8ª MOSTRA ACADÊMICA UNIMEP
26 a 28/10 de 2010
Tema: Desafios da Educação Superior na Agenda do Novo Milênio

18º Congresso de Iniciação Científica

TRATAMENTO DE REGRAS DA ASSOCIAÇÃO MULTIRELACIONAL NA FERRAMENTA DE MINERAÇÃO DE DADOS KIRA

Autor(es)

JONAS RAFAEL ONOFRE

Orientador(es)

MARINA TERESA PIRES VIEIRA

Apoio Financeiro

PIBIC/CNPQ

1. Introdução

A coleta e o acúmulo de dados, em diversos de campos, têm tomado grandes dimensões nos últimos anos. Há uma necessidade imediata de novos processos, técnicas e ferramentas para ajudar os seres humanos a extrair informações úteis (conhecimento) do volume crescente de dados.

O processo de descoberta de conhecimento em bases de dados (*KDD – Knowledge Discovery in Databases*) foi proposto em 1989, por Piatetsky-Shapiro, para enfatizar que conhecimento é o produto final de uma descoberta impulsionada em dados (FAYYAD; PIATETSKY-SHAPIRO; SMITH, 1996).

Segundo Fayyad; Piatetsky-Shapiro; Smith (1996), o processo de descoberta de conhecimento em bases de dados é definido como sendo o processo não trivial de identificação de padrões válidos, novos, potencialmente úteis, compreensíveis e embutidos nos dados.

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

A mineração de dados (*Data Mining*) é considerada a mais importante etapa do processo de KDD. Nela são aplicados algoritmos que buscam extrair conhecimento contido em grandes quantidades de dados, buscando identificar padrões que explicitam ocorrências nos dados, não observadas anteriormente.

Conforme Han; Kamber (2006), trata-se de uma área multidisciplinar de desenvolvimento. O conhecimento obtido pode ser aplicado em situações como tomada de decisões, controle de processo, entre outros.

Existem diversas tarefas de mineração. Algumas das principais são: **regras de associação, classificação e clusterização**.

A tarefa de associação, cujos conceitos foram usados neste trabalho, busca encontrar padrões frequentes, que são aqueles que aparecem frequentemente em um conjunto de dados (HAN; KAMBER, 2006). Por exemplo, em uma base de dados de transações de vendas de um supermercado, é desejável descobrir as associações existentes entre os itens, ou seja, a presença de um item em uma determinada transação irá implicar a presença de outro item dentro da mesma transação (MENDES, 2009).

Uma regra de associação é da forma X => Y. Essa associação estabelece que, se um cliente comprar X, ele também estará propenso a comprar Y (ELMASRI; NAVATHE, 2005).

Como citado em Mendes (2009), geralmente, a mineração de regras de associação pode ser visualizada em um processo composto de duas etapas: primeiro, encontram-se todos os *itemsets* (conjunto de itens) de um conjunto de transações que são frequentes. Esses

itemsets são chamados de *itemsets* frequentes. Segundo, utilizam-se os *itemsets* frequentes, encontrados na primeira etapa, para determinar as regras de associação que existem no banco de dados. As regras são consideradas interessantes se obedecem ao limite mínimo de suporte e confiança estabelecidos pelo usuário.

Conforme Han; Kamber (2006), para as regras encontradas, duas medidas de interesse são utilizadas, **suporte e confiança**. Elas, respectivamente, refletem a utilidade e a certeza da regra descoberta.

A medida de suporte avalia a freqüência com que os itens ocorrem em relação ao total de dados da massa, e é representada pela fórmula:

Suporte (X =>Y) = Ocorrências de X U Y / Total de Ocorrências.

A medida de confiança refere-se a um valor de correspondência entre os itens que compõem uma tarefa de associação, e é representada pela fórmula:

Confiança (X =>Y) = Suporte de (X U Y) / Suporte de X.

