PROGRAMS WITH DATA MINING CAPABILITIES

Ciobanu Dumitru
PhD student, University of Craiova, Craiova, Romania,
ciobanubebedumitru@yahoo.com

Abstract: *The fact that the Internet has become a commodity in the world has created a framework for a new economy. Traditional businesses migrate to this new environment that offers many features and options at relatively low prices. However competitiveness is fierce and successful Internet business is tied to rigorous use of all available information. The information is often hidden in data and for their retrieval is necessary to use software capable of applying data mining algorithms and techniques. In this paper we want to review some of the programs with data mining capabilities currently available in this area.We also propose some classifications of this software to assist those who wish to use such software.*

Keywords: *e-Business, Data Mining Software, MATLAB, Neural Networks.*

1. Introduction

In recent years computers have become more powerful and cheaper and the Internet is already used by a third of population of the globe. Using computers to do all sorts of daily tasks has become a necessity. Internet technology is changing faster, and the pace of it's innovation and adoption is truly staggering. Apart from sharing or transacting data/information, all types of business transactions are frequently done through Internet.

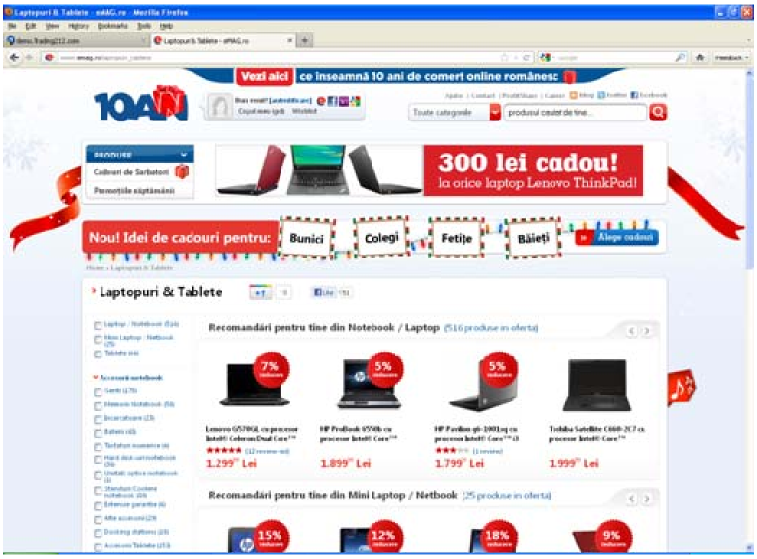

Vezi aici
ce înseamnă 10 ani de comerț online românesc
300 lei cadou!
la orice laptop Lenovo ThinkPad!
Nou! Idei de cadouri pentru:
Bunici
Colegi
Fetițe
Băieți
Laptopuri & Tablete
Recomandări pentru tine din Notebook / Laptop
7%
5%
5%
1.299 Lei
1.899 Lei
1.799 Lei
1.999 Lei
Recomandări pentru tine din Mini Laptop / Netbook
15%
12%
18%
9%

Fig. 1. An online shop

Over time people prefer staying in front of the PC and doing any business transactions for convinience and saving time. The web is now the best medium of doing business. Large companies rethink their business strategy using the web to improve business.

Fig. 1 shows the interface of an online shop, in this case an electronic store, and in Fig. 2 we can see an online business that allows users to perform transactions on the stock exchange.

Business carried on the Web offers the opportunity to potential customers or partners where their products and specific business can be found (Dinucă, 2011b). Business presence through a company web site has several advantages as it breaks the barrier of time and space compared with the existence of a physical office. To differentiate through the Internet economy, winning companies have realized that e-commerce transactions is more than just buying / selling, appropriate strategies are key to improve competitive power.

Knowledge of e-business environment is essential for doing business in this century. New technologies for extracting knowledge from data must be understood and applied. One effective technique used for this purpose is data mining. Data mining is the process of extracting interesting knowledge from data.

For this have developed many software that automates the process of extracting knowledge from data.

Collecting data in various formats, digitization began in the 60s allowing a retrospective analysis of data by computer. In the 80s came relational databases with Structured Query Language (SQL) and application that allows dynamic data analysis. The 90s years are characterized by an explosion of data. To store them it began to use data warehouses. In response to the challenges faced by the community of specialists in database data mining appeared, dealing with massive amounts of data, applying statistical analysis and search techniques specific to artificial intelligence on the data (Dinucă, 2011a).

