Buletinul Ştiinţific al Universităţii Politehnica Timişoara

TRANSACTIONS on ELECTRONICS and COMMUNICATIONS

Volume 59(73), Issue 2, 2014

Standard of Data Mining CRISP-DM

Diana Sipoteanu

Abstract- The need of research studies achievement in different areas (tehnical, medical, sociological, etc) has increased giving rise to a new field called data mining. This article presents an approach over a study of 3D preception over the cogntive charge based on creating and analysis of a multidimensional questionnaire related to this topic. The questionnaire analysis was based on CRISP-DM methodology wich consist of six phases. Results were obtained by using the SAS Enterprise Guide tool options. The aim of this study was to improve the 3D perception by ilustrating 3D disadvantages.

Keywords: data mining, questionnaire, methodology.

1. INTRODUCTION

In the last years, data mining became a highly sought field because of increasing the necessity of research projects and because of the applications complexity from different areas such as telecommunications, biomedicine, financial analysis, etc.

Early data mining applications focused mainly on helping businesses gain a competitive edge. The exploration of DM for businesses, is expanding as e-commerce and e-marketing have become mainstream elements of the retail industry. Emerging application areas include data mining for counterterrorism and mobile (wireless) DM. As generic data mining systems may have limitations in dealing with application-specific problems, we may see a trend toward the development of more application specific data mining systems [2]. Nowadays, with the explosion of information, Data Mining has become one of the top ten emerging technologies that will change the world [5]. "Data Mining is most sought after..." according to Information Week Survey [6].

Data mining can be defined as a science which involves extracting the most important information from a large dataset. Besides this statement, this science can be defined in several ways by scientific community: Data Mining is a decision support process where the users are looking for the interpretation of the data patterns [7]. "Data mining is the process of discovering meaningful new correlations, patterns and trends by shifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques"[5]. Data Mining is an iterative process of analyzing a large set of data which extracts valuable knowledge by the data-analysts that play a central role [7].

DM is a very complex field encompassing a wide variety of topics in computer science and statistics being placed at the intersection of many areas such as data management and databases, machine learning, artificial intelligence, statistics, pattern recognition.

Based on this idea there are several tasks that could be accomplished according to the purpose of each project: Exploratory Data Analysis EDA consisting on interactive or visual techniques, Descriptive Modeling (density estimation, cluster analysis and segmentation, dependency modeling), Predictive Modeling including regression for quantitative variables and classification for categorical variables, Discovering Patterns and Rules, Retrieval by Content [1].

This article will describe what impact can create the 3D perception over the cognitive charge.

The cognitive charge was studied under 3 angles: subjective measures, psychological measures and performance measures based on Cain theory (2007) so this study was made by experimentation of the psychological measures resorted and by achievement of a multidimensional questionnaire.

The subjective measures contain 6 items classified in 3 groups: charge characteristic (mental demand, psychical demand, and temporal demand), behavioral characteristic (performance and sport), and individual characteristic (frustration). The evaluation of the cognitive charge was realized in 2 phases: binary comparison of 6 dimensions achievement evaluated according accomplished task so to each dimension was assigned a weight from 0 to 5.

The main tasks for performance measures is collecting information that allows making assumptions related to the cognitive charge and the double-task protocol which involves errors number computing [4].

Psychological measures involve establishing and verifying the assumptions related to this topic so an analysis of collected information in descriptive terms was required. In addition study of means, variances, and dispersions were made. The field of study related to this type measures is called psychometrics and it is focused on the theory and technique of psychological measurement, Likert scale is commonly used. Different human aptitudes (knowledge, attitudes, abilities, educational measurement, etc) should be measured by constructing instruments and measurement procedures, development and refinement of theoretical approaches. Much of the early theoretical and applied work in psychometrics was undertaken in an attempt to measure intelligence [10].

2. AIMS AND OBJECTIVES

The aim of this project was revealing important information about how people can perceive 3D Cinema sessions and 3D games or other 3D supports.

In this context the objectives to be attained were obtaining a successful analyze of a multidimensional questionnaire supposed to verify assumptions based on scientific studies of this topic: *3D perception can cause tiredness at a cognitive level, 3D perception can modify capabilities, 3D perception causes psychical ailment, 3D perception can cause cognitive tiredness, 3D perception has advantages and disadvantages.*

The main goal of my work was to describe the 3D Cinema part even if the study implied also sessions for 3D video games.

