Data Mining in a System of a Candidate Selection Process

Renata Maria Abrantes Baracho
School of Information Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
renatabaracho@ufmg.br

Márcio Teodoro Dias
School of Information Science
Federal University of Minas Gerais
Belo Horizonte, Brazil
marcio.dias@ifmg.edu.br

Abstract – **Data mining and information retrieval has been research areas increasingly important in the information age where fast, accurate access to information becomes an essential factor in decision-making in various fields of knowledge. The main goal of this study is to analyze the data of candidates to selection processes of Federal Institute from Minas Gerais - IFMG in search of patterns that are not explicit, with the support of a free tool for data mining. This study comes to the research question of what are the relevant information about the profile of candidates who qualify in the selection process of IFMG. Through data mining and analysis of results is possible to define the profile of applicants and the features that influence the selection process. This article presents a study on the database in the candidates selection process of IFMG. One of the requirements of the research includes the study of data mining tools available on the Internet and a test applies in a real situation. Then the article presents an experience of using a free tool for Data Mining. The software chosen for the tests, called Weka, was developed by the University of Waikato, New Zealand. It makes searching for patterns in database of system used to the management of selection processes of the Federal Institute from Minas Gerais. The methodology includes access to the database IFMG, with 9952 records of candidates for the selection process. Five categories were defined as relevant for verification and analysis. Categories include sex, marital status, age, waiver of registration fee and have disabilities or not. We used the attributes that have 100 or more occurrences in the database. The results show that the pattern that occurred more frequently among those classified in absolute numbers was that of male candidates, who had no exemption from the registration fee, less than 30 years old and single, with 166 occurrences. Among all candidates, 718 have these characteristics, so 9.66% of the candidates that fit this pattern was classified. The pattern with the highest percentage of occurrence were male candidates, who had no exemption from the registration fee, less than 30 years old and married. Of the 301 candidates with all these features that participate in the processes of selection, 31 were classified. It represents 10.33% of 301 candidates evaluated. There were 9952 candidates of which 711 were classified, the representing 7.14%. Among the 291 candidates with exemption from the registration fee, only 3 were qualified which represents 1.03%. This result indicates that lower-income applicants had lower performance. The associative algorithm provides a list of up to 100 rules, which were presented in this article the top 15 with a combination of categories, sex, marital status, age, registration fee and deficiencies. The survey allows the discovery of patterns in database performance with different combinations of the categories.**

Information retrieval, information system, database, data mining, Weka, selection process

I. INTRODUCTION

Currently there is growing research on data mining and information retrieval to support the decision-making organizations. There is large databases that can generate direct subsidies for decisions. In this context, the research seeks to use data mining tools to analyze the database. The objective of this research is to analyze the data of applicants to a selection system to define the patterns the profile of classified candidates. This study addresses the research question of what are the relevant information about the backgrounds of the candidates who qualify in the selection process of Federal Institute from Minas Gerais - IFMG.

The Federal Institute from Minas Gerais - IFMG is a federal autarky, created by Law nº 11.892 [1], enacted on December 29, 2008, by Luiz Inácio Lula da Silva who was the President of Brazil. In the article 2 of the Law, Federal Institute are defined as: Institutions of higher, basic and professional education specialized in providing professional and technological education in different modalities of teaching, based on the combination of technical and technology knowledge with their teaching practices, under this Law.

IFMG is now comprised of ten units in different cities of the state of Minas Gerais: Bambui, Betim, Congonhas, Formiga, Governador Valadares, Ouro Branco, Ouro Preto, Ribeirão das Neves, São João Evangelista, Sabará. These units are connected to a central government in Belo Horizonte, state capital of Minas Gerais.

People who work in IFMG are admitted through the selection processes that are performed by the institution itself. This processes can be made to hire permanent or temporary employees, or to hire trainees.

