

VI JORNADA ACADÊMICA 2012
Sustentabilidade e Ética: Oportunidade e desafios na formação profissional
22 a 27 de outubro
Unidade Universitária de Santa Helena de Goiás

ANALISANDO A BASE DE DADOS NA BUSCA DE INFORMAÇÕES PLAUSÍVEIS

Edmar Augusto Yokome

Mestre em Ciência da Computação, professor, Universidade Estadual de Goiás – UnU Santa Helena de Goiás, Via Protestado Joaquim Bueno, 945, edmar.yokome@gmail.com

Resumo: Este trabalho tem por objetivo introduzir o tema mineração de dados, onde será demonstrada sua importância e necessidade. Os principais tópicos a serem abordados são: sua origem e área da computação pertencente, métodos e ferramentas para a elaboração de um projeto de mineração de dados e aplicação da mineração de dados na busca de um padrão em uma base de dados.

Palavra-chave: Business Intelligence, Mineração de Dados, CRISP-DM, Algoritmo Apriori, a ferramentas WEKA e Kira.

INTRODUÇÃO

Atualmente é indiscutível a presença e necessidade de sistemas de informação nos diversos segmentos de negócio, seja: no comércio, na indústria, em instituições financeiras, em segmentos de entretenimento entre outros. Nos últimos anos diante de novas exigências e necessidades os sistemas de informação tem ganhado um nível de maturidade muito grande, como consequência disso tem surgido base de dados enormes e também estas bases armazenam vários tipos de dados, como: números, letras, caracteres especiais, imagens, sons, dados temporais, dentre outros.

A busca constante de levantar informações para tomada de decisões fez nascer uma nova área na computação chamada *Business Intelligence*, que consiste em trabalhar uma base de dados na tentativa de extrair informações úteis para que possam ser utilizados para tomada de decisões. De acordo com Rob e Coronel (2011):

“Business Intelligence (BI) é um termo utilizado para descrever um conjunto amplo, coeso e integrado de ferramentas e processos utilizados para captar, coleta, integrar, armazenar e analisar dados para a geração e a apresentação de informações que deem suporte à tomada de decisões de negócio.” Rob e Coronel (2011, pg 536)

A partir do Business Intelligence é possível construir um Data Warehousing¹ e a partir desse aplicar uma nova área da computação chamada de Mineração de Dados.

A Mineração de Dados surgiu da necessidade de buscar padrões mais efetivos em uma base de dados, onde de acordo com Tan, Steinbach e Kumar (2009) conceituam a mineração de dados como: “uma tecnologia que combina métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados”.

A Mineração de Dados encontra-se nos dias atuais com um bom nível de maturidade, fato que tem como consequências o desenvolvimento de metodologias para a criação de projetos de mineração de dados, por exemplo: as metodologias SEMMA e CRISP-DM, onde cada uma destas metodologias são divididas em etapas pré-definidas para que o minerador de dados documente um projeto de mineração de dados.

¹ De acordo com Santos e Ramos (2009, pg: 78) Data Warehousing: é um repositório construído especificamente para a consolidação da informação da organização num formato válido e consistente, permitindo aos seus utilizadores a análise de dados de uma forma seletiva.

VI JORNADA ACADÊMICA 2012
Sustentabilidade e Ética: Oportunidade e desafios na formação profissional
22 a 27 de outubro
Unidade Universitária de Santa Helena de Goiás

Outro fator que tem comprovado a maturidade da Mineração de Dados são ferramentas para que possa ser aplicada a mesma, no mercado há inúmeras ferramentas tanto proprietária como livre, e desenvolvida por empresas consagradas no mercado bem como ferramentas de cunho científico desenvolvida na academia.

Desta forma o objetivo central deste trabalho é apresentar esta nova área da computação que nos últimos anos vem ganhando espaço cada vez maior tanto na academia quanto em organizações, e assim despertar em seu público interesse nesta área promissora da computação. Será apresentado conceitos introdutórios do tema, bem como sua importância, abrangência, metodologias de desenvolvimento de projetos de mineração de dados, suas tarefas principais, ferramentas que são utilizadas e por fim uma análise dos resultados produzidos.

MATERIAL E MÉTODOS

Para haver a Mineração de Dados é necessário que uma base de dados passe por vários processos, estes processos são chamados de Descoberta de Conhecimento em Base de Dados e envolve todo o processo de coleta dos dados, tratamento dos dados, aplicação da mineração e visualização dos resultados gerados. A Figura 1 apresenta este processo com mais detalhes.