2. CONTRIBUTION

The problem needed to be solved for succeeding in questionnaire analysis was developing a process in data mining that reveals the required information. Therefore a proper methodology and a proper tool for data processing had to be chosen. The purpose of the process was to reduce data so specific techniques were required for getting a smaller dataset with the same or almost the same integrity of information as the initial dataset. As an environmental tool SAS Enterprise Guide Tool has been used and the process was implemented based on CRISP-DM methodology and using as a support the project described in [fr]. The SAS tool was chosen because it has many algorithms already implemented useful for applying different modeling techniques thus the time of DM model development is reduced and another reason is its attractive GUI. It is able to execute the main tasks of DM starting with simple options (bar charts, box plot, scatter plot, histograms, etc) to more complex options (characterize data, query builder, correlation analysis, regression analysis, etc).

My work focused on the second and third phase of a DM model. Consequently a set of tasks had to be followed in concordance with the methodology chosen. After firsts insight of the initial dataset some attributes characteristics were established: name, type, meaning, range. Thus it was noticed that the DM problem type was a combination between data description and summarization type and dependency analysis type. More details will be given in section 4.

Data description and summarization aims at the concise description of characteristics of the data, giving an overview of the structure of the data to the user. It can be a subordinate goal in the process in most of the DM projects and sometimes can be a standalone objective of a DM project. Because of the scant information that user has, related to the nature of the data and precise goal of the analysis, a initial exploratory data analysis should be taken into consideration for understanding the nature of the data and forming potential hypotheses for hidden information and also simple descriptive statistical and visualization techniques are useful for providing first insights into the data.

Dependency analysis consists of finding a model that describes significant dependencies (or associations) between data items or events. Dependencies can be used to predict the value of a data item given information on other data items. Although dependencies can be used for predictive modeling, they are mostly used for understanding. Dependencies can be strict or probabilistic. Appropriate techniques are correlation analysis, regression analysis, association rules, and visualization techniques [3].

3. METHODOLOGY

In this section a summary of CRISP-DM methodology will be presented.

CRISP-DM is a methodology that provides a reference model which emphasizes the life cycle of a data mining project by offering basic support in implementing a proper data mining process.

According to [3], the model consists of sets of tasks described at four levels of abstraction: phase, generic task, specialized task, and process instance.

At the top level, the data mining process is organized into a number of phases; each phase consists of several second-level generic tasks which should be complete and stable as possible.

The CRISP-DM reference model consists of 6 phases of a project with their tasks and the relationships between tasks, such as: business understanding, data understanding, data preparation, modeling, evaluation, deployment.

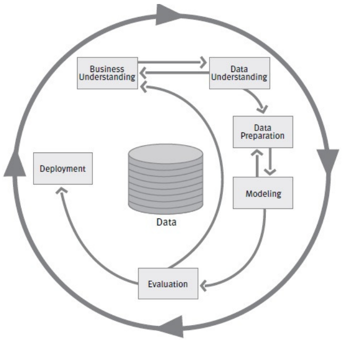


Fig.1. Phases of CRISP-DM reference model

In figure 1 all 6 phases of a DM project are illustrated in a outer circle that symbolize the cyclical nature of DM so it will be necessary to move back and forth between phases. The arrows indicate the most important and frequent dependencies between phases.

According to [3], all 6 phases will be described as it follows and also the DM problem types mentioned in section 2 will be explained.

Business understanding is the initial phase that focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition.

Data understanding it is the second step in creating a DM process, starting with initial data collection and some proceeds supposed to identify data quality problems, discover first insights into the data, and/or detect interesting subsets to form hypotheses regarding hidden information.

Data preparation is focused on constructing the final data set from the initial data set by following certain tasks (table, record, and attribute selection, transformation and cleaning of data) for modeling tools, that should be performed multiple times.

Modeling consists in selecting and applying several modeling techniques with their parameters calibrated to optimal values, returning to the previous phase being necessary in most of the cases because of the specific project requirements.

Evaluation consists in thoroughly evaluate the built process and review the steps executed to create it for detecting any business issue.

Deployment handles with organizing and presenting the knowledge gained for good understanding of the customer.

4. EXPERIMENTAL APROACH

In the first part this section will be presented briefly the practical part described in [4] representing the start point of my work. The second part of this section will be based on the DM process developing. In the last part the results obtained will be seen.