Each Process has one or some requirements, defined in a public notice, to evaluate whether a candidate is able to occupy a post. The candidates who can fulfill these requirements, that are generally related to education and performance in a test, are classified. Among these ones, the best classified candidates will be called to fill the positions, according to the number of vacancies. Those candidates who do not fulfill the requirements specified in the notice are considered disqualified from the selection process. According to the number of places offered in the Notice, a list is generated for ratification of process. Candidates present in the list can still be called to fill the post.

Management of routines related the selection process in IFMG is done with the support of an internally created software, called Recepta. This system was developed with PHP programming language and PostgreSQL Data Base

Management System. Recepta was registered at the National Institute of Industrial Property from Brazil (INPI) and is also available as free software.

In database of the system, which was implemented in March 2010, are stored data relating to registration, candidates' personal data, and whether the candidate is or is not in the list of selection process.

The main objective of this work is to search the database of the selection process to find patterns among the personal characteristics of candidates and their final results, achieving this result with the use of a free tool for data mining.

II. LITERATURE REVIEW

Whereas the text mining involves moving information retrieval. In this research the concept of information retrieval is based on considerations of Lancaster [2] and Hjørland [3].

Lancaster [2] describes retrieval of information how the process of searching within in a collection of documents to identify those dealing with a particular subject.

Hjorland [3] define that one of the problems in retrieval is the definition of access points to a database of electronic documents containing text, images and different media.

Knowledge Discovery in Databases (KDD), is a concept used to describe the exploration of implicit information in large volumes of data, whose technology emerged by necessity and difficulty of exploring large databases (Bigolin, Bogorny, Alvares [4]).

The process KDD can also be defined as "the non-trivial process of identifying patterns that are valid, new, potentially useful and, finally, comprehensible in data" (Fayyad, Piatetsky–Shapiro, Smyth [5]).

Knowledge Discovery in Databases is the result of a process that goes through three main stages: preprocessing, data mining, and the post-processing (Bigolin, Bogorny, Alvares [4]).

In step of preprocessing is necessary to prepare the data so that a mining tool can extract from it the implicit and potentially useful information. This preparation involves the following tasks, in accordance with Bigolin, Bogorny, Alvares [4]:

- determination of the goals of discovery: the problem is clearly defined;
- data cleaning: elimination of noise and inconsistency from data;
- data integration: data from multiple sources can be combined;
- data selection: relevant data for data mining are identified and grouped together, generating a sample of the database;
- data transformation: conversion of data to a format interpretable by the data mining tools.

This stage can demand up to 80% of the total processing time, due to difficulties in integrating heterogeneous databases (Carvalho apud. Manilla [6]).

Data mining is "a step in KDD process that involves applying data analysis and discovery algorithms that, under acceptable limitations of computational efficiency produces a particular enumeration of patterns about the data" (Fayyad, Piatetsky–Shapiro, Smyth [5]).

For EDELSTEIN [7], "Data mining tools find patterns in the data and infer rules from them. Those patterns and rules can be used to guide decision-making and forecast the effect of those decisions".

Post-processing is the step of submission and evaluation of the patterns found in mining process, that are responsible for the identification and analysis of relevant patterns, as well as the definition of the form with the extracted information will be presented. The following figure shows the steps of the KDD process, Figure 1.

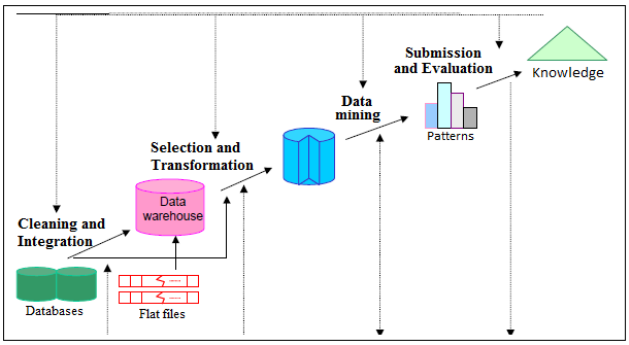

Figure 1. Steps of the KDD process.
Source: BIGOLIN, BOGORNY, ALVARES (2003)

III. METHODOLOGY

The methodology is developed as the goal of finding relevant information on the profile of candidate that can be classified in the selection processes of IFMG. For this work is not important whether the candidate was or was not called to occupy the position to which applied. It depends on the number of vacancies offered by the institution. The most important is whether it was classified.

