CONCEITOS DE *DATA MINING* E O USO DA LINGUAGEM PYTHON PARA A MINERAÇÃO DE DADOS

Thiago Medeiros de Souza (thsouzza@gmail.com)
Valmei Abreu Junior (valmeijr@gmail.com)
João Paulo de Lima Barbosa (joao@barbosa.net.br)

Resumo
Este documento apresenta conceitos de mineração de dados, assim como o significado de descobrimento de conhecimento por dados (KDD) e o termo *data mining*. Também expõe as razões pela adoção da linguagem de programação Python para a utilização dos métodos e técnicas de mineração.

Palavras-chave: *Data Mining*. KDD. Python.

Introdução

A informatização da sociedade tem intensificado a capacidade de gerar e coletar dados de diferentes fontes, além disso as pessoas estão, cada vez mais, inundadas pelo gigantesco número de informações. Dessa forma, este crescimento explosivo de dados cria a necessidade urgente de novas técnicas e ferramentas automatizadas que, inteligentemente, possa transformar essa vasta quantidade de dados em informação útil e conhecimento. Isso aponta para uma promissora fronteira na Ciência da Computação, chamada de *data mining* (mineração de dados) e suas aplicações.

Data mining, também referenciada como descoberta de conhecimento (*knowledge discovery from data - KDD*), é a automatização ou a extração de padrões que representam implícito conhecimento armazenado ou capturado em grandes bancos de dados, *data warehouses*, *Internet*, outros grandes repositórios de informações, ou *data streams*.

Este artigo explora alguns conceitos e técnicas de KDD e *data mining*. A mineração de dados é um campo multidisciplinar, incluindo principalmente as tecnologias de banco de dados, inteligência artificial, estatística, reconhecimento de padrões, sistemas baseados em conhecimento, recuperação da informação, computação de alto desempenho e visualização de dados. Embora muita informação já exista sobre o tema, não há uma padronização e classificação universalmente aceita sobre o assunto, de maneira a facilitar os interessados da área na condução de seus projetos de pesquisa. Uma das justificativas para isso, é justamente essa dimensão de novidade do tema e sua relevância na solução para análise de grandes volumes de dados. Além disso, o material existente sobre *data mining* possui abordagens heterogêneas, dependendo da origem ou do público-alvo a que se destina. O tema é estudado e abordado por profissionais de diversas áreas e cada uma possui aproximações específicas, adequadas para as suas necessidades.

1 Descoberta de conhecimento por dados e *data mining*

Conforme descrito, a mineração de dados é um assunto totalmente interdisciplinar podendo ser definido de diversas maneiras. Até mesmo o termo *data mining* não representa realmente todos os componentes desta área. Han (2012) exemplifica esta questão comentando sobre a mineração de ouro através da extração de rochas e areia, que é chamado de mineração de ouro e não mineração de rochas ou mineração de areia. Analogamente, a mineração de dados deveria se chamar "mineração de conhecimento através de dados", que infelizmente é um termo um tanto longo. Entretanto, uma referência mais curta, como "mineração de conhecimento", pode não enfatizar a mineração de uma enorme quantidade de dados. Apesar disso, a mineração é um termo que caracteriza o processo de encontrar uma pequena quantia de uma preciosa pepita em uma grande quantidade de matéria bruta. Nesse sentido, um termo impróprio contendo ambos *"data"* e *"mining"* se tornou popular e, como consequência, muitos outros nomes similares surgiram: *knowledge mining from data*, *knowledge extraction*, *data/pattern analysis*, *data archaeology* e *data dredging*.

Muitas pessoas tratam a mineração de dados como um sinônimo para outro termo muito popular, descoberta de conhecimento por dados (*knowledge discovery from data*), ou KDD, enquanto outros referenciam *data mining* como apenas uma etapa no processo de descoberta de conhecimento. O processo de KDD é demonstrado através da Figura 1 como uma sequência interativa e iterativa dos seguintes passos:

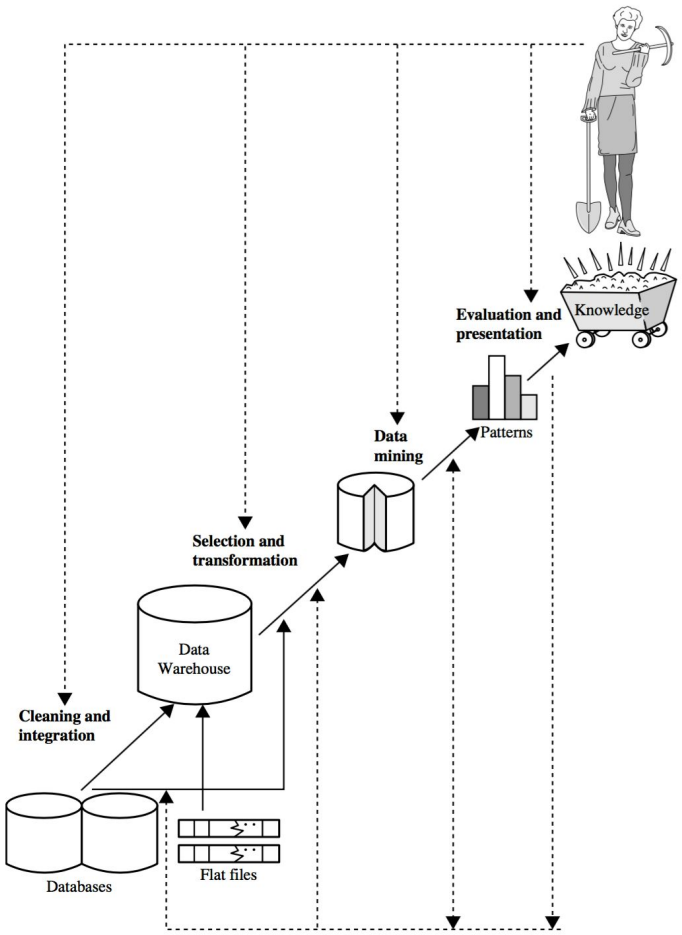

Figura 1 – *Data mining* como uma etapa no processo de descobrimento de conhecimento
Fonte: HAN, 2012

1. *Data cleaning*;
2. *Data integration*;
3. *Data selection*;
4. *Data transformation*;
5. *Data mining*;
6. *Pattern evaluation*;
7. *Knowledge presentation*.

De acordo com Brachnad & Anand (apud FAYYAD, et al., 1996), as etapas são interativas porque envolvem a cooperação da pessoa responsável pela análise de dados, cujo conhecimento sobre o domínio orientará a execução do processo. Por sua vez, a iteração deve-se ao fato de que, com frequência, esse processo não é executado de forma sequencial, mas envolve repetidas seleções de parâmetros e conjunto de dados, aplicações das técnicas de *data mining* e posterior análise dos resultados obtidos, a fim de refinar os conhecimentos extraídos.

Apesar do conceito de *data mining*, na maioria das vezes, ser utilizado pelas indústrias, mídias e centros de pesquisa para se referir ao processo de descobrimento de conhecimento considerado em sua globalidade, o termo *data mining* poderá ser usado também para indicar o quinto estágio do KDD, sendo um processo essencial no descobrindo e extração de padrões de dados. Han (2012), adota uma visão mais abrangente para a funcionalidade de mineração de dados: *data mining* é o processo de descobrimento de padrões interessantes e conhecimentos de um vasto conjunto de dados. A fonte dos dados pode ser banco de dados, *data warehouses*, a *Internet*, outros repositórios de informações, ou dados correntes em sistemas dinâmicos.

2 *Data mining*

Uma das definições, talvez, mais importante de *data mining* foi elabora por Fayyad (1996): "...o processo não-trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis".

Data mining ou mineração de dados, pode ser entendido, então, como o processo de extração de informações, sem conhecimento prévio, de algum conjunto de dados e seu uso para tomada de decisões. A mineração de dados se define através de processos automatizados de captura e análise deste conjunto de dados com a finalidade de extrair algum signficado, podendo descrever características do passado, como também para predizer futuras tendências.

