Técnica de mineração de dados: uma revisão da literatura

Data mining: a literature review

Técnica de mineración de datos: una revisión de la literatura

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

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

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

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] *Doutoranda pelo Programa de Pós Graduação do Departamento de Enfermagem da Universidade de São Paulo – UNIFESP – São Paulo (SP), Brasil; Técnica da Secretaria de Estado de Saúde de Mato Grosso (MT), Brasil.*

[2] *Professora Titular da Universidade Federal de São Paulo - UNIFESP - São Paulo (SP), Brasil.*

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

Artigo recebido em **25/08/2008** e aprovado em **13/04/2009**

INTRODUÇÃO

Nas últimas décadas, em que a maioria das operações e atividades das instituições privadas e públicas são registradas computacionalmente e se acumulam em grandes bases de dados, a técnica da mineração de dados – *Data Mining* (DM) – é uma das alternativas mais eficazes para extrair conhecimento a partir de grandes volumes de dados, descobrindo relações ocultas, padrões e gerando regras para predizer e correlacionar dados, que podem ajudar as instituições nas tomadas de decisões mais rápidas ou, até mesmo, a atingir um maior grau de confiança[1].

Hoje, a informação e o conhecimento são prerrogativas legais, estratégicas e imprescindíveis à busca de maior autonomia nas ações das empresas de saúde, controle social e na tomada de decisão com prazos cada vez mais curtos. Por isso, diversas empresas nacionais e internacionais de produção, consumo, mercado financeiro, instituições de ensino e bibliotecas já adotaram, nas suas rotinas, a mineração de dados para monitorar arrecadações, consumo de clientes, prevenir fraudes e previsão de riscos do mercado, dentre outras[1-4]. No setor saúde, principalmente no público, a aplicação está sendo aceita como uma forma de agilizar a busca de conhecimento. Além do mais, a utilização da mineração de dados nos grandes bancos de dados hospitalares ou até mesmo nos sistemas de informação de saúde pública contribui para descobrir relacionamentos para que possa ser feita uma previsão de tendências futuras baseada no passado, caracteriza melhor o paciente que busca assistência, identifica terapias médicas de sucesso para diferentes doenças e demonstra padrões de novos agravos.

Contudo, há uma apreensão por parte de vários gestores e profissionais de saúde em compreender os dados e em utilizar a informação e conhecimento das bases de dados da saúde para promover a gestão da informação e qualidade de cuidados[5-6]. Isso provavelmente acontece em decorrência de um ritmo alucinante de geração de dados[1], o que produz uma incapacidade natural no ser humano de explorar, extrair e interpretar estes dados para obter conhecimento dessas bases.

Nesse sentido, a informática e as tecnologias voltadas para coleta, armazenamento e disponibilização de dados vêm evoluindo e disponibilizando técnicas, métodos e ferramentas computacionais automatizadas, capazes de auxiliar na extração de informações úteis contidas nesse grande volume de dados complexos[6-7].

No entanto, para atender este novo contexto, a informática em saúde vem se apropriando dessas metodologias da ciência da computação para realizar seus estudos. Dentre elas, a metodologia *Knowledge Discovery in Databases* (KDD), ou seja, descoberta de conhecimento das bases de dados, e a extração ou mineração de dados, que é uma das etapas mais importantes do KDD[1,7].

Como o tema está "pulverizado" nas mais diversas áreas do conhecimento, o presente artigo, teve como objetivo apresentar uma revisão da literatura das principais bases de dados indexadas e alguns livros publicados sobre o assunto, apresentando assim a aplicação da técnica de mineração de dados, conceitos, tarefas e métodos.

