL - TEXT, NTEXT, LONGVARCHAR

From this example we can conclude that the naming of data types and their support in various types of RDBMS are very different. Currently there are a lot of tools for data migration between various types of RDBMS. Here are some of them:

1. SwisSQL Data Migration Tool
2. ESF Database Migration Toolkit
3. DatabaseBridge
4. Cross-Database Studio
5. Data Management Center

From the analysis of these tools we have identified problems and disadvantages of them:

1. Some tools have problems with migration of foreign keys, which are part of relations between database tables. It can be a big problem, because inability of foreign keys migration is the reason to possible data consistency problem in the database. Therefore, the database specialist must create foreign keys manually.
2. There is no possibility to changing parameters or properties of the resulting physical model of the target database during the process of data migration. Process of data migration is invariant, so the database specialist can´t change data types in the target database or optimize logical structure of the target database.
3. Tools can´t be modified or extended – this is caused by commercial character of these tools, but it would be appropriate to modify or expand of the tool for using with another types of RDBMS and their new versions.

3 Problem Solution

For the reasons presented in the previous paragraphs, the main goal of the article is to propose a tool, which will reduce or solve the problems or disadvantages of the current tools for data migration. The proposed tool contains the expert system for decision support during the process of data migration between various types of RDBMS. An expert system proposes and evaluates suitable data types for attributes of database tables of the target database, implemented in target RDBMS. The user then selects final data type, which will be used to construct a physical model of the target database. An expert system also contains knowledge base of IF-THEN rules, which is used for evaluating of the proposed data types. The process to evaluate the suitability of the selected data type using the proposed expert system is shown in the following figure:

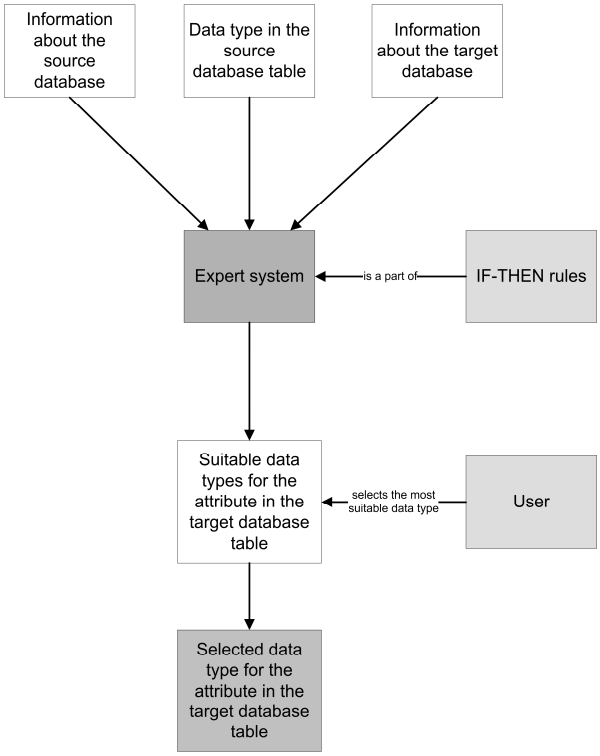

Fig. 1 Scheme of the proposed expert system

From the figure we can conclude the input and output linguistic variables, which are the basis for establishing a knowledge base of IF-THEN rules:
Input linguistic variables:
1. Type of the source RDBMS (SOURCE_RDBMS) – a member of the set of RDBMS {Oracle, MySQL, MSSQL, PostgreSQL}
2. Data type of the attribute in the source database table, supported by selected source RDBMS

(SOURCE_DATA_TYPE) – VARCHAR, CHAR, TINYTEXT, MEDIUMTEXT, LONGTEXT, etc.
3. Type of the target RDBMS (TARGET_RDBMS) – a member of the set of RDBMS {Oracle, MySQL, MSSQL, PostgreSQL}
Output linguistic variable:
1. Data type of the attribute in the target database table, supported by selected target RDBMS (TARGET_DATA_TYPE) – VARCHAR2, NVARCHAR2, TEXT, CLOB, BLOB, etc.

