Revision Article

Data mining: a literature review

Técnica de mineração de dados: uma revisão da literatura

Técnica de mineración de datos: una revisión de la literatura

Noemi Dreyer Galvão[1], Heimar de Fátima Marin[2]

ABSTRACT

The purpose of this study was to conduct a literature review on data mining (DM) technique in the LILACS and SciELO databases and specialized books. A broad data literature search using the words data mining (in English) and/or "mineração de dados" (in Portuguese) and limited to publications between 1999 and 2008, was conducted. The exclusion criteria were the keywords mining industry, mines, mineralogy, and publications that did not describe the methods and the tasks related to data mining. Of 123 publications retrieved, 38 were selected to review. Findings suggest that the existent amount of stored data is titanic and it continue to increase considerably. Thus, the process of knowledge discovery in databases and DM have developed tasks and methods for the retrieval of useful knowledge that may be of interest and necessary for just-in-time decision making in different areas of knowledge.

Keywords: Information storage and retrieval/methods; Pattern recognition, automated/methods; Knowledge bases Medical Informatics

RESUMO

Este artigo teve como objetivo realizar uma revisão da literatura sobre a técnica de mineração de dados (*Data Mining* – DM) nas bases de dados abrangendo o Literatura Latino-Americana e do Caribe em Ciências da Saúde (LILACS), Scientific Eletronic Library Online (SCIELO) e alguns livros sobre o tema. Buscou-se uma coleta ampla utilizando as palavras *data mining* e mineração de dados, abrangendo o período de 1999 a 2008. Como critérios de exclusão foram utilizados os descritores: indústria mineira, minas, mineralogia; foram excluídos artigos que não esclareciam o método e as tarefas relacionadas à mineração de dados. Dos 123 artigos encontrados, 32 foram selecionados. Observou-se que o volume de dados armazenados é gigantesco e continua crescendo exponencialmente. Com isso o processo de Descoberta do Conhecimento em Bases de Dados e DM inclui tarefas e métodos para extração de conhecimento útil, interessante e indispensável na tomada de decisões rápidas nas mais diversas áreas de conhecimento.

Descritores: Armazenamento e recuperação da informação/métodos; Reconhecimento automatizado de padrão/métodos; Bases de conhecimento; Informática médica

RESUMEN

En este artículo se tuvo como objetivo realizar una revisión de la literatura sobre la técnica de *mineración de datos* (*Data Mining* – DM) en las bases de datos que abarcaban la Literatura Latino-Americana y del Caribe en Ciencias de la Salud (LILACS), Scientific Eletronic Library Online (SCIELO) y algunos libros sobre el tema. Se buscó una recolección amplia utilizando las palabras *data mining* y mineración de datos, en el período comprendido entre 1999 a 2008. Como criterios de exclusión fueron utilizados los descriptores: industria minera, minas, mineralogía; se excluyeron artículos que no aclaraban el método y las tareas relacionadas a la mineración de dados. De los 123 artículos encontrados, 32 fueron seleccionados. Se observó que el volumen de datos almacenados es gigantesco y continúa creciendo exponencialmente. Con eso el proceso de Descubrimiento del Conocimiento en Bases de Datos y DM incluye tareas y métodos para la extracción del conocimiento útil, interesante e indispensable para la toma de decisiones rápidas en las más diversas áreas del conocimiento.

Descriptores: Almacenamiento y recuperación de la información/métodos; Reconocimiento de normas patrones automatizadas/métodos; Bases del conocimiento; Informática médica

[1] *PhD research student in the Post Graduation Program of the Nursing Department of the University of São Paulo – UNIFESP – São Paulo (SP), Brazil; Technician of the Health State Secretariat of Mato Grosso (MT), Brazil.*

[2] *Senior Professor of the Federal University of São Paulo - UNIFESP - São Paulo (SP), Brazil.*

Corresponding Author: **Noemi Dreyer Galvão**
R. Veneza, 105 - Jardim Itália - Cuiabá - MT - Brazil
Cep: 78060-810. E-mail: noemigalvao.mt@terra.com.br

Received article **25/08/2008** and accepted **13/04/2009**

INTRODUCTION

In the last decades, in which most of the operations and activities of the public and private institutions are computationally registered and accumulate in large databases, the data mining technique – *Data Mining* (DM) – is one of the most effective alternatives to extract knowledge from the great volume of data, discovering hidden relationships, patterns and generating rules to predict and correlate data, that can help the institutions in faster decision-making or, even reach a bigger degree of confidence[1].

