ISSN 21-76-8498

UTILIZAÇÃO DE INFORMAÇÕES PARA ADQUIRIR CONHECIMENTO ATRAVÉS DA MINERAÇÃO DE DADOS

Allan Carlos Claudino VILLA[1]
Eli CANDIDO JUNIOR[2]

RESUMO: Para continuarem a acirrada disputa no mercado, as empresas estão constantemente buscando maneiras de descobrir mais sobre os seus clientes e fundamentar seus conhecimentos para tomada de decisões. Todo esse conhecimento está armazenado em grandes bases de dados, o que se faz necessário a utilização de tecnologia para auxílio no processo de extração desses conhecimentos. Uma das alternativas é utilizar da mineração de dados que tem por objetivo analisar grandes volumes de informação e descobrir padrões de dados da empresa. Esse artigo aborda os conceitos de Data Mining e uma pequena amostra sobre o poder da extração do conhecimento.

Palavras-chave: Dados. Informação.Conhecimento. Mineração de Dados. KDD.

1 INTRODUÇÃO

Mídias eletrônicas, nuvens, roupas, equipamentos eletrodomésticos. Tudo o que pode se conectar à internet armazena algum tipo de informação, e não é pouca. Calcula-se que a quantia de informação no mundo dobra a cada 20 meses e o tamanho e número de bancos de dados aumentam ainda mais rapidamente, (UFMS, s.d.).

Como forma de obter uma vantagem competitiva, as empresas tendem a organizar e analisar essas informações, com o intuito de adquirir conhecimentos através de dados que foram armazenados durante anos. Para tanto, utilizam-se de tecnologias como o processo de descoberta de conhecimento - KDD (Knowledge Discovery in Databases), que tem como parte das suas etapas a Mineração de Dados. Analisar padrões de dados e até mesmo anomalias, ajudam as empresas a tomarem decisões certas na hora certa.

Com base nessa introdução, esse artigo tem por objetivo mostrar uma breve explanação sobre a mineração de dados, comparar essa técnica com outras já existentes no mercado, demonstrando assim o seu uso efetivo para aquisição de

[1]Discente do 7º termo do curso de Sistemas de Informação do Centro Universitário "Antônio Eufrásio de Toledo" de Presidente Prudente. E-mail: allanvilla@toledoprudente.edu.br
[2]Docente do curso de Sistemas de Informação do Centro Universitário "Antonio Eufrásio de Toledo" de Presidente Prudente e orientador do trabalho. E-mail: eli@toledoprudente.edu.br

conhecimento provido de informações armazenadas por empresas durantes anos, ajudando-as a tomar decisões.

2 O QUE É O DATA MINING ?

Em linhas gerais, data mining ou mineração de dados é definido por Silberschatz(2006, p.496), como sendo um "processo de analisar grandes bancos de dados de forma semi-automática para encontrar padrões úteis". Como a análise dos dados são feitas em grandes bases de bancos de dados, pode-se ocorrer o risco desses padrões permanecerem ignorados a olho humano, e isso poderia dificultar para a empresa analisar os seus dados.

Utilizando-se do KDD, em português conhecido como "*Descoberta de Conhecimento em Banco de Dados*", a mineração de dados tem a capacidade de fornecer uma previsão futura do comportamento de um cliente por exemplo, se ele recém-chegado em uma loja de departamentos gastará mais de 100 dólares em produtos, (Tan, 2009, p.3).

A descoberta de relacionamento entre os dados possibilita fazer essas previsões futuras de tendências baseadas no passado e que auxiliam as empresas para tomada de decisões, gerenciamento de informações, controle de processos entre outras aplicações.

3 EVOLUÇÃO DO DATA MINING

A mineração de dados é uma evolução natural do aumento do uso de bases de dados informatizadas para armazenar dados e fornecer respostas para analistas de negócios, (Alexander, s.d.; s.p.). A tabela a seguir mostra essa evolução.

TABELA 1 - Evolução da Mineração de Dados

Etapa evolutiva	Pergunta de negócio	Tecnologia disponível
Coleta de Dados (1960)	Qual foi a minha receita total nos últimos cinco anos?	Computadores, fitas e discos.
Acesso a Dados (1980)	Quais foram as vendas unitárias na Inglaterra em março do ano passado?	Computadores mais rápidos e baratos com maior capacidade de armazenamento; banco de dados relacional.
Data Warehousing e Apoio a Decisão(1990)	Quais foram as vendas unitárias na Inglaterra em março do ano passado? Movendo-se para Boston?	Computadores mais rápidos e baratos com maior capacidade de armazenamento; On-line analyticalprocessing (OLAP), banco de dados multidimensionais, data warehouses.
Mineração de Dados(atualmente)	O que é provável que aconteça nas vendas da unidade de Boston no próximo mês e por que?	Computadores mais rápidos e baratos com maior capacidade de armazenamento; algoritmos mais avançados.