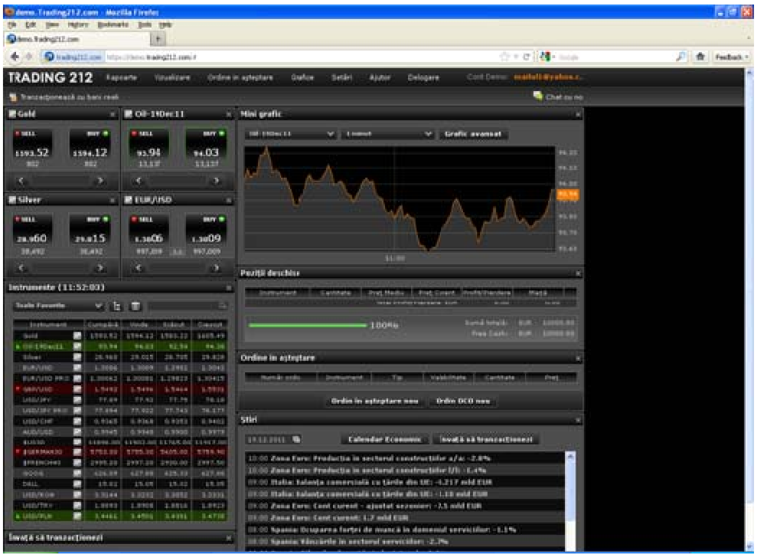


Fig. 2. Another online business

The role of data mining is the extraction of new knowledge, implicit and direct action of large data collections, discovering things that are not obvious from the data, which can not be extracted manually, representing useful information that can improve the current action process.

2. Data mining software

The evolution of data mining techniques is corelated with development of software that applies this techniques and algorithms for extracting knowledge from data. There are a wide variety of software that use data mining to extract knowledge.

Further we will present briefly some of them.

WEKA (Waikato Environment for Knowledge Analysis) is a open source software under the General Public License (GNU), and was developed by the University of Waikato in New Zealand. WEKA includes a large collection of data mining algorithms for learning. In addition to these learning algorithms WEKA contains a variety of instruments that can be used for preprocessing data sets. For examplification of using WEKA have been written many works among which we will remember only two (Scuse & Reutemann, 2007) and (Bouckaert, et al., 2010) that we consider a very good starting point for understanding and using this program.

WebLogMiner is a tool for knowledge discovery from web server logs. This program applies data mining algorithms and techniques to the web logs. WebLogMiner is presented in the paper (Zaine, Xin & Han, 1998).

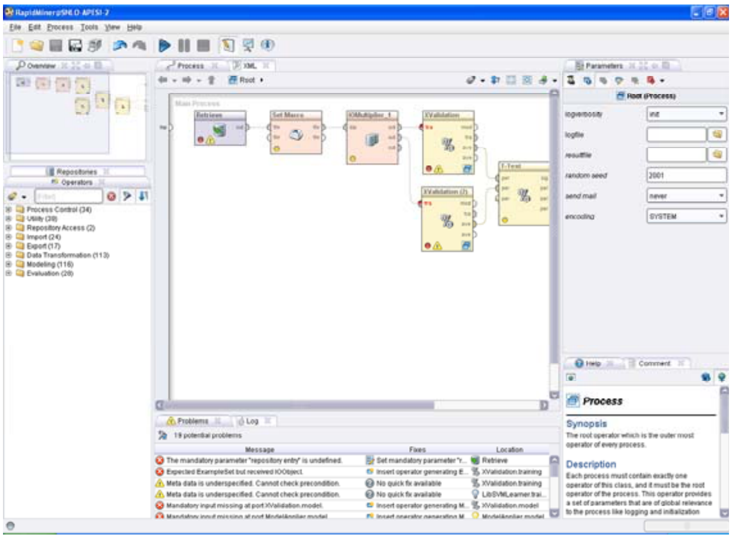
WUM (Web Utilization Miner) is a program used to discover navigation patterns with help of association rules using an extended version of SQL (Spiliopoulou & Faulstich, 1999). In (Etminani, Akbarzadeh & Yanehsari, 2009) is presented WUM using a ant clustering algorithm.