Nowadays, information and knowledge are legal, strategic and indispensable prerogatives in search for greater autonomy in the actions of the health companies, social control and decision-making with time getting shorter and shorter. Because of this, several national and international companies of production, consumption, financial market, teaching institutions and libraries have already adopted in their routines, data mining to monitor funding, client consumption, prevent fraud and foreseeing market risks, among others[1-4]. In the health sector, mainly the public one, the application is being accepted as a way of accelerating the search for knowledge. Besides, the use of data mining in the big hospital databases or even in the information systems of public health contributes to discover relationships so that they can make a prevision of future tendencies based on the past, best characterizes the patient that seeks for assistance, identifies successful medical therapies for different diseases and shows patterns of new injuries.

However, several managers and health professionals are concerned with the understanding of the data and in using the information and knowledge of the health databases to promote the information management and the quality of care[5-6]. This probably occurs due to the fast rhythm of data generation[1], which produces a natural incapacity in the human being to explore, extract and interpret these data to obtain knowledge of these bases.

In this sense, the informatics and the technologies directed to the collection, storage and data availability has been developing and making available techniques, methods and automatic computational tools, capable of helping in the extraction of useful information inside this great volume of complex data[6-7].

However, to attend this new context, the health informatics has been using these methodologies of computing science to accomplish its studies. Among them, the methodology *Knowledge Discovery in Databases* (KDD), that is, discovery of the databases knowledge, and the data mining, which is one of the most important stages of KDD[1,7].

As the theme is "pulverized" in the most diverse areas of knowledge, this article, aimed at presenting a literature review of the main indexed databases and some books published on the subject, thus presenting the use of the technique of data mining, concepts, tasks and methods.

METHODS

This is a bibliographic review study, in the national and international scope. The widest possible collection of data was searched for, using the English term (*data mining*) and in Portuguese (data mining). The reference period was from 1999 to 2008. Databases used in the search for scientific articles: Latin American and Caribbean Literature in Health Sciences (LILACS) and Scientific Electronic Library Online (SciELO), due to the easy access to complete texts for reading, mainly of the method. Five books were selected, used to indicate some concepts not found in the articles. As a criterion for the exclusion of the articles, the key-words related to the mining industry, mines, mineralogy and those articles that did not elucidate the method and the tasks related to data mining were used. Thirty-two (32) quotations were identified in the LILACS databases and only 10 were included in the study. However in the SciELO databases, with a regional index, 91 quotations were found, from which 28 were selected. Of the 38 articles selected, 06 were repeated; so, 32 articles were reviewed and presented to describe the data mining method – DM.

RESULTS

The theme was divided into three topics: Knowledge discovery in databases. Data Mining Tasks and Data Mining Methods.

Knowledge Discovery in databases

Knowledge discovery in databases (KDD) can be defined as a process of obtaining information using data registered in a databank, an implicit, previously unknown, potentially useful and understandable knowledge[1-2,7-8].

The expression Data Mining (DM) first appears, as a synonym of KDD, but it is only one of the stages of the knowledge discovery in databases in the KDD global process. The knowledge that is possible to acquire through the DM has been very useful in the most different areas, such as medicine, finances, commerce, marketing, telecommunications, meteorology, agriculture and cattle raising, bioinformatics, among others[2,7-11].

Data mining is not a trivial process; it consists of the ability to identify, in the data, the valid, new, potentially useful and understandable patterns, involving statistical methods, visualization tools and artificial intelligence techniques[12].

So, the KDD process uses databases concepts, statistical methods, visualization tools and artificial

intelligence techniques, dividing into the following phases: selection, pre-processing, transformation, DM and evaluation/interpretation[1-2,12]. Among these phases, the most important is the data mining, focuses of countless studies in several areas of knowledge[1,7,9-10,13-17], that confirms the assumption of the changing of data into information and later into knowledge, which makes the technique necessary for the making-decision process.

Data mining has several phases: the clear definition of the problem; the selection of all the internal and external data sources and the preparation of data, which includes the pre-processing, data reformatation and analysis of the results taken from the DM process[1,7].

The discovery of knowledge must present the following characteristics: be efficient (accurate), generic (applicable to several kinds of data) and flexible (easily modifiable)[5]. The Dm development process involves tasks, methods and algorithms in order to make possible the extraction of new knowledge[1]. Among the several DM tasks, some that are most used: association, classification, regression, *clusterization* and summarization are emphasized[1-2,8,10].