In the second part was made a search on the Internet to find a free software of data mining to be used. A software to receive an input data and perform data mining for patterns not explicit. During the study found information about several tools such as KDB2000, KNIME, MDR, Orange, Tanagra and Weka. After analyzing the software Weka was chosen because it is widely used, many options for implementing data mining algorithms and be well documented. Their characteristics are described in detail in (Witten and Frank [8]), whose authors are responsible for the implementation of the tool.

The next step was to define what characteristics of the candidates would be considered to verify the association between these characteristics and the results of the candidates in selection processes. After analyzing the characteristics were defined five attributes: sex, marital status, age, the exemption from registration fees and disability. Low-income candidates may request exemption from registration fee in a selection process, being that this request may or may not be accepted.

The attribute in this case seeks to assess the performance of low-income applicants.

The next stage was to define how the data would be prepared for mining. It was decided to use attributes independent of each other and that had 100 or more records in database, because a small sample could lead to errors.

IV. DATA PREPARATION

The data preparation began with cleaning the data, to extract only those who contribute to the aims of research.

During the development of this research had contests in progress but without result, because of this the data subsequent to October 2011 were not extracted.

As described, the processes of selection IFMG include: Tender for Admission, Simplified Process for Substitute Teacher Selection and Selective Processes of Trainees.

The Selection Process of Trainees has a different audience of others, so data for these candidates have different characteristics, especially in relation to age, marital status and income. Data from this type of process were ignored. For this research were considered candidates for Tender for Admission, Simplified Process Selection.

Applications are accepted when the applicant makes the payment of registration fee or is entitled to exemption from this fee. In this study we considered the applications accepted. The data of candidates that did not make tests does not interfere in final result.

In selecting the data are considered those with 100 or more occurrences. Then went to look for attributes that were less than 100 records which will be disregarded. Therefore, the data related to disability were not considered, only 74 candidates were declared to have any special needs. Evaluating the marital status, it was found that only 28 candidates were widowed, so they were grouped in the same category of divorced.

The last step in the data preparation was the elimination of any records that contained incorrect information. It was realized that several candidates make mistakes to fill the date of birth during registration, for example, some candidates whose year of birth was filled in with the value 2011. Were considered only those who have filled the year of birth as 1993 or less, whereas the minimum age to occupy public post in Brazil is 18 years.

Date of birth itself is not an interesting value in search of patterns, because of the large number of distinct occurrences. The candidates were grouped into three age groups according the date of birth: less than 30 years old; between 30 and 45 years old; and more than 45 years old.

After making data preparation including the definition of the five attributes: sex, marital status, age, the exemption from registration fees and disability, these data was exported from the database in CSV format. After the cleaning process, were obtained 9952 records.

V. DATA MINING AND RESULTS

According to research, to data mining software was used Weka (Waikato Environment for Knowledge Analysis), version 3.6. The code was first developed in 1993 at the University of Weikato in New Zealand. It was built with the Java language, supports multiple file extensions as input, one of which is the CSV format.

In the dataset, one of the attributes is considered the class attribute, while the others are attributes predictive. In this research, the class attribute is what shows if the candidate is classified. This attribute is populated with the values yes or no. All other predictive attributes will be evaluated in relation to the class attribute.

In Weka software, the last attribute is considered by default as the class attribute, but to change this order is possible.

In the home screen of the program is the option to open a file with the data to be mined. In this work it was used a file in the CSV format, exported from Recepta system database. After loading the file, Weka's screen displays bar graphics that allow comparing each predictive attribute in relation to the class attribute, as shown in the Figure 2.