Fonte: Alexander, s.d.; s.p.

4 AS FASES DO DATA MINING

São 3 as fases definidas por Zahedi (2008, p.30):

4.1 Exploração

Esta fase geralmente começa com a preparação de dados, que pode envolver a limpeza de dados, transformações de dados, seleção de subconjuntos de registros.

4.1 Construção do modelo e validação

Esta etapa envolve a analisar vários modelos e escolher o melhor com base em seu desempenho preditivo (ou seja , explicar a variabilidade em questão e produzir resultados estáveis em toda amostras). Isto pode soar como uma operação simples, mas, na verdade , às vezes envolve um processo muito elaborado.

4.3 Implantação

Essa etapa final envolve a utilização do modelo selecionado como melhor na etapa anterior e aplicando-o a novos dados, a fim de gerar previsões ou estimativas do resultado esperado .

5 TAREFAS DE MINERAÇÃO DE DADOS

De acordo com Fayyad (1996), o processo de mineração de dados podem executar várias tarefas, as 6 mais comuns são:

5.1 Detecção de anomalias

A identificação de registros de dados incomuns, que podem ser interessantes ou erros de dados que requerem mais investigação.

5.2 Aprendizado da regra de associação (modelagem de dependência)

Procura por relações entre variáveis . Por exemplo, um supermercado pode coletar dados sobre hábitos de compra dos clientes. Usando a aprendizagem de regras de associação, o supermercado pode determinar quais os produtos que são frequentemente comprados juntos e usar essas informações para fins de marketing.

5.3 Clustering

É a tarefa de descobrir os grupos e estruturas nos dados que estão de alguma forma ou de outra "similar", sem o uso de estruturas conhecidas nos dados.

5.4 Classificação

É a tarefa de generalizar a estrutura conhecida para aplicar a novos dados. Por exemplo, um programa de e-mail pode tentar classificar um e-mail como "legítimo" ou como "spam" .

5.5 Regressão

Tentativas para encontrar uma função que modela os dados com o mínimo de erro.

5.6 Resumo

Proporciona uma representação mais compacta do conjunto de dados, incluindo visualização e geração de relatórios.

6 LOCALIZANDO PADRÕES

Como visto anteriormente, a mineração de dados lida com a "descoberta de novas informações em função de padrões ou regras em grandes quantidades de dados", (Elmasri, 2006, p.624). Para tanto, utiliza-se de padrões que são unidades de informações que se repetem. Vamos analisar essa sequência de informações:

Sequência original: ABCXYABCZKABDKCABCTUABEWLABCWO

Observe atentamente essa sequência de letras e tente encontrar alguma coisa relevante. Veja algumas possibilidades:

- Passo 1: A primeira etapa é perceber que existe uma sequência de letras que se repete bastante. Encontramos as sequências "AB" e "ABC" e observamos que elas ocorrem com frequência superior à das outras sequências.
- Passo 2: Após determinarmos as sequências "ABC" e "AB", verificamos que elas segmentam o padrão original em diversas unidades independentes:

"ABCXY"

"ABCZK"

"ABDKC"

"ABCTU"

"ABEWL"

"ABCWO"

- Passo 3: Fazem-se agora induções, que geram algumas representações genéricas dessas unidades:

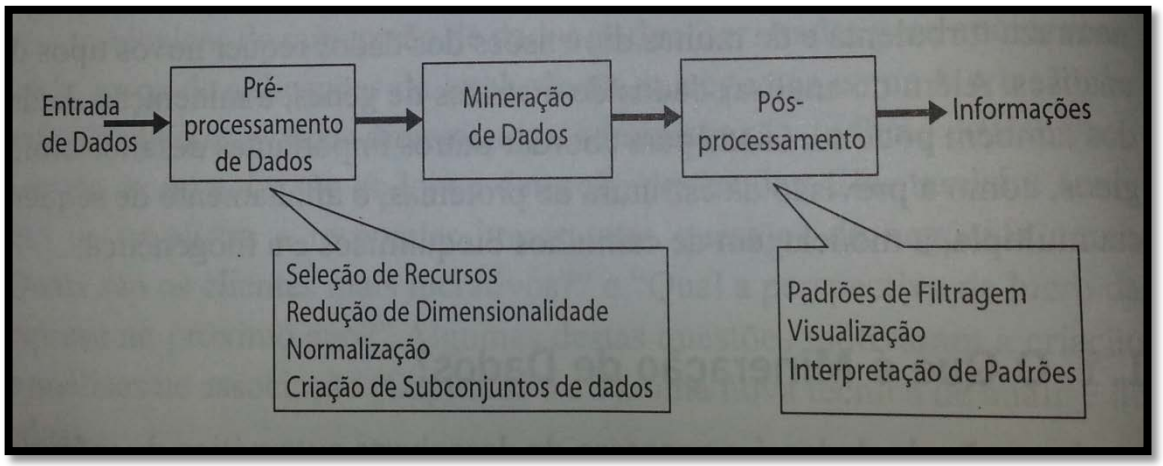
"ABC??" "ABD??" "ABE??" e "AB???", onde '?' representa qualquer letra.

