



18º Congresso de Iniciação Científica

**IMPLEMENTAÇÃO DE BASE DE CONHECIMENTO E MOTOR DE INFERÊNCIA PARA O
PROCESSO DE MINERAÇÃO DE DADOS DA FERRAMENTA KIRA**

Autor(es)

ANDRÉ DE ANDRADE BINDILATTI

Orientador(es)

CECÍLIA SOSA ARIAS PEIXOTO, ANA ESTELA ANTUNES DA SILVA

Apoio Financeiro

PIBIC/CNPQ

1. Introdução

Mineração de dados, ou do inglês, *data mining*, é o processo de revelar conhecimento, tendências ou padrões ocultos a partir de grandes volumes de dados. O interesse nessa tecnologia vem crescendo cada vez mais. Atualmente grandes empresas e organizações vêm armazenando quantidades cada vez maiores de dados, surgindo demanda por métodos mais sofisticados para análise desses dados.

Segundo Han e Kamber (2006), o processo de mineração de dados é uma das etapas de maior importância do **processo de descoberta de conhecimento em base de dados** (do inglês *Knowledge Discovery in Databases*, KDD).

Conforme Elmasri e Navathe (2004), o processo KDD é dividido nas seguintes etapas: seleção de dados; limpeza de dados; transformação dos dados; mineração dos dados; e apresentação do conhecimento obtido. Existe uma boa variedade de **tarefas de mineração de dados**. Durante a etapa de mineração de dados, uma das tarefas de mineração é selecionada para ser utilizada.

Mineração de dados traz consigo uma série de técnicas e algoritmos a serem utilizados sobre conjuntos de dados a fim de se obter conhecimento. A decisão sobre quais conjuntos de técnicas devem ser empregadas a um dado problema, segundo Berry e Linoff (2004), depende da natureza dos dados disponíveis, da preferência e conhecimento daquele que conduz o processo e também da tarefa de mineração a ser empregada.

O processo KDD, que envolve o processo de mineração de dados, trata-se de um processo que exige o conhecimento e atuação da inteligência de um humano especialista. Como foi dito, dentre as várias etapas desse processo, uma delas consiste na escolha de uma tarefa de mineração de dados para aplicar sobre um dado problema.

Essa decisão exige conhecimento sobre o problema e sobre as várias tarefas de mineração de dados existentes (SHARMA; OSEI-BRYSON, 2009). Sendo assim, é proposta a criação de uma **base de conhecimento** tendo como domínio de interesse, o conhecimento sobre as tarefas de mineração de dados e a decisão sobre qual tarefa selecionar. A base de conhecimento desenvolvida para este trabalho contempla as tarefas de associação, classificação e agrupamento.

Sistemas Especialistas (SEs) são sistemas que contam com uma base de conhecimento que concentra conhecimento pertinente a um domínio específico de interesse. Por meio de um mecanismo de inferência, esse conhecimento é aplicado a fim de obter soluções, de forma a emular a capacidade de especialistas humanos.

SEs representam uma abordagem na qual conhecimento específico de domínio é empregado na resolução de problemas pertinentes a este. Segundo Giarratano e Riley (1998), SEs foram um dos primeiros frutos das pesquisas na área da **inteligência artificial (IA)** a alcançarem aplicação efetiva no mundo comercial.

Conforme Giarratano e Riley (1998), o conhecimento de domínio pode ser obtido através de entrevistas com especialistas ou através de livros, artigos e outros tipos de fontes de conhecimento. Uma forma de representação é utilizada para expressar esse conhecimento, de modo a torná-lo acessível para que seja processado em um sistema computacional.

Existem diversas formas de representação de conhecimento. Em geral, são utilizadas: **regras de produção, redes semânticas, frames**, dentre outras (JACKSON, 1990). No contexto em pauta, regras de produção foram utilizadas para a representação do conhecimento.

Um módulo baseado em conhecimento e na tecnologia dos SEs, que conta com um mecanismo de inferência para aplicar o conhecimento sobre as tarefas de mineração de dados, foi então proposto. Tal módulo foi integrado à ferramenta Kira, com a intenção de tornar seu uso mais didático e acessível a usuários que possuem pouco conhecimento sobre as tarefas de mineração de dados.