Data Mining Tasks

In the data mining, the tasks and the algorithms that will be used according to the aims of the study are defined, in order to obtain an answer to the problem[8,18]. The possible tasks of an algorithm of pattern extraction can be gathered in predicted and descriptive activities.

The two main kinds of tasks for prediction are the classification and the regression. The classification consists of the prediction of a categorical variable, that is, to discover an activity that will map a set of registers in a set of predefined variables called classes. This activity can be applied to new registers, so as to foresee the class in which these registers fit. Several algorithms are applied in the classification tasks, but those that appear most are Neural Networks, *Back-Propagation*, Bayesian Classifiers and Genetic Algorithms[2,19].

In the regression, there is a search for linear functions or not, and the variable that is to be predicted consists of a numerical attribute (continuous) present in databases with real values[1-2,20]. In order to implement the regression task, the methods of statistics and Neural Networks are used.

The *clusterization* task is used to separate the registers of databases into subsets or *clusters*, in such a way that the elements of a *cluster* share common properties that serve to distinguish the elements in other *clusters*, aiming at maximizing intra-cluster similarities and minimize *inter-cluster similarities*. Unlike the classification tasks in which the variables are pre-defined, the *clusterization* needs, to identify automatically, the data groups, to which the researcher should attribute the variables[21-22]. The most used algorithms in this task are the *K-Means, KModes, K-Protopypes, K-Medoids, Kohonem*, among others[2,23].

The association task consists of identifying and describing associations among variables in the same item or associations among different items that occur simultaneously, in a frequent way in databases[1-2]. The search for associations among items during the temporal interval is also common[1-2,24-26]. So, the algorithms Apriori and GSP (*Generalized Sequential Patterns)* among others, implement the discovery of association task[27].

The summarization seeks to identify and indicate common characteristics among a set of data. This task is applied in the clusters obtained in the *clusterization task,* with the Inductive Logic and Genetic Algorithms being examples of technologies that can implement the summarization[2].

Data Mining Methods

The methods are technologies that exist, regardless of the data mining context, when applied in the KDD, they produce good results in the health area, changing data into useful knowledge and favoring the health practices based on evidence[28]. There are several methods, but the aim is not to exhaust the subject but to identify the most used. The main technologies are: Neural Networks, Decision Tree, Genetic Algorithms (AGs), *Fuzzy logic* and Statistics[1,5,18,29-30].

The Artificial Neural Networks (RNA) is a computational technique that builds the mathematic model inspired in the human brain for the recognition of images and sounds, with knowledge capacity, generalization, association and abstraction, constituted by parallel systems distributed in compounds of simple processing units[2,24,31].

The processing units are one or more layers interlinked by a great number of connections; in most of the models, these connections are associated to weights, which, after the learning process, store the knowledge acquired by the net[31-32] .

The RNAs have been successfully used to model relations involving complex temporal series in several knowledge areas[31]. The biggest advantage of the RNAs over the conventional methods is that they do not require detailed information about the physical processes of the system that is to be modeled, with it being explicitly described in the mathematical way (enter-exit model) and still for being strong and have a high rate of predictive accuracy[2,24,31-32]. Through repetitive presentations of data to the net, the RNA learns patterns, seeks for relationships and automatically builds models[33].

The Decision Tree is a model graphically represented by branches, similar to a tree, but in the inverted sense; they are also called classification or regression trees, in case the dependent variable be respectively categorical

or numeric[2,29,30,34] .

The model of knowledge that lies in each internal branch of the tree represents a decision about a variable that determines how the data present division to a series of branches (offspring). With this, it describes an association between the attribute and the target variable, that is, the association of each branch with other branches) – offspring created[2,24,29].

The aim of the induction of a Decision Tree is to produce an accurate prediction model or discover the predictive structure of the problem. In the last case, the intention is to understand which of these variables and interactions lead to the phenomenon that is being studied. These two purposes do not eliminate each other, and can appear together in the same study[29,31,34]. Some recent researches have used the induction of a Decision Tree to predict and obtain knowledge[29-30].

The Genetic Algorithms formulate algorithmic optimization strategies inspired in the principles observed in the national evolution and in genetics for the solution of problems. The AGs use the selection operators, crossover and mutation to develop successive solution generations – called reproduction. With the evolution of the algorithm, only the solutions with higher prevision power survive, until they reach an ideal solution[1-2,34].

