ISSN: 2237-826X
3(2), jul./dez. 2014

Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas

Data Mining: Applications, tools, learning types and other subtopics

Deborah Ribeiro Carvalho[1], Marcelo Rosano Dallagassa[2]

[1] *Pontifícia Universidade Católica do Paraná (PUC-PR), Curitiba, PR, Brasil*
[2] *Unimed Paraná, Curitiba, PR, Brasil*

Correspondência para/*Correspondence to*: Deborah Ribeiro Carvalho [ribeiro.carvalho@pucpr.br]

Deborah Ribeiro Carvalho possui graduação em Processamento de Dados pela Universidade Federal do Paraná – UFPR (1979), mestrado em Informática Aplicada pela Pontifícia Universidade Católica do Paraná – PUC-PR (1999), doutorado em Informática Aplicada pela Pontifícia Universidade Católica do Paraná (2002) e doutorado em Computação de Alto Desempenho pela Universidade Federal Do Rio Janeiro (COPPE) (2005). Professor da Pontificia Universidade Católica do Paraná, Programa de Pós-Graduação em Tecnologia Aplicada em Saúde e Professor Colaborador do Mestrado Em Gestao da Informacao (UFPR). Tem experiência na área de Ciência da Computação, atuando principalmente nos seguintes temas: mineração de dados, aquisição de conhecimento, apoio a decisão, pós-processamento dos padrões descobertos, na Saúde.

Marcelo Rosano Dallagassa possui graduação em Engenharia Civil pela Universidade Federal do Paraná – UFPR (1988) e Mestrado em Tecnologia em Saúde pela Pontifícia Universidade Católica do Paraná – PUC-PR (2009). Desde 2001 atua como Analista de Negócios e Especialista da UNIMED Paraná, participando nos projetos; Data Warehouse, Portal BI Unimed PR e Informações Estratégicas. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Apoio a Decisão, atuando principalmente nos temas: Descoberta do Conhecimento em Base de Dados, Data Warehouse, Avaliação de Tecnologias em Saúde, RES – Registro Eletrônico de Saúde e Mineração de Dados. Atua também com Professor de Pós-Graduação nas disciplinas Banco de Dados, Data Warehouse e Data Mining, nas seguintes instituições: FAE, PUC-PR, Universidade Positivo e FESP.

Resumo

Especialistas na área de mineração de dados apresentam conceitos, características, limites e potencialidades da mineração de dados, incluindo indicação de ferramentas disponíveis, relações com a inteligência artificial, e implicações de seu uso na área de business intelligence.

Palavras-chave: Mineração de dados. Ferramentas para mineração de dados. Mineração de dados - uso.

Abstract

Experts in the field of data mining present concepts, features, limitations and possibilities of the data mining process, including the indication of tools available, links to artificial intelligence, and the implications of it's use in business intelligence.

***Keywords**: Data mining. Data mining tools. Data mining use.*

1. Qual a aplicabilidade da mineração de dados na construção do conhecimento organizacional, em especial na área de *business intelligence*?

As organizações, de maneira em geral, acumulam grande volume de dados e informações, em distintos formatos e estrutura, que são constantemente demandados. Entre as estratégias de utilização mais frequentes estão a extração de informação e a de conhecimento. Vale destacar que estas duas formas se complementam, dependendo da situação de contexto do problema de gestão em questão. Se o contexto permitir o estabelecimento prévio de premissas que orientem a extração de informações e estas forem suficientes ao gestor, a mineração de dados teria pouco a contribuir, aliado ao fato de não justificar o custo despendido. Por outro lado, se as estratégias baseadas em premissas não atenderam plenamente às expectativas, a mineração de dados passa a ser uma alternativa a ser considerada.

Vale destacar que algumas vezes a análise das informações geradas a partir de premissas, pode demandar muito tempo e recursos. Este também pode ser uma situação na qual a mineração de dados pode ser considerada para agilizar esse processo.

A mineração de dados deve ser adotada para tornar mais eficiente o apoio à tomada de decisão, elemento essencial para o conceito de *business intelligence*.

São inúmeras as aplicações de mineração de dados utilizadas na área de *business intelligence*, as que identificam perfis e características de clientes conforme as ofertas de produtos, alertas de fraudes, agrupamento de regiões conforme características de vendas, associações de produtos e serviços vinculados aos hábitos de consumo, entre outras.

2. Quais as condições de uso da mineração de dados em organizações de pequeno porte?

Não difere muito das condições de uso da mineração de dados em empresas de grande porte. Em ambos os grupos de organizações é necessário modelar e popular as bases de dados, bem como integrá-las. Ou seja, devem ser garantidas condições de coleta, consistência e armazenamento dos dados para posterior extração de informações. Para estas atividades são necessários profissionais capacitados que façam parte do corpo de colaboradores ou que sejam contratados por demanda específica.