A ferramenta Kira é uma ferramenta de mineração de dados que tem como uma de suas características ser ferramenta didática, facilitando a aprendizagem do processo de descoberta de conhecimento.

O desenvolvimento da ferramenta Kira teve início em um trabalho de dissertação do Mestrado em Ciência da Computação, o qual está tendo continuidade no grupo de pesquisa em mineração de dados do mesmo curso de mestrado (MENDES, 2009; Silva, 2010; VIEIRA et al., 2009).

2. Objetivos

Este trabalho de iniciação científica teve como objetivo, realizar a implementação de um mecanismo de inferência e uma base de conhecimento, constituindo um módulo baseado em conhecimento, para a ferramenta Kira, capaz de fornecer suporte ao usuário da ferramenta na atividade de seleção da tarefa de mineração a ser aplicada durante a etapa de mineração de dados.

Assim, esta pesquisa objetivou também, o estudo sobre SEs, sobre o processo KDD e a etapa de mineração de dados, mais especificamente sobre as tarefas de mineração de dados.

3. Desenvolvimento

A base de conhecimento e o motor de inferência foram desenvolvidos para este trabalho de acordo com o ciclo de vida de prototipagem, baseando-se na prototipagem evolutiva, na qual um protótipo da ferramenta que demonstra as funções mais fundamentais é rapidamente desenvolvido (LUQI, 1989). O protótipo é testado, avaliado e então novas versões do protótipo são desenvolvidas com base nos resultados obtidos.

A base de conhecimento e o motor de inferência foram construídos isoladamente num primeiro momento para posterior inclusão na ferramenta Kira. Essa construção resultou em uma ferramenta *shell*. Esta ferramenta foi utilizada para construção e testes da base de conhecimento a qual foi implementada na ferramenta Kira.

A base de conhecimento adota o paradigma baseado em regras. As regras de produção que compõem a base de conhecimento são armazenadas na forma de arquivos. Sua parte antecedente consiste em uma série de condições lógicas. Uma análise sintática é feita por um mecanismo *parser* construído com o gerador de analisador sintático JavaCC. Conforme as regras são validadas, são estruturadas na forma de grafos E/OU que representam as expressões em tempo de execução.

O motor de inferência compara o antecedente de cada regra com os fatos presentes na memória de trabalho. Quando o antecedente de uma regra é satisfeito, seu consequente dita as alterações a serem realizadas sobre a memória de trabalho. O processo de busca sobre as regras continua em ciclos, até que em uma de suas iterações nenhuma mudança seja realizada sobre a memória de trabalho, indicando a condição de parada.

Uma lista é utilizada para manter traço das regras ativadas durante o processo de inferência. Esta lista é retornada pelo método de inferência como forma de explicar os passos que conduzem as conclusões obtidas no processo. A Figura 1 ilustra, de modo conceitual, este mecanismo.

Uma estrutura de dados do tipo tabela *hash* (hashtable), é utilizada para representar a memória de trabalho, onde cada fato possui um identificador ou símbolo, que é utilizado como chave para a tabela *hash*. Um valor ou estado que caracteriza a instancia do fato é associado a cada chave ou símbolo.

A base de conhecimento é salva em um arquivo. Todo o código necessário para acesso da base de conhecimento e processo de inferência é disponível na forma de uma biblioteca de classes, construída com a linguagem Java. Essa biblioteca pode ser adicionada a outros projetos e utilizada para o processamento de conhecimento, como acontece com a Kira.

O conhecimento de domínio que forma a base de conhecimento é obtido através do estudo de livros e artigos sobre mineração de dados, e então expresso na forma de regras de produção. Este estudo sobre as tarefas foi desenvolvido como tema principal de uma dissertação de mestrado no Curso de Ciência da Computação (SILVA, 2010) e também pelo grupo de mineração de dados do Mestrado do mesmo Curso. A partir deste estudo, uma base de regras foi proposta por este trabalho.

A base de conhecimento é testada comparando respostas fornecidas sobre um conjunto de dados de entrada com as respostas esperadas com base nos estudos e conhecimentos adquiridos sobre mineração de dados. As respostas são analisadas e as regras empregadas são levadas em conta durante os testes. Críticas e aperfeiçoamentos das regras são então iterativamente realizados.

