Towards the identification of Data Mining techniques in the Brazilian governmental context

Francisco José Nardi Filho
Institute of Computing
University of Campinas
Campinas, Sao Paulo, Brazil
francisco.filho@students.ic.unicamp.br

Aracele Garcia de Oliveira Fassbinder
Federal Institute of Education, Science and Technology of the South of Minas Gerais - Campus Muzambinho
Muzambinho, Minas Gerais, Brazil
aracele.garcia@ifsuldeminas.edu.br

Ramon Gustavo Teodoro Marques da Silva
Federal Institute of Education, Science and Technology of the South of Minas Gerais - Campus Muzambinho
Muzambinho, Minas Gerais, Brazil
aracele.garcia@muz.ifsuldeminas.edu.br

Abstract—There are many techniques in the data mining field. However, it is not easy to find in the literature a definition of which technique works better within each application context. Thus, this paper explores the resources of a systematic review to look for the most common data mining techniques within the governmental context, especially working with socioeconomic and educational data sets. Papers were selected, analyzed, and compared with this purpose. The focus is the development of a preliminary theoretical framework in which a public administration can take better investment decisions according to their weaknesses or strengths in those areas. Finally, the review process was able to satisfactorily set the technique and other convenient factors for the forthcoming work.

Keywords—government decision-making; socioeconomic and educational performances; systematic review.

I. INTRODUCTION

The large amount of data that has been generated and stored by companies, the highly competition among them, the availability of robust software for data analysis activities, and the expansion of computational capacity led to the development of the data mining area [1]. In a general way, the data mining process extracts information for decision making and can be applied, virtually, in any field of knowledge. When applied to industries, research or companies, it provides results that help to analyze trends or even dictating the improvement of processes.

When it is used in the governmental context, our research focus, it can give better insights about investment decisions in areas like education, health, labor, and redistribution of wealth, for example, which are lacking incentive, effectiveness or compliance with the administration goals.

According to [6], these insights come from filtered and treated information at operational and management level. This information will provide quick and accurate analysis for the development of administration's strategies, goals, and decision-making. That is, once again the data mining process will lead to this knowledge discovery.

There exist many kinds of data mining techniques, some of which are: classification, regression, clustering, association rules, time series analysis, among others [3]. In order to select the appropriate technique and run the data mining process, the first concern has to be the understanding of the problem to be solved. Every problem is related to the choice of one or more different techniques.

Particularly, this study aims to discover the optimal techniques for socioeconomic and educational data mining in the governmental context. Literature is not generous on this specific issue, since the closest related work, such as [2] and [5], either does not cover the entire area of this study or does not explain the reason for choosing the used data mining techniques.

In order to achieve the goal of this work, first, we created appropriate search arguments for the selection of related studies. Last, we provided a statistical and comparative analysis of the results. This led to adequate techniques, tools and other convenient factors, as application contexts.

The rest of the paper is organized as follows. Section II will present the methodology, including the research strategy, selection of studies, and data extraction. Section III will provide a summary of the results: tools, techniques, and application context. Section IV will recapitulate the main idea behind this systematic review and reinforce the major outcomes of this study.

II. METHODOLOGY

A. Research Strategy

This systematic review follows a process defined by [5] and is based on a survey starting with 937 related results, filtered to 99 published papers, and finally to 12 relevant articles, from 1996 to 2014. The research was carried out between December 10 to December 30, 2014, and sought to answer the following research question: "What data mining techniques have been used in the Brazilian governmental context"? All papers were found using the Google Scholar tool (scholar.google.com.br), which returned results emanating from the various repositories of federal and state universities, besides important online electronic libraries. The line and the bar chart of Figure 1 and Figure 2 show the number of published papers by year and by the repository, respectively.

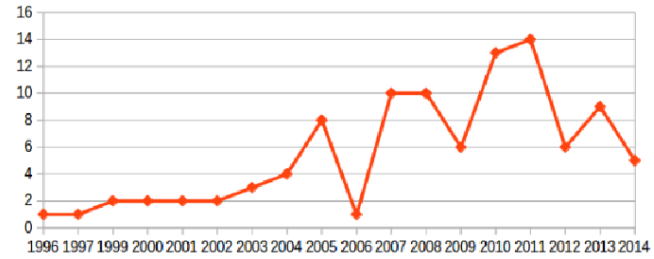

Fig. 1. *The number of published papers by year.*

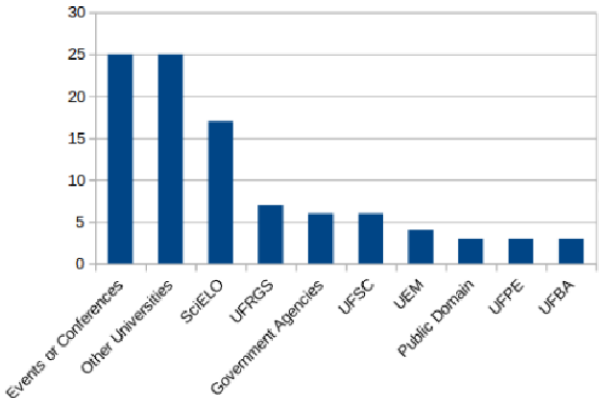

Fig. 2. *The number of published papers by repository.*

B. Studies Selection

Firstly, we considered for inclusion all Brazilian papers, published in Portuguese, available at events or conferences, university repositories or online electronic libraries, as long as they addressed some data mining issue. Another demand was the involvement with the socioeconomic and/or the educational context, applying any data mining technique in order to support municipal, state or federal administration decision-making.