Agregar atividades de mineração de dados vai também exigir que os dados estejam disponíveis e de profissionais que dominem as respectivas técnicas. Profissionais estes que podem fazer parte da equipe funcional ou serem contratados especificamente para esta atividade.

Ou seja, as exigências são muito similares, independentemente do porte da empresa.

3. Como está a disponibilidade de ferramentas de mineração de dados (alternativas em *software* livre/gratuito e alternativas pagas)?

A partir da intensificação recente de pesquisas na área de desenvolvimento de algoritmos, há uma grande oferta de ferramentas para mineração de dados em ambiente livre e gratuita, com código fonte aberto (*General Public Licence*), entre elas: Weka (Witten & Frank, 2000), Mahout (Ingersoll, 2009), Orange data mining (Demsar, Zupan, Leban, & Curk, 2004), Rapid Miner (Hofmann & Klinkenberg, 2013), Tanagra (Rakotomalala, 2005), Keel (Alcala-Fdez et al., 2011) etc.

Em termos de aplicações proprietárias, existem várias soluções e algoritmos de *data mining* incorporadas em ferramentas de *business intelligence*, como por exemplo: Oracle (Tamayo, Berger, Campos, & Yarmus, 2005), Microsoft (Seidman, 2001), SAS (Fernandez, 2003), entre outras.

Porém um dos principais desafios ainda consiste em identificar qual estratégia algorítmica melhor se adequa ao contexto, problema e questão.

4. Como se pode saber se uma base de dados é adequada para uma mineração de dados?

Uma base de dados é adequada para a mineração de dados, a partir da respectiva preparação, conforme as etapas previstas no processo de KDD (*Knowledge Discovery in Database*) proposto por Fayyad, Piatetsky Shapiro, Smyth, e Uthurusam (1996), entre elas: seleção e integração dos dados em um repositório único padronizado, remoção de ruídos e *outliers* etc.

Da mesma forma que para qualquer processamento estatístico, o conjunto de dados deve representar, de forma confiável, o universo (mundo real) a ser investigado, possibilitando assim inferir a situação problema como um todo, seja pela perspectiva de completeza ou complexidade do problema. Um dos "mitos" criados, a partir da motivação inicial das discussões sobre a mineração de dados, era ser uma alternativa para "grandes" bases de dados. Este fato decorreu da própria dificuldade de processamento inerente à descoberta e identificação de informações oportunas ao processo decisório em grandes conjuntos. Mas, dispor de um grande conjunto não constitui requisito obrigatório desde que este represente o universo, por amostragem ou não.

Quando a mineração de dados resultar na descoberta de elementos potencialmente úteis para o apoio a tomada de decisão, pode-se afirmar que a base de dados atendeu às expectativas.

5. O que são "ruídos" nas bases de dados e como eles podem influenciar no resultado obtido em um processo de mineração de dados?

"Ruído" representa conteúdo nas bases de dados que pode prejudicar a qualidade da informação extraída, a partir de qualquer método, seja ele tradicional ou baseado em estratégias mais elaboradas. Destacam-se como ruídos: valores fora do domínio, ausência de valores, inconsistências etc. É importante lembrar que o mundo real é ruidoso, ou seja, se uma base de dados representa uma abstração deste mundo real, esta será ruidosa a despeito dos esforços despendidos para a sua modelagem e respectiva população. Cabe aos profissionais da área de tecnologia da informação minimizar o impacto negativo que estes ruídos possam representar nas informações extraídas e disponibilizadas aos gestores.

Por exemplo, todas as vezes que, ao informar os dados cadastrais se omite ou não se informa corretamente a renda, gera-se um ruído no conjunto de dados.

6. Quão próximos estão os estudos de mineração de dados e aprendizagem de máquina?

O processo KDD como um todo compreende uma série de áreas que se relacionam de forma multidisciplinar, a saber: estatística, banco de dados, inteligência artificial, aprendizado de máquina etc. Ou seja, o processo KDD é construído a partir de conceitos destas diversas áreas. Aprendizagem de máquina é a automação de um processo de aprendizagem.

7. O que é aprendizado preditivo e descritivo? Uma mesma tarefa de mineração de dados pode ser dos dois tipos ao mesmo tempo?

O Aprendizado envolve generalização a partir da experiência e no caso de Aprendizado de Máquina, este busca generalizar a experiência retratada em conjuntos de dados. O aprendizado descritivo analisa os dados e identifica similaridades (agrupamentos) ou associações (regras de associações).

O aprendizado preditivo analisa os dados que representam eventos passados buscando relações entre estes que permitam predizer situações em novos dados futuros, tais como: a classificação para predições de valores discretos e a regressão para predições de valores contínuos.

Como exemplo de aprendizado descritivo, pode-se citar a associação entre eventos demandados por pacientes (de hospitais), que podem indicar possíveis relações de causa-efeito ao ser complementada com a respectivo espaço de tempo entre estes eventos associados.

Como exemplo de aprendizado preditivo, pode-se identificar potenciais clientes interessados em algum novo produto a ser divulgado.