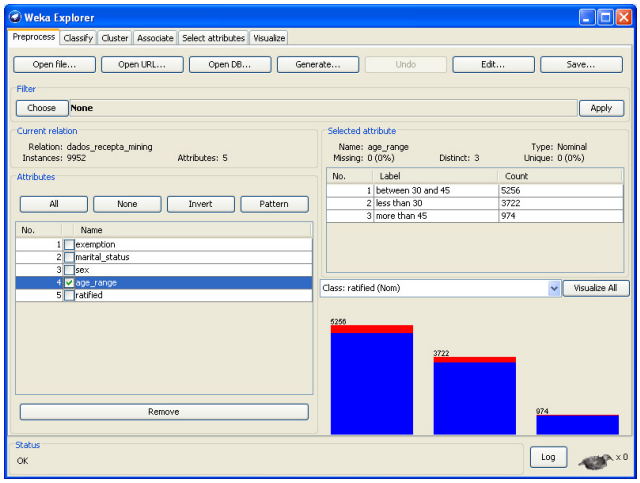

Figure 2. Comparison Graphic of the class attribute with a predictive attribute.

Next, several algorithms available in the software have been tested. Four of them brought results most relevant to the aims of research.

Inside of the classification algorithms, available in tab "Classify", there is a folder called Trees, which contains some algorithms that build decision trees. Below is the result brought by the algorithm "RandomTree", with highlights in areas considered most important for this work (Table 1).

Table 1 – Algorithm "RandomTree"

```
RandomTree
==========

exemption = no
|   sex = m
|   |   age_range = between 30 and 45
|   |   |   marital_status = married : no (898/83)
|   |   |   marital_status = single : no (808/73)
|   |   |   marital_status = divorced or widower : no (68/4)
```

```
|  |  |  marital_status = other : no (55/1)
|  |  age_range = less than 30
|  |  |  marital_status = married : no (300/31)
|  |  |  marital_status = single : no (1718/166)
|  |  |  marital_status = divorced or widower : no (3/1)
|  |  |  marital_status = other : no (18/2)
|  |  age_range = more than 45
|  |  |  marital_status = married : no (205/10)
|  |  |  marital_status = single : no (55/2)
|  |  |  marital_status = divorced or widower : no (31/2)
|  |  |  marital_status = other : no (17/1)
|  sex = f
|  |  age_range = between 30 and 45
|  |  |  marital_status = married : no (1113/59)
|  |  |  marital_status = single : no (985/49)
|  |  |  marital_status = divorced or widower : no
(206/13)
|  |  |  marital_status = other : no (75/3)
|  |  age_range = less than 30
|  |  |  marital_status = married : no (567/45)
|  |  |  marital_status = single : no (2083/140)
|  |  |  marital_status = divorced or widower : no (25/0)
|  |  |  marital_status = other : no (35/3)
|  |  age_range = more than 45
|  |  |  marital_status = married : no (183/8)
|  |  |  marital_status = single : no (110/6)
|  |  |  marital_status = divorced or widower : no (84/3)
|  |  |  marital_status = other : no (19/2)
exemption = yes
|  sex = m
|  |  age_range = between 30 and 45
|  |  |  marital_status = married : no (11/1)
|  |  |  marital_status = single : no (14/0)
|  |  |  marital_status = divorced or widower : no (2/0)
|  |  |  marital_status = other : no (1/0)
|  |  age_range = less than 30
|  |  |  marital_status = married : no (9/1)
|  |  |  marital_status = single : no (73/1)
|  |  |  marital_status = divorced or widower : no (0/0)
|  |  |  marital_status = other : no (1/0)
|  |  age_range = more than 45 : no (3/0)
|  sex = f : no (177/0)

Size of the tree : 48

Correctly Classified Instances     9241          92.8557 %
Incorrectly Classified Instances    711           7.1443 %
```