METODOS

Trata–se de um estudo de revisão bibliográfica, no âmbitos nacional e internacional. Buscou-se uma coleta de dados o mais ampla possível, utilizando o termo em inglês (*data mining*) e em português (mineração de dados). O período de referência foi de 1999 a 2008. Utilizou-se as bases de dados para busca de artigos científicos: Literatura Latino–Americana e do Caribe em ciências da Saúde(LILACS) e Scientific Eletronic Library Online (SciELO), tendo em vista a facilidade dos textos completos para leitura, principalmente do método. Os livros selecionados foram cinco, utilizados para referenciar alguns conceitos não encontrados nos artigos. Como critério de exclusão dos artigos foram utilizados os descritores relacionados a indústria mineira, minas, mineralogia e aqueles artigos que não esclareciam o método e as tarefas relacionadas à mineração de dados. Foram identificadas 32 citações na base de dados LILACS e somente 10 foram incluídas no estudo. Já na base de dados SciELO, com índice regional, encontrou-se 91 citações, das quais 28 foram selecionadas. Dos 38 artigos selecionados, 06 estavam duplicados; portanto, 32 artigos foram revisados e apresentados a seguir para descrever o método de mineração de dados – DM.

RESULTADOS

O tema foi divido em três tópicos: Descoberta de conhecimento em bases de dados, Tarefas do Data Mining e Métodos de Data Mining.

Descoberta de conhecimento em bases de dados

A descoberta de conhecimento em bases de dados (KDD) pode ser definida como o processo de extração de informação a partir de dados registrados numa base de dados, um conhecimento implícito, previamente desconhecido, potencialmente útil e compreensível[1-2,7-8].

A expressão Mineração de Dados (DM) surge inicialmente, como um sinônimo de KDD, mas é apenas uma das etapas da descoberta de conhecimento em bases de dados no processo global do KDD. O conhecimento que se consegue adquirir através da DM tem se mostrado bastante útil nas mais diversas áreas, como medicina, finanças, comércio, marketing, telecomunicações, meteorologia, agropecuária, bioinformáticas, entre outras[2,7-11].

A mineração de dados não é um processo trivial; consiste na habilidade de identificar, nos dados, os padrões válidos, novos, potencialmente úteis e compreensíveis, envolvendo métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial[12].

Assim, o processo de KDD utiliza conceitos de base de dados, métodos estatísticos, ferramentas de visualização e técnicas de inteligência artificial, dividindo-se nas etapas de seleção, pré-processamento, transformação, DM e avaliação/interpretação[1-2,12]. Dentre essas etapas, a mais importante é a mineração de dados, foco de inúmeros estudos em diversas áreas de conhecimento[1,7,9-10,13-17], que comprovam o pressuposto da transformação de dados em informação, e posteriormente em conhecimento, o que torna a técnica imprescindível para o processo de tomada de decisão.

A mineração de dados possui várias etapas: a definição clara do problema; a seleção de todas as fontes internas e externas de dados e a preparação dos dados, que inclui o pré-processamento, reformatação dos dados e análise dos resultados obtidos do processo de DM[1,7]

A descoberta do conhecimento deve apresentar as seguintes características: ser eficiente (acurado), genérica (aplicável a vários tipos de dados) e flexível (facilmente modificável)[5]. O processo de desenvolvimento de DM envolve tarefas, métodos e algoritmos para possibilitar a extração de novos conhecimentos[1]. Entre as várias tarefas de DM, destacam-se algumas que são as mais utilizadas: associação, classificação, regressão, *clusterização* e sumarização[1-2,8,10].

Tarefas da Data Mining

Na mineração de dados, são definidas as tarefas e os algorítimos que serão utilizados de acordo com os objetivos do estudo, a fim de obter uma resposta para o problema[8,18]. As tarefas possíveis de um algorítmo de extração de padrões podem ser agrupadas em atividades preditivas e descritivas.

Os dois principais tipos de tarefas para predição são a classificação e a regressão. A classificação consiste na predição de uma variável categórica, ou seja, descobrir uma função que mapeie um conjunto de registros em um conjunto de variáveis predefinidas, denominadas classes. Tal função pode ser aplicada em novos registros, de forma a prever a classe em que tais registros se enquadram. Vários algorítimos são aplicados na tarefa de classificação, mas os que mais se destacam são as Redes Neurais, *Back-Propagation*, Classificadores Bayesianos e Algorítmos Genéticos[2,19].

