7ª MOSTRA ACADÊMICA UNIMEP
Tema:
CIÊNCIA, TECNOLOGIA E INOVAÇÃO:
A universidade e a construção do futuro

17º Congresso de Iniciação Científica

AVALIAÇÃO DA FERRAMENTA KIRA COMO APLICAÇÃO DO PROCESSO DE KDD E DE TÉCNICAS DE MINERAÇÃO DE DADOS

Autor(es)

MIRELA TEIXEIRA CAZZOLATO

Orientador(es)

MARINA TERESA PIRES VIEIRA

Apoio Financeiro

PIBIC/CNPQ

1. Introdução

Grandes empresas, como bancos e redes de supermercado, coletam diariamente uma grande quantidade de dados. Analisar esses dados sem o auxílio de ferramentas computacionais, devido ao seu volume, torna-se inviável.

Uma maneira automatizada de auxiliar a exploração dessa grande quantidade de dados é o Processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases – KDD*). De acordo com Fayyad, Piatetsky-Shapiro e Smith (1996), a descoberta de conhecimento consiste em um processo não trivial de identificação de padrões contidos nos dados e que sejam válidos, novos, potencialmente úteis e compreensíveis.

Segundo Han e Kamber (2006), o processo de descoberta de conhecimento pode ser dividido em sete etapas: **limpeza**, **integração**, **seleção**, **transformação**, **mineração de dados**, **avaliação dos padrões** e **apresentação do conhecimento**.

A mineração de dados é considerada a etapa mais importante do processo de KDD e tem como objetivo principal transformar grandes quantidades de dados em conhecimento útil. Para que isso ocorra, é utilizada a tarefa de mineração que mais se adeque ao problema tratado.

As tarefas de mineração de dados referem-se aos tipos de padrões que podem ser minerados. Em geral, as tarefas de mineração podem ser classificadas como descritivas e preditivas. As tarefas descritivas caracterizam as propriedades gerais dos dados no banco de dados e as preditivas realizam inferências nos dados correntes para fazer predições. Dentre as principais tarefas de mineração destacam-se a **classificação**, a **clusterização** e as **regras de associação**, discutidas a seguir:

Conforme Han e Kamber (2006), a classificação divide os dados em classes, de acordo com os valores de seus atributos. A tarefa é composta por dois passos principais: primeiramente constrói-se um modelo de classificação que descreve um conjunto pré-determinado de classes de dados; no segundo passo, o modelo é utilizado para classificar os dados desejados.

A clusterização, também conhecida como agrupamento, reune elementos semelhantes em grupos ou classes, de acordo com suas características, baseando-se no princípio de que a semelhança dos elementos de uma mesma classe seja maior que a semelhança entre os elementos de classes diferentes (HAN; KAMBER, 2006). Essa tarefa também é utilizada como pré-processamento para a realização de outras tarefas.

As regras de associação buscam encontrar conjuntos frequentes de itens que ocorram simultaneamente numa base de dados. A descrição formal do problema, segundo Agrawal e Srikant (1994), é a seguinte: seja I = {i1, i2, ..., im} um conjunto de itens e D um conjunto de transações T, onde cada transação T é um conjunto de itens tal que T está contido ou é igual a I. Diz-se que uma transação T contém X, um subconjunto de itens de I, se X está contido ou é igual a T. Uma regra de associação é uma implicação da forma X

=> Y, onde X e Y estão contidos em I e a intersecção de X com Y é vazia. A regra X => Y ocorre no conjunto D com confiança c se c% das transações em D que contém X também contém Y. A regra X => Y tem suporte s no conjunto de transações D se s% das transações em D contém X união com Y. Em outras palavras, o suporte da regra é a razão entre o número de transações em D que contém X e Y, e o número total de transações de D. A confiança é a razão entre o número de transações que contém X e Y, e o número de transações que contém X.

É comum a utilização de ferramentas que auxiliem a realização do processo de descoberta de conhecimento. Atualmente existem diversas ferramentas disponíveis para a execução de diferentes tarefas de mineração, mas elas não eliminam a necessidade do usuário possuir um grande grau de conhecimento sobre o processo de KDD e suas etapas.

A ferramenta Kira surge como uma opção de abstrair boa parte do conhecimento exigido do analista de dados para a execução do processo de KDD. Ela foi desenvolvida com o intuito de ensinar passo a passo o processo de descoberta de conhecimento e a aplicação da etapa de mineração de dados.