The algorithm "RamdomTree" constructs a decision tree by filling out the class attribute with the value that appeared most frequently, informing in the end of each line how many instances are not in accordance with this model. For example, in the tenth row of result, the model predicts that if the candidate is male, has no exemption from the registration fee, has less than 30 years of age and is single, the trend is that isn't qualified, based on what happened in most situations However, the result shows how many times the test was performed (1718) and how often the class attribute was different from that described in the model (166). That is, the 1718 candidates with these characteristics, only 166 were classified. In this tree in all cases the class attribute will be considered as "no", because most candidates could not (the total percentage of qualified applicants was 7.14%). This Randon tree allows check which the numeric pattern that was repeated over. In 166 cases a male candidate, who had no exemption from the registration fee, has less than 30 years old and is single managed to qualify. Of the total 9,952 candidates analyzed, 9241 were not classified. Among 711 qualified candidates, 166 candidates has pattern 1 (male candidate, the exemption from the registration fee, has less than 30 years old, single). In numerical terms is the pattern 1 that appears most frequently. In percentage terms (10.33%) of the 300 candidates 31 were classified presenting the pattern 2 (male candidate, the exemption from the registration fee, has less than 30 years old, married), managed to qualify. That is, the pattern 2 (10.33%) in percentage terms represents a superior result to the pattern 1 (7.14%).

In the end, the algorithm shows the Correctly Classified Instances with 9241 events (92.8557%) and Incorrectly Classified Instances with 711 occurrences (7.1443%).The model defines the classifier attribute as no. The classifier attribute tends to be no. When the classifier attribute is yes the program describes how misclassification because it did not follow the trend indicated by the model. The percentage of Incorrectly Classified Instances refers to cases where the class attribute is yes. The misclassification shows how many times the model was not followed.

Another classification algorithm was tested "AdaBoostM1", which is in the folder "Meta". The first rows of the result are transcribed below (Table 2).

Table 2 – Algorithm "AdaBoostM1"

```
AdaBoostM1: Base classifiers and their weights:

Decision Stump

Classifications

sex = m : no
sex != m : no
sex is missing : no

Class distributions

sex = m
no                              yes
0.9116550116550116      0.08834498834498834
sex != m
no                              yes
0.9415400918403392      0.0584599081596609
sex is missing
no                              yes
0.9286575562700965      0.07134244372990353
```

This algorithm compared each predictive attribute with the class attribute, informing the percentage of occurrence of each value of class attribute. In our test, it presents information quite

interesting in sex attribute. The last line of the example presents the total, without considering the sex of the candidate. As stated previously, in general, about 7.14% of the candidates were classified in selection processes. Approximately 8.83% of candidates classified are men and 5.84% are women.

The last classification algorithm used was the "NaiveBayesSimple", located in folder "Bayes". Below is described the result of the test (Table 3).

Table 3 – Algorithm "NaiveBayesSimple"

Naive Bayes (simple)

Class no: P(C) = 0.92857143

Attribute exemption
no yes
0.96873648 0.03126352

Attribute marital_status
married single divorced or widower other
0.33290071 0.59928618 0.04391088 0.02390223

Attribute sex
m f
0.42319342 0.57680658

Attribute age_range
between 30 and 45 less than 30 more than 45
0.43374797 0.49302326 0.07322877

Class yes: P(C) = 0.07142857

Attribute exemption
no yes
0.99438202 0.00561798

Attribute marital_status
married single divorced or widower other
0.33473389 0.61344538 0.03361345 0.01820728

Attribute sex
m f
0.53370787 0.46629213

Attribute age_range
between 30 and 45 less than 30 more than 45
0.40252454 0.5483871 0.04908836

