USING CLUSTERING METHODS IN DATA DERIVED CUSTOMER SEGMENTATION

Krzysztof Świder
Rzeszów University of Technology,
W. Pola 2, 35-959 Rzeszów, POLAND
kswider@prz-rzeszow.pl

ABSTRACT

The significant issue of typical mass-customization activity is the problem of understanding the customers and their needs. Many successful customer relationship management (CRM) projects resulted in customer-centric strategy of the company, gathering numerous *data* about customer behavior and finally an extensive use of data analysis, e.g. data mining, techniques to discover the useful *knowledge*. The sophisticated methods are widely exploited to enable an organization to analyze the existing CRM data and identify patterns to predict customer behaviors including purchase preference as well as their demographic and other characteristics. The process of collection, storage, processing and permanent analysis of large amounts of data to provide the knowledge about customers is often called Business Intelligence (BI). The typical BI tool usually provides data mining part with clustering option, which, among others, enables discovering characteristic groups of the customers directed by data. The process is usually referred to as *customer segmentation*. The aim of the paper is to introduce the clustering techniques to the mass-customization community in rather informal and intuitive way. Additionally some illustrating examples will be presented to show the significant capabilities of modern BI software in generating clusters from data.

KEYWORDS

Customer Relationship Management, Business Intelligence, clustering, customer segmentation

1. INTRODUCTION

In the last few decades, corporate strategy primarily concentrated on internal processes and supply chain operations. *Becoming more competitive* was synonymous of downsizing, cost-cutting, optimizing production, reducing inventories and other efforts to wring more from less. These strategies drove the widespread adoption of enterprise resource planning (ERP) systems. However some recent developments in strategic thinking led to a new approach characterized by intensified focus on the customer [Siebel, 2000]. More and more of present-day commercial companies are realizing the importance of creating and sustaining the highest levels of customer satisfaction. They are applying sophisticated information technology to identify, acquire, and retain the most profitable customers. By gearing marketing and sales activities as closely as possible to its customer's requirements, a company can increase the rate at which it converts customer contacts into sales. The principal idea here is to generate profitable *customer knowledge*, which can be used during all customer interactions. Customers are also increasingly coming to expect and demand higher levels of personal service. In an ideal situation a company would like to be in a position to treat each customer as an individual and to maximize customer satisfaction and profitability. In order to offer the products and services that meet the individual needs of their customers, organizations need to find effective ways to increase the value of customer interactions. Customer relationship management (CRM) systems are normally able to facilitate communications and summarize their history. Steps forward are *analytical* CRM (aCRM), which include investigative tools to analyze existing CRM data. New analysis techniques like data mining offer the opportunity to exploit valuable information within huge amounts of customer data. With this information a company is able to improve useful knowledge about customers, including their behavior, characteristic groups etc. This knowledge could make a considerable difference to the way the business is run and the way the companies interact with their customers. In the real world one-to-one interaction is in general not economically possible and to address this challenges most organizations attempt to group their customers into relatively small number of market segments. By using data mining the segments can be derived directly from data as

opposed to using pre-defined business definitions. The data driven approach provides a natural grouping of the customers based on customer attributes. For example an attempt to identify groups of customers with similar characteristics can be based on the types of products they purchase. Grouping or segmenting customers in this way makes it possible to develop production, sales and marketing strategies that are targeted on the group rather then the individual.
The present work is generally dedicated the way the knowledge about customers is discovered rather then how to deploy it in a specific business. An example for retail will be used to demonstrate the sort of issues that the typical data mining process may deal with.

2. DATA MINING BASICS

An important objective of modern information systems aimed for commercial companies is to generate profitable customer knowledge, which could be used for decision support purposes. Data mining is rapidly growing research area relating to use of artificial intelligent and machine learning methods in advanced data analysis. The data is usually stored in large repositories including relational databases, data warehouses, text files, web resources etc. and the result of mining procedure is a kind of knowledge base called *mining model* containing useful patterns and characteristics derived from data (Fig.1).