Another program from this area is **WebSIFT** that uses clustering, statistical analysis and association rules (Cooley, Tan & Srivastava, 2000).

RapidMiner (formerly Yale) is an environment for machine learning and data mining processes. A modular operator concept allows the design of complex nested operator chains for a huge number of learning problems. The data handling is transparent to the operators. They do not have to cope with the actual data format or different data views - the RapidMiner core takes care of the necessary transformations. Today, RapidMiner is the world-wide leading open-source data mining solution and is widely used by researchers and companies (RapidMiner 4.4. User Guide, 2009).

RapidMiner has embedded WEKA and released a very popular free version that led to the widespread use of the program.

In Fig. 3 we can see the main window of RapidMiner. It shows that is working with modules, as in the case of Simulink from Matlab, which are assembled and customized according to user needs.



Root (Process)

logverbosity	init
logfile	
resultfile	
random seed	2001
send mail	never
encoding	SYSTEM

Problems Log

19 potential problems

Message	Fixes	Location
The mandatory parameter "repository entry" is undefined.	Set mandatory parameter "r...	Retrieve
Expected ExampleSet but received IOObject.	Insert operator generating E...	XValidation.training
Meta data is underspecified. Cannot check precondition.	No quick fix available	XValidation.training
Meta data is underspecified. Cannot check precondition.	No quick fix available	LibSVMLearner.trai...
Mandatory input missing at port XValidation.model.	Insert operator generating M...	XValidation.model
Mandatory input missing at port ModelApplier.model	Insert operator generating M...	ModelApplier.model

Process

Synopsis

The root operator which is the outer most operator of every process.

Description

Each process must contain exactly one operator of this class, and it must be the root operator of the process. This operator provides a set of parameters that are of global relevance to the process like logging and initialization

Fig. 3 The interface of RapidMiner software

Web Data Extractor is a utility that allows you to automatically extract specific information from web pages in a few simple steps. With this application, you will be able to get lists of meta-tags, e-mails, phone and fax numbers, etc. and store them in different formats for future use. Extract targeted company contact data (email, phone, fax) from web for responsible business to business communication. Extract url, meta tag (title, keyword) for website promotion, search directory creation, web research.

BulkVerifier is an multi-threaded application for checking e-mail addresses and domain availability. Quickly and easily verifies email addresses, huge domains and clean up your mailing / domain list. BulkVerifier can process both plain list of e-mail addresses / domains where each line contains one item and files of more complex structure like CSV file where lines represents multi-field records of the same structure (i. e. containing the same fields separated with the same delimiter). For example, you can export a worksheet of an MS Excel file to check availability of e-mail addresses/domains listed there.

LisMotor is an application for e-mail lists management. To ensure the efficient and high-performance direct e-mail marketing you will need more than just e-mail addresses and a mailing application. ListMotor operates with input and output data stored in simple text files or csv files, thus ensuring the fastest, most effective and clear results. The amount of data which ListMotor is able to process is limited only by volumes of your hard drives.

Target Email Collector (TEC) is an email extractor, an ideal tool for email marketing. It can extract emails from various sources including Search Engines, WebSite, Web Directories, Web Groups, List of URLs and save the extracted data in text file of user hard drive. Easy to use, multi-threaded, full support for proxy and has numerous options and filters to control extraction.

Whois Extractor extract domain owner address, phone, fax, email, dns, date value from whois. Whois Extractor extracts domain information from global whois database source. It extracts Domain, Registrant, Admin Name, Address, City, State, Zip, Country, Phone, Fax, Email, NameServer, Domain Created Date, Updated Date, Domain Expired Date. The program auto saves all extracted data in csv/text file with success, error, log text.

Deep Log Analyzer is an web analytics solution for small and medium size websites. It analyze web site visitors' behavior and get complete website usage statistics in several easy steps.

SmarterStats is a Web log analytics tool that delivers accurate and detailed website statistics for sites on Windows and Linux Web servers. The program is capable of delivering detailed, accurate and relevant statistics for thousands of websites. It provides a wide range of reports about visitor’s navigation, entry/exit pages, sites/urls/links that refer web traffic to you, monthly/weekly/daily/hourly traffic, search queries, geographical analysis, search engine spiders, user browsers and operating systems, web server errors and much more.

WebGrab is a simple and convenient way to save and organize Web Pages for off-line viewing. Using WebGrab you can grab files and images from a Web Server and store them anywhere on your local hard drive. WebGrab is also a tool for maintaining your Web Site.

