Rapid Miner E-Commerce

DANIEL HUNYADI
"Lucian Blaga” University of Sibiu, Romania
daniel.hunyadi@ulbsibiu.ro

Abstract: The aim of the present paper is to contribute to the existing pool of studies in the field of information technology, at international level, and to integrate these into economic applications. Integrating and using data mining techniques in economy opens new perspectives regarding real-time data manipulation between all decisions factors from companies, achieving new advantages for current companies also on global level, obtaining a superior economical efficiency of all production items involved in the process.

Key-Words: association rules, data mining, E-commerce platform, Rapid Miner

1 Introduction

Programming for companies has evolved a lot lately. This evolution is based especially on the Web. Web allows companies to automate and integrate their business.

Organizations also realize the necessity of using new systems, capable to mine the benefit derived from a close collaboration between partners and suppliers, which can add value by using their experience and know-how. Systems must be redefined and extended in order to integrate Internet, CRM (Customer Relationship Management) system and the new e-commerce models B2B (Business-to-Business) and B2C (Business-to-Consumer).

In many ways, Internet is the most efficient marketing tool which has ever existed. It allows companies to interact to the market on a very high level. The companies can establish a bilateral relationship difficult to imagine in “Production Era”. Internet helps in eliminating “supposing” in detecting client’s wishes. But it cannot indicate its own objectives.

As nowadays economy is a mixture between old and new, the biggest companies are hybrids between old and new. Although marketing competences are still available, new competences and knowledge add a good potential to the development. In the same way, the consumers can be also seen as hybrids. They are tacking advantage by shopping online but still visiting the shops for having a human interact and “shopping experience”.

Online sales are an extension of today trade. The costs are lower and relationship between clients is better. Online sales also make the collaboration between companies easier: reduce marketing and delivery costs, sustain marketing strategy of the company, and conquer new markets more easily. Nearby future, electronic trade will impact strongly on companies’ competition. Furthermore the electronic trade is not restricted by borders but it depends on computer networks and banks infrastructure needed for online payments. Online trades allow even the smallest companies to trade the products around the world. The trade positively influences the activity of small and medium companies, but also helps clients by offering them a great variety of options.

The topics dealt with in this thesis are of utmost actuality for the studied field. Considering the efforts of translating the information society into reality, a series of technologies have been studied updated and implemented which have deep impact on the manner of dealing of consortium companies with business issues on the global market.

2 Data Mining

Data mining is a strong and modern instrument of the information and communication technology, used to extract useful, however unknown information. The tool automates the discovery process of relations and combinations in raw data, the results being then able to be set in an automated support system for decisions.

The data mining methods result from classical statistic calculations, from the administration of databases and from artificial intelligence. They do not replace traditional statistical methods, being more regarded as extensions of graphical and statistical techniques. The results of the data mining methods must be systematically subjected to human supervision because software applications lack
human intuition to distinguish between relevant and irrelevant information.

Data mining involves the application of techniques that transform data into information. At the same time data mining represents the critical interface between synthetic knowledge or machine generated patterns and semantically knowledge required by man for reasoning about the real world.

Data mining is a new and rapidly growing field. It draws ideas and resources from multiple disciplines, including machine learning, statistics, database research, high performance computing and commerce. This explains the dynamic, multifaceted and rapidly evolving nature of the data mining discipline. While there is a broad consensus that the abstract goal of data mining is to discover new and useful information in databases this is where the consensus ends and the means of achieving this goal are as diverse as the communities contributing. The foundations of all data mining methods, however, are in mathematics. Any moderately sized treatment of data mining techniques necessarily has to be selective and maybe biased towards a particular approach. Despite this, we hope that the following discussion will provide useful information for readers wishing to get some understanding of ideas and challenges underlying a selection of data mining techniques. This selection includes some of the most widely used data mining problems like frequent item sets and association rule mining.