4. Resultado e Discussão

Uma ferramenta *shell* foi desenvolvida como resultado da adição de uma interface de usuário, aos mecanismos de inferência e representação de conhecimento desenvolvidos em função dos objetivos deste trabalho.

O processo de construção de uma base de conhecimento inclui um processo contínuo de refinamento e aprimoramentos. A inserção de novos conhecimentos, bem como a manutenção e teste do conhecimento já presente na base é uma necessidade constante. A *shell* pode ser definida como a ferramenta que oferece suporte para esse processo.

A base de conhecimento implementada, é representada na forma de um conjunto ordenado de regras de produção. As regras representam conhecimento para a decisão entre quais tarefas de mineração de dados o usuário da ferramenta Kira deverá escolher, ou seja, a execução das regras aponta para a tarefa mais adequada ao problema do usuário.

O conhecimento expresso nesta base visa modelar o conhecimento do domínio de mineração de dados, com a intenção de permitir, por meio de um mecanismo de inferência, a capacidade de responder ao usuário qual tarefa de mineração é mais indicada a se aplicar, levando em conta aspectos como:

- Quais expectativas para com o conhecimento resultante do processo de mineração.
- Domínio de interesse no qual o processo de mineração é aplicado.
- Se os dados estão dispostos como transações.
- Consciência do objetivo ou deficiência que se objetiva resolver com o processo de mineração.

De modo heurístico, as regras tratam características sobre o problema, conforme sua relevância e influência, de modo a revelar a tendência do uso de cada uma das tarefas de mineração sobre o problema abordado.

Desta forma, ao término do processo de inferência, o sistema indica com base nos aspectos citados, o quanto se conclui viável o uso de cada tarefa, podendo em alguns casos não ser capaz de indicar respostas; tal como especialistas humanos por vezes não podem decifrar de imediato a solução para certo problema peculiar, ou pelo fato das informações disponíveis possuírem certa inconsistência.

A ferramenta Kira guia o usuário ao longo do processo de descoberta de conhecimento. Na etapa de mineração, o usuário se depara com a escolha de qual tarefa aplicar sobre seus dados. Atualmente, o usuário pode selecionar entre classificação, associação e agrupamento.

O motor de inferência em conjunto com a base de conhecimento desenvolvidos definem o módulo de conhecimento integrado com a ferramenta Kira. A Figura 2 exibe a tela onde o usuário tem contato com o módulo especialista. Um conjunto de questões é realizado ao usuário para colher os dados sobre o problema.

Os códigos e classes desenvolvidos em Java para a implementação do mecanismo de inferência são incorporados na ferramenta Kira na forma de uma biblioteca de classes. Esta biblioteca fornece funcionalidades que permitem carregar os arquivos contendo as questões e a base de regras, além de permitir a execução do processo de inferência e outras funções importantes.

O usuário responde as questões como uma forma de descrever seu “problema”. O motor de inferência processa estes dados de entrada e aplica o conhecimento presente na base de conhecimento. Ao final do processo, uma conclusão é apresentada ao usuário. O usuário pode então optar por selecionar a tarefa sugerida pelo sistema com maior indicação, ou selecionar qualquer outra.

A lista de regras utilizadas no processo de inferência é apresentada ao usuário como forma de explicar o raciocínio que levou a tal conclusão. Isto pode tanto ser usado para a inspeção do conhecimento, contudo, também pode servir como uma espécie de tutor para o usuário.

5. Considerações Finais

O módulo especialista foi efetivamente implementado na ferramenta Kira por meio de uma biblioteca de classes desenvolvida durante o projeto, e de uma base de conhecimento. A base de conhecimento pôde ser construída com auxílio de uma ferramenta *shell*, desenvolvida como consequência da pesquisa realizada; o mecanismo de inferência e estruturas para a representação do conhecimento necessárias para o projeto, passaram a formar uma ferramenta *shell* com a simples adição de uma interface de usuário.

A inclusão do módulo especialista permite que usuários que não detêm conhecimento sobre as várias tarefas de mineração, possam mesmo assim, usufruir da ferramenta Kira e das possibilidades de aplicação das tarefas de mineração de dados. Outro aspecto importante é que dado o fato de que o módulo especialista exibe as regras utilizadas no processo de inferência, o usuário tem a possibilidade de absorver conhecimento sobre o domínio.