TextPipe Pro is used by IT personnel to repair hyperlinks when a server gets renamed, by SMEs and large organisations when they change their contact details/name/etc, by translators to apply massive search/replace lists.

NetTools Spider is a multi-functional Internet spidering utility that supports: website downloading, offline browsing, link checking, and real-time web mining. You can download, search and data mine websites. You can navigate through downloaded web sites much faster than possible if they were viewed online. You can search thousands of web sites and only download the files that contain the words you're looking for. With its web mining features, the possible uses for NetTools Spider are numerous.

CacheBack rebuild cached web pages and examine Internet histories for Internet Explorer. View cached web pages and pictures in a single consolidated thumbnail gallery. Import pictures and movies directly from local hard disks using GrabMedia data mining tool. Categorize, group, bookmark and/or exclude any quantity of pictures or movies from your case. Eliminate hundreds of thousands of images from analysis, in seconds, using filtering technology and Photograph Aspect Ratio Differential algorithm.

Web Scraper Lite eliminate cut and paste. Web spider / web crawler using web data extraction / screen scraping technology. Use the web extract for web data mining of contact lists, product catalogs, govt. databases, real estate listings.

Bget is a professional web data extraction software that will extract the data from any type of websites through flexible rules, therefore what you can see through the browser is also what you can get. Features of Bget: flexible, extendable, efficient, fast, stable, humanized.

Data Record Extractor Extract data records from web and local text file. Data Record Extractor is a tool for extracting regular text to data records. It supports updating the page number and replacing query parameter in the URL with database field value, quite suites for extracting record from dynamic web pages, search results, text bits and more. Data Record Extractor supports database Access, you can create a table via database management module first, then save the extraction result to it, also you can save as CSV/HTML/XML and other custom formats.

The list of software with data mining capabilities is very large, besides those outlined above we enumerate: CART, SPSS, Clementine, SAS, Oracle Data Mining, MATLAB and many others.

3. MATLAB features for data minnig

A special series of programs used for data mining are MATEMATICA, MAPLE, MATLAB that offer programming opportunities. Because this facility they are widely used in academic centers. Perhaps the most used of these is MATLAB on which we focus attention in this section.

A numerical analyst called Cleve Moler wrote the first version of MATLAB in the 1970s. It has since evolved into a successful commercial software package. MATLAB has evolved:

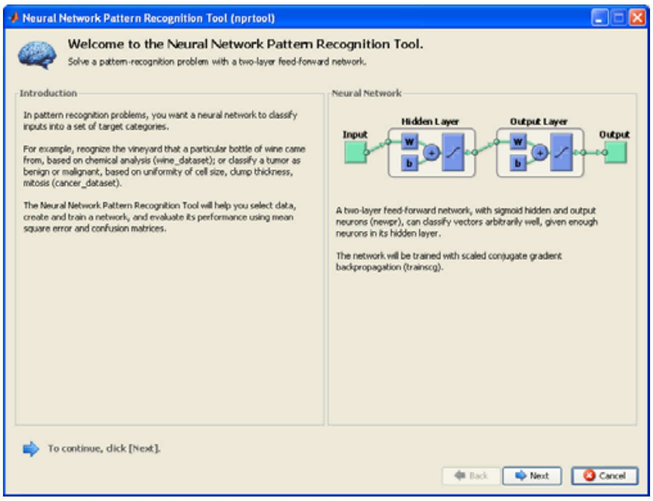
- at the university where is the standard package for introductory and advanced courses in mathematics, engineering and science,
- in industry, where is used for high yield research, development and production.

MATLAB® is a high performance language for computer-aided design. MATLAB is both a programming language and development system that integrates computation, visualization and programming easy-to-use, problems and solutions to these problems are expressed in an accessible mathematical language.

The name MATLAB came from **Mat**rix **lab**oratory and the producing firm is The MathWorks, Inc., SUA.

MATLAB allows development of a family of applications under the form of toolboxs. **Toolboxs** are collections of MATLAB functions (M-files) that extend the MATLAB environment to solve particular classes of problems These toolboxs allow learning and application of specialized technology in various fields. Toolboxs are available for areas such as digital signal processing, automatic control systems, neural networks, fuzzy logic, wavelet, simulation (SIMULINK), identification, statistics, create maps, image processing, etc. Further summarize the Neural Network Toolbox facilities.