Data mining techniques are used to find patterns, structure or regularities and singularities in large and growing data sets. A necessary property of algorithms which are capable of handling large and growing datasets is their scalability or linear complexity with respect to the data size. Scalability in the data mining literature means a time (and space) complexity which is proportional to the size of the data set, i.e., \( O(n) \) if \( n \) is the number of records of a data set. The proportionality "constant" may actually grow slightly as well and complexities like \( O(n \log(n)) \) are usually also acceptable.

Patterns in the database are described by relations between the attributes. In a sense, a relational database itself defines a pattern. However, the size of the relations of the data base makes it impossible to use them directly for further predictions or decisions. On the other hand, these relations only provide information about the available observations and cannot be directly applied to future observations. Methods which are able to generalize their results to future observations are investigated in statistics and machine learning.

The variables or attributes are mainly assumed to be either continuous or categorical. However, more general data types are frequently analyzed in data mining [4]. The techniques discussed here are not based on sampling and access every item in the full data set. However, this does not imply that sampling is unimportant in data mining, in fact, it is often the only way to deal with very large data sets.

Most of the established companies have accumulated masses of data from their customers for decades. With the e-commerce applications growing rapidly, the companies will have a significant amount of data in months not in years. Data Mining, also known as Knowledge Discovery in Databases (KDD), is to find trends, patterns, correlations, anomalies in these databases which can help us to make accurate future decisions.

Data mining deals with the processing of large and complex data. Robust tools are required to recover weak signals. These tools require highly efficient algorithms which scale with data size and complexity. Association rule discovery is one of the most popular and successful tools in data mining. Efficient algorithms are available. The developments in association rule discovery combine concepts and insights from probability and combinatory.

### 3 Association Rules

Large amounts of data have been collected routinely in the course of day-to-day management in business, administration, banking, the delivery of social and health services, environmental protection, policing and in politics. This data is primarily used for accounting and management of the customer base. However, it is also one of the major assets to the owner as it contains a wealth of knowledge about the customers which can assist in the development of marketing strategies, political campaigns, policies and product quality control. Data mining techniques help process this data which is often huge, constantly growing and complex. The discovered patterns point to underlying mechanisms which help understand the customers and can give leads to better customer satisfaction and relations.

While data mining had been studied since before 1988, it was the introduction of association rule mining in 1993 by Agrawal, Imielinski and Swami [2] and the publication in 1995 of an efficient algorithm by Agrawal and Srikant [3] and, independently, by Mannila, Toivonen and Verkamo [9] which initiated a wealth of research and development activity. This research has been dealing with efficiency, applications, the interface with data.
access, and the relation with other concepts like prediction and has strengthened the young discipline and helped establish it as an important and exciting research area in computer science and data processing.

An association rule is an implication or if-then-rule which is supported by data. The motivation given in [2] for the development of association rules is market basket analysis which deals with the contents of point-of-sale transactions of large retailers. A typical association rule resulting from such a study could be “90 percent of all customers who buy bread and butter also buy milk”. While such insights into customer behavior may also be obtained through customer surveys, the analysis of the transactional data has the advantage of being much cheaper and covering all current customers. The disadvantage compared to customer surveys is in the limitation of the given transactional data set. For example, point-of-sale data typically does not contain any information about personal interests, age and occupation of customers.

Understanding the customer is core to business and ultimately may lead to higher profits through better customer relations, customer retention, better product placements, product development but also fraud detection. While originating from retail, association rule discovery has also been applied to other business data sets including: credit card transactions, telecommunication service purchases, banking services, insurance claims and medical patient histories.

However, the usefulness of association rule mining is not limited to business applications. It has also been applied in genomics and text (web page) analysis.

In these and many other areas, association rule mining has lead to new insights and new business opportunities. Of course the concept of a market basket needs to be generalized for these applications. For example, a market basket is replaced by the collection of medical services received by a patient during an episode of care, the subsequence of a sequence of amino acids of a protein or the set of words or concepts used in a web page. Thus when applying association rule mining to new areas one faces two core questions:

- what are the “items” and
- what are the “market baskets”.

The answer of these questions is facilitated if one has an abstract mathematical notion of items and market baskets.