For the creating of knowledge base and evaluation of data types, the LFLC tool will be used, which is described in [5]. LFLC tool was verified on various implementations [7], [8], [9].
For the definition of fuzzy sets for specific linguistic variables we can use difference functions. An example of the usage of triangular function is the definition of fuzzy sets for the types of the source RDBMS, where each type of RDBMS is represented by triangular function:

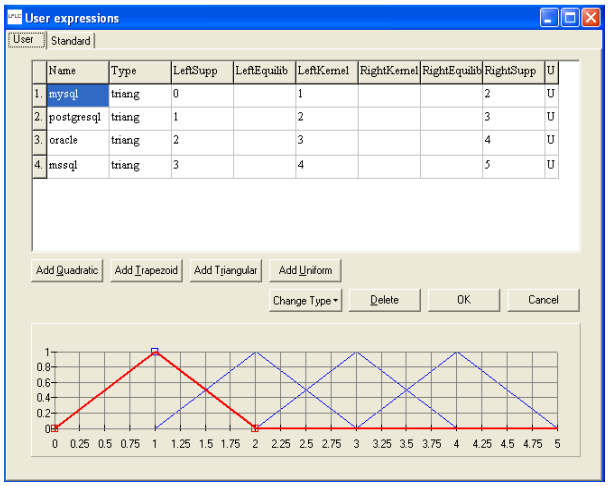

Fig. 2 Definition of fuzzy sets in LFLC tool

In the input data file, which is used by expert system, we fill the appropriate RDBMS using numbers:
1 – MySQL, 2 – PostgreSQL, 3 – Oracle, etc.

Based on the input and output linguistic variables we can create knowledge base of IF-THEN rules. Several examples of IF-THEN rules are below:

```
IF (SOURCE_RDBMS IS MYSQL) AND
(SOURCE_DATA_TYPE IS VARCHAR) AND
(TARGET_RDBMS IS ORACLE) THEN
(RATE OF VARCHAR2 IS EXTREMELY BIG)

IF (SOURCE_RDBMS IS MYSQL) AND
(SOURCE_DATA_TYPE IS TEXT) AND
(TARGET_RDBMS IS ORACLE) THEN
(RATE OF VARCHAR2 IS SMALL)

IF (SOURCE_RDBMS IS MYSQL) AND
(SOURCE_DATA_TYPE IS CHAR) AND
(TARGET_RDBMS IS ORACLE) THEN
(RATE OF BLOB IS VERY SMALL)
```

For clarity and practicability of the proposed expert system, we will divide knowledge base into smaller knowledge bases. Each knowledge base will represent one data type supported by the target RDBMS and contains all IF-THEN rules, which can be used for evaluating of data type, which is represented by knowledge base. The knowledge bases of the selected data types, supported by the target RDBMS Oracle can be created in LFLC tool as the .rb files:

- varchar2.rb
- nvarchar2.rb
- text.rb
- blob.rb
- integer.rb
- smallint.rb

We use the hierachical knowledge base, which is composed by specific knowledge bases for data types supported by the target RDBMS, for the proposal and evaluation of suitable data types. Evaluation process using a hierarchical knowledge base is shown in the following figure:

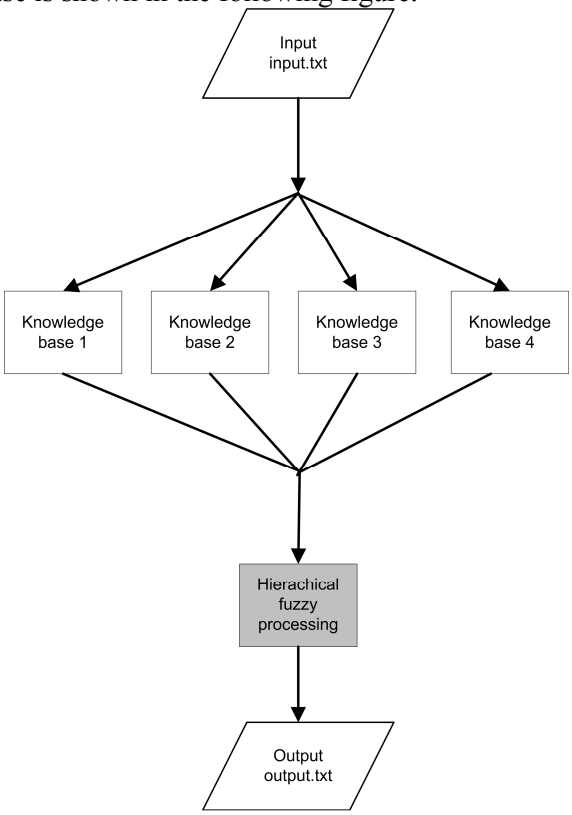

Fig. 3 Hierarchical knowledge base

ISBN: 978-1-61804-118-0

A hierarchical knowledge base for the proposal of suitable data types for the target database implemented in the target RDBMS Oracle with selecting the data type VARCHAR in the source database, implemented in the source RDBMS MySQL, is shown below:

- varchar2.rb
- nvarchar2.rb
- char.rb
- long.rb
- blob.rb

As we can see, the hierarchical knowledge base consists of appropriate knowledge bases for evaluating character data types supported by RDBMS Oracle.

Evaluation of suitable data type is represented as the number of fuzzy set A on universe U, which is understood as special function:

where the values are the degrees of membership of an element in the fuzzy set A [6].

In our case values determine suitability of the data type for an attribute in a database table of the target database. As the value is higher, the proposed data type is more suitable for attribute in a database table of the target database.

An example of the evaluation of suitable data types for the data type VARCHAR in source database is shown below. The target RDBMS is Oracle:

Table 1 Evaluation of suitable data types

SOURCE RDBMS	SOURCE DATA TYPE	TARGET DATA_TYPE	RATE
MySQL	VARCHAR	VARCHAR2	0.99
MySQL		NVARCHAR 2	0.99
MySQL		CHAR	0.97
MySQL		LONG	0.05
MySQL		BLOB	0.03

.

4 Results

The proposed expert system will be verified by created tool for data migration between various types of RDBMS, which contains the expert system. Process of data migration provided by the proposed tool is divided to five steps, which are explained below. The selected database for data migration is used to storing information about faculties, departments, students and subjects. Database consists of these database tables:

1. faculty
2. department
3. student
4. subject

4.1 Specification of the source and target RDBMS

Firstly we need to specify the source and target RDBMS. This is very important, because from the source RDBMS we will load data types of the source database tables, and with the selected target RDBMS we can choose appropriate hierarchical knowledge bases for successful proposal and evaluation of suitable data types for target database tables.

In our case we choose MySQL as the source RDBMS of the source database and Oracle as the target RDBMS of the target database.

Then we must ensure connection to the source RDBMS. In our tool we connect to the source database via JDBC driver. Finally we choose the type of the target RDBMS from prepared set of RDBMS. The process of connecting to the source RDBMS and choosing the target RDBMS is shown in the following figure:

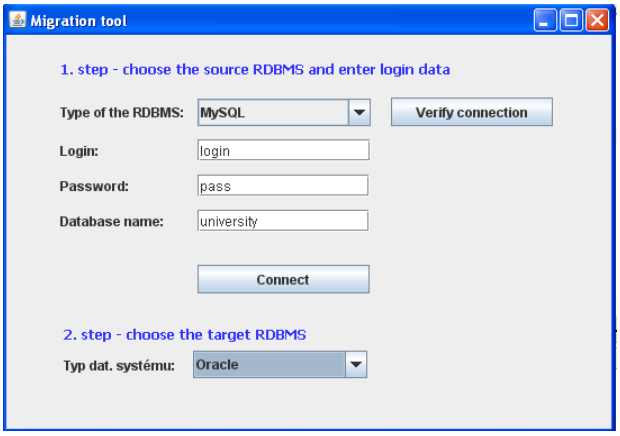

Fig. 4 Specification of the source and target RDBMS

4.2 Loading the logical structure of the source database

In the second step we need to load the logical structure of the source database.

We need to get these parts of the source database:

a) Database tables and their attributes
b) Data types of attributes
c) Relationships between database tables

ISBN: 978-1-61804-118-0

We can load these parts of the source database implemented in RDBMS MySQL by the predefined methods in java.sql package:

```
DatabaseMetaData meta =
con.getMetaData();
tables =
meta.getTables(null, null, null,
new String[]{"TABLE"});
while (rs.next())
{
  String table_name =
rs2.getString("TABLE_NAME");
  ResultSet rsPrimaryKeys =
  meta.getPrimaryKeys(null, null,
  table_name);
   while (rsPrimaryKeys.next())
   {
     columnPrimaryKey =
     rsPrimaryKeys.getString("COLUMN
     _NAME");
   }
  ResultSet rsColumns =
  meta.getColumns(null, null,
  table_name, null);
   while (rsColumns.next())
   {
     String columnName =

  rsColumns.getString("COLUMN_NAME"
);
     String columnType =

  rsColumns.getString("TYPE_NAME");
     String DataType =

  rsColumns.getString("DATA_TYPE");
     int columnSize =

  rsColumns.getInt("COLUMN_SIZE");

   }
}
```

Logical structure of the source database is shown in the following figures, in first figure there are database tables, attributes and data types, in second figure there are relationships between database tables:

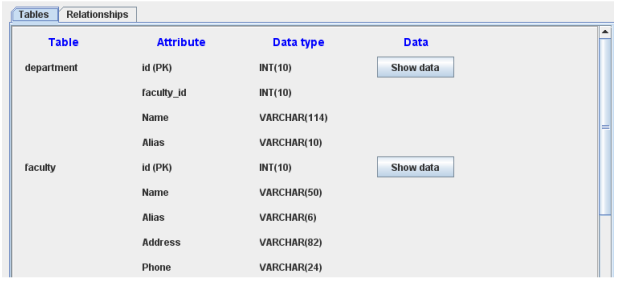

Fig. 5 Database tables of the selected database

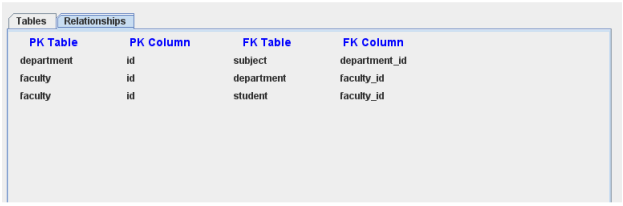

Fig. 6 Relationships between database tables in the selected database

4.3 Proposal of suitable data types by expert system

In this step the expert system proposes appropriate data types for the target database tables and their attributes. Firstly the tool is looking for suitable hierarchical bases for each attribute in the source RDBMS. Then the expert system evaluates appropriate data types proposed for the specific attributes in the target database tables. Finally the tool shows an evaluated and sorted list of suitable data types for each attribute in the target database. The following figure shows attributes of the target database tables and predefined data types for each attribute:

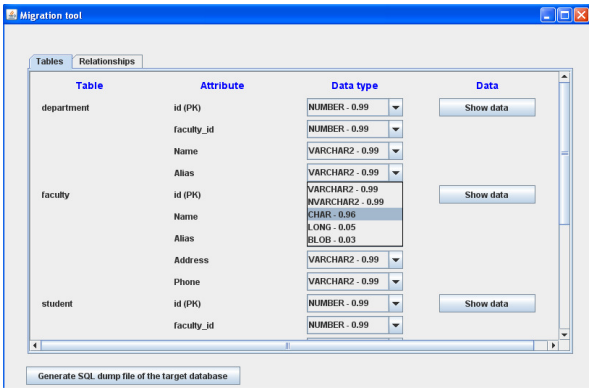

Fig. 7 Suitable data types for the target database

4.4 Selection of suitable data types

The database specialist chooses the most suitable data type for each attribute in database tables of the target database or he can use predefined data types, which are proposed by expert system as the first

suitable data types for the specific attribute. The following figure shows possibility of choosing various character data types for the specific attribute:

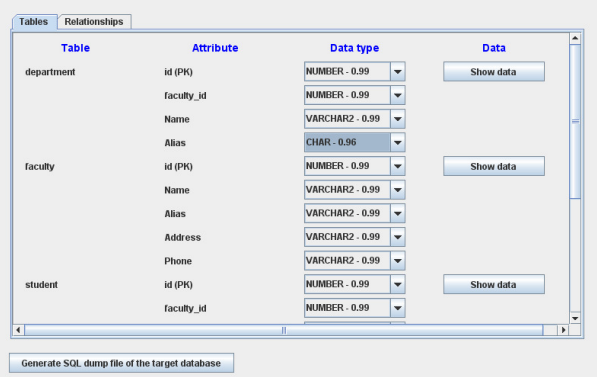

Fig. 8 Selecting the most suitable data type for the specific attribute

4.5 Generating SQL dump file for creating the target database in the RDBMS

Finally the tool generates SQL dump file consists of complete physical model of the target database (database tables, attributes, primary and foreign keys) and data contained in database tables. SQL dump file can then be imported to RDBMS Oracle.

The following figure presents the database table *department*, which was migrated from the source RDBMS MySQL to the target RDBMS Oracle as the completion of the data migration process:

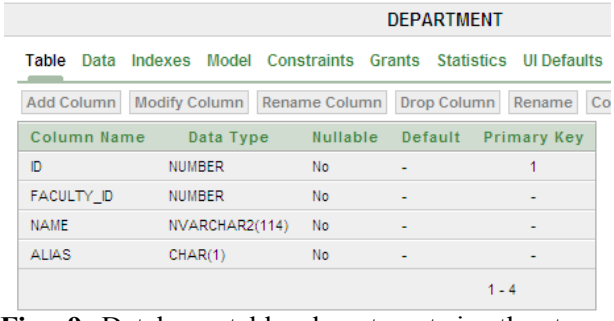

Fig. 9 Database table department in the target RDBMS

5 Conclusion

The article analyzed current state in the area of database migration between various types of RDBMS and identified problems and disadvantages of current tools for migration of databases and their data. Then we proposed the expert system for decision support during the migration process. An expert system proposes and evaluates suitable data types for attributes of database tables in the target database. After this the database specialist chooses the most suitable data types for specified attributes in the target database. Proposed expert system was verified by created tool, which can migrate database and their data from the source RDBMS to the target RDBMS. Finally, we presented created tool on the process of data migration of the selected database migrated from MySQL to Oracle. An expert system and its knowledge base can be easily extended for using in various types of RDBMS.

6 Acknowledgment

Presented topic is also a part of internal grant SGS10/PřF/2012, called Fuzzy modeling tools for analysis and design of information systems, at Department of Infomatics and Computers, University of Ostrava.

References:

[1] Morris J., *Practical data migration*, The British Computer Society, 2009.

[2] Atzeni P., De Antonellis V., *Relational Database Theory*, The Benjamin/Cummings Publishing Company, Inc., 1993.

[3] Harrington L. Jan., *Relational database design and implementation*, Elsevier Inc., 2009.

[4] ISO, ANSI: Framework (SQL/Framework), International Organization for Standardization, 2003.

[5] Habiballa H., Novák V., Dvořák A., Pavliska V., Using software package LFLC 2000, *2nd International Conference Aplimat 2003*, 2003, pp. 355-358.

[6] Klimeš C., Model of adaptation under indeterminacy, *Kybernetika*, Vol.47, No.3, 2011, pp. 355 – 368.

[7] Bartoš J. Walek B., Smolka P., Procházka J., Klimeš C., Fuzzy modeling tools for information system testing, *17th International Conference on Soft Computing Mendel 2011*, Brno, 2011, pp. 154-161.

[8] Walek B., Bartoš J., Smolka P., Masár J., Procházka J., Klimeš C., Creating component model of information system under uncertainty, *18th International Conference on Soft Computing Mendel 2012*, Brno, 2012, pp. 221-226.

[9] Walek B., Bartos J., Klimes C., Prochazka J., Smolka P., Masar J., Pesl M., Fuzzy modeling tool for creating a component model of information system, *International Conference on Fuzzy Systems and Neural Computing 2012*, Paris, 2012, pp. 732-735.