Insights offered by data-mining when analyzing media space data

Maja Skrjanc, Marko Grobelnik, and Darko Zupanic
Jozef Stefan Institut, Jamova 39, Ljubljana, Slovenia
Maja.Skrjanc@ijs.si, marko.grobelnik@ijs.si, darko.zupanic@ijs.si

Keywords: data mining, media space data, data analysis

Received: June 30, 2001

Media space consists from many different factors fighting for the attention of customer population in a certain environment. Common problem in bigger environments (or countries) is that datasets describing complete media space is hard or almost impossible to get since the detailed picture is too complex or to expensive to compose. However, this is not the case in environments, which are smaller, and is therefore easier to collect the data. We have access to the data entirely describing the media space of population of 2 million people. Because of the language and economy this media space functioning relatively independently from different factors, specially outside the country. The data was collected by Media Research Institute Mediana. The database consists of 8000 questionnaires, gathered in 1998. The sample and the questionnaires were made by comparable research standards. In this paper we will discuss different type of questions, which might become in a great assistants in unfolding the media groups, audience fluctuation and profound understanding of happenings in media space, as well as for their predictions.

1 Introduction

New emerging technologies enable more transparent communication between the media and the audience. As the response time for information feedback is becoming shorter, the general public can be more involved in the process of shaping the media. For the same reason measuring the media impact on the public is becoming much easier. Information about the media space, dynamics, interactions between the public and media are regularly monitored, collected and analyzed. Those type of information and knowledge raise different kinds og questions, which were also addressed in several other analysis [9,10]. Knowledge or new information extracted from the data is a value added information, which in this highly competitive environment represents a crucial factor.

One of the possible approaches to get additional value from the gathered information is the use of the data mining techniques, which can not only contribute to the deeper data analysis, but can also create new additional services providing new insights into the data. Although Slovenia is a small media space is also very specific because of the language, it is not an exception in comparison to some other media environments.

2 The Data

In this section we will describe the contents, structure and quality of the analyzed data set.

For the purpose of our analysis we took one of the Mediana's data sets, describing the whole Slovenian media space. The data was gathered by comparable international research standard. The data set consists of about 8000 questionnaires tracking all of the important reading, listening, and watching media in Slovenia. Each questionnaire consists of about 1100 questions collected in groups: about the person's relation to the certain media, person's activities, interests, life style, income, property, and demographic data. The relation of a person to the certain media is checked in more details with

Since 1992 Media Research Institute, Mediana (http://www.irm-mediana.si/) follows all printed, TV and radio media in Slovenia. They are trying to unfold and explain Slovenia's media image by collecting all kind of data about the media and analyze them with simple statistical methods. As a part of the Sol-Eu-Net project (http://SolEuNet.ijs.si/) we came into the position to analyze the data with more sophisticated data mining methods and present our outcomes to Mediana.

In this article we will answer to some selected questions from the large space of interesting questions, arising from the need for better understanding of the media space.
In the first section we will describe the quality, structure and contents of the data set. The second section will define the selected questions. In the third section we will describe our experiments by the methods we used to answer the questions. The answers are supported by many comprehensible examples of the rules and trees. At the end, in the fourth section we will describe the Mediana response and show some directions for the future work.

different questions testing the several aspects of using the media.

The first page of the poll contains general questions about followed by 19 pages of specific questions. The most of the questions are asked in the way that the answer consists from a grading scale with 2, 5 or 9 levels. The data set is presented in a spreadsheet table, were each questionnaire represents a row and each question represents a column. In general, Mediana's dataset if of rather high quality meaning that we didn't have much work cleaning and transforming the data.

3 Questions

Mediana didn't give us any specific questions; they just gave us a challenge to find something interesting in their data, which might be of any interest to them. Therefore, we had to think of some questions, which Mediana would find interesting in a way that the answers or techniques would represent a possibility for offering an additional commercial service. The requirements for the answers and resulting models to the selected questions were comprehensibility especially in the comparison to the classical statistical methods, which they have already used.

