INFO- 2061
Primljeno / Received: 2012-04-22

UDK : 001:658.7:316.77
Izvorni znanstveni rad / Original Scientific Paper

KNOWLEDGE DISCOVERY PROCESS FOR BUILDING CUSTOMER PROFILES

PROCES OTKRIVANJA ZNANJA ZA IZGRADNJU PROFILA KUPACA

Brano Markić
Faculty of Economics, University of Mostar, Mostar, Bosnia and Herzegovina
Ekonomski fakultet, Sveučilište u Mostaru, Mostar, Bosna i Hercegovina

Abstract
The knowledge about customer preferences and behavior is fundamental for personalization of products and service. Personalization products and services are possible only if we have enough knowledge of who customers are, how they are similar among, how they behave. Knowledge discovery is process of transforming data into knowledge by adequate algorithms and software tools. In the paper is developed an approach that uses data in the form of transactional databases to construct accurate individual profiles. In developed data model are integrated transactional data and rules describing customer's behavior. The rules are extracted from transactional data and cover individual customer behavior as well as the common behavior of all customers in the market segment. There are two rules types: first, for describing individual customer behavior and second, for the market behavior. Knowledge discovery plays a crucial role as an enabler to the organizations to integrate effective analytical data mining methods for prediction, classification, cluster, anomaly detection with data management and information visualization. Knowledge discovery is oriented to learning. In the process of learning we are implementing the functions of R language and this tool has shown satisfactory application and development power.

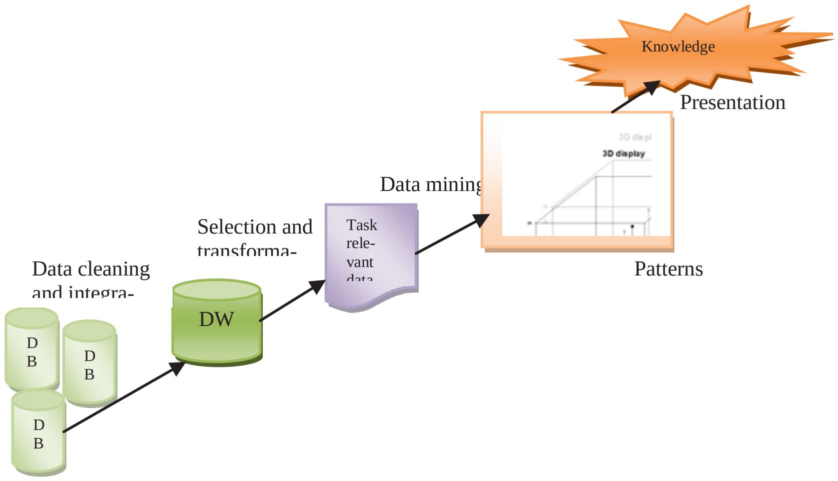
Sažetak
Znanje o ponašanju kupaca i njihovim preferencijama je ključno za personalizaciju proizvoda i usluga. Personalizacija proizvoda i usluga je moguća samo ako imamo dovoljno znanja o tome tko su kupci, koliko su međusobno slični, kako se ponašaju. Otkrivanje znanja je proces transformacije podataka u znanje primjenom odgovarajućih algoritamai softverskih alata. U radu je razvije pristup koji koristi podatke u obliku transakcijeke baze podataka kako bi se izgradio individualni profil. Pravila se ekstarhiraju iz transakcijskih baza podatakai otkrivaju pojedinačno ponašanje kupaca ali i njihovo zajedničko ponašanje unutar nekog klastera. Zato postoje pravila koje otkrivaju pojedinačno ponašanje kupca ali i cijelog tržišta (klastera). Otkrivanje znanja je tako orijentirano učenju. U procesu učenja primjenili smo fnkcije jezika R i taj softverski alat je pokazao zadovaljavajuću aplikacijsku i razvojnu moć.

1. INTRODUCTION

Knowledge discovery in data is a complex process of extracting hidden relationships that exist among data and cannot be uncovered by implementing trivial (simple) sequence of steps, by simple algorithm. Therefore we stress that the relationships among data are hidden and can be discovered by application adequate algorithms. In theoretical sense the steps in knowledge discovery process are prescribed and defined /**1**/.