A arquitetura da Kira, como pode ser observada na Figura 1, é composta por três módulos principais: **módulo de apoio à origem**, **módulo de apoio à preparação** e **módulo de apoio à análise**. Segundo Mendes (2009), o módulo de apoio à preparação de dados é composto por submódulos que orientam as etapas de integração, limpeza, seleção e transformação dos dados. O módulo de apoio à análise é formado por: apoio à mineração e apoio à avaliação. Cada módulo disponibiliza diversas facilidades para auxiliar o usuário tanto para preparar os dados como para aplicar os algoritmos de mineração e avaliar os resultados obtidos ao longo do processo de descoberta de conhecimento.

Na utilização da Kira, inicialmente o usuário deve escolher a fonte de dados a ser utilizada na etapa de mineração. Após escolhida, conforme dito por Mendes (2009), o usuário é orientado passo a passo a fornecer determinadas informações, escolher e moldar os dados de interesse, aplicar o algoritmo minerador e analisar o resultado obtido.

Ao término de cada etapa a ferramenta Kira orienta ao usuário qual a próxima etapa do processo a ser realizada, instruindo-o a executar satisfatoriamente o processo envolvido na mineração de dados.

2. Objetivos

Este trabalho teve como objetivo avaliar a metodologia adotada na ferramenta Kira, para a aplicação do processo de KDD, quanto aos aspectos de facilidade de uso, objetividade das instruções para auxiliar as etapas do processo de descoberta de conhecimento e atendimento dos objetivos propostos pela ferramenta, isto é, conseguir ensinar a utilizar na prática os conceitos envolvidos no processo de mineração de dados. Com base em estudos realizados sobre o processo de KDD, procurou-se avaliar a ferramenta Kira por meio de testes com diferentes conjuntos de dados, de modo a observar seu comportamento e elaborar um conjunto de sugestões para sua melhoria.

3. Desenvolvimento

Foi obtido prévio conhecimento sobre a ferramenta Kira em (MENDES, 2009), e sobre o Processo de Descoberta de Conhecimento em Bases de Dados, de modo que possibilitasse a utilização da ferramenta.

Em (MENDES, 2009) é descrito o funcionamento da Kira utilizando o banco de dados que acompanha sua instalação, como padrão; dessa forma, no primeiro contato com a ferramenta, essa mesma base de dados foi utilizada diversas vezes, facilitando inicialmente o uso da ferramenta. Essa base contém um conjunto de informações de alunos inscritos para um congresso de tecnologia realizado em 2007 na cidade de Mococa – SP.

Após o primeiro contato, a ferramenta foi utilizada com uma base de dados diferente, a *Census-Income Database*, extraída de (HEITTICH; BAY, 1999). Assim como foi feito com a base de dados padrão, a *Census-Income Database* foi utilizada diversas vezes na Kira, permitindo que se pudesse avaliar a ferramenta com detalhes e elaborar sugestões para sua melhoria. Essa base de dados contém uma série de dados demográficos e de emprego dos anos de 1994 e 1995.

4. Resultado e Discussão

O conhecimento obtido com os estudos realizados foi suficiente para que se pudesse alcançar os objetivos propostos. Os estudos sobre

a ferramenta Kira (MENDES, 2009), possibilitaram o uso da ferramenta sem maiores problemas. Além disso, com a utilização da Kira obteve-se uma boa compreensão do Processo de KDD na prática, pois se trata de uma ferramenta intuitiva, de fácil manuseio.
Como previsto, ao longo do projeto de pesquisa, com a utilização da Kira, foram elaboradas sugestões de melhorias para a ferramenta. Conforme já citado, elas foram elaboradas com base na facilidade de uso, na objetividade das instruções para auxiliar as diversas etapas do processo de KDD e no atendimento da ferramenta aos objetivos propostos.
Um exemplo de alteração sugerida é descrito a seguir:
Problema: A ferramenta disponibiliza uma ajuda lateral, com o intuito de ajudar o usuário durante o processo de descoberta de conhecimento. Porém, na ajuda lateral da etapa de seleção dos dados, a descrição do item "Restrições" necessita de correção. Está descrita da seguinte forma: "Todas as tabelas pertencer ao banco de dados que foi selecionado. Para selecionar outro banco de dados basta clicar no item sobre o nome do banco de dados".
Sugestão de solução: não estão especificadas quais tabelas devem pertencer ao banco de dados, além da frase estar confusa. Dessa forma, há a necessidade de correção da descrição do item "Restrições", de modo a torná-la mais clara.
As sugestões elaboradas durante o projeto foram acatadas e estão sendo incorporadas na ferramenta.