For our analysis, we selected the following questions:

- ⇒ Which printed media are also read by the readers of a selected newspaper/magazine?
- ⇒ What are the properties of readers/listeners/watchers of the specific media?
- ⇒ Which of these properties distinguish between the readers of different newspapers?
- ⇒ What are typical groups of people according to their personal characteristics?
- ⇒ Which media are similar?

4 Experiments

4.1 Methods

To answer the selected questions we used the following methods:

- o Correlation based clustering of the attributes
- o Association rules (Apriori) [4]
- o Decision trees (C4.5 , C5) [5]
- o Clustering (K-Means) [7]
- o Kohonen Network [6]

Our goal was to find some relations within the data, which are not obvious at the first sight and are comprehensible to Mediana. Due to that, our results do not optimize accuracy, but comprehensibility. Usually, our interpretation of the results does not rely only on the particular rule, but it is generalized over a group of similar rules. By that, we can offer better interpretation and generalization.

4.1.1 Clustering of the attributes

In this section we discuss the relation between the attributes.

To determine the dependencies between the attributes we used clustering, where the distance was the correlation coefficient between the attributes. As the result we got clusters of attributes, which are correlated with the correlation coefficient above a certain threshold (0.5). Some of the resulting groups collect different attributes describing several aspects of the same media. For the most of the groups we were able to provide very comprehensive explanation. Some other groups consist of the attributes having very obvious relationships, like the group from EXAMPLE 1. This group is composed of the attributes dealing with the same media. Questions, which correspond to this particular attributes are: (1)Did you read magazine Golf in the last year (BMERead_Golf)? (2)How many issues did you read in last 12 months (BMEIssues_Golf)? (3)How long ago did you read your last issue (*BMELastRead_Golf)*?

EXAMPLE 1

Correlations between different aspects of the same media.

Attributtes:	BMERead_Golf
	BMEIssues_Golf
	BMELastRead_Golf

Another type of clusters represents attributes with the high correlations between region and all or most of the editions of the same newspaper company. In particular, case in the EXAMPLE 2, the newspaper company Vecer is a local newspaper company. *Vecer* is the main daily newspaper and *Vecerov Cetrtek, Vecer Televizija in Radio, Vecer v Soboto* are its supplements. We can see that similar groupings as in the EXAMPLE 1 joined with some demographic attributes like region, community and local community. They emphasize the local influence of this media.

EXAMPLE 2

Attributtes:	*REGION*
	COMMUNITY
	LOCAL_COMMUNITY
	QUESTIONER
	DERead_Vecer
	DESRead_Vecerov Cetrtek
	DESRead_Vecer.Televizija,Radio
	DESRead_Vecer v Soboto
	DEIssues_Vecer
	DESIssues_Vecerov Cetrtek
	DESIssues_Vecer,Televizija,and Radio
	DESIssues_Vecer v Soboto
	DELastRead_Vecer
	DESLastRead_Vecerov Cetrtek
	DESLastRead_Vecer Televizija, Radio

DESLastRead_Vecer v Soboto

Another type of clusters stress out the correlations between the attributes describing the person's age and spare time activities, which are part of a life style group of questions. EXAMPLE 3 presents the group of attributes describing the persons age and type of spare time activities: How often are you going to the cinema (*spareTime_Cinema*), Do you study in your spare time (*spareTime_Study),* How often are you listening to CDs, LPs *(spareTime_LP/CD)?* Are you speaking English *(languagesUnderstanding_English)?* When was the last time you were in the cinema (*cinema*)? This type of clusters also point out that part of the person's life style is age dependent.