The process starts with data stored in relational or objects oriented databases. At the first view to the Figure 1., it would be wrong to make conclusion that knowledge discovery is a linear process consisting of a few sequence steps: data cleaning, selection, implementing data mining algorithm and finally is produced knowledge in the form of patterns and models. Although we must divide the complex task into parts because it is only reasonable way of analyzing complex systems, in each step is necessary to monitor the whole knowledge discovery process. Timely most required step is data cleaning and preprocessing and may take 60% of the whole effort. In this step we have to know and understand the nature and objectives of knowledge discovery in business context.

Figure 1. Knowledge discovery process



If the objective is classification then various model are at disposals such as: rule induction, neural networks, case based reasoning, logistic regression, Bayesian networks, and decision trees. Data mining algorithm is only one step in the whole process with the goal to discover and identify truly interesting patterns. The data mining algorithm is applicable to any data repository such as relational and objects oriented databases, data warehouse, transactional databases, and text. Finally, the patterns are evaluating and the knowledge is presenting /**2**/.

Knowledge discovery is learning process. In business context the discovered knowledge and skills allow us to direct the system to set goals and predict their behavior in the future. Today dominates two types of learning: supervised and unsupervised learning.

2. THE INTEGRATION OF CLUSTER ANALYSE AND ASSOCIATION RULES FOR CUSTOMER PROFILES

The data model is a set of relations (or classes), attributes and relations among tables (classes) and

its attributes. In semantic sense data model use two basic types: factual-uncover who the customer

is- and transactional- uncover what the customer does.

"The data in the database are dynamic. Change over time, unlike the data warehouse, which is time-dependent, but fixed and subject-oriented information "/**3**/. There is a chain of interrelated and dependent activities. The first is the selection of data, and then their filtering and cleaning (a process known as extraction (E), transformation (T) and filling (L)) /**4**/.

Data are stored in relational database tables:

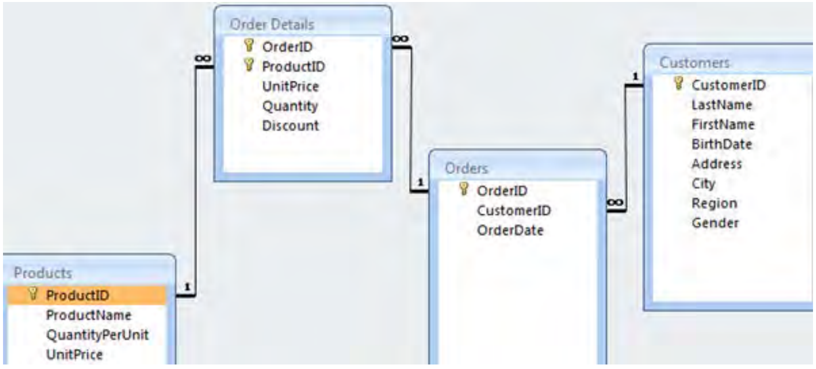
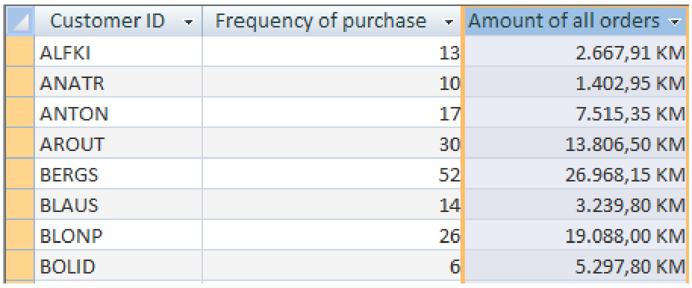


Figure 2.: Data model for discovering customer profiles

There are four tables: Products, Orders, Order Details and Customers. The factual data about customers are stored in the table Customers and includes the attributes: CustomerID (identification number assigned to each customer and this attribute is primary key), LastName, FirstName, BirthDate, City, Region and Gender. The transactional data are stored in three tables: Orders, Orders Details and Products. Customer behavioral variables are hidden in these three tables. We will first calculate frequency of purchase for each customer (how many times during the given period the customer buys the products), then the amount of orders. The next SQL query selects the required data from database:

```
SELECT Customers.CustomerID, Count(Orders.OrderID) AS [Frequency of purchase], SUM(([Order Details].UnitPrice)*([Order Details].Quantity)) AS [Amount of all orders]
FROM (Customers INNER JOIN Orders ON Customers.[CustomerID] = Orders.[CustomerID]) INNER JOIN [Order Details] ON Orders.[OrderID] = [Order Details].[OrderID]
GROUP BY Customers.CustomerID;
```

The result of the query will be the next data set:



Customer ID	Frequency of purchase	Amount of all orders
ALFKI	13	2.667,91 KM
ANATR	10	1.402,95 KM
ANTON	17	7.515,35 KM
AROUT	30	13.806,50 KM
BERGS	52	26.968,15 KM
BLAUS	14	3.239,80 KM
BLONP	26	19.088,00 KM
BOLID	6	5.297,80 KM

Table 1.: Data set for clustering

The next step is clustering the given data set using two variables: frequency of purchase and amount of all orders for given customer during the monitoring time period.