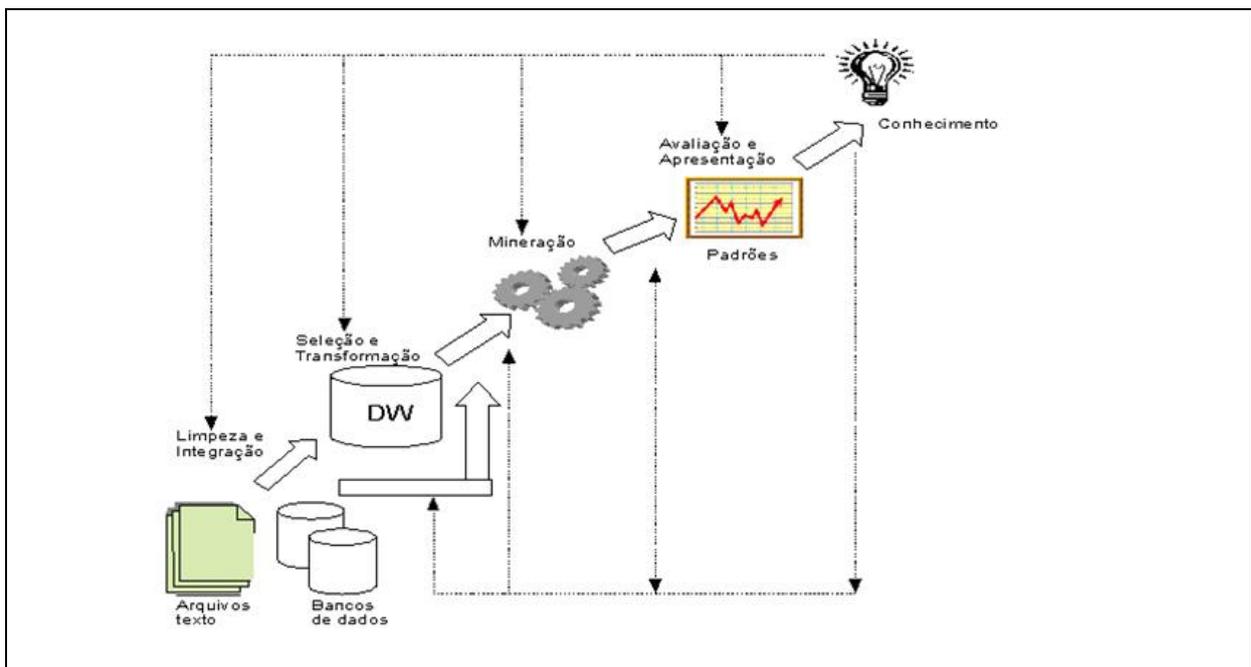


Figura 1: Descoberta de Conhecimento em Base de Dados
Fonte: Han & Kamber (2006, pg: 6, tradução nossa)

Ao analisar a Figura 1 é notado que o termo geral para se descobrir padrões em uma base de dados é chamado “Descoberta de Conhecimento em Base de Dados” onde a Mineração de Dados é uma das etapas deste processo.

As etapas anteriores a Mineração de Dados são chamadas de pré-processamento, onde consiste no emprego da eliminação ou tratamento dos dados para então aplicar o processo de Mineração de dados. As fases que estão após a Mineração de dados são chamadas

VI JORNADA ACADÊMICA 2012
Sustentabilidade e Ética: Oportunidade e desafios na formação profissional
22 a 27 de outubro
Unidade Universitária de Santa Helena de Goiás

de pós-processamento que, constituem em apresentar interfaces visuais para uma melhor representação dos resultados gerados a partir da mineração de dados.

A mineração de dados se divide em três tarefas bases e a partir destas tarefas é possível desenvolver outras tarefas fazendo uma combinação entre estas três principais, além do fato de haver uma quarta chamada detecção de anomalias que são regras que fogem aos padrões encontrados. A Figura 2 apresenta estas três tarefas bases.

