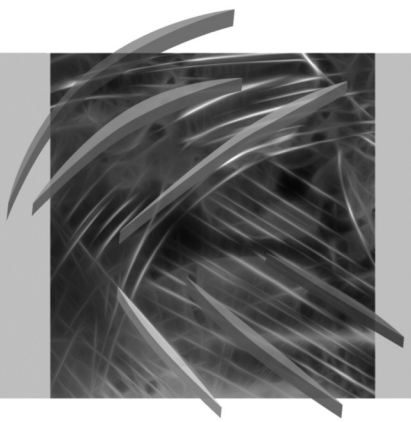
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

232

Knowledge Acquisition and Management



edited by
Małgorzata Nycz
Mieczysław Lech Owoc

Publishing House of Wrocław University of Economics
Wrocław 2011

Reviewers: Grzegorz Bartoszewicz, Witold Chmielarz, Halina Kwaśnicka,
Antoni Ligęza, Stanisław Stanek

Copy-editing: Marcin Orszulak

Layout: Barbara Łopusiewicz

Proof-reading: Barbara Łopusiewicz

Typesetting: Beata Mazur

Cover design: Beata Dębska

This publication is available at www.ibuk.pl

Abstracts of published papers are available in the international database
The Central European Journal of Social Sciences and Humanities
http://cejsh.icm.edu.pl and in The Central and Eastern European Online Library
www.ceeol.com as well as in the annotated bibliography of economic issues BazEkon
http://kangur.uek.krakow.pl/bazy_ae/bazekon/nowy/index.php

Information on submitting and reviewing papers is available
on the Publishing House's website www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192
ISBN 978-83-7695-200-0

The original version: printed

Printing: Printing House TOTEM

Contents

Presentations

Streszczenia

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 232
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS
Knowledge Acquisition and Management ISSN 1899-3192

Maciej Pondel
Wrocław University of Economics

DATA MINING WITH MICROSOFT SQL SERVER 2008

Summary: Data mining is becoming a more and more popular way of processing data in business. Business processes generate tremendous data volumes that contain applicative knowledge describing business. If we are able to acquire this knowledge, we can reach the competitive advantage. Major producers of the software supporting business have among their products also data mining tools. This paper focuses on showing the data mining process and the implementation of the process in Microsoft SQL Server 2008.

Keywords: data mining, association rules, MS SQL Server.

1. Introduction

Current business life generates a huge amount of data. Data is stored in databases to support everyday business processes and provide operational activities in a company. Data is also required for a company to provide analytic processes. In decision-making processes we need information about the processes in a company (about our products, customers, services, and transactions that we have already carried out). Such information can be provided in the form of reports. The reports can be shipped in static paper form or by special IT system enabling to generate interactive reports that have the form of pivot tables, charts, or any other forms enabling a user to build his or her own query. Such systems are very often build as a user interface for browsing data warehouse or any other analytical database. A more sophisticated way of analysing data is the processes of data mining.

Data mining is defined as the process of discovering patterns in data. The process must be automatic or (more usually) semi-automatic. The patterns discovered must be meaningful in that they lead to some advantage usually an economic advantage [Witten, Frank 2005]. David Hand says that data mining is "the discovery of interesting, unexpected, or valuable structures in large data sets" [Hand 2005].

Using data mining, we can solve specific types of problems. We do not use it to get numeric value of benefits or costs concerning our product, service, or group of clients. We receive such information from classical reports or from data warehouses. We use data mining when we want to acquire the rules that describe business processes in our company. Being aware of those rules, we can change processes to make them more profitable.

The most popular examples of data mining problem in business are analysis of the former customer's behaviour to identify the characteristics of those who:
– can be attracted by our product or service,
– may resign from our services and move to our competitor,
– should be treated as risky for claim in insurance,
– are risky to go bankrupt and not be able to pay off a credit.

Obviously, there are many more possibilities of taking advantage of the data mining in business.