3. CLUSTERING AS UNSUPERVISED LEARNING TECHNIQUES

Clustering is the process of dividing a set of data on a predefined number of clusters so that the similarity between the elements of the cluster is the largest and difference (distance) between clusters is large as possible. There are different types of clustering. For example, agglomerative hierarchical clustering in the beginning joins each one point (object, record, row, tuple) to a cluster. In the next step, agglomerative hierarchical clustering collects the closest pair of clusters (those who are closest) to a new cluster. Collection procedure continues until the last one (or k) clusters /**5**/.

Divisive clustering begins with a cluster which is containing all points (objects, records). The following steps of divisive clustering are continuation the splitting process until each cluster becomes a single

point. The paper applies the k-means clustering procedure /**6**/.

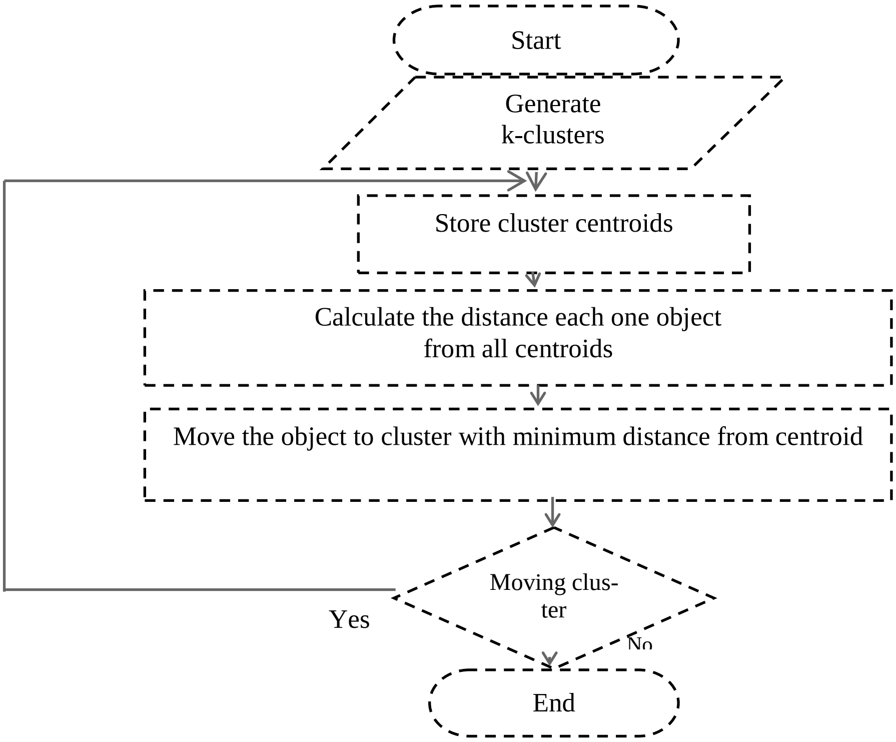

Start
Generate k-clusters
Store cluster centroids
Calculate the distance each one object from all centroids
Move the object to cluster with minimum distance from centroid
Moving clus- ter
Yes
No
End

Figure 3. : Block diagram of the k-means algorithm

The algorithm k-means clustering is a series of next steps:

1. Randomly select k-clusters

2. Determine the center (centroid) for each cluster

3. Repeat:

Determine the distance of objects from the center of the cluster and assign object to the nearest   cluster.

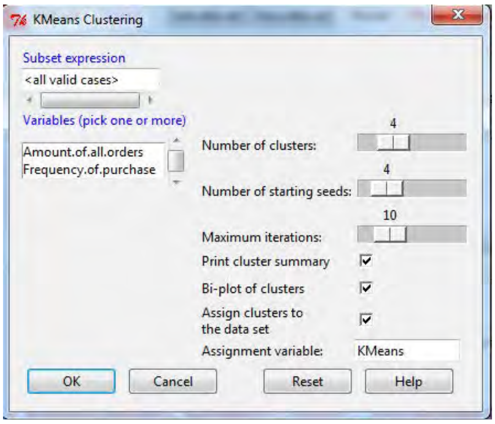
Recalculate the center of the cluster.