Em geral, a natureza do problema em questão define a natureza do aprendizado a ser adotado na experimentação de mineração de dados, ou seja, em geral os problemas são preditivos ou descritivos. Porém muitas vezes as duas estratégias podem ser adotadas de forma complementar. Por exemplo, o sistema aprende sobre as preferências históricas de compras e avaliações e a partir deste aprendizado realiza recomendações de novos produtos. O aprendizado para esta recomendação pode ser obtido pela hibridização destes dois tipos de aprendizados.

8. Quais as relações, se existentes, entre mineração de dados e inteligência artificial?

A inteligência artificial discute estratégias para emular o comportamento humano, da natureza etc., como estratégia de solução para problemas computacionais. Neste sentido, se a mineração de dados busca descobrir generalizações nos dados, como uma instancia de aprendizado, a inteligência artificial apoia na pesquisa e modelagem destas estratégias para emulação. Ou seja, a inteligência artificial contribui na modelagem do "coração" dos algoritmos que descobrem padrões a partir dos dados.

Como exemplo de aplicações de inteligência artificial que estejam diretamente relacionadas com a mineração de dados, pode-se citar o desenvolvimento da robótica móvel para apoiar o processo produtivo das indústrias.

E, como exemplo da mineração de dados, cita-se a possibilidade de avaliar se um futuro cliente é um potencial adimplente ou inadimplente.

9. Como podemos identificar, e justificar, a necessidade de um "cientista de dados" no mercado de trabalho com informação?

A partir da quantidade de dados e informações, nas mais diversas áreas do conhecimento, gestores e profissionais envolvidos em tomada de decisão demandam por novas possibilidades de otimizar este processo. A integração de dados oriundos de diversas fontes, como redes sociais e bases de dados próprias ou mesmo públicas, possibilita compreender melhor todas as variáveis envolvidas.

Nesse contexto, profissionais com competência para explorar estes dados e informações, desenvolver modelos matemáticos, identificar novas oportunidades que melhor atendem às necessidades de gestão do negócio, são essenciais.

Este profissional pode ser o cientista de dados, o qual trabalha em três espaços: área do problema, o da matemática e o de tecnologia da informação. Sua função é transformar os dados disponíveis em elementos de apoio a decisão.

REFERÊNCIAS

Alcala-Fdez, J., Fernandez, A., Luengo, J., Derrac J., Garcia, S., Sanchez, S., & Herrera F. (2011). KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *J. of Mult.-Valued Logic & Soft Computing*, 17, 255–287. Retirado de http://sci2s.ugr.es/publications/ficheros/2010-JMVLSC-Alcala_Fdez-KEEL-dataset.pdf

Demsar, J., Zupan, B., Leban, G., & Curk, T. (2004). Orange: From experimental machine learning to interactive data mining. *8th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 537-539. doi: 10.1007/978-3-540-30116-5_58

Fayyad, U. M., Piatetsky Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in Knowledge Discovery and Data Mining*. California, USA: AAAI, MIT.

Fernandez, G. (2003). *Data mining using SAS application*. London: Chapman & Hall.

Hofmann, M., & Klinkenberg, R. (2013). *RapidMiner: Data mining use cases and business analytics applications*. Retirado de https://books.google.com/books?isbn=1482205491

Ingersoll, G. (2009). *Introducing Apache Mahout Scalable, commercial-friendly machine learning for building intelligent applications*. Retirado de http://www.ibm.com/developerworks/java/library/j-mahout/j-mahout-pdf.pdf

Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes. *Proceedings of EGC RNTI-E-3*, 2th, 697-702. Retirado de http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html

Seidman, C. (2001). *Data mining with Microsoft SQL Server 2000 technical reference*. Redmond: Microsoft.

Tamayo, P., Berger, C., Campos, M., Yarmus, J., Milenova, B., Mozes, A., ... , & Myczkowski, J. (2005). Oracle data mining. In Maimon, O., & Rokach, L. (Eds.). *Data Mining and Knowledge Discovery Handbook* (1315-1329). New York: Springer. doi: 10.1007/0-387-25465-X_63

Witten I. H., & Frank E. (2000). *Machine learning algorithms in Java*. Retirado de http://www.cs.waikato.ac.nz/ml/weka/

Como citar esta entrevista (ABNT):

CARVALHO, D. R.; DALLAGASSA, M. R. Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. **AtoZ**: novas práticas em informação e conhecimento, Curitiba, v. 3, n. 2, p. 82-86, jul./dez. 2014. Disponível em: <http://www.atoz.ufpr.br>. Acesso em:

***How to cite this interview* (APA):**

Carvalho, D. R., & Dallagassa, M. R. (2014). Mineração de dados: aplicações, ferramentas, tipos de aprendizado e outros subtemas. *AtoZ: novas práticas em informação e conhecimento*, 3(2), 82-86 Retrieved from http://www.atoz.ufpr.br