Microsoft SQL Server is a classical database server that plays also an analytical role because it is equipped with business intelligence extensions called Analysis Services. This is the tool for multidimensional analysis (OLAP Server) and for data mining. This paper shows the scope of data mining options and describes the data mining process on example data.

2. Data mining process

We must know that data mining is not a magical option which we apply to the existing data and as a result we get valuable knowledge that help us to increase our business efficiency. Edelstein says that "if you've got terabytes of data and you're relying on data mining to find interesting things in there for you, you've got lost before you've even begun" [after Beck 1997]. To conduct the data mining process, we need to formulate the problem regarding the structure and the content of the data that we possess. We also need to understand the business very well to be able to interpret results.

MS SQL Server provides a user with tools that support the whole process of data mining. The process consists of the following steps [MacLennan 2009; Owoc 2003]:

1) business problem formation;
2) data collection;
3) data cleaning and transformation;
4) model building;
5) model assessment;
6) reporting and prediction;
7) application integration;
8) model management.

The first step of the data mining process is to **formulate a business problem**. For some decision problems, reporting or OLAP may be sufficient. When we apply the data mining solution, it is possible that we will not receive back any valuable knowledge. Fortunately, research studies show that average return of investment (ROI) for data mining projects is 150%.

Data collection is necessary because data are stored in a company in many different databases. When we want to analyse them, they appear very often to be

incomplete. They need to be supplemented by the data from another internal database or by the data acquired somewhere outside (like demographic or geographic data).

In the whole data mining project, data **cleaning and transformation** is the longest and most resource-consuming step. Sometimes irrelevancy and noise can appear in the data that we want to use in data mining. Such problems should be solved to make our data useful. Among the problems with data we can specify:

– missing values;
– outliers;
– too detailed data.

Missing values may be caused by different reasons such us joining data from two sources where in one of them some attributes are not required or some records are not present. Sometimes our IT system gathering data do not force a user to complete all the attributes and they are left blank. There may be also some technical problems with reading data especially in some historic volumes that were stored in old systems or on some old or failing data carrier. We can resolve the problem of missing values in two ways:

– discard the whole record with missing values (it can influence the model that we are building if there are a huge number of discarded records);
– replace missing values by another value (for example, last value, most popular one, mean value, or value calculated in some other way).

Outliers are abnormal data that can be real data or errors. They can influence the quality of our mining model. The best solution in most cases is to get rid of such instances. We can afford to remove them because these are extremely untypical values and regard only a very small amount of records. For example, analyzing the behaviour of our customers, we can skip the small amount of them with extremely high income knowing that we also skip probably mistaken records.

Sometimes, dealing with data describing every single transaction, when amount of the transactions is too high (every phone call or ever visited web page), we need to aggregate data before any further analysis. We should choose the most important parameters describing customer behaviour and count their values basing on detailed information that we have. Data mining processes applied on very detailed data can consume too much time and resources, but sometimes the structure of detailed data is improper for data mining algorithms. For example, when we investigate the number of transactions and their value done by a customer in every month, we should not apply data mining on the structure storing every single transaction. First of all, monthly aggregations must be calculated.

Besides data cleaning, to start the data mining process, sometimes we have to perform data transformations. We do it because:

– we would like to change data structure to adapt it to chosen data mining algorithm;
– we need to modify the meaning of some values.

We modifying the meaning of values in two cases:

1) When we want to discretise numeric values. Having continuous data may not give us expected results that is why we want to bin the values into the buckets (like changing the specific city population to predefined groups).

2) When discrete data has more distinct values then required. In this case we can merge distinct values into groups. We can for example join postal codes to the group of codes from one city.

All the data transformation can be done by SQL commands that transform data from source tables to target structures. There is also a tool called SQL Server Integration Services (SSIS). It is ETL class tool that provides user functionality of data extraction, transformations, and loading. The idea of this tool is to build a control flow that allows for an execution of various tasks in sequential order, loops or in a flow determined by the result of previous tasks. The example flow is presented in Figure 1, showing the designer of Integration Services.