Repeat until the objects are moving from one cluster to another (provided the completion of the algorithm).

For cluster analysis we are using R language. "R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environment for data mining, statistical analyses, and data visualization" /7/. Data are located in a relational table that is extracted from operational databases (financial accounting). K-means clustering is implemented in R language. ***R Commander*** uses graphical user interface and it's very easy for learning and quick application.

Statement: Dataset <- sqlQuery(channel = 4, select * from [Sheet1$])

selects all records from Sheet1$ in Dataset.


KMeans Clustering

Subset expression
<all valid cases>

Variables (pick one or more)
Amount.of.all.orders
Frequency.of.purchase

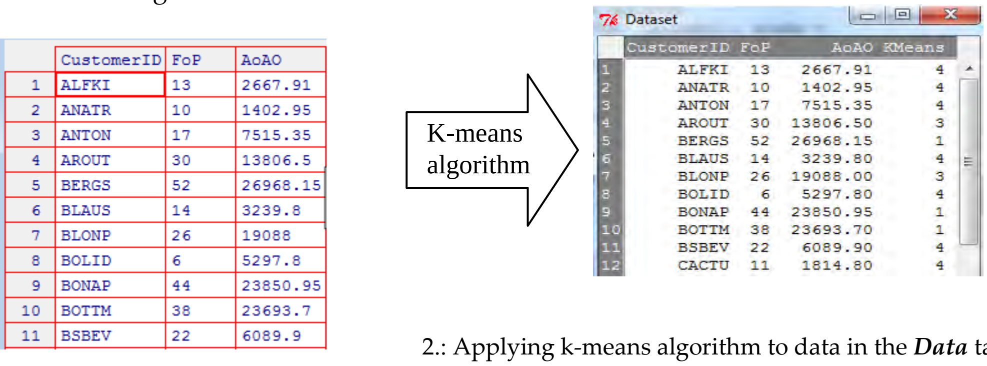
Number of clusters: 4
Number of starting seeds: 4
Maximum iterations: 10
Print cluster summary ☑
Bi-plot of clusters ☑
Assign clusters to the data set ☑
Assignment variable: KMeans

Statement

.cluster <- KMeans(model.matrix(~-1 + Amount.of.all.orders + Frequency.of.purchase, Dataset), centers = 4, iter.max = 10, num.seeds = 4) assigns to variable ***.cluster*** the clustering results. There are four clusters with corresponding centers (centroids) for attributes: Amount of all orders and Freaquency of purchase:

```
.cluster$size # Cluster Sizes
> .cluster$size # Cluster Sizes
[1] 15  3 25 46
> .cluster$centers # Cluster Centroids
  new.x.AoAO
[1]new.x.FoP
1  31818.709  42.86667
2 115464.487 101.33333
3  14565.625  26.04000
4   3604.382  12.19565
> .cluster$withinss # Within Cluster Sum of Squares
[1] 1394917889    9083184  326966494  237621111
```

After applying k-means algorithm, objects (customers) are joined to one of the four clusters what illustrates the following table:



	CustomerID	FoP	AoAO
1	ALFKI	13	2667.91
2	ANATR	10	1402.95
3	ANTON	17	7515.35
4	AROUT	30	13806.5
5	BERGS	52	26968.15
6	BLAUS	14	3239.8
7	BLONP	26	19088
8	BOLID	6	5297.8
9	BONAP	44	23850.95
10	BOTTM	38	23693.7
11	BSBEV	22	6089.9

	CustomerID	FoP	AoAO	KMeans
1	ALFKI	13	2667.91	4
2	ANATR	10	1402.95	4
3	ANTON	17	7515.35	4
4	AROUT	30	13806.50	3
5	BERGS	52	26968.15	1
6	BLAUS	14	3239.80	4
7	BLONP	26	19088.00	3
8	BOLID	6	5297.80	4
9	BONAP	44	23850.95	1
10	BOTTM	38	23693.70	1
11	BSBEV	22	6089.90	4
12	CACTU	11	1814.80	4

Table 2.: Applying k-means algorithm to data in the ***Data*** table

Using the programming language R /**8**/ it is possible to visualize the results of k-means algorithm in two dimensional system