EXAMPLE 3

Attributes: *spareTime_Cinema*
spareTime_Study
cinema
basicDescriptions_BirthYear
delivered_Age
spareTime_LP/CD
languagesUnderstanding_English

An interesting but not unexpected type of clusters represents EXAMPLE 4. It presents the correlations between spare time activities and items possession. The person who is interested in the science (*mediaInterest_Science*) and the computer science (*mediaInterest_ComputerScience*) and is in a spare time is using a computer (*spareTime_Computers*) has very likely a computer at home (*householdItemsHas_PersonalComputer*), as well as modem, internet access, video, CD ROM (*householdItemsHas_Modem, householdItemsHas_InternetAtHome, householdItemsHas_Video/ComputerGames, householdItemsHas_CD ROM2).*

EXAMPLE 4

Attributtes: *spareTime_Computers*
mediaInterest_ComputerScience
householdItemsHas_PersonalComputer
mediaInterest_Science
householdItemsHas_CD ROM
householdItemsHas_Video/Computer Games
propertiesHas_PersonalComputer
householdItemsHas_Modem
householdItemsHas_Internet at Home

Similar type of clusters as the one from EXAMPLE 2 is EXAMPLE 5. It also includes correlations from EXAMPLE 1. Here we can observe that all media from a certain province are highly correlated. Dolenjski List, TV Novo Mesto, Radio Krka, Studio D, Radio Sraka are media from the province Dolenjska. People obviously like to keep track of local events, which are covered mostly from local media.

EXAMPLE 5

Attributtes: *WERead_Dolenjski list*
WEIssues_Dolenjski list
WELastRead_Dolenjski list
televisionsWatched_TV Novo mesto
televisionsLastWatch_TV Novo mesto
televisionsDays_TV Novo mesto
radiosLastListen_Radio Krka, Novo mesto
radiosLastListen_Studio D
radiosDays_Radio Krka, Novo mesto
radiosDays_Studio D
radiosListened_Radio Krka, Novo mesto
radiosListened_Radio Sraka - Novo mesto
radiosListened_Studio D

4.1.2 Association rules

Association rules [4] are part of the standard selection of data mining techniques. We used them (as implemented in [1] and [2]) to answer the two questions: (1)how readers of one daily newspaper are related to the readers of other newspapers and (2)what is the relation to all other attributes.

These questions were tested on the whole set of attributes, except the question about relations between different newspapers, which was tested only on the chosen newspapers attributes.

Surprisingly, the answers of both tests were almost identical, which gave us an idea, that reading certain newspaper is very much dependent on reading some other newspapers. The exceptions were two local daily newspapers, which are very region dependant.

Lets look at some examples of association rules. EXAMPLE 6 and EXAMPLE 7 present the comprehensible interpretation of the rules. Attributes in the EXAMPLE 6 and the EXAMPLE 7 correspond to the question about the reading certain magazines and newspapers in the last year or in the last six months. The number at the end of the left side of the rule represents number of covered examples (support) and numbers on the right side of the rule represent the number of covered examples (confidence).

As the results we got rules, which uncover the relations between different publications. These relations were very interesting – particular because of the nature of the topics these publications mainly cover. In the EXAMPLE 6 we got the rule, which associate readers of the biggest Slovenian daily newspaper Delo, with readers of magazines (Marketing magazin, Finance, Razgledi, Denar, Vip), which are covering mainly economics and marketing topics.

In the EXAMPLE 7, the rules show the connection between the readers of Slovenske Novice and the publications, which are known more as a 'yellow press' (Sara, Ljubezenske zgodbe, Omama). They cover mostly

romantic and erotic topics. This is not surprising since Slovenske Novice is known as a kind of yellow press daily newspaper. It is the most read newspaper in Slovenia.

EXAMPLE 6.
Interpretation: "The majority of the readers of any of the following publications:
Marketing magazin, Finance, Razgledi, Denar, and Vip are also readers of Delo."
Rules: DERead_Delo

1. MERead_Marketing magazin (sup.=116) ==> DERead_Delo (conf.=0.82)
2. WERead_Finance (sup.=223) ==> DERead_Delo (conf.=0.81)
3. BWERead_Razgledi (sup.=201) ==> DERead_Delo (conf.=0.78)
4. BWERead_Denar (sup.=197) ==> DERead_Delo (conf.=0.76)
5. MERead_Vip (sup.=181) ==> DERead_Delo (conf.=0.74)