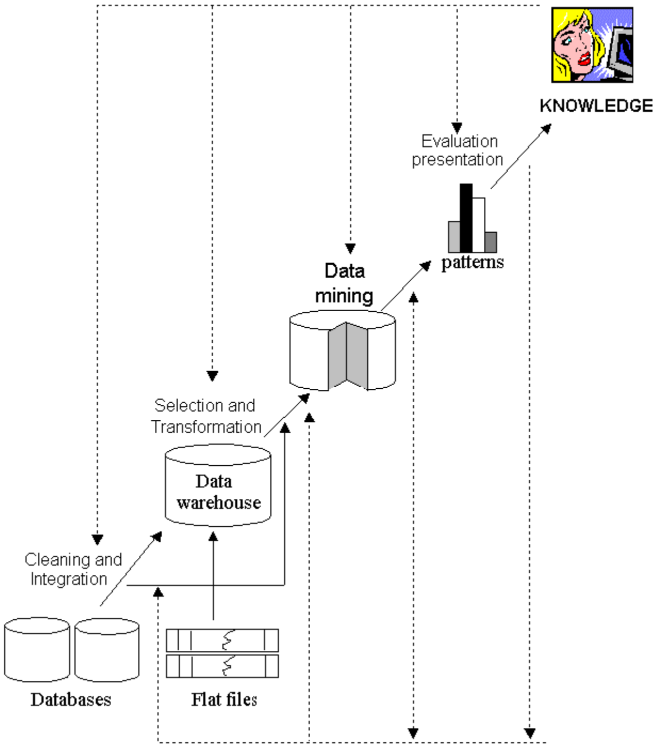


Fig. 1. Data mining as a step in knowledge discovery process [Han, Kamber, 2001]

The term ‘knowledge discovery’ is usually employed to describe the whole process of extraction of knowledge from data while ‘data mining’ is normally used to the *discovery* stage of the knowledge discovery process. Generally speaking, relationships and patterns between data elements can be extracted using a number of techniques, including:

- query tools,
- visualization,
- On-Line Analytical Processing (OLAP),
- case-based learning,
- decision trees,
- association rules,

- clustering,
- neural networks,
- statistical techniques,
- fuzzy-logic,
- genetic algorithms.

Data mining is not so much a single technique as the idea that there is more knowledge hidden in the data that shows itself on the surface [Adriaans, Zantige, 1996]. Some of the above techniques including decision trees, association rules and clustering are commonly used in up to date data mining tools.

Data mining became useful over the past decade in business to gain more information and to find new ways and ideas to improve a business. Today technology supports the insertion of data mining findings in the company processes of interaction between different business users.

2.1 The mining process

At present, data mining is no longer a set of stand-alone techniques. It was also realized that stand-alone data mining tools and workbenches are difficult to sell. They are better as part of a solution and are often integrated with relational databases, as well as with business-oriented and e-commerce applications [Bargoin et al., 2002]. The traditional method of mining process versus the application-oriented approach will be characterized in the next two subsections.

2.1.1 The generic data mining method

With the traditional approach to data mining an expert uses a workbench to run preprocessing steps, mining algorithms, and visualization of the mining models. The generic data mining method is displayed in Fig.2.

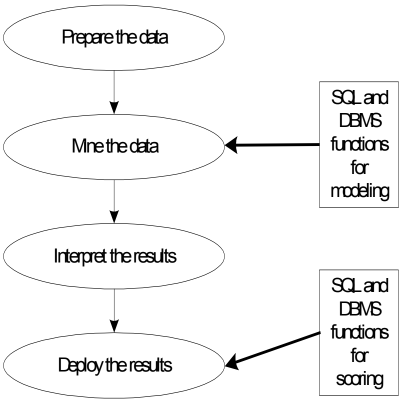

Prepare the data
Mine the data
Interpret the results
Deploy the results
SQL and DBMS functions for modeling
SQL and DBMS functions for scoring

Fig. 2. The generic data mining method [Bargoin et al., 2002]

The process starts with defining business issue in a precise statement. The next step refers to data preparation, which normally includes: data model definition, sourcing data from available repositories, preparing data and evaluating quality of the data. Mining the data means to choose the mining function and to run it. The next stage includes interpreting the results and detecting new information. Finally the results and the new knowledge are to deploy into particular business. Effective and appropriate deployment of mining results becomes the problem, if the right skills and technology are not placed at the right moment in the overall process. *Scoring* denotes here the use of existing data mining models based on historical data and applying these models to new data.

The power users, as data mining analysts, typically need specialized data warehouses (*data marts*) for doing specialized analysis on preprocessed data. Typically, the time to move from steps 1 to step 5 can take several weeks to several months, depending on the maturity of the data warehouse or operational data stores. The goal is to find new and interesting patterns in the data and somehow use the gained knowledge in business decisions. The integration of mining results into the operational business is usually done in an ad-hoc manner.

