

# FERRAMENTA DE SUPORTE AO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS RELACIONAIS.

Fernando Takeshi Oyama, Carlos Roberto Valêncio – Ciência da Computação – Bacharelado em Ciência da Computação – Departamento de Ciência da Computação e Estatística – Instituto de Biociências, Letras e Ciências Exatas – Campus de São José do Rio Preto.

A evolução da tecnologia de informação, bastante evidente nas últimas décadas, tem disponibilizado uma diversidade de ferramentas atendendo as mais diferentes áreas de aplicação, tais como administração de negócios, engenharias, ciências, entre outras. Em meio a esse avanço, as tecnologias de captação e armazenamento de dados têm uma posição de destaque, pois a sua utilização permitiu que a coleta e a manipulação da volumosa quantidade de dados gerada por essas aplicações pudessem ser efetivadas. Porém, o acesso e a possibilidade de armazenamento dos dados, bem como a disponibilidade dessas ferramentas, acabaram criando grandes repositórios não-analisáveis, caracterizando um cenário no qual se tem riqueza em dados, mas pobreza em informações.

A fim de preencher essa lacuna tecnológica, surgiu o processo de Descoberta de Conhecimento em Bases de Dados (KDD *Knowledge Discovery in Databases*), como uma área de estudos multidisciplinar que visa o desenvolvimento de ferramentas e técnicas para a transformação de dados brutos em informações úteis. O KDD é um amplo processo que envolve, basicamente, as etapas de limpeza e integração dos dados, seleção dos dados, *data mining*, avaliação dos padrões e visualização do conhecimento. Dentre elas se destaca a etapa de *data mining*, o cerne do processo de descoberta de conhecimento, na qual são efetivamente aplicados os métodos para a extração de padrões úteis e implícitos da base de dados.

Uma dificuldade observada nos métodos típicos de *data mining* é a necessidade dos dados estarem armazenados em uma única tabela, o que, em geral, não ocorre nas aplicações atuais. Diante desse fato, o presente trabalho tem como objetivo a construção de uma ferramenta que forneça o suporte necessário às principais etapas do KDD em bases de dados relacionais. Com isso, objetiva-se a aplicação do processo diretamente sobre múltiplas tabelas ou relações, sem a necessidade da operação de junção de diversas tabelas para uma única.

A ferramenta proposta pode ser estruturada em camadas com funcionalidades bem definidas:

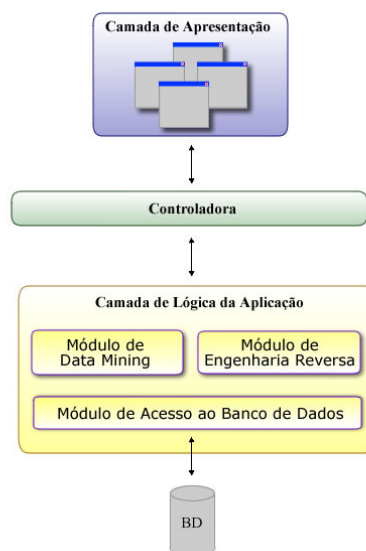


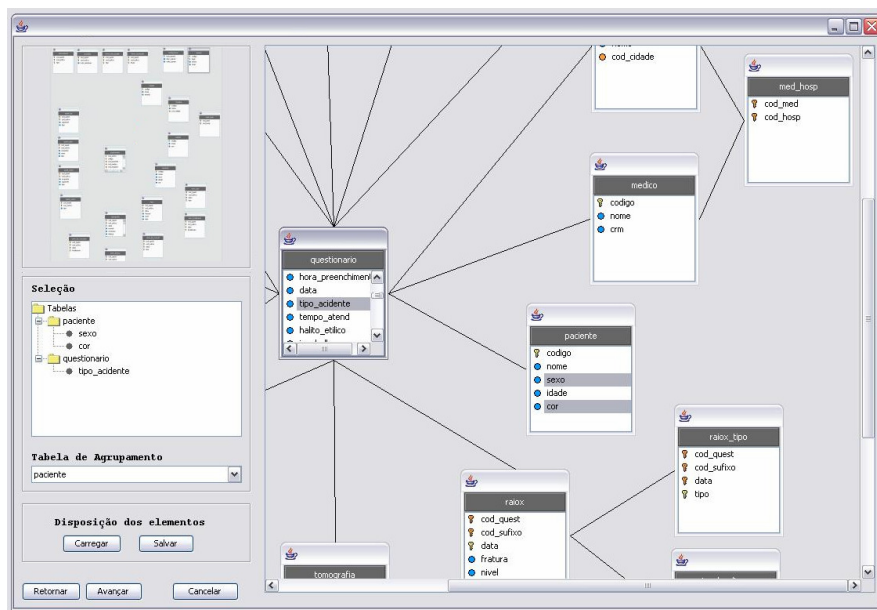
Figura 1 Estruturação do sistema

A “Camada de apresentação” é composta pelos objetos relacionados à interface do sistema, tais como os formulários e é por meio desses componentes que o usuário interage com o sistema no processo de KDD. Disso decorre a preocupação em atrelar características de um bom projeto de interface a essa camada, tais como usabilidade, entrada de dados consistentes, mensagens de erros esclarecedoras etc.

A camada “Controladora”, por sua vez, desempenha a função de interligar a “Camada de apresentação” e a “Camada de lógica da aplicação”. Tal camada é importante para garantir um fraco acoplamento entre a interface e os módulos funcionais.

Por último, a “Camada de lógica da aplicação” reúne os algoritmos, as classes e os demais objetos que formam o domínio da aplicação e que são responsáveis pela funcionalidade do sistema, por meio da disponibilidade de técnicas e métodos para as camadas superiores.

A primeira das etapas do KDD suportada pela ferramenta é a seleção de dados, na qual o usuário define os atributos que são relevantes à análise. Para isso, é disponibilizada uma interface gráfica que permite a visualização do esquema da base de dados (tabelas, atributos, relacionamentos etc.). Tal funcionalidade é realizada com a utilização do processo de engenharia reversa da base de dados, o qual recupera as informações estruturais existentes nos metadados do gerenciador de banco de dados.



**Figura 2** Etapa de seleção de dados

A etapa seguinte corresponde à aplicação de uma técnica de *data mining* para a extração de padrões. Tais técnicas são alimentadas com as informações estruturais provenientes da engenharia reversa da base de dados e também com o conjunto de tabelas e atributos previamente selecionados pelo usuário na etapa de seleção de dados. Na figura a seguir, apresenta-se a interface da ferramenta que permite a escolha do algoritmo de *data mining* e o ajuste das medidas de interesse.



**Figura 3** Etapa de *data mining*

Após a execução do algoritmo de *data mining*, a ferramenta apresenta um conjunto de resultados obtidos no processo de extração de conhecimento. No caso do algoritmo utilizado, há a geração de regras de associação fortes, ou seja, regras que satisfazem simultaneamente os valores das medidas de interesse propostas pelo usuário.

Observamos que a prospecção de conhecimento em bases de dados relacionais oferece um resultado bastante satisfatório, uma vez que obtemos um desempenho computacional aceitável, ao mesmo tempo em que foi possível extrair valiosos conhecimentos inicialmente ocultos no conjunto de dados. Além disso, o KDD é visto como uma área de estudo promissora devido à abundância de “matéria-prima” (dados) – resultado da era da informação digital a qual o mundo está submetido – e das inúmeras aplicações que fazem uso da extração de conhecimento, tais como astronomia, marketing, investimentos, detecção de fraudes, telecomunicações, bioinformática etc.