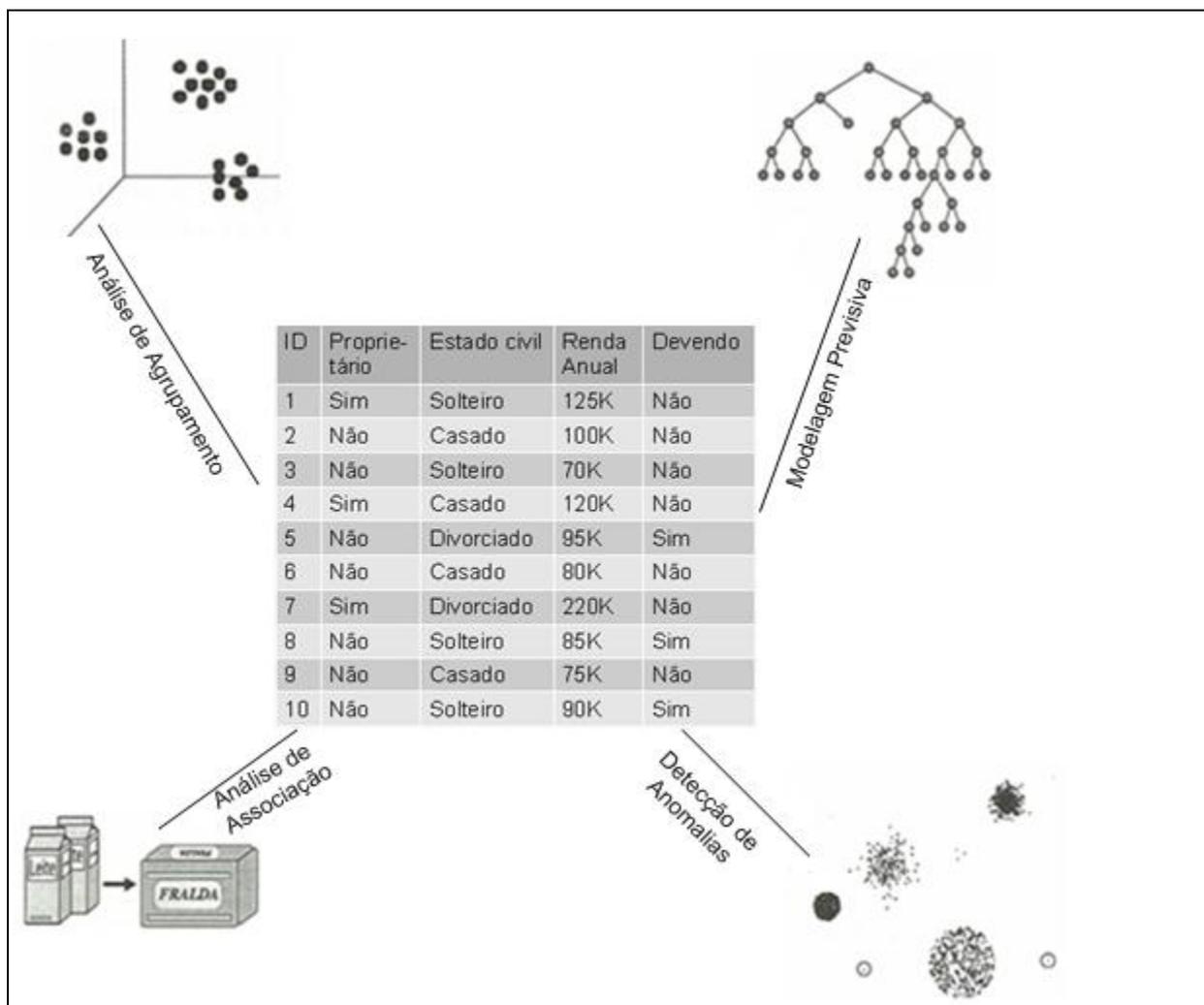


Figura 2: AS TAREFAS CENTRAIS DA MD

Fonte: Tan et al. (2009, pág: 9)

Ao analisar a Figura 2 é notado que as tarefas centrais da Mineração de dados são: Agrupamento, Associação e Clusterização². Na tarefa de Agrupamento consiste em agrupar os dados de acordo com uma característica pré-estabelecida, na tarefa de Associação tem o objetivo de associar os campos de uma determinada base de dados um com os outros campos e a Clusterização tem como objetivo criar uma árvore onde são apresentados a inter-relação entre os campos de uma base de dados. Na figura ainda há mais um item chamado anomalias que consiste em dados que fugiram um padrão pré-estabelecido.

Atualmente devido a complexidade cada vez maiores em se desenvolver projetos, seja de software, construção de casas, elaboração de um plano de negócio, dentre outros há

² Forma de classificar os dados em formato de árvores

uma grande necessidade de planejar e registrar o que foi feito e o que será, assim surgiu as metodologias. Na Mineração de Dados não é diferente, há várias metodologias, neste trabalho será abordado em específico a metodologia CRISP-DM para a elaboração de projetos de Mineração de Dados. Esta metodologia consiste em etapas pré-definidas a serem seguidas. A Figura 3 apresenta as etapas dessa metodologia.