Diversos métodos são usados em *data mining* para encontrar respostas ou extrair conhecimento interessante. Pode se obter os mesmos por meio dos seguintes métodos:

- Classificação: associa ou classifica um item a uma ou várias classes. Os objetivos dessa técnica envolvem a descrição gráfica ou algébrica das características diferenciais das observações de várias populações. A ideia principal é derivar uma regra que possa ser usada para classificar, de forma otimizada, uma nova observação a uma classe já rotulada;
- Modelos de Relacionamento entre Variáveis: associa um item a uma ou mais variáveis de predição de valores reais, conhecidas como variáveis independentes ou exploratórias. Nesta etapa se destacam algumas técnicas estatísticas como regressão linear simples, múltipla e modelos lineares por

transformações, com o objetivo de verificar o relacionamento funcional entre duas variáveis quantitativas, ou seja, constatar se há uma relação funcional entre X e Y;

- Análise de Agrupamento (*Cluster*): associa um item a uma ou várias classes (ou *clusters*). Os *clusters* são definidos por meio do agrupamento de dados baseados em modelos probabilísticos ou medidas de similaridade. Analisar *clusters* é uma técnica com o objetivo de detectar a existência de diferentes grupos dentro de um determinado conjunto de dados e, caso exista, determinar quais são eles;
- Sumarização: determina uma descrição compacta para um determinado subconjunto, por exemplos, medidas de posição e variabilidade. Nesta etapa se aplica algumas funções mais sofisticadas envolvendo técnicas de visualização e a determinação de relações funcionais entre variáveis. Estas funções é usada para a geração automatizada de relatórios, sendo responsáveis pela descrição compacta de um conjunto de dados;
- Modelo de Dependência: descreve dependências significativas entre variáveis. Estes modelos existem em dois níveis: estruturado e quantitativo. O nível estruturado demonstra, através de gráficos, quais variáveis são localmente dependentes. O nível quantitativo especifica o grau de dependência, utilizando alguma escala numérica;
- Regras de Associação: determinam relações entre campos de um banco de dados. Esta relação é a derivação de correlações multivariadas que permitam auxiliar as tomadas de decisão. Medidas estatísticas como correlação e testes de hipóteses apropriados revelam a frequência de uma regra no universo dos dados minerados;
- Análise de Séries Temporais: determina características sequênciais, como dados com dependência no tempo. Tem como objetivo modelar o estado do processo extraindo e registrando desvios e tendências no tempo. As séries são compostas por quatro padrões: tendência, variações cíclicas, variações sazonais e variações irregulares. Existem vários modelos estatísticos que podem ser aplicados a essas situações.

A maioria destes métodos são baseados em técnicas de *machine learning* (aprendizado de máquina), reconhecimento de padrões e estatística. Estas técnicas vão desde estatística multivariada, como análise de agrupamentos e regressões, até modelos mais atuais de aprendizagem, como redes neurais, lógica difusa e algoritmos genéticos.

Devido a vários métodos estatísticos que são aplicados no processo de *data mining*, Fayyad (1996) mostra uma relevância da estatística para o processo de extração de conhecimentos, ao afirmar que essa ciência provê uma linguagem e uma estrutura para quantificar a incerteza resultante quando se tenta deduzir padrões de uma amostra a partir de uma população.

3 Python como ferramenta para a mineração de dados

Para a análise e a interação de dados, computação exploratória e visualização de dados, a linguagem de programação Python vai, inevitavalmente, ser comparada a muitas outras, tanto no domínio de software livre, como também com linguagens e ferramentas comerciais, como R, MATLAB, SAS, Stata e outros. Atualmente, o Python possui bibliotecas

que se tornaram fortes alternativas para a tarefa de manipulação de dados. Combinado com o poder de programação que a linguagem tem, é uma excelente escolha como única linguagem para a construção de aplicações centradas em dados.

Em muitas organizações, é comum realizar pesquisas, prototipar e testar novas ideias utilizando mais de um domínio específico de linguagem computacional como MATLAB ou R e, posteriormente, estas ideias viram parte de um sistema de produção maior, escrito, por exemplo, em Java, C#, ou C++. O que se percebe é que Python não é somente uma linguagem adequada para a pesquisa e prototipagem, mas também para o desenvolvimento de sistemas. (MCKINNEY, 2013)

Devido a esta solução de apenas uma única linguagem, as organizações podem se beneficiar tendo cientistas e tecnólogos usando o mesmo conjunto de ferramentas programáticas, portanto Python é a ferramenta escolhida pela maioria desses profissionais. Essa escolha se deve, não somente à alta produtividade que a linguagem fornece, como também ao fato de ela ser uma ferramenta comum a diferentes times e organizações. (KALDERO, 2015)