Na regressão, busca-se funções lineares ou não, sendo que a variável a ser predita consiste em um atributo numérico (contínuo) presente em banco de dados com valores reais[1-2,20]. Para implementar a tarefa de regressão, utilizam-se os métodos da estatística e de Redes Neurais.

A tarefa de *clusterização* é utilizada para separar os registros de uma base de dados em subconjuntos ou *clusters (agrupamentos)*, de tal forma que os elementos de um *cluster* compartilhem propriedades comuns, que servem para distinguir os elementos em outros *clusters*, tendo como objetivo maximizar similaridade *intra-cluster* e minimizar similaridade *inter-cluster*. Diferente da tarefa de classificação, em que as variáveis são predefinidas, a *clusterização* precisa, automaticamente, identificar os grupos de dados, aos quais o pesquisador deverá atribuir as variáveis[21-22]. Os algorítimos mais utilizados nessa tarefa são os *K-Means, KModes, K-Protopypes, K-Medoids, Kohonem*, dentre outros[2,23].

A tarefa de associação consiste em identificar e descrever associações entre variáveis no mesmo item ou associações entre itens diferentes que ocorram simultaneamente, de forma freqüente em banco de dado[1-2]. É também comum a procura de associações entre itens durante um intervalo temporal[1-2,24-26]. Portanto, os algorítimos Apriori e GSP (*Generalized Sequential Patterns - Padrão Sequencial Geral*), dentre outros, implementam a tarefa de descoberta de associações[27].

A sumarização procura identificar e indicar características comuns entre um conjunto de dados. Essa tarefa é aplicada nos agrupamentos obtidos na tarefa de *clusterização, s*endo a Lógica Indutiva e Algorítmos Genéticos exemplos de tecnologias que podem implementar a sumarização[2].

Métodos de Data Mining

Os métodos são tecnologias existentes, independente do contexto mineração de dados, uma vez que, aplicados na KDD, produzem bons resultados na área da saúde, transformando dados em conhecimento útil e favorecendo as práticas de saúde baseadas em evidências [28]. São vários métodos existentes, mas o objetivo não é esgotar o assunto e sim identificar os mais utilizados. As principais tecnologias são: Rede Neurais, Árvore de Decisão, Algorítmos Genéticos (AGs), Lógica Nebulosa (*Fuzzy logic*) e Estatística[1,5,18,29-30].

A Rede Neural Artificial (RNA) é uma técnica computacional que constrói modelo matemático inspirado em cérebro humano para reconhecimento de imagens e sons, com capacidade de aprendizado, generalização, associação e abstração, constituído por sistemas paralelos distribuídos em compostos de unidades simples de processamento[2,24,31].

As unidades de processamento são uma ou mais camadas interligadas por um grande número de conexões; na maioria dos modelos, essas conexões estão associadas a pesos, os quais, após o processo de aprendizagem, armazenam o conhecimento adquirido pela rede[31-32] .

As RNAs têm sido utilizadas com sucesso para modelar relações envolvendo séries temporais complexas

em várias áreas do conhecimento[31]. A maior vantagem das RNAs sobre os métodos convencionais é que elas não requerem informação detalhada sobre os processos físicos do sistema a ser modelado, sendo este descrito explicitamente na forma matemática (modelo de entrada-saída) e ainda por ser robusta e ter uma alta taxa de acurácia preditiva[2,24,31-32]. Por meio de repetidas apresentações dos dados à rede, a RNA aprende padrões, procura relacionamentos e constrói modelos automaticamente[33].

A Árvore de Decisão é um modelo representado graficamente por nós e galhos, parecido com uma árvore, mas no sentido invertido; também são chamadas de árvores de classificação ou de regressão, caso a variável dependente seja categórica ou numérica, respectivamente[2,29,30,34] .