4.1. Questionnaire achievement

An online questionnaire with multiple resources was obtained as a result of the study implementation, psychological measures and it was based on theory of Wickens which allows adding or removing elements according study context. The study implementation consisted in creating a 3D test called “3D delta easy” to illustrate subjects 3D perception capacity by using as a support 3D video game and 3D fovea platform. O session of 30 minutes game playing was made and extra tasks of type object “recognition” were given to each participant. Measures of EEG, of electrocardiogram, eyes tracking and measures of body temperature were taken to illustrate fluctuations differences of brain waves. Time reaction of the subjects were made using two channels visual and audio and also the number of errors was recorded. At the end of the 3D test, the subjects had to complete a questionnaire for defining a sociological standard of the group.

The online questionnaire was acquired in *3 phases* based on previous studies on the specific viewing 3D movies from Cinema and on questionnaires offered by ophthalmologists and opticians to evaluate the impact of 3D: *introductory phase* (question 1 ÷ question 4), *thoroughgoing phase* (question 5 ÷ question 15), *conclusion phase* (question 16 ÷ question 17). In this context screening questions were implemented to eliminate observations outside of the sample. Therefore completion of questionnaire and exploratory discussions were required to include each questions into categories.

This questionnaire was structured in 4 parts: "Partie 1", "Partie 2", "Conclusion", "Talon sociologique".

The part called "Partie 1" contains the first **10** questions revealing important information about 3D Cinema part, including phase1 and a part of phase 2 .

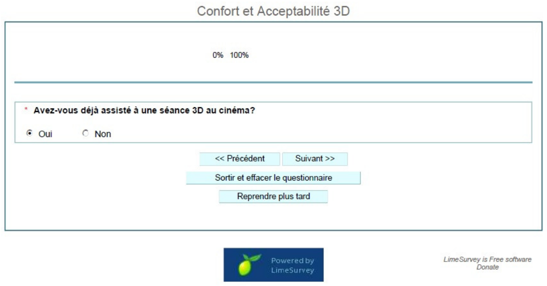
The part called "Partie 2" contains the next **5** questions (11÷ 15) revealing information about other 3D supports (games, etc), it corresponds to the second phase.

The part called "Conclusion" contains the next **3** questions (16÷ 18), it corresponds to the third phase and it offers information about how interested are people for participating to a future 3D Cinema and other 3D supports experience.

The part called "Talon sociologique" contains the next **8** questions and it offers information about people who are having glasses or lens, their age, their genre, time of completion entire questionnaire, time of completion each part and each question.

The remained variables are generated automatically Figure 2 illustrates how is displayed a page when a subject wants to complete the online questionnaire. There 4 pages per total. If the response will be “Oui” it will display the rest of the questions from “Partie1” otherwise it will skip the next 2 pages and it will pass directly to the last page with "Talon sociologique".

All the answers to it were stocked in a document representing the initial data set and called “results-survey_20130626_entetes.xls”.



Confort et Acceptabilité 3D

0% 100%

* Avez-vous déjà assisté à une séance 3D au cinéma?

● Oui ○ Non

Fig.2. A page from online questionnaire

4.2. DM process implementation

In this section is presented a method of developing a DM process based on CRIP-DM methodology and solving a problem, using the results stocked in the initial data set and the SAS Enterprise Guide Tool. Necessary tasks for implementing every phase (data understanding, data preparation, and modeling) will be described.

Data understanding DU involves performing its main tasks: initial data collection; data description, data exploration; data quality verification.

Initial data collection was made by importing the initial dataset in SAS tool which contains 227 observations and 121 variables (75 numeric and 46 character variables).

Describe data consisted in understanding the meaning of each variable; accessibility and validity of attributes verification; type and range variables verification; analysis of data properties; correlations analysis. *The data type* was established based on level of measurement theory presented in [w]. According to this the initial data set contains 40 ordinal variables, 29 binary variables, 4 nominal variables, 10 interval variables, 6 ratio variables, 32 unary variables. For defining *data range* of each variable simple tool options were used: bar charts and summary statistics. *Bar charts* for variables supposed to be binary to detect correct answers (“Oui” or “Non”) and wrong answer (“N/A”) number. *Summary statistics* (max, min, range) were computed only for numeric variables because SAS tool doesn't allow using the summary statistics block (also correlations, summary tables, displaying dispersions, etc) for other types of variable (character). *Data properties analysis* was based on „characterize data” SAS option. It was applied once on entire population, that made it a faster solution comparing with bar charts and summary statistics for every attribute. It consists on computing the frequency count and the percent of total frequency of each binary and text attribute; computing summary statistics of each numeric attribute such as the number of the observations, number of the missing values, sum of the values, min, max, mean, standard mean displaying plots. Based on the previous task from DU a *correlation analysis* was made using „correlations” SAS option and Pearson’s and Spearman’s correlation coeficient which was supposed to illustrate the linear dependency between variables, respectively the non-linear dependency. It were made correlations between ordinal variables; binary variables, ordinal and binary variables, variable 3 "Dèrniere page vue" and variable 5 "Date de la dernière action"; variable 9 "Avez-vous déjà assisté à une sea" and variable 53 "Avez-vous eu la possibilité d'ut"; variables from section 4.E. from "assumptions.doc" file and variables from "Talon sociologique" part.