No final desse processo, toda a sequência original foi substituída por regras genéricas indutivas, o que simplificou (reduziu) a informação original a algumas expressões simples. Esta explicação é um dos pontos essenciais da mineração de dados, como se pode fazer para extrair certos padrões de dados brutos. Contudo, mais importante do que simplesmente obter essa redução de informação, esse processo nos permite gerar formas de predizer futuras ocorrências de padrões, (Wikipédia).

7 Knowledge Discovery in Database x Data Mining

Descoberta de conhecimentos em banco de dados ou KDD (Knowledge Discovery in Database) é o "processo geral de conversão de dados brutos em informações úteis", (Tan, 2009, p.4). A figura 1 representa esse processo que consiste de uma série de passos de transformação, do pré-processamento dos dados até o pós-processamento dos resultados da mineração dos dados.

FIGURA 1 - O processo de descoberta de conhecimento em banco de dados (KDD)



Fonte: Tan, 2009, p.4

Com o crescimento das informações nas empresas, especialmente em áreas de negócios, o KDD tornou-se um processo muito importante para converter esta grande riqueza de dados para inteligência de negócios já que a extração manual de padrões tornou-se aparentemente impossível nas últimas décadas.

Ragel (2011), define KDD:

> "é um campoda ciência da computação, o que inclui as ferramentase teoriaspara ajudaros seres humanosem extrairinformação útil epreviamente desconhecida(ou seja,o conhecimento) degrandes coleções dedados digitalizados. KDDconsiste em várias etapas, e mineração de dados é uma dela.[3]" (Ragel, 2011)

Embora os termos KDD e Data Mining possam ser vistos como sinônimos, podem ser visto como conceitos ainda um pouco diferentes relacionados. Enquanto o KDD é um processo global de extração de conhecimentos a partir dos dados, a mineração de dados é um passo dentro do processo de KDD, que trata da identificação de padrões em dados. "Data Miningéapenas a aplicaçãode um algoritmoespecífico, baseadono objetivogeral do processo deKDD"[4], Ragel (2011).

8 WEB MINING X DATA MINING

Web Mining é definido por Krugler et al, como sendo o "uso de técnicas de mineração de dados para descobrir e extrair informações de maneira automática a partir de documentos e serviços da web.[5]" Essas informações podem ser atividades na rede, como logs de servidores e registros de navegação; gráficos com dados entre páginas, pessoas ou outra informação; conteúdo da web achado de dados nas páginas da internet ou dentro de documentos que estão disponíveis online.

Em comparação à mineração de dados, há pelo menos 3 principais diferenças a considerar:

---

[3] Traduçãolivre. *"KDD (Knowledge Discovery in Databases) is a field of computer science, which includes the tools and theories to help humans in extracting useful and previously unknown information (i.e. knowledge) from large collections of digitized data. KDD consists of several steps, and Data Mining is one of them. "*

[4] Traduçãolivre. *"DataMining is only the application of a specific algorithm based on the overall goal of the KDD process."*

[5] Traduçãolivre. *"Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services."*

- Escala: Enquanto um processamento de 1 milhão de registros de um banco de dados é um trabalho complexo para a mineração de dados, na web, até mesmo 10 milhões de páginas não seriam um grande número.
- Acesso: Enquanto no processo de data mining em uma empresa, o acesso à informação é privada e geralmente requer acesso para leitura, na web os dados são públicos e raramente pedem direito de acesso.
- Estrutura: Uma tarefa tradicional de mineração de dados obtém informações a partir de um banco de dados que fornece um nível de estrutura explícita. Já na web, o processamento é em cima de dados não estruturados ou semi-estruturados a partir das páginas da internet.

9 BIG DATA X DATA MINING

Big data é um termo para qualquer coleção de conjuntos de dados tão grande e complexo que se torna difícil para processar usando ferramentas de gerenciamento de banco de dados em mão ou aplicações de processamento de dados tradicionais. E como ja vimos anteriormente, a mineração de dados refere-se à atividade de passar por grandes conjuntos de dados para procurar informações relevantes ou pertinentes, nos caso os padrões. Em resumo, big data é o grande ativo e a mineração de dados é o manipulador desses ativos para proporcionar resultados benéficos a empresa, (Techopedia).