2.1.2 Integrating with applications

The modern approach employs the integration of mining, where the focus shifts towards the deployment of mining in business applications. An important requirement is to integrate mining functions, such as modeling and scoring with the relational database management system. This allows for a faster Return On Investment, improving the ability to evaluate results – an important quality of successful data mining application. It also advocates the need for applying the modeling function to any data preparation source, and scoring of new transaction, without needing in-depth knowledge of the mining techniques. The audience includes the end users in the line of business who need an easy-to-use interface to the mining results and a tight integration with the existing environment (Fig.3).

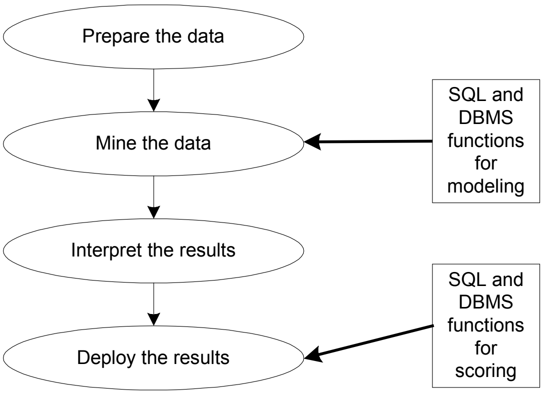

Prepare the data
Mine the data
Interpret the results
Deploy the results
SQL and DBMS functions for modeling
SQL and DBMS functions for scoring

Fig. 3. Focus on modeling and deployment of scores [Bargoin et al., 2002]

A power user is still required to design and build optimized and efficient data mining models. Simple predictive models anticipation of customer behavior, and automatic optimization of business processes by means of data mining become more important than some general knowledge discovery. When data mining becomes part of a solution, the actual user of the mining functions is a developer who packages the mining results for the end-user community. This person has the database application development skills. They provide the models and scores back to the data warehouse with aggregated data, specialized data marts, and operational data that hold data at the transactional level deployment.

2.2 Data mining functionalities

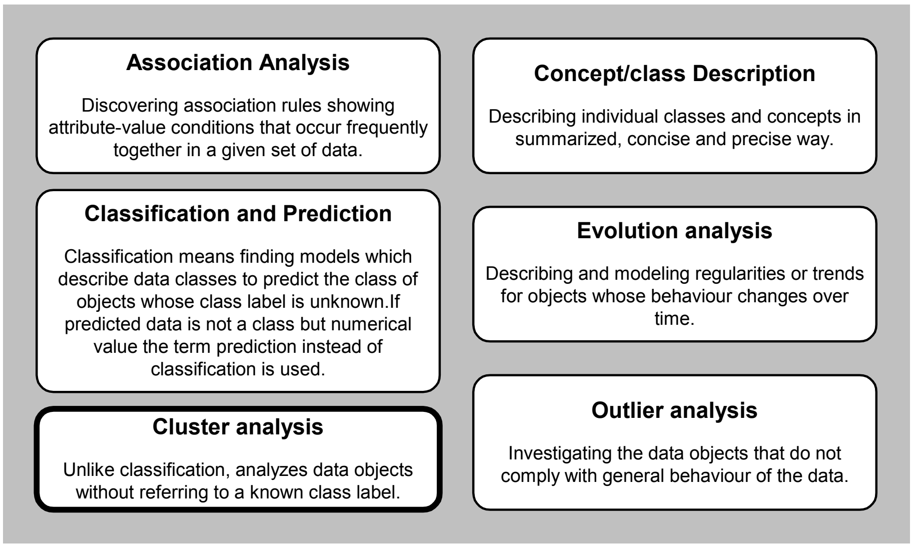
Data mining systems can be classified according to various criteria. On the whole data mining tasks can be classified into the following two categories:

- descriptive,
- predictive.

Descriptive mining tasks exemplify the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

The other possible classification criteria for data mining systems are: the kinds of databases mined (e.g. relational, object-relational, object-oriented, data warehouses, etc.), the kinds of techniques utilized, the application adapted (e.g. finance, telecommunications, health care, etc.) and the kinds of knowledge (patterns) mined.