Neural Network Toolbox extend MATLAB with tools for designing, implementing, visualizing and simulating neural networks. Neural networks are very important in applications where formal analysis would be difficult or impossible, such as pattern recognition or non-linear system identification and control. Neural Network Toolbox provides comprehensive support for many network paradigms, as well as graphical interface that allows you to design and manage networks. Modular design, open and extensible of tools set significantly simplify the creation and personalization of functions and networks



Neural Network Pattern Recognition Tool (nprtool)

Welcome to the Neural Network Pattern Recognition Tool.

Solve a pattern-recognition problem with a two-layer feed-forward network.

Introduction

In pattern recognition problems, you want a neural network to classify inputs into a set of target categories.

For example, reognize the vineyard that a particular bottle of wine came from, based on chemical analysis (wine_dataset); or classify a tumor as benign or malignant, based on uniformity of cell size, clump thickness, mitosis (cancer_dataset).

The Neural Network Pattern Recognition Tool will help you select data, create and train a network, and evaluate its performance using mean square error and confusion matrices.

Neural Network

A two-layer feed-forward network, with sigmoid hidden and output neurons (newpr), can classify vectors arbitrarily well, given enough neurons in its hidden layer.

The network will be trained with scaled conjugate gradient backpropagation (trainscg).

To continue, click [Next].

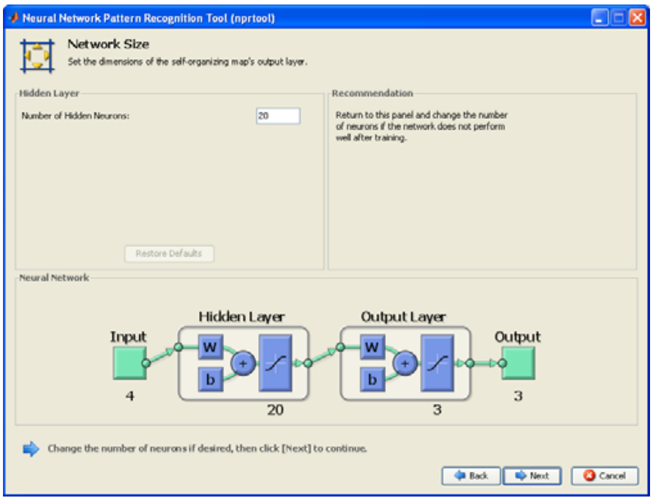
Fig. 4. Creating a Neural Network with Matlab

Key fatures:

- Graphical interface for creating, training and simulation of neural networks,

- Assistant (wizard) for matching, pattern recognition and clustering,
- Support for all major network architectures,
- Complex set of features for training and learning,
- Dynamic Learning Networks,
- Simulink blocks for building neural networks and advanced blocks for applications for control systems,
- Support for automatic generation of Simulink blocks from the object type neural networks,
- Modular representation of networks that allows an unlimited number of levels of input settings,
- Features for pre-processing and post-processing,
- View of functions and graphical interface for viewing network performance and monitoring of training.

In Fig. 4. and Fig. 5. are presented Neural Network Pattern Recognition Tool of Neural Network Toolbox that perform recognition models. All stages can be viewed and can intervene to change the parameters. Operations are made easy by performing only a few clicks and filling boxes dialog.



Neural Network Pattern Recognition Tool (nprtool)

Network Size

Set the dimensions of the self-organizing map's output layer.

Hidden Layer

Number of Hidden Neurons: 20

Recommendation

Return to this panel and change the number of neurons if the network does not perform well after training.

Change the number of neurons if desired, then click [Next] to continue.

Fig. 5. Customizing the Neural Network

SIMULINK is a software package for modeling, simulating and analyzing dynamic systems. The following type of systems can be modeled: linear and nonlinear, continuous, discrete, hybrid, multiple sampling periods.

SIMULINK provides a graphical user interface (GUI) for creating models as diagrams constructed from blocks, based on techniques click-and-drag using the mouse. Thus, drawing diagrams is simple and intuitive, almost as simple as drawing these diagrams directly on paper. In addition, it avoids laborious mathematical formulation (dynamic systems are usually described by differential equations or difference).

4. Classifications of data mining software

In this section we propose several classifications of software that can perform data mining operations.