Another very used method is the (*Fuzzy logic*), a mathematic theory that allows a modeling in a way close to reasoning, imitating the human ability of making decisions in environments of uncertainties and inaccuracy. With this, intelligent systems of control and decision support can be built[34].

The *fuzzy* logic can be used mainly in two forms: one is to represent the classic logic extension for a more flexible one, aiming at making formal inaccurate concepts and the other is where *fuzzy* sets are applied to several theories and technologies to process inaccurate information, such as in decision making processes[2].

Finally, statistics, one of the most traditional techniques, provides models for analyses and data interpretation. The most used models are Bayesian Networks, Discriminate Analysis, Exploratory Data Analysis, among others. The basic statistical principle concerns the way in which the probability of an event is estimated based on two kinds of knowledge[1-2,35].

The Bayesian Networks[24] appeared recently as a powerful data mining technique that offers graphic representations of probabilistic distributions derived from the counting of data occurrence in a certain set, representing a relationship of variables.

Finally, with all the concepts and information on the subject, we can affirm along with some authors[1,6,10,36-37], that informatics, its technologies and tools, such as DM, brought great advantages for the areas that operate voluminous databases.

FINAL CONSIDERATIONS

The process of knowledge extraction can bring a valuable reward to the health area with the identification of the pattern of new diseases directing a fast decision making and useful knowledge in several sectors. Yet, it is worth highlighting that for each objective proposed tasks and specific materials must be applied. The tools do not replace the need for a previous and deep knowledge of the exploitation domain, by the researchers, but, on the other hand, it must be highlighted that the KDD and DM are in constant evolution and application in health. For instance, such resources are being used in a study carried out by the authors about data mining of public health and public security, for the construction of a set of data, aiming at establishing associations and predicting a prevention model.

This article, attempted to provide information that can give support to the discussions and doubts in the health area in relation to the use of great databases in the health science areas to extract knowledge of the great databases that exist in the health science area, in an attempt to delimit the knowledge as a support for the actions and decision making, thus intervening in the public health problems.

REFERENCES

1. Cardoso ONP, Machado RTM. Gestão do conhecimento usando data mining: estudo de caso na Universidade Federal de Lavras. Rev Adm Pública. 2008;42(3):495-528.
2. Goldschmidt R, Passos E. Data mining: um guia prático, conceitos, técnicas, ferramentas, orientações e aplicações. São Paulo: Elsevier; 2005.
3. Marcano Aular YJ, Talavera Pereira R. Minería de datos como soporte a la toma de decisiones empresariales. Opcion. 2007;23(52):104-18.
4. Araujo Júnior RH, Tarapanoff K. Precisão no processo de busca e recuperação da informação: uso da mineração de textos. Ci Inf. 2006;35(3):236-47.
5. Steiner MTA, Soma NY, Shimizu T, Nievola JC, Steiner Neto PJ. Abordagem de um problema médico por meio do processo de KDD com ênfase à análise exploratória dos dados. Gest Prod. 2006;13(2):325-37.
6. Costa Lda F. Bioinformatics: perspectives for the future. Genet Mol Res. 2004;3(4):564-74.
7. Quoniam L, Tarapanoff K, Araújo Júnior RH, Alvares L. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. Ci Inf. 2001;30(2):20-8.
8. Matos G, Chalmeta R, Coltell O. Metodología para la extracción del conocimiento empresarial a partir de los datos. Inf Tecnol. 2006;17(2):81-8.
9. Naãs IA, Queiroz MPG, Moura DJ, Brunassi LA. Estimativa de estro em vacas leiteiras utilizando métodos quantitativos preditivos. Ciênc Rural. 2008;38(8):2383-7.
10. Febles Rodríguez JP, González Pérez A. Aplicación de la