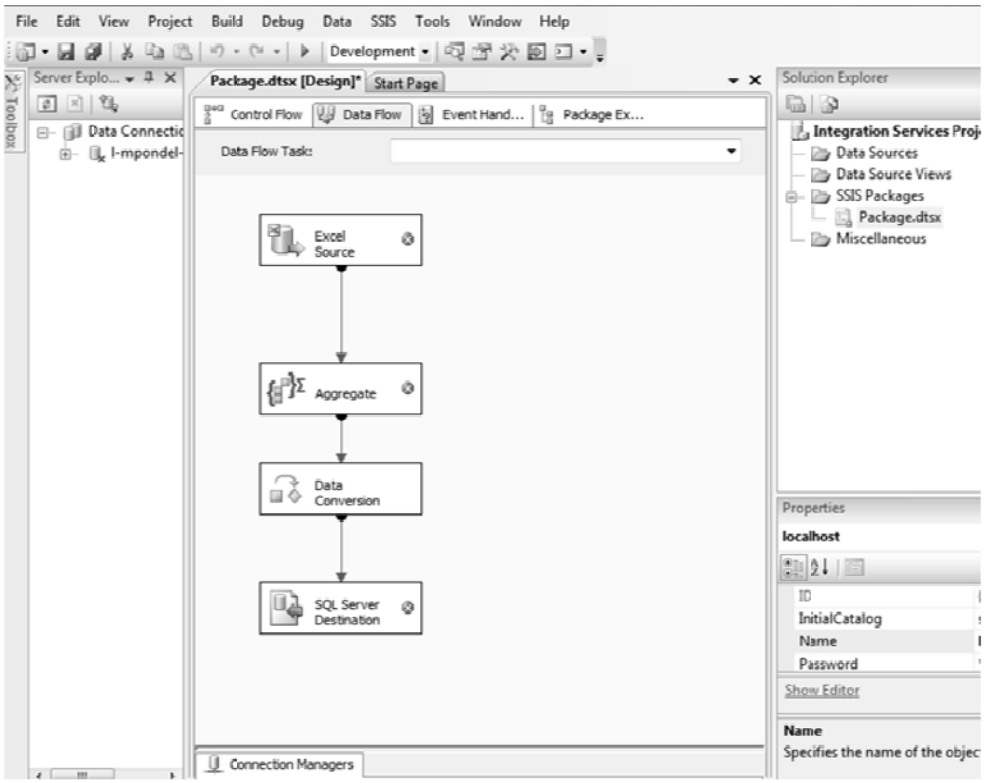


Figure 1. SSIS Designer

Source: author's own study.

The next and most crucial step of data mining process is **building data mining model**. The model is based on the chosen data mining task and on the chosen algorithm. In MS SQL Server 2008 there are implemented algorithms providing following data mining tasks:

1) classification;
2) clustering;
3) association;
4) regression;
5) forecasting;
6) sequence analysis;
7) deviation analysis.

We can divide those data mining tasks into two groups called:

– predictive modelling;
– pattern discovery.

In predictive modelling, called also supervised prediction or supervised learning, we try to identify relationships between the input values and the target. Input values are attributes describing the case and the target value is the result of classification.

Pattern discovery also known as unsupervised learning or unsupervised classification investigates relationships between input values or finds similarities between them. There is no specific target identified.

The accuracy of the built model can depend on the nature of source data and on the parameters of a chosen algorithm. For some data mining tasks, there are many algorithms solving a problem but in a different way. Good practice is to build more than one model solving the same problem and compare results. If we have for example a classification problem, we can build a decision tree, naïve bayes, and neural network. Every algorithm has also individual parameters that can be tuned to optimiseamodel. A model built in this step will be used in next steps to predict the value of the target (supervised learning) or to analyse it (unsupervised learning).