10 Exemplos reais de uso do Data Mining

10.1 Vestibular PUC-RJ

Utilizando as técnicas da mineração de dados, um programa de obtenção de conhecimento depois de examinar milhares de alunos forneceu a seguinte regra: se o candidato é do sexo feminino, trabalha e teve aprovação com boas notas no vestibular, então não efetivava a matrícula. Estranho, ninguém havia pensado nisso. Mas uma reflexão justifica a regra oferecida pelo programa: de acordo com os costumes do Rio de Janeiro, uma mulher em idade de vestibular, se trabalha é porque precisa, e neste caso deve ter feito inscrição para ingressar na universidade pública gratuita. Se teve boas notas provavelmente foi aprovada na

universidade pública onde efetivará matrícula. Claro que há exceções: pessoas que moram em frente à PUC, pessoas mais velhas, de alto poder aquisitivo e que voltaram a estudar por outras razões que ter uma profissão, etc.. Mas a grande maioria obedece à regra anunciada. (Wikipédia).

10.2 Walmart

Embora recente, a história do data mining já tem casos bem conhecidos. O mais divulgado é o da cadeia americana Wal-Mart, que identificou um hábito curioso dos consumidores. Há cinco anos, ao procurar eventuais relações entre o volume de vendas e os dias da semana, o software de data mining apontou que, às sextas-feiras, as vendas de cervejas cresciam na mesma proporção que as de fraldas. Crianças bebendo cerveja? Não, uma investigação mais detalhada revelou que, ao comprar fraldas para seus bebês, os pais aproveitavam para abastecer o estoque de cerveja para o final de semana. (UFMS).

12 CONCLUSÃO

Tendo em vista a auto competitividade entre as empresas, é difícil de pensar em não usar a mineração de dados que é umas das etapas do processo de KDD - Knowledge Discovery in Database, para a obtenção de conhecimentos e padrões em grandes bases de dados. Padrões de dados ou até mesmo anomalias, como vimos anteriormente, podem estar escondidas a olho humano, mas as tecnologias hoje em dia estão ai para fomentar esses dados às empresas para que as mesmas utilizem dessa informação para tomadas de decisão.

12 REFERÊNCIAS

ALEXANDER, D. ***Data Mining.*** Disponível em: <http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/#3>. Acesso em: 16 abr. 2014.

BRAGA, L. P. V. ***Introdução a mineração de dados:*** **2º edição ampliada e revisada.** Rio de Janeiro: E-Papers Serviços Editoriais, 2005. 212 p.

ELMASRI, R.; NAVATHE, S. B. ***Sistemas de banco de dados*****.** 4. ed. São Paulo: Pearson Addison Wesley, 2006. 744p.

KRUGLER, K.; SHCNEIDER C.; MAGOTRA V. ***What is web mining?***.Disponível em: <http://www.scaleunlimited.com/about/web-mining/>. Acesso em: 16 abr. 2014.

RAGEL, R. ***Difference Between KDD and Data mining.***2011. Disponível em: <http://www.differencebetween.com/difference-between-kdd-and-vs-data-mining/>. Acesso em: 16 abr. 2014.

RUSSELL, M. A. ***Mineração de dados da web social:*** **Análise de dados do Facebook, Twitter, Linkedin e outros sites de mídia social.** São Paulo, SP: Novatec Editora Ltda., 2011. 357 p.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. ***Sistema de banco de dados.*** Rio de Janeiro: Elsevier, Campus, 2006.

TAN, P.; STEINBACH, M.; KUMAR, V. ***Introdução ao DATAMINING:*** **Mineração de Dados.** Rio de Janeiro: Editora Ciência Moderna, 2009. 900 p.

TECHOPEDIA.***What is the difference between big data and data mining***. Disponível em: <http://www.techopedia.com/7/29678/technology-trends/what-is-the-difference-between-big-data-and-data-mining>. Acesso em: 16 abr. 2014.

UFMS. ***Data Mining***. Disponível em: <http://www.dct.ufms.br/~mzanusso/Data_Mining.htm>. Acesso em: 20 maio 2014.

WIKIPÉDIA. ***Data Mining***. Disponível em: <http://en.wikipedia.org/wiki/Data_Mining> Acesso em: 16 abr. 2014.

WIKIPÉDIA. ***Data Mining***. Disponível em: <http://pt.wikipedia.org/wiki/Mineraração_de_dados> Acesso em: 16 abr. 2014.

ZAHEDI, Hamed. ***Data Mining to Examine the treatment of osteomyelitis.*** University of Esfhan, Iran, 2000.