The efficiency of the algorithms will depend on the particular characteristics of the data sets. An important feature of the retailer data sets is that they contain a very large number of items (tens of thousands) but every market basket typically contains only a small subset.

Association rule discovery has originated in market basket analysis. Here the object is a market basket of items purchased by a customer. While many features may be of interest in market basket analysis, the main features studied are the types of items in the market basket.

The market basket example is just one incidence where association rule discovery is used. In general, it is used whenever the objects are sets of items, and, more generally, a collection of properties of the objects, statements which are either true or false.

4 RMEC

Using Java Server Pages (JSP) technology, we have implemented a web application called Rapid Miner E-Commerce (identified by RMEC), built on client-server architecture, comprising:

- the client (also identified by RMEC Client) - a general online store designed for any type of products,
- application server (RMEC Server) - virtual store management console that provides management services and data processing services through Rapid Miner tab.

RapidMiner is an environment developed for data mining processes. It contains a collection of modular operators enabling the complex processing of a large number of data mining problems. These operators do not have to deal with the current format or the different representations of data, the RapidMiner core taking care of all the necessary changes. Data handling is done transparently and allows the access to several types of data sources. Due to the modular concept of the operators, one operator is usually enough to be replaced so as to assess its performances, the rest of the process remaining unchanged. This is an important feature both for scientific research and for the optimization of real world applications [10].

Communication between client and server we made it through the web services, as shown in Figure 1.
There are two services exposed by the server: security service and service logic processing. Security Service is in ro.rmec.server.ws.SecurityWebService and methods are contained in Table 1.

Table 1. SecurityWebService

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>authenticateUser</td>
<td>username, password</td>
</tr>
<tr>
<td>clearAuthentication</td>
<td>username</td>
</tr>
</tbody>
</table>

Service logic processing is in ro.rmec.server.ws.ShopWebService and methods are contained in Table 2.

Table 2. ShopWebService

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>getCategories</td>
<td></td>
</tr>
<tr>
<td>getProducts</td>
<td></td>
</tr>
<tr>
<td>getManufacturers</td>
<td></td>
</tr>
<tr>
<td>getProduct</td>
<td>productId</td>
</tr>
<tr>
<td>getProductsByCategory</td>
<td>categoryId</td>
</tr>
<tr>
<td>getProductsByManufacturer</td>
<td>manufacturerId</td>
</tr>
<tr>
<td>getTax</td>
<td>taxId</td>
</tr>
<tr>
<td>saveOrder</td>
<td>order</td>
</tr>
<tr>
<td>saveBasketProduct</td>
<td>basketProduct</td>
</tr>
<tr>
<td>saveBasket</td>
<td>basket</td>
</tr>
<tr>
<td>saveOrderByProduct</td>
<td>orderProduct</td>
</tr>
<tr>
<td>saveOrderTotal</td>
<td>orderTotal</td>
</tr>
<tr>
<td>saveCustomer</td>
<td>user</td>
</tr>
<tr>
<td>saveCustomerAddress</td>
<td>address</td>
</tr>
</tbody>
</table>

4.1 RMEC – Client

Being developed as a Service Oriented Architecture (SOA), we divide application into services and modules specialized in execution and settlement of certain tasks. Client architecture is shown in Figure 2.

4.2 RMEC – Server

RMEC Server is a web application which is achieved through virtual store management and processing of data mining.

Server provides performance pages dedicated identity management (categories, products, manufacturers, invoices, customers) as well as statistical processing of transactions.

We built an application respecting multi-layer architecture, modular and service-oriented. This architecture gives flexibility, robustness and ease of modification, updating and improvement. Server architecture is shown in Figure 3.
For statistical processing, the application uses Rapid Miner module, which is integrated as external application in a server. This module calls process implemented in RapidMiner, it executes and displays the results in web page. Data transactions to be processed must be exported in external files, because Rapid Miner does not know to process the data stored on a web server. It creates a file with extension .aml containing information about the format and type of data to be saved and a file with extension .dat which will save the desired data.

These files are then taken by RapidMiner process, data are processed, and finally the process will return the desired results in a file with extension .res. RapidMiner module will retrieve the data and using Ajax technology will parse the results and displays them on the web page.