After a model is build, we have to make **assessment**. We need to determine its accuracy and examine how or if it is worth applying in the real business. The accuracy is calculated during the process of model validating and testing. Usually the data set that we posses should be split into three sets (learning, validating, and testing). A learning set is used to build a model, validating one is used to build the second model. The validating model is compared to the original one. The model should be applied also in the cases coming from testing set. This way we can compare the original value of the target with the one predicted by a model.

The next step is to apply the model on current data. This step is **called reporting and prediction**. The results of applying the model should be delivering information to executives who make adecision. MS SQL Server is equipped with reporting mechanism called Reporting Services, which can generate reports directly from mining results. Also the model is stored in the tables so they can be explored by reporting tools.

Mining models may be automatically implemented in enterprise IT systems. This step is called **application integration** because prediction or pattern discovery

may become a part of the business process supported by CRM, ERP, or any other system. For example, we can perform client segmentation as a feature of CRM system or select customers to be offered some product or service. Our ERP system can prepare forecasts of the production basing on the mining model.

In some cases, patterns found during modelling is very stable and can be used for some time without any modifications. But most often patterns changes dynamically. In marketing, new appearing products may determine new rules to be discovered. That is why data miners must perform **model management**. In some cases we should build new models and asses them. If they are not accurate enough, the whole process must be repeated (starting from data collection, cleaning, and transformations). Sometimes, however, model rebuilding can be an automated step. SQL Server Integration Services have an option of automated model processing that can be scheduled and run automatically.

3. Available algorithms

Developers are allowed to prepare their own mining algorithms and apply them in MS SQL Server 2008. Obviously, there are Microsoft algorithms provided in MS SQL Server 2008.They are:

1) Microsoft Naïve Bayes;
2) Microsoft Decision Trees Algorithm;
3) Microsoft Time Series Algorithm;
4) Microsoft Clustering;
5) Microsoft Sequence Clustering;
6) Microsoft Association Rules;
7) Microsoft Neural Network;
8) Microsoft Logistic Regression.

The kind of data mining tasks and matching algorithms are presented in Table 1.

Table 1. Algorithms and data mining tasks

Data mining algorithm	Data mining task
Microsoft Naïve Bayes	Classification
Microsoft Decision Trees Algorithm	Classification, association
Microsoft Time Series Algorithm	Forecasting future series based on history
Microsoft Clustering	Clustering, anomaly detection
Microsoft Sequence Clustering	Clustering, anomaly detection id sequence data
Microsoft Association Rules	Market basket analysis (association rules)
Microsoft Neural Network	Classification
Microsoft Logistic Regression	Classification

Source: author's own study.

4. Data mining tools

There are three main approaches to building data mining models. We can build the model using:
- SQL Server Business Intelligence Development Studio;
- Query Language for performing Data Mining operations – Data Mining Extensions to SQL (DMX);
- SQL Server 2008 Data Mining Add-Ins for Microsoft Office 2007.

SQL Server Business Intelligence Development Studio is more a developer tool that is based on programming environment of Microsoft – Visual Studio. It gives us the whole data mining functionality. To perform a data mining task, we create a data mining project. In this tool, the data miner has possibility of:
- defining a data source (which is normally the SQL table or a set of SQL tables);
- manging source data by creating relationships diagram;
- exploring data using pivot tables and pivot charts;
- building a model on chosen data set with a chosen data mining technique and algorithm;
- exploring a model with Mining Model Editor;
- assessing a model with a set of Accuracy Charts.

SQL Server Business Intelligence Development Studio is also a tool in which we build Integration Services projects.

Next option of building data mining models is to create them using DMX queries. This language gives us the whole functionality of building and browsing a model. Obviously, it is the most difficult option to manage data mining.