É importante ressaltar que o processo de aquisição de conhecimento não é uma tarefa trivial, o que restringe de certa forma o emprego de SEs. Além disto, o processo de verificação e teste de base de conhecimento não é simples de aplicar, tal como em sistemas convencionais.

Contudo, SEs correspondem a uma tecnologia interessante e que pode resultar em boas aplicações sobre domínios para os quais soluções algorítmicas sejam inexistentes ou inviáveis. O uso de conhecimento específico de domínio, que por meio de um sistema, é aplicado de forma a emular a capacidade de um especialista humano, trata-se de uma abordagem que permite a solução de vasta variedade de problemas.

A base de conhecimento expressa conhecimento sobre as várias tarefas de mineração, e a partir da resposta de uma série de questões sobre os dados, intenções e expectativas do usuário sobre o domínio e o resultado do processo de mineração, é capaz de oferecer apoio a sua decisão sobre qual tarefa de mineração utilizar.

Referências Bibliográficas

BERRY, Michael J. A.; LINOFF, Gordon F.. *Data Mining Techniques*. Second ed. Canada: Wiley Publishing, 2004.

ELMASRI, Ramez; Navathe, Shamkant B.. *Fundamentals of Database Systems*. Fourth ed. Nova York: Pearson Education, 2004.

GIARRATANO, Joseph; RILEY, Gary. *Expert Systems Principles and Programming*. Third ed. PSW Pub., 1998.

HAN, Jiawei; KAMBER, Micheline. *Data Mining - Concepts and Techniques*. Second ed. Nova York: Morgan Kaufmann, 2006.

JACKSON, Peter. *Introduction to Expert Systems*. Second ed. Addison Wesley Longman Limited. 1990.

LUQI. *Software Evolution Through Rapid Prototyping Computer*. vol. 22, n. 5, p. 13-25, May 1989.

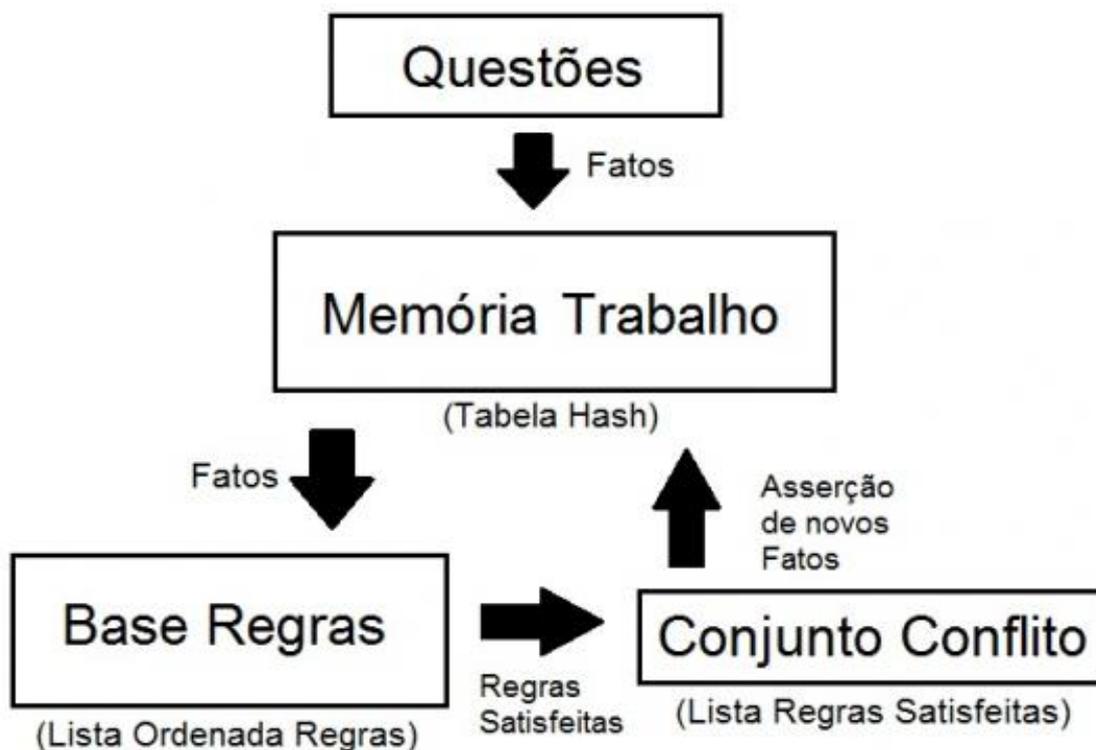
MENDES, Eduardo F. *Automatização da Técnica de Mineração de Dados Auxiliado por Guias*. Dissertação de Mestrado. Universidade Metodista de Piracicaba. Piracicaba. São Paulo, 2009.