Data exploration covered tasks as detailed properties analysis of most important attributes; subpopulation characteristics identification; main purpose of DM process clarification (section 2); assumptions verification performing basic analysis (“assumption.doc”). The first task highlited basic variables properties using “summary statistics” and “characterize data” options: median (ordinal); mode (nominal); mean, standard deviation, covariance (interval and ratio) in order to verify the basic hypothesis of the study ***the major disadvantage of 3D is visual tiredness***. In this context were created certain subpopulations: subjects that completed the entire questionnaire; subjects that completed the entire questionnaire and participated to a 3D Cinema session; subjects of genre masculine; subjects of genre feminine; subjects that completed the entire questionnaire and responded with “N/A” at several questions.

Data preparation DP, the main tasks to be performed are: select data, clean data, construct data, integrate data and format data. Based on this, the final data set was achieved in 6 steps. The first 4 steps correspond to 4 filtrations by variable 3 „derniere page vue” , variable 41, variable 79 “sexe” and variable 9, obtained when using “Filter and Sort” SAS option. Another possible filtration could be that wich is made by variable 53, related to the participants who used other 3D supports. Since my work interest was to reveal only the results for 3D Cinema part, all variables from „Partie 2” were removed (53 ÷ 75). *At first step*, to variable 3 it was assigned the value 4 meaning that only those who answered to the questions till the last page were retained. Thus it was obtained a sample with 170 observations and 87 variables. *At the second step* the new data set was sampled thus it was assigned value „N/A” to variable 41 based on noticing that the same percent (8,82%) of errors aparition was obtained for questions with wrong answer (question nr.8, nr.9, nr.11, nr.16) after first filtration so any variable correponding to those questions could be chosen. The variable 41 indicates if the subjects felt a loss of quality during 3D cinema session or not. The sample obtained had 15 observations and 87 variables. *The third filtration* was divided in 2 filtrations: one with assignment of “masculin” value to variable 79 and another one with assignment of “feminin”. After each case different samples were obtained with the same number of variables but different number of observations, 99 respectively, 71 and the frequency of error apparition was reduced (4.04%, respectively 15.49). *At the fourth step* it was assigned value „Oui” to variable 9, related to the particpants who assisted to a 3D Cinema session or not and the begining of „Partie 1”. Therefore a new sampling was made from the first sample obtained. *The fifth step* was done in parralel with every filtrations. Each dataset obtained was described by characterize data blocks and several variables were eliminated.

Based on this steps a *data select* was done. In this case the most important informations were chosen based on the following variables selection cryterias:

variable is unary; reduced number of answers comparing with number of observations; the same properties, but different name; variables from „Partie 2”, gives irrelevant information for 3D perception; hypothesis from „assumption.doc”. *The last step* implied changing data type so all character variables from dataset obtained at step 4 were converted into numeric, using „Query Bilder” SAS option. A final sample was achieved containing 155 observations and 83 variables, dataset supposed to be introduced in modeling tool.

4.3. Experimental results

Based on [3], [4], and on approach method described in previous section, a DM process was implemented properly so it was obtained a schematic representation of it consisting in several blocks: import data block, filter and sort blocks, characterize data blocks, query builder and correlations block. It wasn't possible to capture the entire process so I aded below a diagram with a part of it consisting in different blocks for realizing the filtrations step 2 and for describing the file information (figure 3).

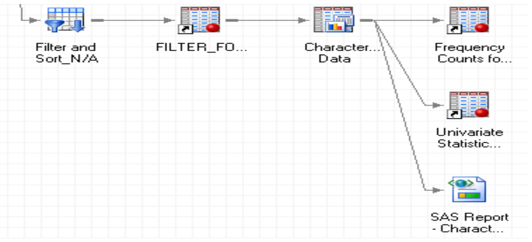
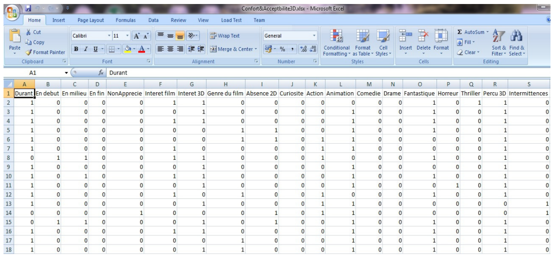


Fig.3. SAS diagram for second filtration

After each data preparation step new results were acquired which were stocked in a different excel document each time, using export option of SAS tool. This fact will be represented in the following figure (nr.4) including a part of the final dataset achieved.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Durant	En debut	En milieu	En fin	NonApprecie	Interet film	Interet 3D	Genre du film	Absence 2D	Curiosite	Action	Animation	Comedie	Drame	Fantastique	Horreur	Thriller	Percu 3D	Intermittences
2	1	0	0	0	0	1	1	0	0	0	0	0	0	0	1	0	1	1	0
3	1	0	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
4	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0
5	1	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	1	0
6	1	0	0	0	0	0	0	1	1	0	1	1	0	0	0	0	0	1	0
7	1	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	0	1	0
8	0	1	1	0	0	1	1	0	0	0	1	0	0	0	1	0	0	1	0
9	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0
10	1	0	1	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0
11	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0
12	1	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	0	1	0
13	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0	0	0	1
14	0	0	0	0	1	0	0	0	1	0	1	1	0	0	0	0	0	0	1
15	0	1	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0	1	0
16	1	0	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1	0
17	1	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	1	0
18	1	0	0	0	0	0	1	1	0	0	0	1	0	0	1	0	0	1	0

Fig. 4. Results of final DM process

“Treating” dependency analysis problems led to only 3 higher correlations between 5 variables: *‘’Une vision double’’,‘’Un rétrécissement’’,‘’Des vertiges’’, ‘’Une sensation de désorientation’’, ‘’Des nausées’’, the values are in the range (0.7÷0.9).* ‘’Des vertiges’’ is correlated with ‘’ Une sensation de désorientation’’ having the value 0.784 for Pearson’s correlation coefficient and 0.806 for Spearman’s correlation coefficient, and also with ‘’ Des nausées’’ having 0. 801 value for Spearman correlation coefficient. ‘’Une vision double’’ is correlated with ‘’Un rétrécissement’’ having 0.724 value for Spearman’s correlation coefficient. In figure 5 an example of highly correlated variables (“plus realism” and “immersion”) is illustrated; both variables represent gains of 3D Cinema viewing, meaning realism increasing, respectively, increasing of immersion.

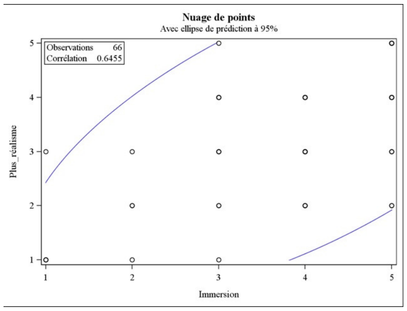


Fig.5. Correlation analysis

The study required formulating and verifying of some assumptions for all the questionnaires parts. Based on subjects answers and on process development approach, the results in comparison with the ones from [4] will be seen in the table below.

Nr. Assumption	Name variable	Type variables	V1	V2
1	Assiste	binary	91.8 %	
2	Durant	binary	58.71 %	70.68 %
3	Curiosite, Interet film	Binary binary	36.77 %	44.82 %
4	Animation	binary	64.51 %	75.8 %
5	Percu 3D	binary	73.54 %	89.65 %
6	Difficultes	binary	63.22 %	59.65 %
7	Perte qualite	binary	66.45 %	31%
8	relief de certaines, effects speciaux, + imersion	ordinal ordinal ordinal	4 4 4	- 3.7 3.63
9	Hausse du prix	ordinal	5	4.37
10	Fatique visuelle	ordinal	4	3.5

Table1. Results of questionnaire analysis

The statements concerning 3D Cinema perception are: 1. participants went to a 3D Cinema session before; 2. the 3D Cinema session was totally appreciated by participants; 3. the reasons for choosing a 3D Cinema session is curiosity and interest on movie; 4. the subject's favorite 3D Cinema viewing type is animation; 5. 3D viewing during session was perceived by participants; 6. adapting to 3D without difficulties was fulfilled; 7. a loss of quality of 3D during the session was not felt by participants, 8. the advantages retrieved by the 3D film are certain scenes of actions highlighting, special effects highlighting and an increased immersion offered; 9. the disadvantage retrieved by the 3D film was increasing 3D session price; 10. the negative sensation given by the 3D Cinema viewing was visual tiredness.