This algorithm first separated the class attribute, and then compared it with each attribute predictive. The class attribute refers to the exemption from the registration fee. The table shows those candidates who failed the exemption from the registration fee (0.92857143) and those who got the exemption (0.07142857). This result provides an interesting analysis, when comparing the percentages in each attribute. According to Table 3, the age range has attribute values near between 30 and 45 (0.43374797) and less than 30 (0.49302326) representing a majority for the "Class no". The "Class yes" also presents values close between 30 and 45 (0.40252454) and less than 30 (0.5483871) The marital status attribute to the "Class no" has values close to married (0.33290071) and single (0.59928618) and the "Class yes" features were similar between married (0.33473389) and single (0.61344538). The results show that age and marital status did not significantly affect the probability of classification, since the values "yes" and "no" are very close. However, in the attributes sex and exemption from the registration fee this difference is large. There is 3.12% of candidates with exemption from the registration fee among those who were disqualified and 0.56% among those who were qualified. In other words, candidates with lower income had performance far below the overall average. Among the 291 candidates with exemption from the registration fee, only 3 were qualified which represents 1.03%.

In relation to associative algorithms, the most relevant among the tested was "PreditiveApriori", which lists a maximum of 100 found best rules. Below are listed the 15 main rules (Table 4).

Table 4 – Algorithm "PreditiveApriori" - 15 main rules

Best rules found:

1. age_range=between 30 and 45 ratified=yes 286 ==> exemption=no 285 acc:(0.99491)
2. marital_status=single sex=f ratified=yes 195 ==> exemption=no 195 acc:(0.99488)
3. sex=f age_range=less than 30 ratified=yes 188 ==> exemption=no 188 acc:(0.99486)
4. exemption=yes sex=f 177 ==> ratified=no 177 acc:(0.99483)
5. marital_status=single sex=m ratified=yes 242 ==> exemption=no 241 acc:(0.99477)
6. exemption=yes marital_status=single 218 ==> ratified=no 217 acc:(0.99463)
7. marital_status=married sex=f ratified=yes 112 ==> exemption=no 112 acc:(0.99428)
8. marital_status=married sex=f age_range=more than 45 184 ==> exemption=no 183 acc:(0.99424)
9. marital_status=single age_range=more than 45 166 ==> exemption=no 165 acc:(0.99389)
10. marital_status=married ratified=yes 238 ==> exemption=no 236 acc:(0.99378)
11. sex=m age_range=more than 45 ratified=no 296 ==> exemption=no 293 acc:(0.99318)
12. exemption=yes 291 ==> ratified=no 288 acc:(0.993)
13. marital_status=married sex=m age_range=more than 45 207 ==> exemption=no 205 acc:(0.99274)
14. sex=m age_range=less than 30 ratified=yes 202 ==> exemption=no 200 acc:(0.99251)
15. marital_status=other age_range=more than 45 36 ==> exemption=no 36 acc:(0.98762)
exemption=no sex=f 81 acc:(0.7181)

In the results above are described associations that appeared most often, with the percentage corresponding to that situation. Attribute of exemption from the registration fee appears several times, since the vast majority of candidates have not obtained this exemption. The rule number 4, that combined (sex=female and exemption=yes) the two attributes that had lower performance. In the analyzed data, were 177 female candidates who had exemption from registration fee, i.e., has low income, and none of them is in the list of qualified. The rule number 6 shows us that low-income single candidates (marital_status=single and exemption=yes) had very low performance, because of 218 candidates in this situation, 217 were disqualified.

The software also lets you select which attributes will be considered in the test. In the example below were taken into account only the attributes sex and marital status, in association algorithm "PreditiveApriori" (Table 5).

Table 5 – Algorithm "PreditiveApriori" - attributes sex and marital status

Best rules found:

1. marital_status=other 232 ==> sex=f 140 acc:(0.59134)
2. marital_status=divorced or widower 428 ==> sex=f 323 acc:(0.45237)
3. sex=m 4290 ==> marital_status=single 2668 acc:(0.42982)
4. marital_status=other 232 ==> sex=m 92 acc:(0.41665)
5. sex=f 5662 ==> marital_status=single 3309 acc:(0.39666)
6. marital_status=married 3315 ==> sex=f 1890 acc:(0.37453)
7. marital_status=single 5977 ==> sex=f 3309 acc:(0.34185)
8. marital_status=single 5977 ==> sex=m 2668 acc:(0.31202)
9. marital_status=married 3315 ==> sex=m 1425 acc:(0.28479)
10. sex=f 5662 ==> marital_status=married 1890 acc:(0.2438)
11. sex=m 4290 ==> marital_status=married 1425 acc:(0.24006)
12. marital_status=divorced or widower 428 ==> sex=m 105 acc:(0.14226)
13. sex=f 5662 ==> marital_status=divorced or widower 323 acc:(0.02419)
14. sex=m 4290 ==> marital_status=divorced or widower 105 acc:(0.01379)
15. sex=f 5662 ==> marital_status=other 140 acc:(0.01376)
16. sex=m 4290 ==> marital_status=other 92 acc:(0.01373)

Assessing the Rule 2, we see that among those divorced or widowed women who participated in the selection process, the number of women exceeds the number of men, because they are 323 in total of 428, or 75.47%, this particular attribute .

Table 5 allows better understand the candidates profile who participated in the selection process.

Rule 3 shows that of 4,290 men surveyed, 2,668 are single which represent 62.19% of the total. Rule 5 shows that among the 5,662 women, 3,309 are single, thus 58.44%. Most candidates are single, this index shows that the number of single men more than single women. Rules 10 and 11 show the number of married men and women. Married men are 1225 (33.21%) and married women are 1890 (33.38%). That is, the percentage of married people is almost the same considering men and women.

VI. CONCLUSION

The use of the software allowed the discovery of patterns in the database, noting that the performance was different according to selection attributes. Considering the sex attribute, it was revealed that the men fared better than women. The poor candidates had a result less than the other candidates. Another interesting pattern was discovered in the relationship between marital status and sex. The evaluation results showed that most people widowed or divorced who participated in the selection processes are female. In relation to the married people, the percentage of men and women is almost the same.The software Weka has many mining algorithms, which may be relevant to other databases. The usability of the tool is good, and is compatible with a widely used file standard, the csv standard. As a suggestion for a future work, evaluate the relationship between other factors and classification, as the relationship between physical disabilities and classification or between physical disabilities and income.

REFERENCES

[1] Brasil. Lei Nº 11.892, DE 29 DE DEZEMBRO DE 2008. Disponível em:<http://www.planalto.gov.br/ccivil_03/_Ato2007-2010/2008/Lei/L11892.htm>. Acesso em 26/11/2011.

[2] F. LANCASTER, A. WARNER, "Information retrieval today". Arlington: Information Resources Press, 1993.

[3] B. HJORLAND, "The concept of subject in information science. Journal of Documentation", [S. l.], 19928. Lancaster, F.W., Information retrieval today. 47p. (1993).

[4] BIGOLIN, N. M. ; BOGORNY, V. ; ALVARES, L. O. "Uma Linguagem de Consulta para Mineração de Dados em Banco de Dados Geográficos Orientado a Objetos". In: XXIX Conferencia Latinoamericana, 2003, La Paz. XXIX Conferencia Latinoamericana, 2003. v. 1. p. 23-35.

[5] FAYYAD, U.M.,PIATETSKY–SHAPIRO, G, SMYTH, P. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". Proceeding of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, august, 1996.

[6] CARVALHO, D. R.; LOPES, L. R. ; BUENO, M. ; ALVES NETO, W. "Ferramenta de Pré e Pós-processamento para Data Mining". In: XII Seminário de Computação, 2003, Blumenau. XII Seminário de Computação, 2003.

[7] EDELSTEIN, H. " Technology how to: Mining Data Warehouses". Information Week, 8 Jan., 1996.
http://www.informationweek.com/561/61oldat.htm

[8] WITTEN, I. H.; FRANK, E. "Data Mining: Practical Machine Learning Tools and Techniques", 2nd edition, San Francisco, Morgan Kaufmann Publishers, 2005.