A pattern discovered during data mining process represents knowledge if it is easily understood by humans and potentially useful. Measures of pattern interestingness can be used to guide the discovery process. The kinds of patterns discovered in data mining tasks are referred to as data mining functionalities. There are several types of data mining functionalities connected with specific nature of patterns they discover. (Fig.4)



Association Analysis

Discovering association rules showing attribute-value conditions that occur frequently together in a given set of data.

Concept/class Description

Describing individual classes and concepts in summarized, concise and precise way.

Classification and Prediction

Classification means finding models which describe data classes to predict the class of objects whose class label is unknown.If predicted data is not a class but numerical value the term prediction instead of classification is used.

Evolution analysis

Describing and modeling regularities or trends for objects whose behaviour changes over time.

Cluster analysis

Unlike classification, analyzes data objects without referring to a known class label.

Outlier analysis

Investigating the data objects that do not comply with general behaviour of the data.

Fig. 4. Data mining functionalities

In some cases users may not know which kinds of patterns may be interesting for them, therefore it is important to have a data mining system that can mine multiple kinds of patterns. Some of the functionalities shown in Fig.4 have their 'standard' area of applications. For example for association rules this is so called *market basket analysis* and for cluster analysis it is to identify homogenous subpopulations of customers. Cluster analysis will be outlined briefly in the next subsection.

2.3 The principle of clustering method

The cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering (Fig.5).

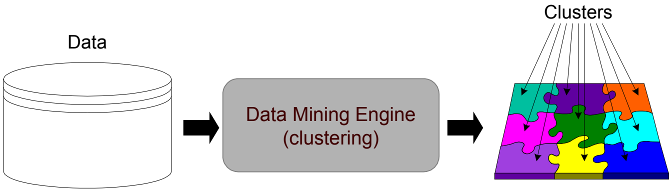


Fig. 5. The principal idea of clustering

Clustering is a dynamic field of research in data mining. A number of clustering algorithms have been developed using the following methods:

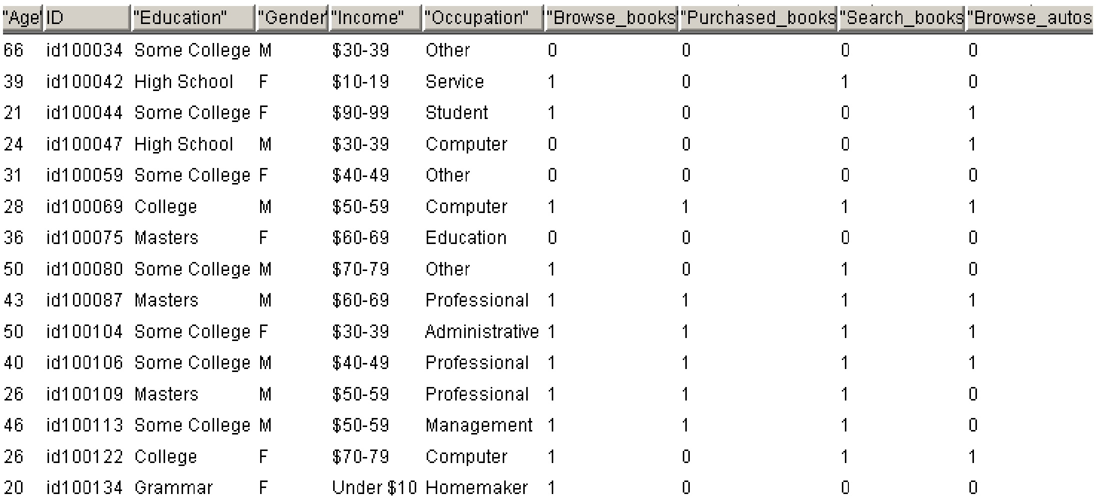
- partitioning,
- hierarchical decomposition,
- density-based,
- grid-based,
- model-based.

Unlike classification, cluster analysis does not apply the class labels in the training data. The labels are not known in advance and clustering can be used rather to generate them.
Apart from market or customer segmentation, cluster analysis has another applications, e.g.: pattern recognition, biological studies, spatial data analysis and others.