The most common way of data mining for the majority of users is using MS Excel ADD-In that allows connecting to SQL Server and performing data mining. This is a tool more for an analyst than a developer. It combines Excel features such us simplicity with advanced options of MS SQL Server Data mining. We provide data for data mining in Excel and from the Excel user interface we launch data mining operations. Excel is also the user interface of browsing the results of data mining.

5. Examples of data mining

For sample data mining we will choose an association rules mining task. The input data contains a set of hobbies of customers. We have data in transactional format (every row is described by a customer id and a customer hobby). Our task is to find associations between hobbies. Some sample records are presented in Table 2.

We will use Microsoft association rules to identify rules between different hobbies.

Table 2. Sample record describing hobbies of customers

CustomerID	Hobby
877687	Business and investing
877687	Travel
877687	News and media
877687	Kids and family
877687	Camping and hiking
877687	Science and technology
877723	Other
877723	Computer
877723	Travel

Source: author's own study.

The result of data mining is presented in Figure 2.

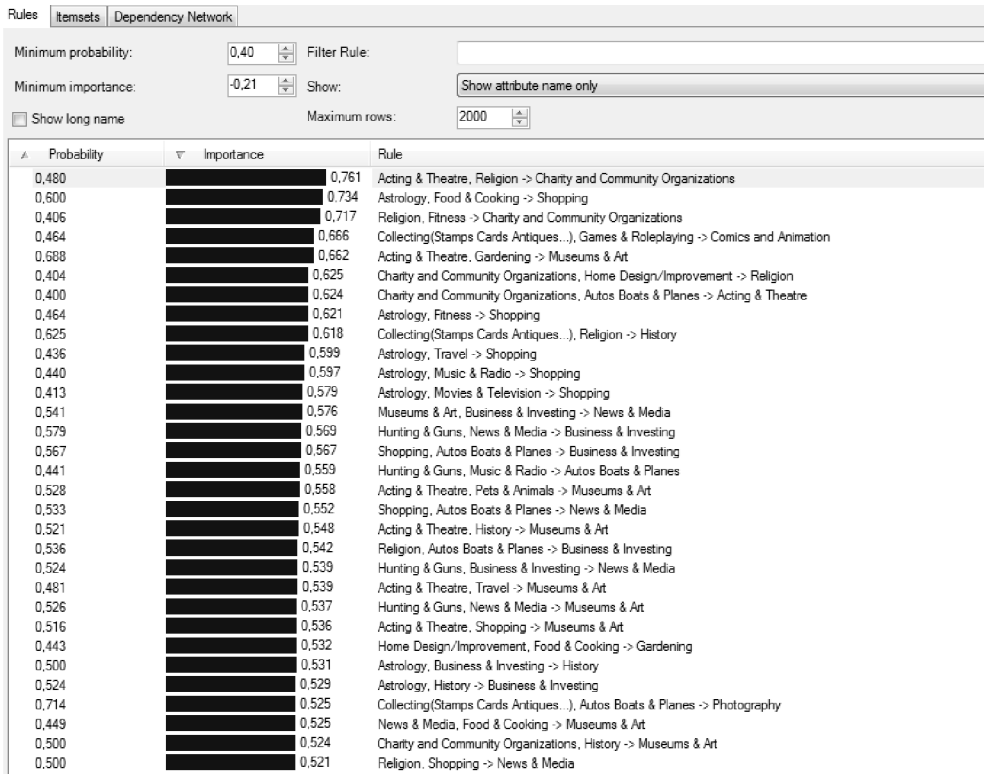


Figure 2. Part of the result of data mining

Source: author's own study.