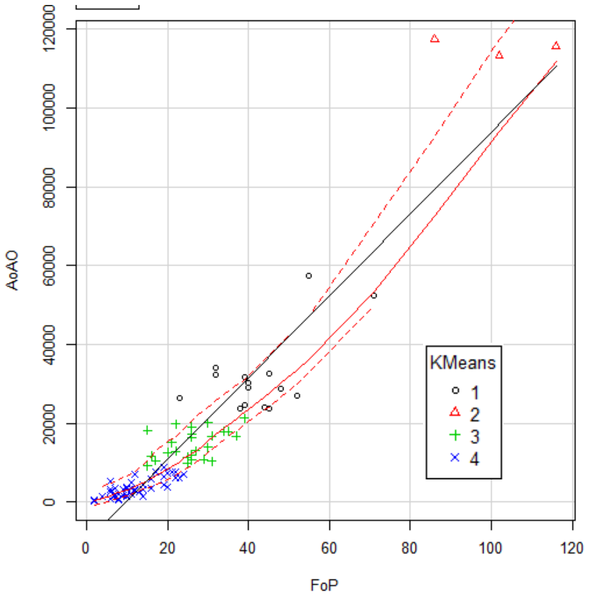


Figure 4.: The visualization of the results of clustering objects

[1] AoAO is abbriviation from Amount of All Orders and FoP stands from Frequency of Purchase.

Customers are now clustered into four clusters and their behavior may be analyzed using association rules. For each cluster will be implemented association rules algorithm because is reasonable suppose that the customers in the same cluster are behaving on similar way.

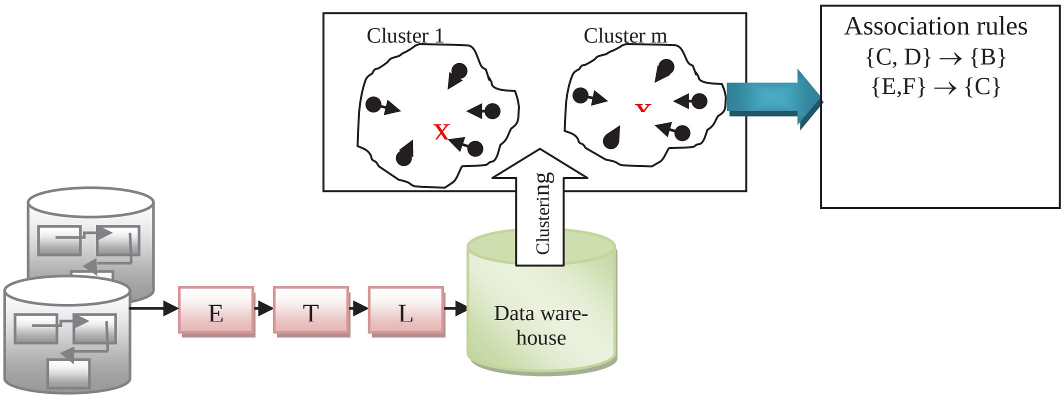


Figure 5.: The model of integration cluster analyses and association rules

The result of selecting the data often must be further transformed and processed by simply aggregating. Algorithm will depend on the given task. For discovering the customer profiles will be implemented cluster analysis (k-means algorithm) and association rules. Cluster analysis (k-means algorithm) and association rules are implemented by R language. Association rules algorithm often generate many irrelevant rules that are subsequently rejected during the validation process. Domain expert has to specify constraints on the types of rules of interest before the rule discovery stage and reduce the number of discovered rules that are irrelevant.

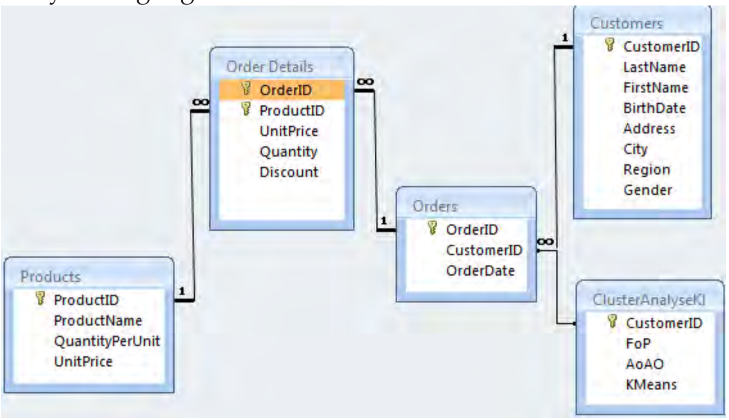


Figure 6.: Data model for implementation association rules

Therefore we will select only the customers which belong to the same cluster and analyze their purchase behavior.