Differences between my work and [4], the statement 3 from [4] was different because the higher percent of "Oui" answers was obtained for "par interet pour la 3D" variable and for statements 8,9, 10 means were computed. I computed median because according to [8] is the most adequate measure of central tendency.

5. REMARKS

5.1. Abbreviations and acronyms

CRISP-DM- = Cross-Industry Standard Process for Data Mining

SAS- Statistical Analysis System

DP = Data Preprocessing

DU = Data Understanding

Method = set of procedures used for accomplish a task

Methodology = instance of a process

Modeling = step from the CRISP-DM process, that succeeds the Data Preprocessing step.

Phase = each distinct step from a process

Process = sequence of tasks that led to a result

User = in this paper, the term is used as a DM end-user

Level of measurement= type of data that arise in the theory of scales types required in statistics and quantitative research

Likert scale= psychometric scale involved in the research of questionnaires

Ordinal scale= type of data that has order, but the interval between measurements is not meaningful

Nominal scale= type of data having no order and thus it only gives names or labels to various categories

Interval scale= contains type of data having meaningful interval measurements but there is no true starting point

Ratio scale= contains type of data having the highest level of measurement, there is a starting point

5.2. Conclusions and Further recommendations

The target for this questionnaire was to highlight some negative and positive factors with a big influence over the 3D Cinema perception "felt" by the subjects. In this case information about the personal experience of 3D viewing, 3D perception and adaptation, the gain and the disadvantage were obtained. A DM process was implemented focusing on second and third step described in CRISP-DM methodology thus the questionnaire analysis was succeeded, and on resolving the 2 DM problems. Data description and summarization problem confirmed 3D Cinema main disadvantage that visual tiredness is the only criteria for physic perturbation though correlation analysis revealed the fact that the variable corresponding to visual tiredness is not highly correlated to any variables. Only moderately correlations were obtained for variable "Fatigue visuelle" with 4 other variables: "Une migraine" (0.607), "Fatigue" (0.572), "Des picotements" (0.577), "Une vision troublée" (0.525).

For further work the process obtained could be used for modeling using an adequate technique such as PCA, regression analysis, etc and information from different studies involved in recommending the most helpful methodologies and tools. Finding the best solutions for achieving the proper model will improve considerably the life cycle of DM process. At modeling level, the hierarchical model proposed by Felleman and Van Essen (1991) could be taken into consideration, supposing the process operates on successive levels over the visual vision, also an analysis of variances ANOVA could be made to compare the groups means, supposing the most relevant variable has a normal probability distribution.

REFERENCES

[1] Jiawei Han, Micheline Kamber, *Data Mining Concepts and Technique*. Illinois : University of Illinois at Urbana-Champaign

[2] Hand, David J. *Principles of Data Mining*.

[3] Pete Chapman, Julian Clinton, Randy Kerber,Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth,*CRISP-DM 1.0*. USA : SPSS Inc, 2000.

[4] Pfleger, Delphine, *Etude de l'impact de la 3D en jeu vidéo sur la charge cognitive: un test réalisé avec la version P.C du F.P.S "Just For Cause"*. Brest : Institut Mines-Télecom, Université Télécom-Bretagne, 2011-2012

[5] Mateyaschuk J. : *The 1999 National IT Salary Survey: Pay up, Information Week*, 1999

[6] Emerging *Technologies That Will Change the World*, Technology review, Published by MIT, 2001.

[7] Fayyad U., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. : *Advances in Knowledge Discovery and Data Mining, MIT Press*, 1996.

[8] https://en.wikipedia.org/wiki/Level_of_measurement

[9] https://en.wikipedia.org/wiki/Central_tendency

[10] https://en.wikipedia.org/wiki/Psychometrics.

[11] http://www.andrews.edu/~calkins/math/edrm611/edrm05.html