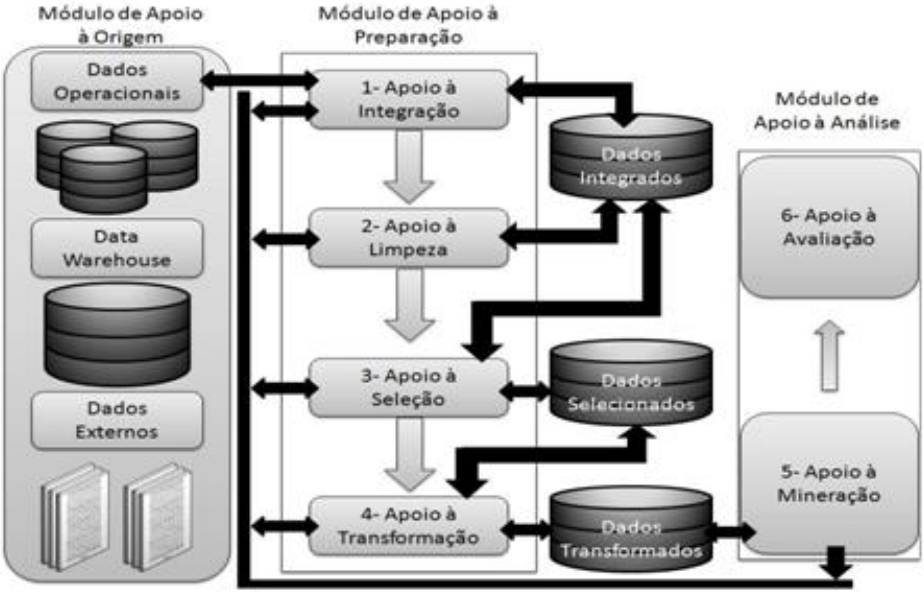
5. Considerações Finais

Durante o projeto foi utilizada a ferramenta Kira sem um grande conhecimento sobre o processo de KDD e suas etapas, de modo a avaliar o quão didática seria a ferramenta. Dessa forma, observou-se que a ferramenta contribui satisfatoriamente ao usuário com pouco conhecimento para a realização do processo de descoberta de conhecimento. A Kira mostrou-se intuitiva, tanto em relação à sua interface como à sua utilização na aplicação do processo de KDD, apoiando o usuário a cada etapa.
Existem outros projetos em andamento que visam aperfeiçoar a ferramenta Kira com:
• Tratamento de regras da associação multirelacional;
• Incorporação da tarefa de classificação;
• Modelagem e criação de base de conhecimento para o Processo de Mineração de Dados.

Referências Bibliográficas

AGRAWAL, R.; SRIKANT, R. ***Fast algorithms for mining association rules***. In Proc. Of the Int'l Conf. on Very Large Databases, Santiago de Chile, Chile, 1994.
FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. ***From Data Mining to Knowledge Discovery: An Overview***. In: Advances in Knowledge Discovery and Data Mining, AAAI Press/ The MIT Press, MIT, Cambridge, Massachusets, England, 1996.
HAN, J.; KAMBER, M. Data Mining - Concepts and Techniques. 2a edição. Nova York: Morgan Kaufmann, 2006.
HETTICH, S.; BAY, S. D. (1999). *The UCI KDD Archive* . Irvine, CA: University of California, Department of Information and Computer Science. Acesso em: 10 de ago. 2009.
MENDES, E. F. **Automatização da técnica de mineração de dados auxiliada por guias**. 2009. 115 f. Dissertação (Programa de Mestrado em Ciência da Computação) – Faculdade de Ciências Exatas e da Natureza, Universidade Metodista de Piracicaba, Piracicaba, 2009.
RIBEIRO, M. X. **Mineração de Dados em Múltiplas Tabelas Fato de um Data Warehouse**. 2004. 130 f. Dissertação (Programa de Pós-Graduação em Ciência da Computação) – Centro de Ciências Exatas e Tecnologia, Universidade Federal de São Carlos, São Carlos, 2004.

Anexos


Módulo de Apoio à Origem
Módulo de Apoio à Preparação
Módulo de Apoio à Análise
Dados Operacionais
Data Warehouse
Dados Externos
1- Apoio à Integração
2- Apoio à Limpeza
3- Apoio à Seleção
4- Apoio à Transformação
Dados Integrados
Dados Selecionados
Dados Transformados
6- Apoio à Avaliação
5- Apoio à Mineração