Python é uma linguagem de programação livre e multiplataforma, possui uma excelente documentação e está sobre o cuidado de uma enorme comunidade, através da qual é possível obter ajuda e melhores soluções para problemas durante a codificação. Tem, como grande vantagem, a facilidade de aprendizado, porque foi desenvolvida para ser simples e descomplicada. É uma linguagem interpretada, dinâmicamente tipada, com grande precisão e sintaxe eficiente. Janert (2011), afirma que Python possui grande popularidade para analisar dados devido ao enorme poder que suas bibliotecas principais possuem (*NumPy*, *SciPy*, *pandas*, *matplotlib*, *IPython*). A linguagem apresenta alta produtividade para prototipação, desenvolvimento de sistemas menores e reaproveitáveis.

4 Caso de uso

Uma ferramenta de busca, por exemplo *Google*, recebe centenas de milhares de buscas a cada dia. Cada busca pode ser visualizada como uma transação onde o usuário descreve a informação que procura. Os padrões encontrados em buscas de usuários que provê conhecimento útil do que a leitura de dados de itens separados. Por exemplo, *Flu Trends*[1] da *Google* utiliza termos de busca específicos para indicar ocorrências de gripe. Este serviço é capaz de encontrar uma relação entre o número de pessoas que procuram informações no *Google* sobre o assunto e o número de pessoas que realmente possuem os sintomas da influenza. Um padrão surge quando ambas as buscas de gripe são reunidos. Usando o agregador de busca de dados do *Google*, *Flu Trends* consegue estimar as ocorrências de gripe com duas semanas de antecedência do que os sistemas tradicionais. (GINSBERG, et al., 2009) Este caso exemplifica como a mineração de dados pode transformar um grande conjunto de dados em conhecimento para ajudar um fator global.

Considerações Finais

O uso dos métodos de *data mining*, associados aos processos do KDD, permite às organizações trabalhar com informações implícitas e, a partir delas, buscar padrões de

[1] *Google Flu Trends* é um serviço online operado pelo *Google*. Este serviço provê estimativas sobre atividades de influenza em mais de 25 países. Através da junção de buscas, é possível fazer predições quase precisas sobre ocorrências de gripe. Este projeto se iniciou em 2008 e teve atualizações em 2009, 2013 e 2014, com a finalidade de ajudar a prever os surtos de gripes. https://www.google.org/flutrends/about/

consumo e tendências de mercado, bem como mudar estratégias adotadas, a fim de gerar lucro significativo ou minimizar perdas.

Apesar da relevância do *data mining* e de seus benefícios, o papel do analista, gestores e pesquisadores não deve ser desconsiderado, mesmo com o uso das melhores técnicas de mineração de dados, já que a aplicação e a análise adequada desses dependem de bons profissionais.

Referências

BRACHNAD, R.J. & ANAND, T. **The process of knowledge discovery in databases.** Em: FAYYAD, U.M. et al. Advances in Knowledge Discovery in Data Mining. Menlo Park: AAAI Press, 1996.

FAYYAD, U., et al. **From data mining to knowledge discovery in databases.** 1996. Acesso em 23 de outubro de 2015. Disponível em: <http://www.csd.uwo.ca/faculty/ling/cs435/fayyad.pdf>.

GINSBERG J, et al. **Detecting influenza epidemics using search engine query data.** Nature, 457:1012–1014, Fevereiro de 2009.

HAN, J., et al. **Data mining: concepts and techniques.** 3. ed. Elsevier Inc, 2012.

JANERT K. P. **Data Analysis with Open Source Tools.** O'Reilly, 2011.

KALDERO, N. **Why is Python a language of choice for data scientists?** 2015. Acesso em 27 outubro de 2015. Disponível em: <http://qr.ae/RkIeiB>.

MCKINNEY W. **Python for Data Analysis.** O'Reilly, 2013.

Bibliografia Consultada

HAND, D. J. **Data Mining: statistics and more?** The American Statistician, England, 1998. Acesso em 26 de outubro de 2015. Disponível em: <http://kdd.org/exploration_files/hand.pdf>

HETLAND L. M. **Python Algorithms - Mastering Basic Algorithms in the Python Language.** 2. ed. Apress, 2014.