3. AN EXAMPLE CLUSTER ANALYSIS

3.1 The data to analyze

Although a vast amount of data is presently stored in various data repositories, it is not always easy to get the right data to experiment with. In many cases the only way out is to use one of the numerous Web resources publishing various kinds of data. To obtain the required cluster models we used a real data reporting purchase history available via Internet. Primarily all the data was recorded in a single text file with elements separated with *tab* symbols. We have performed elementary preparation e.g.: invalid records were removed and attribute values were standardized if necessary. Additionally the *tab* separators were replaced by the regular width of columns (Fig.6).



"Age	ID	"Education"	"Gender	"Income"	"Occupation"	"Browse_books	"Purchased_books	"Search_books	"Browse_autos
66	id100034	Some College	M	$30-39	Other	0	0	0	0
39	id100042	High School	F	$10-19	Service	1	0	1	0
21	id100044	Some College	F	$90-99	Student	1	0	0	1
24	id100047	High School	M	$30-39	Computer	0	0	0	1
31	id100059	Some College	F	$40-49	Other	0	0	0	0
28	id100069	College	M	$50-59	Computer	1	1	1	1
36	id100075	Masters	F	$60-69	Education	0	0	0	0
50	id100080	Some College	M	$70-79	Other	1	0	1	0
43	id100087	Masters	M	$60-69	Professional	1	1	1	1
50	id100104	Some College	F	$30-39	Administrative	1	1	1	1
40	id100106	Some College	M	$40-49	Professional	1	1	1	1
26	id100109	Masters	M	$50-59	Professional	1	1	1	0
46	id100113	Some College	M	$50-59	Management	1	1	1	0
26	id100122	College	F	$70-79	Computer	1	0	1	1
20	id100134	Grammar	F	Under $10	Homemaker	1	0	0	0

Fig. 6. The category of data to analyze

The data contains of 2148 records describing customers of web stores. Each record includes general information about a customer (*Age*, *Gende*r, *Education*, *Geographical Location*, *Income*, *Occupation*) together with purchase history divided into following three sections: products browsed (*Browse_autos*, *Browse_books,* etc.), products purchased (*Purchased _autos*, *Purchased _books,* etc.) and products searched (*Search _autos*, *Search _books,* etc.). The products data refers to *books*, *software*, *musi*c, *hardware*, *travel*, *apparel*, *video*, *flowers*, *magazines*, *electronics*, *quotes*, *concerts*, *recreation and jewelry*. Each product field contains 1 or 0 informing that the customer browsed (purchased, searched) the product or did not do it.

3.2. Segmentation based on products purchased

The customers expect the highest quality of personal services. In order to achieve this, an attempt will be made to identify groups of customers with similar characteristics e.g. based on the products they purchase. The grouping (segmentation) of customers allows the selling targeted to the groups not to individual customers. At the same time each group member may have an impression to be treated individually.

Given a set of mining parameters the clustering process was carried out using IBM BI software. the results of segmentation for the considered population of customers are shown in Fig.7. The four clusters numbered in the right upper corner were generated containing 28, 26, 25 and 22 percent of population respectively. The valuable quality of each cluster is a possibility to compare the characteristics of the particular cluster (the inside circle) with the same characteristics for the whole population (the outside ring). These characteristics refer to the groups of variables exemplifying purchasing (not purchasing) some products.