EXAMPLE 7.
Interpretation: "The majority of the readers of any of the following publications:
Sara, Ljubezenske zgodbe, Dolenjski list, Omama, and Delavska enotnost
are also readers of Slovenske novice."
Rules: DERead_Slovenske novice

1. BWERead_Sara (sup.=332) ==> DERead_Slovenske novice (conf.=0.64)
2. WERead_Ljubezenske zgodbe (sup.=283) ==> DERead_Slovenske novice (conf.=0.61)
3. WERead_Dolenjski list (sup.=520) ==> DERead_Slovenske novice (conf.=0.6)
4. MERead_Omama (sup.=154) ==> DERead_Slovenske novice (conf.=0.58)
5. WERead_Delavska enotnost (sup.=177) ==> DERead_Slovenske novice (conf.=0.58)

4.1.3 Decision Trees

Decision Trees [5] are also part of the standard repertoire of the data mining and machine learning techniques. We used C4.5 (as implemented in [1]) and C5.0 (as implemented in [2]) to describe the characteristics of the readers reading certain daily newspaper. Another question that we tried to answer was how readers of one daily newspaper differ from the readers of the other daily newspaper.

With the decision trees we get the most natural and understandable interpretation for these problems. We try to identify typical description characteristics of the readers, including their life style, life statements, capability of trademarks recognition, their interests, etc. After putting together the rules and their interpretations, we got description of typical reader for every daily newspaper. Parts of descriptions correspond to the characteristics, Mediana already did with statistical methods, but some of the descriptions include very interesting personal characteristics. At first sight these rules seem unreasonable, but after some more careful analysis and discussions with the experts a perfectly reasonable explanation could be found. Those rules represent the most valuable results. We could observe that effect especially in the EXAMPLE 9, where we describe readers of the newspaper Slovenske novice.

While testing we run algorithm on several groups of attributes, which describe the person's life style, statements, activities, interests, recognition of trade marks, statements. Our priority was comprehensibility of the trees. Since some newspapers have a very few readers, we had to adopt the thresholds and parameters, to get the right level of the comprehensibility.

We presented the most promising trees in a form of rules. We took those rules, which have the accuracy above the certain threshold (0.65). On this set of rules we based our interpretations and each interpretation is based on several different rules.
Examples 8-10 will present several interpretations and example of one the rules, this particular interpretation is based on. On the right side of the rule is class (True, False) and the first number stands for the number of cases and the second number for the accuracy.

EXAMPLE 8: Description of readers of daily newspaper Delo

Typical reader of Delo reads newspapers several times per week, has higher level of education recognizes certain trademarks of newspapers/magazines, cars, bears, washing powders, he or her is tracking information about the manufactures, shopping, information about inland news, likes to watch TV and videocassettes.
The results fit well on our intuitive presumption.

Several rules, which are the basis for the description above:
Interpretation: Read newspapers several times per week.
Rule:
if (reading daily newsp. at all)>5) and (Tracking topical subjects at home)>3) and (Recognition of trade marks of daily newspapers >2) and (Recognition of trade marks of daily newspapers <=4) ➔ *T (1051, 0.84)*

Interpretation: Recognize trademarks of newspapers/magazines, cars, bears, washing powders,...
Rule:
if (Recognition of trade marks of daily newspapers >2) and (Recognition of trade marks of daily newspapers <=4) ➔ *T (2333, 0.672)*

Interpretation: Higher levels of education.
Rule:
if (reading daily newsp. at all >1) and (spend evenings at home>1) and (interest at animal world>1) and

(recognition of trade marks of washing powder<=40) and (recognition of trade marks of magazines<=270) and ((Education)>5) ➔ *T (242.9, 0.855)*

Interpretation: Watching TV and videocassettes.
Rule:
if (watching. videocasset.<=6) and (theater<= 4) and (Inf. about manufactures, shopping>4) and (recognition of trade marks of daily newspapers>0) and (recognition of trade marks of daily newspapers<=4) and ((Time of getting up<=11) ➔ *T (108.8, 0.845)*

Interpretation: Interested in inland news.
Rule:
if (newspapers weekly_read)>5) and (Topical subjects at home >3) ➔ *T (1051, 0.84)*

EXAMPLE 9: Description of readers of daily newspaper Slovenske Novice

Typical reader of Slovenske Novice is a regular reader of newspapers/magazines and likes to sit in coffeehouses, bars and sweetshops.: Recognize trademarks of newspapers/magazines, commercials for newspapers /magazines. They recognize less trademarks as readers of Delo newspaper and they also typically recognize different trademarks than readers of Delo newspaper, also reads Slovenski Delničar (magazine that covers economical topics), Jana (magazine tracking topics, which are more feminine), Kaj, Vroči Kaj (yellow press, erotic contents). If he or she is speaking Croatian then also probably reads Kaj magazine.

The most interesting statement is the one saying that readers of Slovenske Novice like to sit in coffeehouses, bars, and sweetshops. This rule looks very strange, but just at first sight. Slovenske Novice is namely the newspaper that has the highest number of readers in Slovenia, but Delo newspaper has the largest edition. Slovenske Novice has the second largest edition. How is this possible? If look closer at the bars, coffeehouses or sweetshops, we could find in the most cases Slovenske Novice newspaper on the table. So, when you are in this in kind of places, besides drinking coffee, or eating sweets, you also read Slovenske Novice.
The statements concerning trademarks could be profitably used for marketing planning.

Interpretation: Regular readers of newspapers/magazines and like to remain sitting for a while in coffeehouses, bars, sweetshops,...
Rule:
if (reading daily newsp. weekly)>4) and (Visiting coffeshops more then weekly<=6) and (interest at animal world>1) and (reading SportNovosti ==F) ➔ *T (331, 0.889)*

Interpretation: Recognize trademarks of newspapers/magazines, commercials for newspapers/magazines,...
Rule:
if (Remembering Commercials for daily newspaper >0) and (recognition of trade marks of Daily newspaper>12) and (recognition of Trade marks of Daily newspaper<=16) ➔ *T (624, 0.887)*

EXAMPLE 10: Description of readers of daily newspaper Ekipa

Typical readers of Ekipa also read Sportske novosti (newspaper which cover exactly the same topics as Ekipa, except that it is in Croatian language), they visit sport events, are interested in motoring, they like to experiment with novelties. Those characteristics are intuitive if we know, that Ekipa newspaper is dedicated to the sports, especially to team sports. Unexpected were characteristics, that they are very tidy and that they have mostly one child between 7 and 14 years old.

Interpretation: Tidy.
Like to experiment with novelties.
Read Sportske novosti.
Have at most one child between 7 and 14 years old.
Rule:
if (importance of tidyness, cleanness >3) and (trying new things >2) and (new challenges <=4) and (liking new things <=3) and (children 14 years old<=1) and (SportNovosti==T) ➔ *T (71, 0.712)*

All the examples until now are dealing with question of describing typical readers of a certain newspaper. How did we attach the other question, which try to distinguish between the readers of the two largest newspapers in Slovenia? We select only those readers who read either Delo or Slovenske Novice. The class for the learning became which newspaper they read. The decision tree divided the readers into the two distinctive groups. Tree is optimized for comprehensibility.

EXAMPLE 11: How do readers from Delo differ from readers of Slovenske Novice? See Figure 1