Our data model is now extended with new table ClysterAnalyseKL. New data model is data source for data mining algorithm association rules.

4. ASSOCIATION RULES

Data mining algorithms are finite set of steps that find out pattern in large data set. Patterns may be described by rules IF....THEN, decision tables, neural networks, genetic algorithms, linear and nonlinear models. There is not a general acceptable and good data mining algorithm applicable to all situation and decision problems.

Generally speaking knowledge discovery in databases may be applicable to any data repository. Data applicable for association rules must be in transactional form.

Basically, data mining systems can identify frequent sets in transactional databases and perform *market basket data analysis.*

The goal of algorithm a priori is extracting rules of the form A-->B, where A i B are itemsets.

Rule is an implication expression. Data must be in transaction databases where does exists one table with attributes transaction_id (unique identifies the record) and transation items (items which belong to one transaction).

It is obviously that such relation is not in the first normal form. Therefore is necessary prepare data from transactional databases for implementation a priory algorithm. Before we present experimental results of implementation algorithm apriori in R language is necessary to analyze the steps of the algorithm.

The mining of association rules has two main steps:

1. Find out the itemset (the products in transactions) with support greater or equal to specified minimal support factor ***s***.

2. Use that itemset (itemset with support greater or equal to s) for generating association rules with confidence factor ***c*** (strong association rules) .

Data may be in the form of transactional databases or transformed into matrix /**9**/.

transaction_Id	transaction_itemset
1	A,B,C
2	A,B,C,D,E
3	A,C,D
4	A,C,D,E
5	A,B,C,D

Items transaction_Id	A	B	C	D	E
T_1	1	1	1	0	0
T_2	1	1	1	1	1
T_3	1	0	1	1	0
T_4	1	0	1	1	1
T_5	1	1	1	1	0

Figure 7.: Data in the transaction database form
Figure 8.: Data in the matrix form

Association rule is an implication expression of the form X → Y, where X and Y are item sets.
Example (Figure 8.):
{C, D} → {B}

There are two main parameters as evaluation metrics: support (s) and confidence (c). Support (s) is fraction of transactions that contain both X and Y. The second parameter, confidence (c) measures how often items in Y appear in transactions that contain X.

Example:

$$s = \frac{s(\mathrm{C,D,B})}{|\mathrm{T}|} = \frac{2}{5} = 0.4 \quad \text{and}$$

$$c = \frac{s(\mathrm{C,D,B})}{s(C,\mathrm{D})} = \frac{2}{4} = 0.5$$

Data mining algorithm association rules has the next steps:

Step 1. Generate frequent itemsets of length 1.

Step 2. Repeat until no new frequent itemsets are identified

a) Generate length (k+1) candidate itemsets from length k frequent itemsets

b) Prune candidate itemsets containing subsets of length k that are infrequent

c) Count the support of each candidate by scanning the DB

d) Eliminate candidates that are infrequent, leaving only those that are frequent.

Given a set of transactions T, the goal of association rule mining is to find all rules having thresholds
support ≥ *minsup* and confidence ≥ *minconf*.

List all possible association rules, compute the support and confidence for each rule and choose rules that pass the *minsup* and *minconf* thresholds is computationally prohibitive.

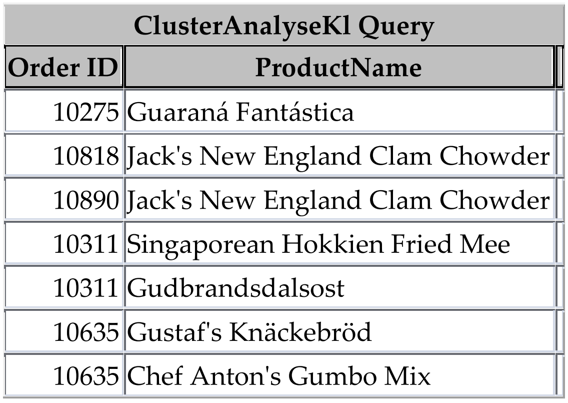
The total number of possible association rules is given by formula

$$R = \sum_{k=1}^{d-1}\left[\binom{d}{k} \times \sum_{j=1}^{d-k}\binom{d-k}{j}\right] = 3^d - 2^{d+1} + 1$$

where d is the number of unique items.