Thus, by the nature of software licenses required, they can be free (WEKA) or commercial as are the most programs presented earlier. Some software have releases with low features that can be used without purchasing a license, remember here RapidMiner and Matlab programs for students and the majority of software makes possible the use for a trial period.

After programming facilities offered we can split them in two categories:

- With programming capabilities such as Matlab, Maple, Mathematics, Rapidminer and others.
- Widouth programming capabilities.

After the nature of the data it can run analysis programs can be divided into programs that are used on unstructured data such as for example web pages or acting on structured data from databases.

Another classification can be based on preprocessing possibilities offered:

- Those that do not offer possibilities for data preprocessing in which case is necessary to use another program to provide data in a particular format requested by the program .

- Those providing data preprocessing capabilities in which case the user can perform various operations to prepare the data.

After how they can be used distinguish the category of those that can be used online and category of those that are used offline.

Such classifications are necessary for those who wish to use such programs to determine which program is most useful to company needs.

5. Conclusions

E-business helps organizations conduct business online but also connects the organization with all its external and internal components of the value chain: the components supply chain, logistics providers, distributors, service providers and buyers. E-business creates integrated network of relationships with channels, end users, suppliers and rivals that were not possible before. E-business solutions make customers available in almost all industries. Customers can buy non-stop and companies can offer self-service applications and distribution of products and services to customer. E-business offers new forms of market conditions which radically change the game.

The Internet and e-business is an important key to future success of any organization, offering huge opportunities and market outlets worldwide. E-business projects are doomed to failure because of misunderstanding the new rules of e-economy environment, thus failing to step into the competition. To differentiate into the Internet economy, companies must realize that winning e-business means more than simple transactions of purchase/sale, the appropriate strategies being the key to improving competitive power. This can be done by using software with data mining capabilities and other statistical analysis on historical data from e-business activities.

There are now many commercial and freeware software packages that provide statistics about web sites, including number of page views, hits, traffic patterns by day-of-week or hour-of-day, etc. These tools help ensure the correct operation of web sites (e.g., they may identify page not found errors) and can aid in identifying basic trends, such as traffic growth over time, or patterns such as differences between weekday and weekend traffic.

With growing pressure to make e-commerce sites more profitable, however, additional analyses are usually requested.

The software is connection between theory and results, so they must respond to various needs of people and business and permanently improve their facilities.

We have reviewed verry shortly some of the programs used to cope with data mining analisys and proposed some classifications of this software.

References

1. Claudia Elena Dinucă, Using Web Mining in E-Commerce Applications, Annals of the University “Constantin Brâncuşi” of Târgu Jiu, Economic Series, Issue 3/2011, “Academica Brâncuşi” Publisher, ISSN 1844-7007, pp. 65-75, 2011a.
2. Claudia Elena Dinucă: E-Business, a new way of trading in virtual environment based on information technology, Annals of the “Ovidius” University, Economic Sciences Series Volume XI, Issue 1 /2011, ISSN 1582-9383, pp. 613-617, 2011b.
3. David Scuse, Peter Reutemann, WEKA Experimenter Tutorial for Version 3-5-5, University of Waikato, 2007.
4. Remco R. Bouckaert, Eibe Frank, Mark A. Hall, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H.Witten, WEKA - Experiences with a Java Open-Source Project, Journal of Machine Learning Research 11, 2533-2541, 2010.
5. Osmar R. Zaine, Man Xin, Jiawei Han, Discovering Web Acces Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, In Proceedings on Advances in Digital Libraries Conference (ADL'98), 1998.
6. M. Spiliopoulou, L.C. Faulstich, WUM: A Web Utilization Miner, Proceeding of EDBT Workshop on the Web and Data Bases (WebDB’98), Springer Verlag, pp. 109-115, 1999.
7. Kobra Etminani, Mohammad-R. Akbarzadeh-T, Noorali Raeeji Yanehsari, Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method, IFSA-EUSFLAT, 2009.
8. Cooley R., Tan P. N., Srivastava J. Discovery of Interesting Usage Patterns from Web Data, Proceeding WEBKDD '99 Revised Papers from the International Workshop on Web Usage Analysis and User Profiling Springer-Verlag London, UK, 2000.
9. http://www.rapidminer.com/ RapidMiner 4.4: User Guide, Operator Reference, Developer Tutorial, 2009.