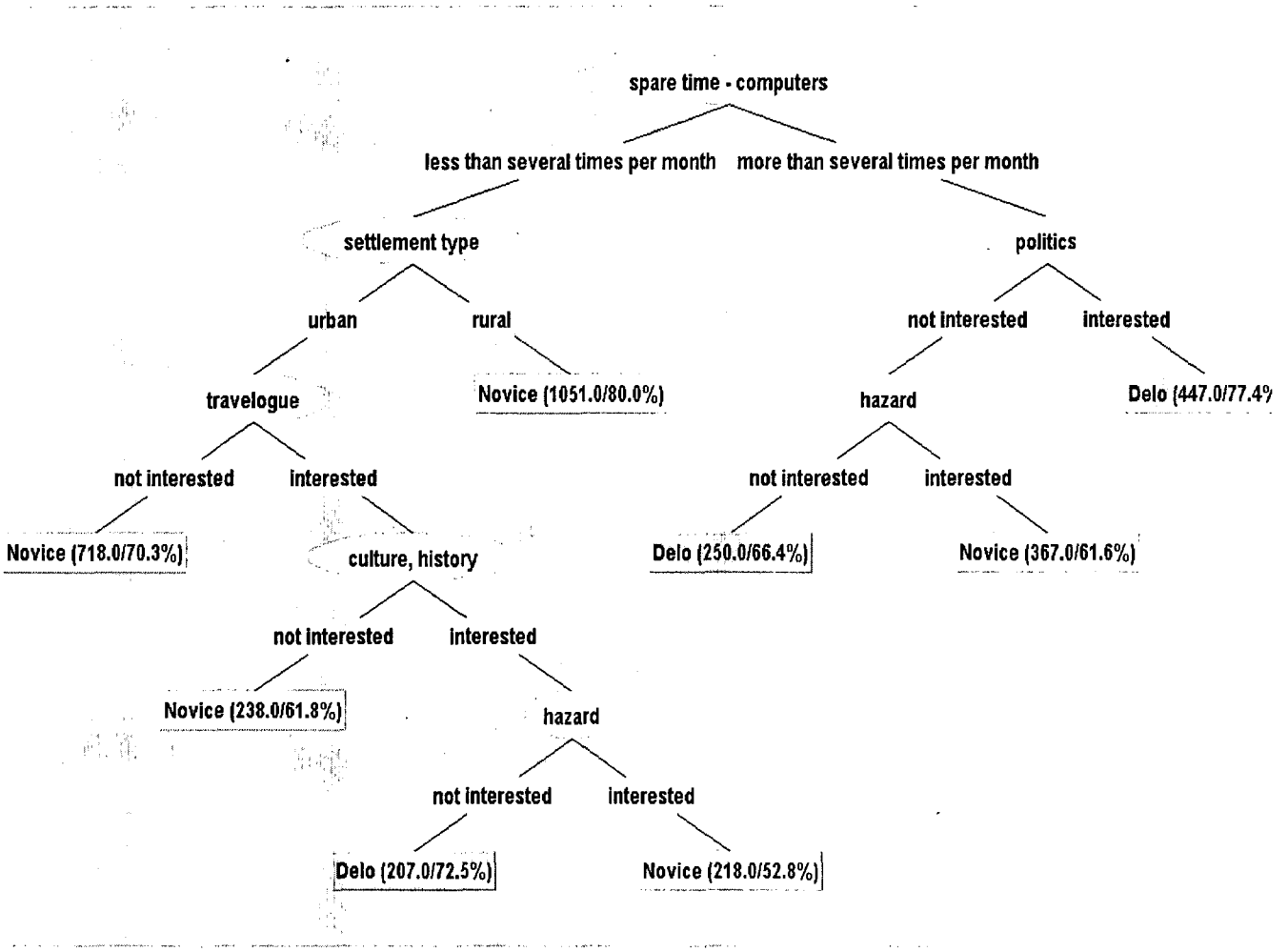

Figure 1: Decision tree for readers of Slovenske Novice and Delo

4.1.4 Clustering

One of our interests was also the identification of different groups of people and to describe their characteristics, regardless on their interest for the media.

First, to determine how many different distinctive groups could be find, we run Kohonen Network algorithm (as implemented in [2]). The result of Kohonen Networks was used as a K parameter in the K-means algorithm[7]. We run the K-means algorithm on 4 groups of attributes (life viewpoints, media topics, spare time activities, demographic properties like age, education, sex). As the result we got four clusters., consisting of 2550, 2124, 1385 and 1894 examples. For the description of cluster's characteristics we use C.5 algorithm, with the additional constrained of 200 examples per leaf. So, each cluster represents separate learning problem, where the target cluster (the cluster we try to describe) represents a positive class and the other three rep[resents a negative class value.

We got rather comprehensive trees with the following interpretation:

- 1st group consists of people younger than 30 years, they are not interested in topics like family, breeding, partnership;
  or
  younger than 20 years, they are interested in topics like living exciting, interested in novelties, films, challenges,...
  We could describe them in short as inspiring young people.
- 2nd group consists of passive people, they don't like challenges, are not interested in entertainment, science, techniques, economics, their main satisfaction is family, they like life without major changes.
  They could be described as inactive older people.
- 3th group consists of people with higher level of education, occupied by computers; mostly they are older than 30 years and listen to music; they

are interested in most of the topics, they have classic taste, they are occupied by their children, promotions, novelties, challenges are important to them, they follow media quite often.
We could describe them as ambitious people.

- 4th group form older people, occupied by handicraft, they are not interested in sport, but they are interested in most of the other topics, they also like to get know with novelties, accepting as challenge.
They can be referred as active older people.

Although most of the results impressed Mediana, the clustering results were the most exciting for them, since they performed the same experiment on the same data with the classical statistical analysis [3]. It took then several steps to define the criteria for the groups they were pleased with. It took their time, resources and expert knowledge. Our results, which were practical the same, were gained in a very short time and without any prior expert knowledge.

5 Conclusions

In the paper we presented an experience with the data mining analysis of the real world data describing the complete media space in Slovenia. Since the agency (private institute Mediana) providing us with the data didn't specify the goals and tasks we should follow, we decided for a set of tasks seemed to us as interesting. The dataset is a collection of approx. 8000 questionnaires each having approx. 1100 questions covering all kinds of topics about the personal interests and relationships of a questioned person to all important Slovenian media (newspapers, radio and TV) as well as it's personal interest, lifestyle, social status etc. For the analysis phase we decided to use several techniques with the some main goal to enable deeper understanding about the dataset.

Since the number of attributes was fairly large (above 1100) we decided first to find highly correlated groups of attributes to give some insight into the structure of the questionnaire. Next, we created with the algorithm Apriori the association rules discovering the relationships in reading habits for the people reading more than one newspaper. Using decision tree learning we enlightened personal characteristics of the people reading certain newspapers and finally with clustering (K-Means) we split the people answering the questionnaires in to several groups according to the attributes describing their personalities and lifestyle.

Most of the results were very useful for the Mediana Institute to get additional insights into their own data. Some of the results have also potential to become additional commercial services offered by Mediana.

Acknowledgement

This work was supported by the EU project Sol-Eu-Net, IST-1999-11495, and the Slovenian Ministry of Science, Education, and Sport.

References

[1] Ian H. Witten, Eibe Frank (1999), *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, and the implementation of Weka system, Morgan Kaufmann

[2] Clementine system (http://www.spss.com/clementine/)

[3] Jasna Zdovc, (2000) Segmentation of audience by style (in Slovene), MSc Thesis University of Ljubljana, Slovenia

[4] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo, Fast discovery of association rules. In U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy (eds.) Advances in Knowledge Discovery and Data Mining, AAAI Press/The MIT Press, pp. 307-328, 1996.

[5] Ross Quinlan (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers, Inc.

[6] T. Kohonen (1984) Self-Organization and Associative Memory, Springer-Verlag

[7] Selim, S.Z. and Ismail, M.A. (1984). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6, 81-87.

[8] Arabie,P.,and Hubert, L., (1995) Advances in cluster analysis relevant to marketing research. In From Data to Knowledge, W. Gaul and D. Pfeifer, Eds. Springer, Berlin, pp. 3--19.

[9] Leo Bogart (1989), Press and Public: who reads what, when, where, and why in American newspapers, Lawrence Erlbaum Associates, Publishers.

[10] Women and Elections '99" and "Elections in Croatia 2000 - 20 % is (not) Enough" , brochure of Women's Information and Documentation Center, Zagreb, Croatia (http://www.zinfo.hr/engleski/research.htm)