minería de datos en la bioinformática. ACIMED. 2002;10(2):69-76.
11. Jones PBC. The commercialization of bioinformatics. Electron J Biotechnol. 2000;3(2):33-4.
12. Fayyad UM, Shapiro GP, Smyth P, Uthurusamy R. Advances in knowledge discovery and data mining. Menlo Park, Calif.: AAAI Press: MIT Press; c1996; 611p.
13. Calzadilla Fernández Castro O, Jiménez López G, González Delgado BE, Ávila Pérez J. Aplicación de la minería de datos al Sistema Cubano de Farmacovigilancia. Rev Cuba Farm. 2007;41(3):1-5.
14. Botta Ferret E, Cabrera Gato JE. Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital. ACIMED [Internet]. 2007;16(4). Disponível em: http://scielo.sld.cu/pdf/aci/v16n4/aci051007.pdf
15. Wickert E, Marcondes J, Lemos MV, Lemos EGM. Nitrogen assimilation in Citrus based on CitEST data mining. Genet Mol Biol. 2007;30(3 Suppl):810-8.
16. Mahalakshmi V, Ortiz R. Plant genomics and agriculture: from model organisms to crops, the role of data mining for gene discovery. Electron J Biotechnol. 2001;4(3): 169-78.
17. Prati RC, Monard MC, Carvalho ACPLF. Looking for exceptions on knowledge rules induced from HIV cleavage data set. Genet Mol Biol. 2004;27(4):637-43.
18. Rodríguez Perojo K, Ronda León R. El web como sistema de información. ACIMED [Internet]. 2006;14(1). Disponível em:http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352006000100008&lng=es&nrm=iso
19. Pereira GC, Coutinho R, Ebecken NFF. Data mining for environmental analysis and diagnostic: a case study of upwelling ecosystem of Arraial do Cabo. Braz J Oceanogr. 2008;56(1):1-12.
20. Pereira BB. Estatística em psiquiatria. Rev Bras Psiquiatr. 2001;23(3):168-70.
21. Telles GP, Braga MDV, Dias Z, Lin TL, Quitzau JAA, Silva FR, Meidanis J. Bioinformatics of the sugarcane EST project. Genet Mol Biol. 2001;24(1/4):9-15.
22. Zhu D, Porter A, Cunningham S, Carlisie J, Nayak A. A process for mining science & technology documents databases, illustrated for the case of "knowledge discovery and data mining". Ci Inf. 1999;28(1):7-14
23. Scarpel RA, Milioni AZ. Otimização na formação de agrupamentos em problemas de composição de especialistas. Pesqui Oper. 2007;27(1):85-104.
24. Abbott PA, Lee SM. Data mining and knowledge discovery. In: Saba VK, Mccormick KA. Essentials of nursing informatics. 4th ed. New York: McGraw-Hill Medical Pub. Division; c2006.
25. Horng JT, Cho WF. Predicting regulatory elements in repetitive sequences using transcription factor binding sites. Electron J Biotechnol. 2000;3(3):6-7.
26. Pôssas B, Meira Júnior W, Carvalho M, Resende R. Using quantitative information for efficient association rule generation. J Braz Comp Soc. 2000;29(4):19-25.
27. Cavique L. Graph-based structures for the market baskets analysis. Inv Op. 2004;24(2):233-46.
28. Rodrigues RJ. Information systems: the key to evidence-based health practice. Bull World Health Organ. 2000;78(11):1344-51.
29. Meira CAA, Rodrigues LHA, Moraes SA. Análise da epidemia da ferrugem do cafeeiro com árvore de decisão. Trop Plant Pathol. 2008;33(2):114-24.
30. Vale MM, Moura DJ, Nääs IA, Oliveira SRM, Rodrigues LHA. Data mining to estimate broiler mortality when exposed to heat wave. Sci Agric (Piracicaba, Braz). 2008;65(3):223-9.
31. Kovács ZL. Redes neurais artificiais: fundamentos e aplicações. 3a ed. rev. São Paulo: Livraria da Física; 2002.
32. Tarapanoff K, Araújo Júnior RH, Cormier PMJ. Sociedade da informação e inteligência em unidades de informação. Ci Inf. 2000;29(3):91-100.
33. Costa JAF, Andrade Netto ML. Segmentação de mapas auto-organizáveis com espaço de saída 3-D. Sba Controle & Automação. 2007;18(2):150-62.
34. Han J, Kamber M. Data mining: concepts and techniques. 2nd ed. Amsterdam; Boston: Elsevier: Morgan Kaufmann; c2006.
35. Lee BS, Snapp RR, Musick R, Critchlow T. Metadata models for ad hoc queries on terabyte-scale scientific simulations. J Braz Comp Soc. 2002;8(1):5-15.
36. Carazzolle MF, Formighieri EF, Digiampietri LA, Araujo MRR, Costa GL, Pereira GAG. Gene projects: a genome web tool for ongoing mining and annotation applied to CitEST. Genet Mol Biol. 2007;30(3 Suppl):1030-6.
37. Castillo Zayas YM, Leiva Mederos AA. La minería de texto: perspectiva metodológica para la realización de resúmenes documentales. ACIMED [Internet]. 2007;15(5).http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352007000500014&lng=es&nrm=iso&tlng=es