6º Congresso de Pós-Graduação

METODOLOGIA E AUTOMATIZAÇÃO DO PROCESSO DE APLICAÇÃO DA MINERAÇÃO DE DADOS

Autor(es)

EDUARDO FERNANDO MENDES

Orientador(es)

MARINA TERESA PIRES VIEIRA

1. Introdução

Nos últimos anos tem ocorrido o crescimento rápido do volume e da dimensão das bases de dados, causado, principalmente, pela queda no custo de armazenamento dos dados.

A análise desse grande volume de dados em busca de informações e conhecimento começou a ficar humanamente impossível utilizando os sistemas existentes.

Para atender as novas exigências identificou-se então a necessidade de evolução das tecnologias. A criação de métodos e ferramentas para automatizar o processo torna-se evidente. Neste cenário surge, no final da década de 1980, um novo ramo da computação, a descoberta de conhecimento em bases de dados. Segundo Fayyad, Piatetsky-Shapiro, Smith (1996) o processo de descoberta de conhecimento em bases de dados (KDD - Knowledge Discovery in Databases) é definido como sendo um processo interativo e iterativo, não trivial de identificação de padrões válidos, novos, potencialmente úteis, compreensíveis e embutidos nos dados, envolvendo numerosos passos, com muitas decisões sendo feitas pelo usuário.

Segundo Han e Kamber (2006) o conjunto de atividades do processo de descoberta de conhecimento em bases de dados é composto de sete etapas: limpeza dos dados, integração dos dados, seleção dos dados, transformação dos dados, mineração dos dados, avaliação dos padrões, apresentação do conhecimento.

O KDD é considerado um padrão, permitindo a realização do processo de forma mais rápida, confiável e com melhor controle gerencial.

Para a execução do processo KDD o analista de dados precisa ter muito conhecimento sobre como preparar os dados e as técnicas de mineração. Este, talvez, seja o problema da mineração de dados não ter sido adotada em grande escala pelas empresas.

Na tentativa de tornar a descoberta de conhecimento em base de dados realidade e à medida que o mercado mostrou interesse pela mineração de dados torna-se necessária a criação de uma nova abordagem para demonstrar o quanto a mineração era suficientemente madura para ser adaptada como parte do processo de negócio.

Pensando nisso, em meados da década de 90 empresas reuniram seus esforços para a criação de um processo padrão de mineração de dados denominado CRISP-DM (Cross-Industry Standard Process for Data Mining). Cada empresa contribuiu com suas experiências no contexto da aplicação da mineração de dados para o desenvolvimento de um padrão que pudesse ser aplicado a qualquer tipo de problema.

Segundo Chapman (2000) a metodologia CRISP-DM é descrita em termos de um modelo hierárquico de processos, que consiste num conjunto de tarefas representadas por quatro níveis de abstração (do mais geral para o mais específico): Fases, Tarefas Genéricas, Tarefas Especializadas e Instâncias de Processos.

O modelo do processo de mineração de dados proporciona uma visão do ciclo de vida do projeto, contendo as fases de um projeto, suas respectivas tarefas e as relações entre essas tarefas. O ciclo de vida de um projeto de mineração de dados consiste de seis fases: Entendimento do Negócio (Business Understanding), Entendimento dos Dados (Data Understanding), Preparação dos Dados (Data Preparation), Modelagem (Modeling), Avaliação (Evaluation) e Aplicação (Deployment).

Para execução da fase de modelagem é necessário que o analista de dados tenha conhecimento sobre as técnicas de mineração de dados e o formato que cada uma delas exige. A utilização da metodologia não garante resultados; é uma forma de disciplinar o processo de mineração de dados.

A mineração de dados (DM - Data Mining), tem atraído muito a atenção da indústria da informação e da sociedade como um todo. Sendo considerada a etapa mais importante do processo KDD, muitas vezes os dois conceitos se confundem e são utilizados de forma errada. Segundo Fayyad, Piatetsky-Shapiro, Smith (1996) o termo KDD refere-se a todo o processo de descoberta de conhecimentos úteis em base de dados, e a mineração de dados refere-se particularmente a uma etapa deste processo.

Mineração de dados é a aplicação de um algoritmo específico para a extração de padrões nos dados.