O modelo de conhecimento que tem em cada nó (galho) interno da árvore representa uma decisão sobre uma variável que determina como os dados apresentam partição por uma série de galhos (nós filhos). Com isso, descreve uma associação entre o atributo e variável alvo, ou seja, associação de cada galho com outro(s) galho(s) – filhos gerados[2,24,29].

A finalidade da indução de uma Árvore de Decisão é produzir um modelo de predição preciso ou descobrir a estrutura preditiva do problema. No último caso, a intenção é compreender quais variáveis e interações dessas conduzem ao fenômeno estudado. Esses dois propósitos não são excludentes, podendo aparecer juntos em um mesmo estudo[29,31,34]. Algumas pesquisas recentes têm utilizado a indução de Árvore de Decisão para predizer e obter conhecimento[29-30].

Já os Algorítmos Genéticos formulam estratégias de otimização algorítmica inspiradas nos princípios observados na evolução natural e na genética, para solução de problemas. Os AGs usam os operadores de seleção, cruzamento e mutação para desenvolver sucessivas gerações de soluções – chamado de reprodução. Com a evolução do algorítmo, somente as soluções com maior poder de previsão sobrevivem, até convergirem numa solução ideal[1-2,34].

Outro método muito utilizado é a Lógica Nebulosa (*Fuzzy logic*), uma teoria matemática que permite uma modelagem do modo aproximado de raciocínio, imitando a habilidade humana de tomar decisões em ambientes de incertezas e imprecisão. Com isso, pode-se construir sistemas inteligentes de controle e suporte à decisão[34].

A lógica *fuzzy* pode ser utilizada principalmente de duas formas: uma é representar a extensão da lógica clássica para uma mais flexível com objetivo de formalizar conceitos imprecisos e a outra é onde se aplicam conjuntos *fuzzy* à diversas teorias e tecnologias para processar informações imprecisas como por exemplo, em processos de tomada de decisão[2].

Por fim, a estatística, uma das técnicas mais tradicionais, fornece modelos para análise e interpretação de dados. Os modelos mais utilizados são Redes Bayesianas, Análise Discriminante, Análise Exploratória de dados, dentre outros. O princípio estatístico de base concerne à maneira pela qual se estima a probabilidade de um evento a partir de dois tipos de conhecimento[1-2,35].

As Redes Bayesianas[24] emergiram em anos recentes como uma poderosa tecnologia de mineração de dados, que fornecem representações gráficas de distribuições probabilísticas derivadas da contagem da ocorrência dos dados num determinado conjunto, representando um relacionamento de variáveis.

Por fim, diante dos conceitos e informações sobre o assunto, pode-se afirmar em conjunto com alguns autores[1,6,10,36-37], que a informática, suas tecnologias e ferramentas, como DM, trouxeram grandes vantagens para as áreas que operam bancos de dados volumosos.

CONSIDERAÇÕES FINAIS

O processo de extração de conhecimento pode trazer uma recompensa valiosa à área da saúde com identificação dos padrões de novas doenças direcionando uma rápida tomada de decisão e conhecimento útil nos diversos setores. Contudo, vale destacar que para cada objetivo proposto deve-se aplicar tarefas e métodos específicos. As ferramentas não substituem a necessidade de um conhecimento prévio e profundo do domínio de exploração, por parte dos pesquisadores, mas, por outro lado, deve-se ressaltar que a KDD e DM estão em constante evolução e aplicação na saúde. Por exemplo, tais recursos estão sendo aplicados em um estudo conduzido pelas autoras sobre a mineração de dados da saúde pública e da segurança pública, para construção de um conjunto de dados, visando estabelecer associações e predizer um modelo de prevenção.

Neste artigo procurou-se fornecer informações que possam sustentar as discussões e os questionamentos na área da saúde quanto à utilização da mineração de dados para extrair conhecimentos das grandes bases de dados existentes na área de ciências da saúde, na tentativa de procurar demarcar o conhecimento como apoio para as ações e tomada de decisão, intervindo assim nos problemas de saúde pública.

REFERÊNCIAS

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