We used the following search argument, which returned 937 results: (Mineração de dados OR data mining OR inteligência empresarial OR business intelligence OR descoberta de conhecimento OR KDD) AND (socioeconômico OR socioeducacional OR educacional OR econômico OR social) AND (IDH OR IBGE OR IPEA OR INEP OR censo OR censitário) AND (governamental OR governo). To clarify, IBGE, IPEA, and INEP are Brazilian government agencies who hold relevant data majorly about demographics, economics, and education, respectively.

We read titles from these studies, and abstracts from the articles with the most meaningful titles. Hence, we selected 99 papers with the following exclusion criteria: language (only Portuguese), duplicity (only an article with the same title), search argument term (at least one search argument term), and summary (papers regarding to data mining techniques, in general, or involving socioeconomic/educational data in the governmental context). Finally, by taking as exclusion criterion the content itself, through full article readings, we obtained 12 relevant papers to our systematic review.

C. Data Extraction

Firstly, we considered for inclusion all Brazilian papers, published in Portuguese, available at events or conferences, university repositories or online electronic libraries, as long as they addressed some data mining issue. Another demand was the involvement with the socioeconomic and/or the educational context, applying some data mining technique in order to support municipal, state or federal administration decision making.

III. RESULTS AND DISCUSSION

This study aimed to answer the question: "What data mining techniques have been used in the Brazilian governmental context"? Through the reading of the selected papers, we realized that, for the sake of synthesizing their main ideas and answering the previous question, we needed to identify common cores of information. These are related to the Knowledge Discovery in Databases (KDD) process [3], and can be grouped as follows: i) socioeconomic and educational databases, ii) used data mining tool, iii) applied data mining technique, and iv) the application context of the study.

In this work, the data sets are composed of spreadsheets containing census from the Brazilian Institute of Geography and Statistics (IBGE). They are organized by the Institute for Applied Economic Research (IPEA) according to areas of relevance and can be found at http://goo.gl/K9rXJF.

Examples of areas of relevance are our target areas (social, economic and educational), among others (basic sanitation, health services etc). In possession of one or more available spreadsheets, firstly, we selected the relevant census data involved directly or indirectly with the social, economic and/or educational area. Lastly, we filtered the results according to the target city(ies), state(s) or region(s).

Additionally, by doing a statistical and comparative analysis in all results, we accomplished the further main evidence represented by pizza charts of Figure 3, Figure 4, and Figure 5. They provide a general, but useful guidance for some future work in governmental data mining.

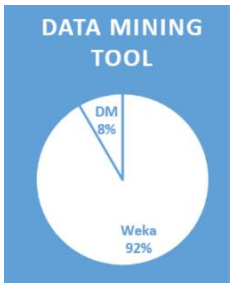

Fig. 3. *Most recurrent data mining tools.*

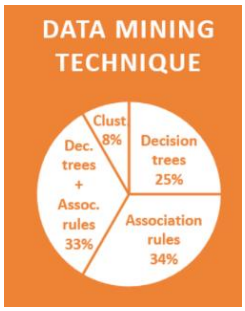

Fig. 4. *Most recurrent data mining techniques.*

Fig. 5. *Most recurrent application contexts.*

The results show, essentially, the predominance of Waikato Environment for Knowledge Analysis (Weka) as an adequate software to run the process of data mining (Fig. 3), the prevalent use of association rules and decision trees to extract information and support decision making in the Brazilian governmental context (Fig. 4), as well as applicability of these resources in diverse application contexts (Fig. 5).

IV. CONCLUSION

In pursuance of identifying the prevailing data mining techniques which could be applied in the Brazilian governmental context, this work followed a systematic review process, from the selection to the comparison of state-of-art papers. In addition to identifying these techniques (association rules and decision trees), the work was able to highlight the most common used tool (Weka) and confirm its applicability in that context. A greater understanding can be achieved through the link https://goo.gl/sGjT14.

Our main contribution is, thereby, this entire analysis. Thus, we expect this paper to be used as a resource for some future work in governmental data mining, taking advantage of our suggestions and conclusions. Particularly, to one that intends to help the public administration taking better investment decisions according to their socioeconomic and educational conditions. Another forthcoming work may involve the use of tools, techniques and application contexts highlighted in this paper to find additional useful knowledge on the IBGE database.

REFERENCES

[1] Berry, M. J. A. and Linoff, G. S. (2004) “Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management”. Wiley.

[2] Coradine, L. M. L. C., Lachtermacher, G. and Coelho, P. S. S. (2007). “Determinação de Fatores Críticos para o IDH-M a partir de Técnicas de Mineração de Dados”. In XXXIX SBPO, p. 821-832.

[3] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). "From data mining to knowledge discovery in databases". In AI magazine, v. 17, n. 3, p. 37.

[4] Gomes, J. C., Levy, A. and Lachtermacher, G. (2004). Segmentação do censo educacional 2000 utilizando técnicas de mineração de dados. In XXXVI SBPO, p. 820–831.

[5] Kitchenham, B. (2004). "Procedures for performing systematic reviews". Keele, UK, Keele University, In Keele University Technical Report TR/SE-0401, v. 33, p. 1-26.

[6] Moraes, A. F. de. (2003). "Um modelo representativo de conhecimento para aplicação da mineração de dados no cadastro técnico urbano". UFSC.