We will now prepare the data for implementation association rules in R language. First, we extract orders and products for cluster denoted as cluster "3".

```
SELECT   Orders.OrderID , [Order Details].ProductName
FROM (ClusterAnalyseKl INNER JOIN Orders
ON ClusterAnalyseKl.[CustomerID] = Orders.[CustomerID]) INNER JOIN [Order Details] ON Orders.[OrderID] = [Order Details].[OrderID]
ORDER BY ClusterAnalyseKl.KMeans =3;
```



ClusterAnalyseKl Query	
Order ID	**ProductName**
10275	Guaraná Fantástica
10818	Jack's New England Clam Chowder
10890	Jack's New England Clam Chowder
10311	Singaporean Hokkien Fried Mee
10311	Gudbrandsdalsost
10635	Gustaf's Knäckebröd
10635	Chef Anton's Gumbo Mix

Figure 9.: The market segment identified as cluster 3

The result of SQL query is the table with two columns: OrderID and Product name.

After data filtering and pre-processing we have to remove noise, outliers, missing fields, time sequence information and choose data mining tasks: classifications, segmentations, deviation detections, link analysis. In our example the knowledge discovery process revolves around discovery paradigm. It is association rules, an unsupervised learning approach for exploratory data analysis.

We are interested only to build association rules for nine products:

"Mozzarella di Giovanni" , "Queso Cabrales" , "Singaporean Hokkien Fried Mee" , "Manjimup Dried Apples" , "Ravioli Angelo" , "Chang" , "Pavlova" , "Schoggi Schokolade" , "Sir Rodney's Scones".

Therefore we make a new query that will extract only orders with this nine products. In SQL is enough to write the next code:

```
SELECT AR.OrderID, AR.ProductName
FROM AR
GROUP BY AR.OrderID, AR.ProductName
HAVING (((AR.ProductName)="Mozzarella di Giovanni" Or (AR.ProductName)="Queso Cabrales" Or (AR.ProductName)="Singaporean Hokkien Fried Mee" Or (AR.ProductName)="Manjimup Dried Apples" Or (AR.ProductName)="Ravioli Angelo" Or (AR.ProductName)="Chang" Or (AR.ProductName)="Pavlova" Or (AR.ProductName)="Schoggi Schokolade" Or (AR.ProductName)="Sir Rodney's Scones"));
```

After running this query we get the table with two columns OrderID and ProductName and the table is imported in Excel and saved as ban.xls.

In R language[2] are prepared the functions for implementing association rules. Using the next sequence of statements is relatively easy to perform the analysis and produce the association rules based on desired level of satisfaction and confidence.

```
> library(arules)
> library(RODBC)
> file.choose()
[1] "C:\\Users\\Brano\\Documents\\ban.xls"
```

[2] R started in the early 1990's as a project by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. Assumptions for using R language are that the user has a basic knowledge of statistics, econometrics, modeling, etc. An interesting of language R is that the user can create their own packages (package) and distribute them to other users. The package may contain a dataset, functions, reformulation of existing functions and more.

Data are stored in Excel file ban.xls. Package RODBC provides an interface to database sources supporting an ODBC interface. This is very widely available, and allows the same R code to access different database systems. In our example we access to data stored inExcel.

```
> conExcel<-odbcConnectExcel("C:\\Users\\Brano\\Documents\\ban.xls")
> data<-sqlFetch(conExcel,"Sheet1")
```

In data frame with name **data**, are stored data from file ban.xls. The function sqlFetch() enables to get data from a worksheet (Sheet1).

Split the data frame into two columns, one being identifier (OrderID) and other being an item (ProductName) contained in the data frame named dtra.

```
> dtra<-as(split(data$ProductName, data$OrderID),"transactions")
> data
   Orders_OrderID          ProductName
1      10250         Manjimup Dried Apples
2      10257         Schoggi Schokolade
3      10264         Chang
4      10269         Mozzarella di Giovanni
5      10272          Mozzarella di Giovanni
6      10298          Chang
7      10309          Singaporean Hokkien Fried Mee
8      10325           Mozzarella di Giovanni
9      10327          Chang
10      10327         Queso Cabrales
11      10332         Singaporean Hokkien Fried Mee
12      10335         Chang
13      10335         Manjimup Dried Apples
14      10337          Mozzarella di Giovanni
```

```
> dtra
transactions in sparse format with
 71 transactions (rows) and
 9 items (columns)
```

We can now build the model using dataset dtra:

```
> evalq({model<-apriori(dtra,parameter=list(support=0.01,confidence=0.01))})
```