Para determinar se uma determinada tarefa de mineração de dados gerou regras úteis utilizam-se medidas, conhecidas como medidas de interesse, para determinar o nível de interesse de uma determinada regra gerada. As regras geradas pelas tarefas precisam estar dentro do valor mínimo estabelecido para as medidas de interesse. Caso uma regra seja encontrada dentro dos padrões mínimos fornecidos pelo analista de dados é considerada como uma regra de interesse (HAN, KAMBER, 2006).

De acordo com Han e Kamber (2006), as tarefas de associação, classificação e agrupamento estão entre as mais importantes da mineração de dados.

Uma vez que as tarefas e aplicações de mineração de dados são amplas e diversas, várias ferramentas foram surgindo com interfaces flexíveis e interativas.

As ferramentas disponíveis para auxiliar o analista de dados não conseguem eliminar o alto grau de conhecimento exigido para realizar o processo de mineração. Desta forma fica limitada a adoção desta tecnologia em grande escala.

Mesmo, tendo disponível algumas ferramentas para automatizá-lo, as mesmas, não conseguiram eliminar a falta de conhecimento do analista de dados.

2. Objetivos

O objetivo do trabalho é desenvolver uma metodologia de apoio ao processo de mineração de dados, dando subsídios suficientes para orientar o analista de dados em como executar o processo de mineração de dados. Após o desenvolvimento da metodologia será implementado um módulo que será parte de uma ferramenta para auxílio, ao analista de dados, à aplicação do processo de descoberta de conhecimento em bases de dados.

Seguem algumas questões a serem consideradas para a definição da metodologia.

As tarefas de mineração de dados que serão tratadas pelo módulo são: regras de associação, classificação e agrupamento. Para as regras de associação, que consistem em encontrar padrões em um conjunto de dados, o analista de dados será orientado a preparar adequadamente os dados conforme a necessidade dos algoritmos de mineração, a definir os valores de suporte e confiança que melhor represente os objetivos de mineração e a avaliar as regras candidatas após a execução do algoritmo.

Para a tarefa de classificação, que consiste em predizer o valor que um determinado atributo assumirá, o analista de dados será orientado a preparar adequadamente os dados conforme a necessidade dos algoritmos de mineração, a selecionar o atributo classe do conjunto de dados disponíveis, a definir a distribuição dos dados em base de dados de treinamento e base de dados de teste. O analista de dados também deverá saber como medir a eficiência do algoritmo, acurácia preditiva, sobre a base de dados de teste.

Para a tarefa de agrupamento, que consiste em organizar um conjunto de dados em classes similares, o

analista de dados será orientado a preparar adequadamente os dados conforme a necessidade dos algoritmos de mineração e será orientado a definir o melhor método de agrupamento a ser utilizado.

3. Desenvolvimento

Para alcançar os objetivos do trabalho estão sendo realizados estudos sobre o processo de KDD, metodologia CRISP-DM e o processo de mineração de dados. Após essa etapa, tendo conhecimento sobre os conceitos necessários, será definida uma metodologia de apoio ao processo de mineração de dados. Após o desenvolvimento da metodologia será implementado um módulo que será parte de uma ferramenta para auxílio, ao analista de dados, à aplicação do processo de descoberta de conhecimento em bases de dados.

Essa ferramenta deverá orientar o analista de dados em como realizar cada etapa do processo de KDD. A ferramenta é composta de três módulos principais que são (figura 1): Módulo de Apoio à Origem, Módulo de Apoio à Preparação, Módulo de Apoio à Análise. O Módulo de Apoio à Preparação é formado por: Apoio à Integração, Apoio à Limpeza, Apoio à Seleção e Apoio à Transformação. O Módulo de Apoio à Análise é formado por: Apoio à Mineração e Apoio à Avaliação. Cada módulo tem disponível uma série de facilidades que irão auxiliar o analista de dados a preparar os dados, aplicar os algoritmos de mineração e avaliar os resultados obtidos.

Os recursos computacionais utilizados para o desenvolvimento e teste do módulo de apoio à análise da ferramenta são:

- Linguagem de programação: Java 2 SDK – Standard Edition, versão 1.6.0.1.
- Plataforma de desenvolvimento Java - Eclipse SDK – Versão 3.3.1.1.
- Sistema gerenciador de banco de dados – Firebird – Versão 2.0.4.
- Base de dados de testes de uma aplicação real – inscrição de alunos para um congresso de tecnologia – composta por mais de 2000 transações.