Application modules are:
- **RMEC Server Core** – the main module of the application, mediate communication between modules;
- **Domain** – module that contains all other models and entities used in application;
- **Util** – utility module containing classes / methods helpful in the execution of tasks and processes;
- **External** - RapidMiner application integration module, which contains classes for RapidMiner application;
- **Web** – used in publication of logical models and parsing in HTML format;
- **Persistence Services** – module that contains all persistence services. This made communication with the database. The module can support communication with any database;
- **Business Services** – module of logical level, which mediates communication between the Presentation Tier (Web module) and the Data Tier (module Persistence Services), coordinates the entire application, executes commands and processes;
- **Web Services** – module provide Server communication with its customers, through Web Services. This module communicates directly with the main module, being independent of the remaining modules. Is built service oriented and the exposure operation is independent of their implementation.

Web Application RMEC Server contains a tab RapidMiner in which crowds are options to display frequent items sets and generated association rules. Association rules can be displayed in tabular form or list form.

As input data are used transactions obtained from the web application RMEC Client. At execution of RapidMiner process, data are taken from table rmec-basketball-products and are exported to file database.dat.

Frequent items sets and association rules obtained on test data collected by RMEC client application can be found in Figure 4 and Figure 5.

```
Figure 3. Server Architecture

For statistical processing, the application uses Rapid Miner module, which is integrated as external application in a server. This module calls process implemented in RapidMiner, it executes and displays the results in web page. Data transactions to be processed must be exported in external files, because Rapid Miner does not know to process the data stored on a web server. It creates a file with extension .aml containing information about the format and type of data to be saved and a file with extension .dat which will save the desired data.

These files are then taken by RapidMiner process, data are processed, and finally the process will return the desired results in a file with extension .res. RapidMiner module will retrieve the data and using Ajax technology will parse the results and displays them on the web page.

Application modules are:
- **RMEC Server Core** – the main module of the application, mediate communication between modules;
- **Domain** – module that contains all other models and entities used in application;
- **Util** – utility module containing classes / methods helpful in the execution of tasks and processes;
- **External** - RapidMiner application integration module, which contains classes for RapidMiner application;
- **Web** – used in publication of logical models and parsing in HTML format;
- **Persistence Services** – module that contains all persistence services. This made communication with the database. The module can support communication with any database;
- **Business Services** – module of logical level, which mediates communication between the Presentation Tier (Web module) and the Data Tier (module Persistence Services), coordinates the entire application, executes commands and processes;
- **Web Services** – module provide Server communication with its customers, through Web Services. This module communicates directly with the main module, being independent of the remaining modules. Is built service oriented and the exposure operation is independent of their implementation.

Web Application RMEC Server contains a tab RapidMiner in which crowds are options to display frequent items sets and generated association rules. Association rules can be displayed in tabular form or list form.

As input data are used transactions obtained from the web application RMEC Client. At execution of RapidMiner process, data are taken from table rmec-basketball-products and are exported to file database.dat.

Frequent items sets and association rules obtained on test data collected by RMEC client application can be found in Figure 4 and Figure 5.

```

```

```

```

```

```

```

```

```
5 Conclusion

An E-commerce Web application has been implemented, called Rapid Miner E-Commerce. The design for this Web application started from the idea of functional integration of the e-commerce application with a strong and flexible data mining processor like RapidMiner.

The software has client-server architecture. The client is a virtual shop, where any user can trade one or many products, and server is a console administrating the virtual shop. The server also offers data handling services using Rapid Minder. Communication between client and server is assured by web services. There are two services server is responsible of: security and logical working data.

The software is design as Service Oriented Architecture (SOA), module oriented, having modules specialized in running and solving different tasks. This architecture creates a flexible software, easy to modify, distribute and improve.

Since this type of data (a supermarket transactions or transactions of real e-commerce applications) is confidential, I have not had a chance to run the application on a real database. I managed to generate frequent item sets and association rules only with test data collected by RMEC Client Application.

References:


