Software system for managing and analyzing data using open source technologies

Student: Marinescu Dorin Alexandru (dorinalexandru.marinescu@gmail.com)
Coordinator: Cristian Mihăescu, PhD., (mihaescu@software.ucv.ro)
Presentation date: July, 2012

Project Goal

This application has two main goals: **data management** and **intelligent data analisys**.

Data management is given by the capability of transfering database records into XML files. For each table from the database will be created one XML file, which contains the corresponding records and all the details necessary for the reverse operation: the restoring of the database using the entries from the XML file.

The application also has an **intelligent data analisys** module, realized using WEKA platform and the **KMeans clustering algorithm**, which can send suggestions to persons registered in the system in order to improve communication between them. The data needed for this module is gathered from the database (all the messages sent by a user from the system). Using this data, the system generates the three attributes necessary for generating three clusters: the users ID, the number of messages sent by each user, the average of number of characters of the messages sent by each user. This module also analyze the problem of choosing the optimum number of clusters, by generating a chart. The optimum number of clusters is given by the point from the elbow of the curve.

Involved Technologies

XML	Data representation: records from database are transfered to XML files using DOM (Document Object Model) parsing, and the database is restored using the entries from XML files using SAX (Simple API for XML) parsing.
Reflection API	Is used to maintain the generic nature of the application.
Object Relational Mapping (Hibernate)	Is used for representing the records from the database as Java objects and to perform operations on the database.
JUnit	Is used for testing the data integrity and the accuracy of the operations. The results are sent by e-mail.
ANT	Is used for scheduling the unit tests to perform at specific moments of time.
Weka	Is used for applying the KMeans clustering algorithm on the data from the database.

Project Output

Data Management Tool	Converts the records from the database into XML files and restores the database using data from generated XML files and contains a view of the main entities necessary for a faculty administration.
Data Analysis Tool	Iused for obtaining three clusters for grouping the users depending on the three attributes mentioned above and drawing one chart in order to determine the optimum number of clusters.