4. Resultado e Discussão

Após o desenvolvimento da metodologia, o resultado a ser alcançado será a implementação da metodologia em uma ferramenta. Com a sua devida automatização pretende-se reduzir a grande quantidade de detalhes sobre mineração de dados que o analista de dados precisa conhecer, assim como reduzir o tempo de execução da etapa de mineração de dados. A pesquisa busca alcançar os seguintes resultados:

- Redução do alto grau de conhecimento exigido pelo analista de dados para tornar realidade a etapa de mineração de dados;
- Auxílio ao analista de dados na escolha da tarefa de mineração de dados conforme os objetivos da mineração;
- Auxílio ao analista de dados na preparação dos dados conforme a tarefa de mineração de dados selecionada;
- Auxílio ao analista de dados na parametrização do algoritmo minerador;
- Redução do tempo gasto com a etapa de mineração de dados.

5. Considerações Finais

A ferramenta encontra-se parcialmente implementada, com previsão de um protótipo pronto para testes em novembro de 2008.

Referências Bibliográficas

CHAPMAN, P. CRISP-DM 1.0: Step-By-Step Data Mining Guide. [S.I.]: 2000. Disponível em:

CRISP-DM Consortium. CRISP-DM – Cross Industry Standard Process for Data Mining. Disponível no site CRISP-DM (2000). URL: http://www.crispdm.org/.

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.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMITH, P. Knowledge Discovery and Data Mining: Towards a Unifying Framework. In Proceedings of the Second International Conference on Data Mining and Knowldege Discovery, AAAI Press, Menlo Park, US; 1996.

HAN, J., FU, Y., WANG, W., CHIANG, J. DBMiner: A System for Mining Knowledge in Large Relational Databases. In: Proceedings of the International Conference on Knowledge Discovery in Databases, Portland, 1996.

HAN, J.; KAMBER, M. Data Mining - Concepts and Techniques. 2a edição. Nova York: Morgan Kaufmann, 2006.

Anexos

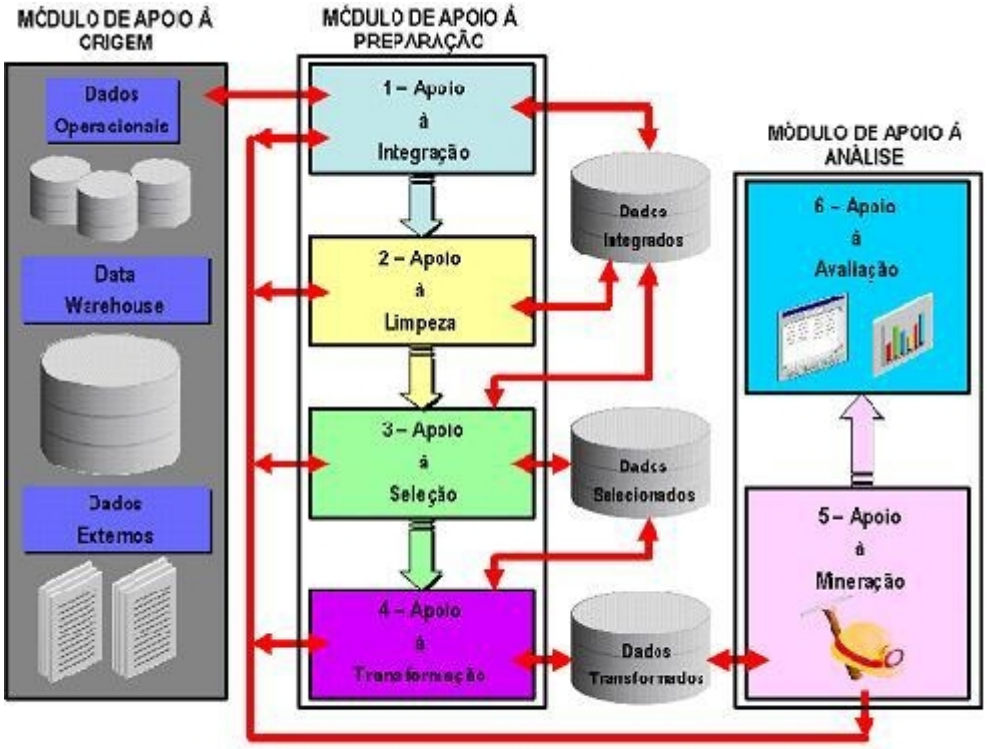

FIGURA 1 – ARQUITETURA DA FERRAMENTA