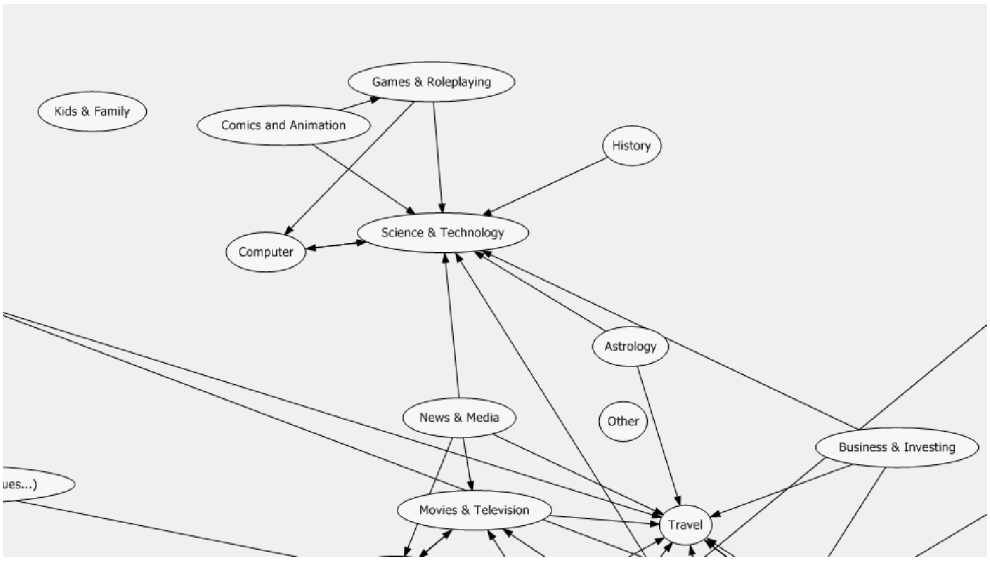


Figure 3. Dependency network showing association rules

Source: author's own study.

As we see the result of association rules is a set of rules generated by an algorithm. Each rule contains two indicators: probability and importance. The rule have a following form: "If A, then B". In this case A is a logical expression containing A1 and A2. Probability, sometimes called support, is probability that all the items in the rule (A, B) occur together among all transactions. Importance, also called confidence, is a conditional probability that all the items in the rule occur together in a set of transactions where A occurs. We can also observe those rules on Dependency Network. It is a graphical interface showing all hobbies and the dependencies between them.

6. Summing-up

SQL server is a commonly used database server that also contains business intelligence functionality. There are quite interesting data mining functions implemented in it what makes the whole solution worth considering when an enterprise looks for a data mining platform. Future research studies of the author will focus on comparing SQL Server 2008 data mining options with other leading data mining solutions such as SAS Enterprise Miner and Oracle Data mining (ODM).

References

Witten I., Frank E. (2005), *Data Mining Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers (Elsevier), San Francisco.

Hand D. (2005), What you get is what you want? Some dangers of black box data mining, [in:] *M2005 Conference Proceedings*, Cary, NC: SAS Institute, The MIT Press, Cambridge.

Beck A. (1997), *Herb Edelstein Discusses the Usefulness of Data Mining*, http://www.tgc.com/dsstar/971014/100007.html.

MacLennan J., Tang Z., Crivat B. (2009), *Data Mining with Microsoft SQL Server 2008*, Wiley Publishing, Indianapolis.

Owoc M., Hauke K., Pondel M. (2003), Building data mining models in the Oracle 9i Environment, [in:] *Informing Science*, Pori.

DRĄŻENIE DANYCH W MS SQL SERVER 2008

Streszczenie: Drążenie danych staje się coraz bardziej popularnym sposobem przetwarzania danych w firmie. Procesy biznesowe generują ogromne ilości danych, które zawierają przydatną wiedzę opisującą biznes. Jeśli jesteśmy w stanie wydobyć tą wiedzę z danych, wówczas możemy osiągnąć przewagę konkurencyjną. Najwięksi producenci oprogramowania wspierającego biznes mają wśród swoich produktów także narzędzia drążenia danych. W tym artykule przedstawiony jest proces eksploracji danych i jego implementacja w Microsoft SQL Server 2008.

Słowa kluczowe: drążenie danych, reguły asocjacyjne, MS SQL Server.