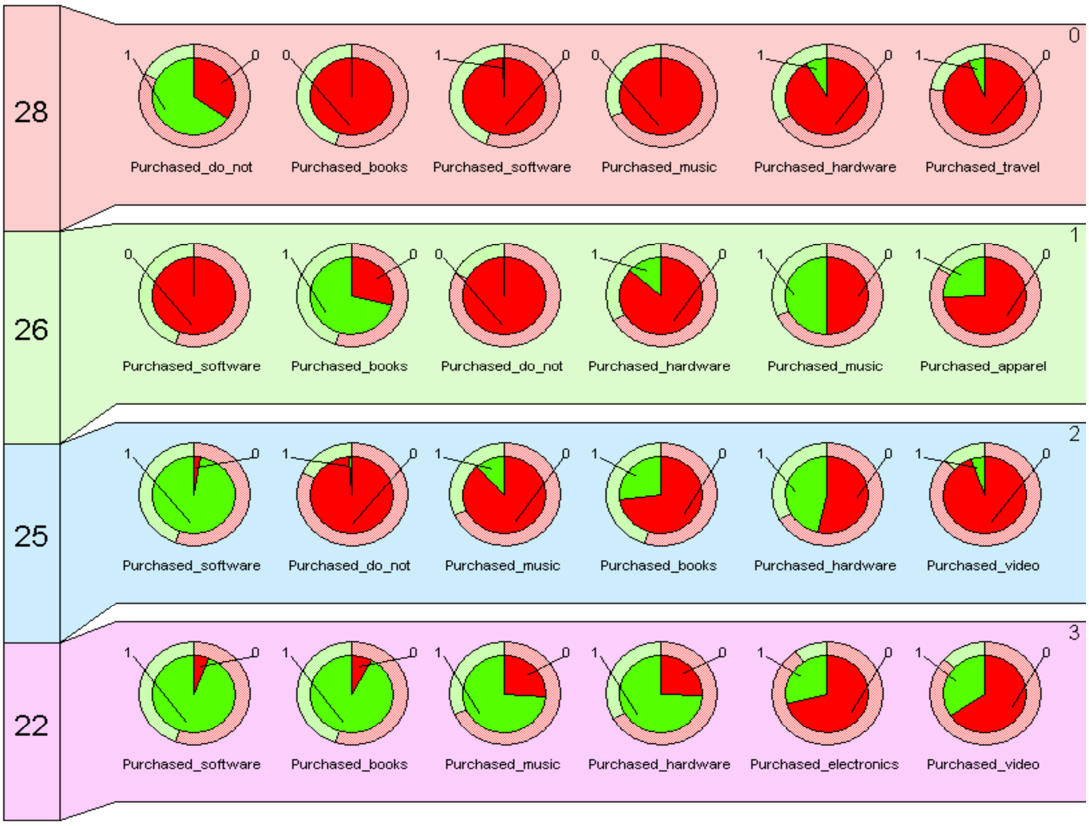


Fig. 7. Customer population partitioning on the basis of purchase activity

It is noteworthy that the variables characterizing each segment apear in different order in this segment. This is not a coincidence and the position of a variable shows its relative interestingness for a given segment. The IBM Inteligent Miner for Data used in this example allows further analysis of the clusters in Fig.7. For example rolling up the clusters 0 and 1 we obtain the information shown in Fig.8. and Fig 9 respectively.

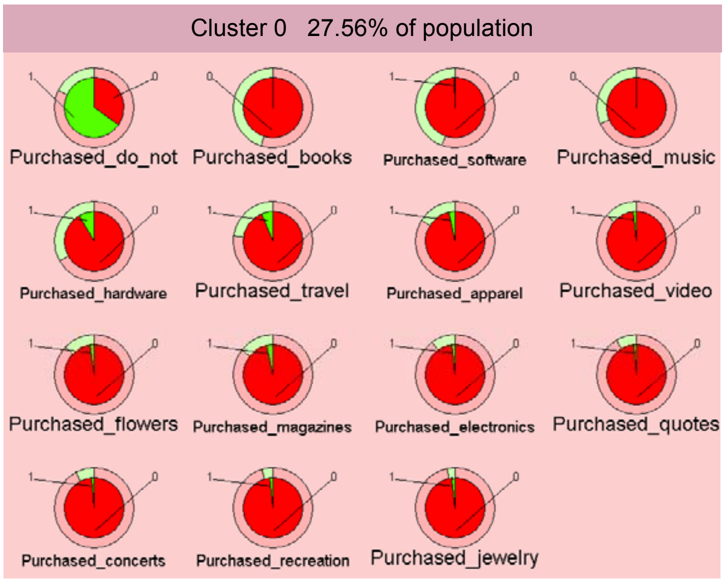


Fig. 8. The customers of the cluster 0

The first and most important variable characterizing the Cluster 0 in Fig.8 is *Purchase_do_not* showing the rather high rate of customers of this segment who did not ordered any product so far. Thus the general description of this subpopulation is rather obvious: it describes a community who is relatively inactive by ordering the products.