SILVA, Lidia M.. *Modelagem e teste de uma base de conhecimento de instrução de mineração de dados relacionais com ênfase na tarefa de classificação*. Dissertação de Mestrado. Universidade Metodista de Piracicaba. São Paulo, 2010.

SHARMA, Sumana; OSEI-BRYSON, Kweku-Muata. *Role of Human Intelligence in Domain Driven Data Mining*. In: CAO, L.; YUP, S.; ZHANG, H.. *Data Mining for Business Applications*. First ed. New York: Springer Science+Business Media, 2009.

VIEIRA, Marina T. P.; SILVA, Ana E. A.; PEIXOTO, Cecilia S. A.; MENDES, Eduardo F.; GOMIDE, Rodrigo S.. *Kira: A Tool Based on Guides and Domain Knowledge to Instruct Data Mining*. In: IADIS, Proceedings of the IADS International Conference Applied Computing, p. 12-16, 2009.

Anexos



Kira - Tarefa

Marque sim ou não para as perguntas. O programa decidirá qual melhor tarefa a se seguir.

Fontes de Dados:

+ Novo +

Dados Fontes

Projetos:

+ Novo +

Compras

Etapas de Mineração:

- 1. Entendimento do Negócio
 - 1.1. Problema
 - 1.1.1. Objetivo
 - 2. Identificação da Tarefa de Mineração
 - 2.1. Tarefa**
 - 3. Preparação dos Dados
 - 3.1. Integração
 - 3.1.1. Valores Ausentes
 - 3.1.2. Suavizar o Ruído
 - 3.1.3. Identificar Dependências
 - 3.2. Limpeza
 - 3.3. Seleção
 - 3.4. Transformação
 - 4. Análise
 - 4.1. Mineração
 - 4.2. Avaliação

resultado obtido durante o processo de mineração será aplicado a novos dados?
 sim não

Um dos atributos existentes no seu conjunto de dados está diretamente ligado ao problema ou interesse pela mineração? No caso de não existir o atributo, há possibilidade de criá-lo?
 sim não

Deseja explorar a possível semelhança que pode haver dentro de dados em seu conjunto?
 sim não

O resultado obtido durante o processo de mineração será aplicado a um problema atual resolver visando melhorar algo?
 sim não

As tuplas do seu conjunto de dados possuem alguma forma de categoria e você deseja entender melhor os fatores que proporcionam estas coisas?
 sim não

A base de dados contém dados referentes à compra de clientes com diversos itens e o objetivo do processo de mineração envolve a análise de vendas?
 sim não

Não possui um objetivo definido ou expectativa sobre o conhecimento a ser produzido durante o processo de mineração deseja encontrar correlações?
 sim não

As tuplas do seu conjunto de dados representam transações?
 sim não

Atributos das tuplas do seu conjunto de dados são em-base por valores numéricos?
 sim não

Identificação da Tarefa de Mineração

Tem por objetivo orientar a análise de dados com algumas questões para selecionar aquelas que melhor representam o objetivo de mineração de dados.

Dados Solicitados: O analista de dados deverá preencher os seguintes dados:
 Qual a descrição da tarefa de mineração? Informe a descrição da tarefa de mineração de dados. Esta descrição será utilizada para diferenciar as diversas tarefas que podem ser realizadas para um mesmo objetivo.
 Dentro as frases abaixo, selecionar aquela que melhor representa a tarefa de mineração de dados que deseja executar para atingir o objetivo da mineração de dados. Selecionar a frase que melhor representa a sua necessidade de mineração.
 Responder: Apenas uma frase pode ser selecionada.

Incluir: Inclui uma nova tarefa de mineração de dados.
Localizar: Localiza todas as tarefas de mineração de dados associadas a um determinado objetivo.
Salvar: Salva os dados.
Excluir: Exclui a tarefa de mineração de dados.
Cancelar: Cancela as alterações realizadas.