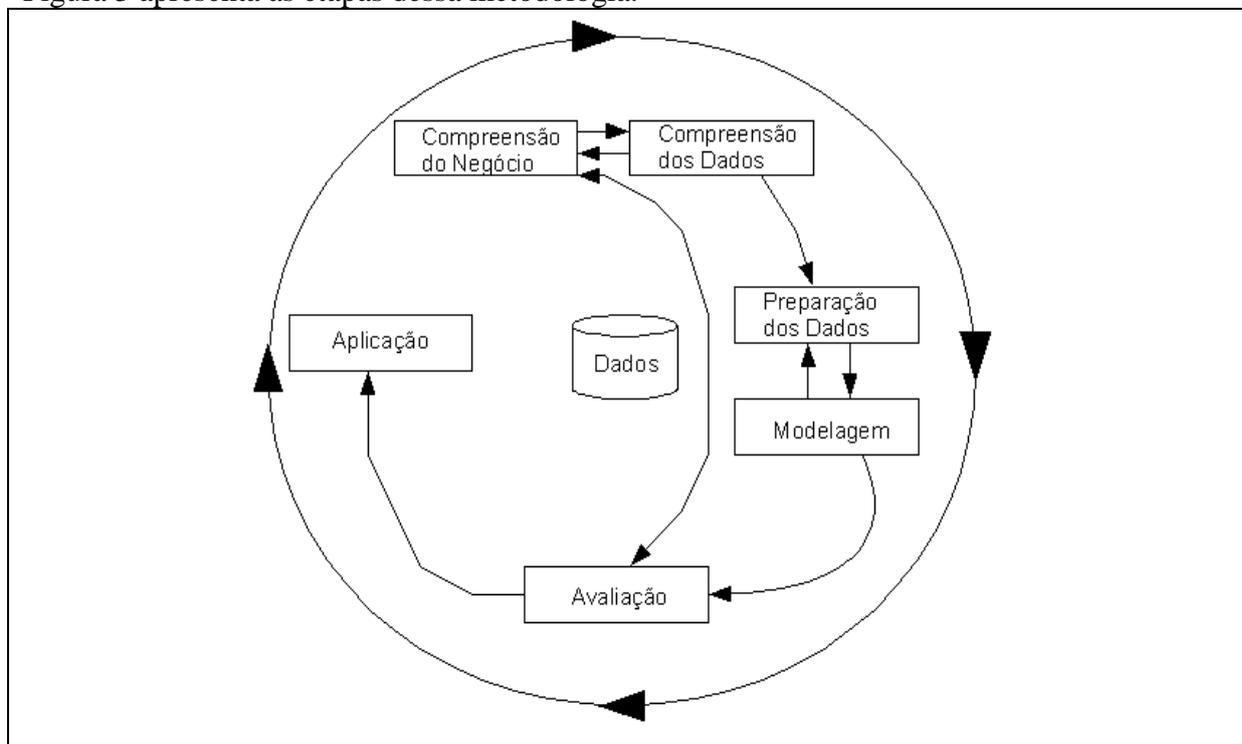


Figura 3: FASE DO MODELO DE REFERÊNCIA DO CRISP-DM
Chapman et al. (2000, pg. 13, tradução nossa)

Ao elaborar e documentar um projeto de mineração de dados é possível então utilizar uma ferramenta de mineração de dados, as ferramentas que serão apresentadas neste trabalho são as ferramentas: Weka³ e Kira⁴.

RESULTADOS E DISCUSSÕES

Com a apresentação deste trabalho será possível aos seus participantes interessar-se por uma área da computação que está em franca expansão e não tem sido tão utilizada por falta de profissionais que possam trabalhar com esta tecnologia, pois hoje em dia são raras as notícias de instituições que promovem cursos profissionalizantes sobre este tema.

CONCLUSÕES

O tema mineração de dados é um tema relativamente novo na área da Ciência da Computação, porém tem um bom nível de maturidade. Este trabalho teve como objetivo abordar esta área apresentando conceitos, ferramentas e metodologias que possam ser trabalhados este tema.

REFERÊNCIAS

³ <http://www.cs.waikato.ac.nz/ml/weka/>

⁴ Programa de Mestrado de Ciência da Computação da UNIMEP (2009)

VI JORNADA ACADÊMICA 2012
Sustentabilidade e Ética: Oportunidade e desafios na formação profissional
22 a 27 de outubro
Unidade Universitária de Santa Helena de Goiás

CHAPMAN, P.; CLINTON, J.; KERBER, R.; KHABAZA, T.; REINARTZ, T.; SHEARER, C. CRISP-DM 1.0 Step-by-step Data Mining Guide. 2000. Disponível em: <<http://www.crisp-dm.org/download.htm>>. Acesso em: 08 set. 2012.

HAN, J.; KAMBER, M. *Data Mining: Concepts and Techniques*. Ed: Elsevier. Second Edition, 2006.

SANTOS, M. Y. ; RAMOS, R. Business Intelligence – Tecnologias da Informação na Gestão de Conhecimento, FCA, 2ª Edição, 2009.

TAN, P.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining (Mineração de Dados)*, Ciência Moderna, 2009.

ROB, Peter; CORONEL, Carlos. Sistemas de Banco de Dados: Projeto, Implementação e Administração. Cengage Learning. Tradução da 8ª Edição, São Paulo, 2011.