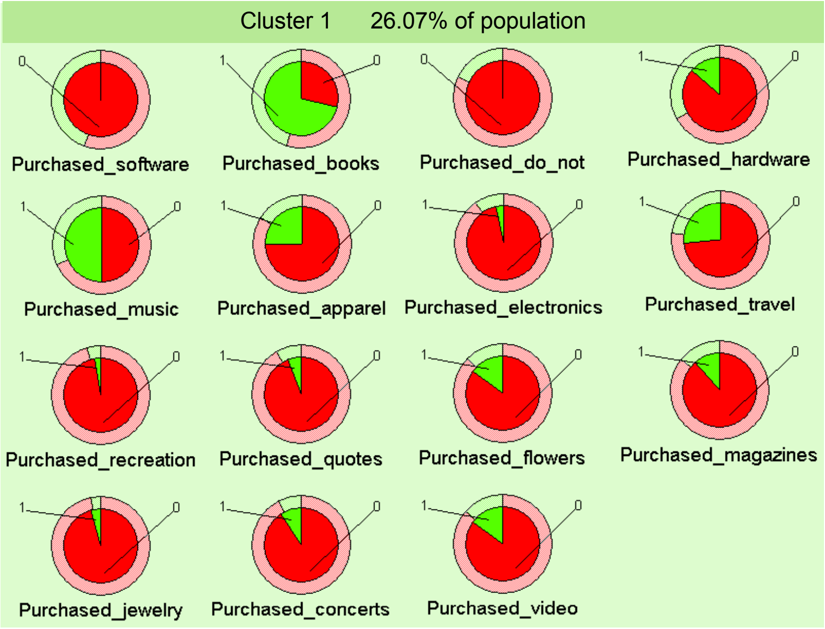


Fig. 9. The customers of the cluster 1

Unlike Cluster 0 the Cluster 1 describes the subpopulation much more inclined to buy products via Internet. Almost everyone has ordered something in the past but there is a clearly selected group of products the customers of this segment especially like to purchase. First of all these include: *books*, *music*, *travel* and *video*, i.e. *entertainment*. The noteworthy observation is that a group of customers described by this cluster relatively rarely buys hardware and practically doesn’t purchase software.

4. CONCLUSIONS

Data mining is no longer thought of as a set of stand-alone techniques far from business applications. Today many applications, such as Web analytics, Web personalization, e-commerce and CRM require integrated data mining functions.

The subject of this paper is closely connected with clustering analysis – a widely known data mining technique. In typical business applications clustering can help marketers discover distinct groups in their customer bases and characterize customer groups based on purchasing patterns. The informal introduction to data mining and clustering was completed by presentation of a set of clusters build for example data. The results are encouraging and show the prospective way of future research, which is intended to focus on developing methods and tools for effective use of mining models for mass-customization purposes.

ACKNOWLEDGEMENTS

I would like to thank all my students who have worked or are the currently working with me on data mining related research. Especially I would like to express my sincere appreciation to Justyna Michalczyk and Justyna Nowak who contributed to this work by preparing examples of mining models with use of the IBM Business Intelligence tools. I also wish to thank the IBM Poland Ltd. for providing the necessary software resources.

REFERENCES

Adriaans, Zantige (1996), *Data Mining*, Addison Wesley.

Bargoin et al. (2002), *Enhance Your Business Applications. Simple Integration of Advanced Data Mining Functions*, IBM International Technical Support Organization.
Available via: http://www.redbooks.ibm.com/redbooks/pdfs/sg246879.pdf

Bargoin et al. (2001), *Mining Your own Business in Retail. Using Intelligent Miner for Data*, IBM International Technical Support Organization.
Available via: http://www.redbooks.ibm.com/redbooks/pdfs/sg246271.pdf

DataDistilleries (2002), *Maximizing the Value of Customer Interactions*, Business White Paper, DataDistilleries B.V., The Netherlands.

Han, Kamber (2001), *Data mining. Concepts and Techniques*, Morgan Kaufmann, Publishers.

Michalczyk, Nowak (2003), *Examples Data Mining Application in Retail (MSc. Thesis in Polish)*, Rzeszów University of Technology, Poland.

Pyle (1999), *Data Preparation for Data Mining*, Morgan Kaufmann Publishers.

Siebel (2000), *eBusiness. Managing the Demand Chain*, Business White Paper, Siebel Systems Inc.

Świder (2003), *Knowledge Discovery in Customer Relationship Management (Conference Paper in Polish)*, Zamojskie studia i materiały, seria Informatyka, zeszyt 1(10), ss.129-137, Zamość, Poland.

Świder (2001), *Applications of data mining techniques in customer-centric commerce*, Proceedings of the 3rd Conference IBIS'01/ A. Lawrynowicz (Ed.), Malmo-Copenhagen 28-30 Sep. 2001, pp.165-170.

Witten, Frank (2000), *Data mining*, Morgan Kaufmann Publishers.