The rules can be extracted and ordered using the function inspect(). In the following code block we use only first eight association rules.

```
> inspect(head(sort(model, by = "confidence"), 8))
 lhs                         rhs                          support      confidence    lift
1 {Schoggi Schokolade} => {Sir Rodney's Scones}          0.01408451    0.3333333  2.6296296
2 {Ravioli Angelo}      => {Manjimup Dried Apples}   0.01408451    0.3333333   2.1515152
3 {}                    => {Pavlova}                 0.18309859    0.1830986   1.0000000
4 {}                    => {Chang}                   0.18309859    0.1830986   1.0000000
5 {}                    => {Queso Cabrales}          0.16901408    0.1690141   1.0000000
6 {}                    => {Manjimup Dried Apples}   0.15492958     0.1549296   1.0000000
7 {Pavlova}             => {Chang}                    0.02816901    0.1538462   0.8402367
```

```
8 {Chang}                    => {Pavlova}                       0.02816901    0.1538462   0.8402367
9  {}                   => {Mozzarella di Giovanni}          0.12676056     0.12676056   1.0000000
10 {}                    => {Singaporean Hokkien Fried Mee} 0.12676056   0.12676056   1.0000000
11 {}                    => {Sir Rodney's Scones}           0.12676056      0.12676056   1.0000000
12 {Sir Rodney's Scones}         => {Schoggi Schokolade}      0.01408451    0.11111111   2.6296296
13 {Mozzarella di Giovanni}       => {Pavlova}                  0.01408451    0.11111111   0.6068376
14 {Singaporean Hokkien Fried Mee} => {Manjimup Dried Apples}  0.01408451 0.11111111  0.7171717
15 {Sir Rodney's Scones}         => {Chang}                     0.01408451    0.11111111   0.6068376
16 {Manjimup Dried Apples}        => {Ravioli Angelo}            0.01408451    0.09090909    2.1515152
```

We notice that the rules seven and eight are symmetric, that means everyone who purchases one of these products always also purchases the other.

CONCLUSION

Knowledge discovery is a process that requires an understanding of the business domain and business data. The nature of business domain determines the data preprocessing and creating target data set. Then we choose adequate data mining algorithm. The final goal is in this learning process extract new knowledge. In this paper, we have shown how knowledge discovery can be integrated into a marketing knowledge management framework. Major problem is to filter, sort, process, analyze and manage data in order to extract the information relevant to customer profiles. Learning ability and knowledge are two preconditions of survival and development. The paper clearly shows the integration of cluster analysis and a priori algorithm using R language. Cluster analyze is implemented on the given data set and result was the segmentation of customers. Similar customers are included in the same clusters. Then we discovered the behavior of customers inside the clusters using a priori algorithm. Cluster analysis and association rules have shown a strong implementation power in discovering customer's profiles. The quality of learning and discovering the hidden relationships among variables of data mining algorithm is dependent the quality of data and its amount. In the paper primary goal was to generate the conceptual framework of knowledge discovery and its practical value and consequences. The complexity of experimental results is reduced using simplicity and high applicable power of R language and its packages.

Notes

/1/ Markic, B. (2011) Customer segmentation by integrating unsupervised and supervised learning, Proceeding from international conference Economic Theory and Practice: Meeting the New Challenges, Faculty of Economics University of Mostar, october 2011, Mostar.

/2/ Markic, B. (2011) Integrating Theory and Practice: Knowledge Discovery and Unsupervised Learning, Meeting of Management Departments the Faculties of Economics, Faculty of Economics University of Split, September 2011.

/3/ John Maindonald and John Braun, (2003) *Data Analysis and Graphics Using R - An Example-Based Approach,* Cambridge University Press.

/4/ Markic, B. (2011) Integrating Theory and Practice: Knowledge Discovery and Unsupervised Learning, Meeting of Management Departments the Faculties of Economics, Faculty of Economics University of Split, September 2011.

/5/ Markic, B. (2011) Customer segmentation by integrating unsupervised and supervised learning, Proceeding from international conference Economic Theory and Practice: Meeting the New Challenges, Faculty of Economics University of Mostar, october 2011, Mostar.

/6/ Ibidem

/7/ Markic, B. (2011) Integrating Theory and Practice: Knowledge Discovery and Unsupervised Learning, Meeting of Management Departments the Faculties of Economics, Faculty of Economics University of Split, September 2011.

/8/ John Maindonald and John Braun, (2003) *Data Analysis and Graphics Using R - An Example-Based Approach,* Cambridge University Press.

/9/ Markic, B. (2011) Integrating Theory and Practice: Knowledge Discovery and Unsupervised Learning, Meeting of Management Departments the Faculties of Economics, Faculty of Economics University of Split, September 2011.