Conforme citado em Pizzi, Ribeiro, Vieira (2005), as técnicas de mineração de dados tradicionais processam os dados que estejam armazenados em uma única tabela. Se os dados envolvidos pertencem a tabelas distintas, é necessário que eles sejam transferidos para uma única tabela para que os algoritmos possam ser utilizados. O processo de transferência dos dados é uma operação de alto custo e pode levar à perda de informações, além de poder resultar em uma grande quantidade de dados replicados. Outro problema causado pela junção das tabelas é a grande quantidade de dados resultante que pode afetar o desempenho do algoritmo de mineração ou mesmo tornar o processo impraticável.

O presente trabalho usa os resultados dos trabalhos de Ribeiro (2004) e Garcia (2008), visando sua incorporação na ferramenta Kira (MENDES, 2009). Esses trabalhos são apresentados a seguir.

Abordagem de Ribeiro: Em Ribeiro (2004), foram determinados novos conceitos para regras de associação para gerar regras mais confiáveis, como o conceito de bloco, segmento e peso de um item. Sua abordagem se baseia em manter as relações separadas entre si e aplicar o algoritmo Connection, proposto pela mesma autora. As relações consideradas nesse trabalho são tabelas fato de um data warehouse e as medidas de interesse como suporte e confiança foram alteradas para que os padrões gerados possam representar melhor a verdadeira relação entre os itens das múltiplas relações.

Abordagem de Garcia: Com base no trabalho de Ribeiro (2004), Garcia (2008) criou o algoritmo ConnectionBlock, tratando-se de uma variação da abordagem adotada no algoritmo Connection. Para isso criou uma nova contagem de suporte e confiança.

O algoritmo ConnectionBlock endereça o problema de encontrar regras de associação entre tabelas que não são explicitamente relacionadas, mas que têm influência entre si devido à semântica entre os dados envolvidos, isto é, elas são semanticamente relacionadas, no sentido que a informação em uma ou mais tabelas pode afetar a informação em outras tabelas. Por exemplo, pode-se encontrar esse tipo de relacionamento semântico entre informações de cartão de crédito e empréstimo de contas bancárias e entre conceitos de atividades e de provas de estudantes, em um banco de dados acadêmico.

A exemplo do algoritmo Connection, o ConnectionBlock foi desenvolvido com base no algoritmo FP-Growth. Para lidar com os dados das várias tabelas em conjunto o algoritmo agrupa esses dados em blocos e calcula o suporte e confiança dos blocos. O algoritmo ConnectionBlock usa uma estrutura chamada MFPtree que é uma extensão da FP-tree usada pelo FP-Growth. Cada nó da MFP-tree corresponde a um item freqüente e cada ramo corresponde a um itemset encontrado em uma ou mais transações dos blocos de uma tabela.

Ferramenta Kira: Essa ferramenta tem como objetivo ensinar como preparar os dados, como escolher a tarefa de mineração adequada e como analisar os resultados obtidos. Ela foi criada com o intuito de abstrair boa parte do conhecimento exigido do usuário para executar o processo envolvido na mineração de dados.

Sua arquitetura é composta por três módulos principais: **Módulo de Apoio à Origem, Módulo de Apoio à Preparação e Módulo de Apoio à Análise**.

No **módulo de apoio à origem** são identificadas as fontes de dados que serão utilizadas.

No **módulo de apoio à preparação** são executadas todas as atividades de preparação de dados, que contempla a integração, limpeza, seleção e transformação dos dados.

No **módulo de apoio à análise** são executadas as etapas referentes à mineração de dados e análise dos resultados obtidos. São exibidos os dados selecionados e transformados e executado o algoritmo minerador, mostrando em seguida os resultados obtidos, auxiliando o usuário em sua análise.

2. Objetivos

A mineração de dados é uma área muito extensa e, existe uma carência de desenvolvimento de técnicas para minerar vários tipos de conhecimento envolvendo múltiplas tabelas. Este projeto foca a mineração de regras de associação multirelacional, visando sua incorporação na ferramenta Kira. Em metodologia, para atingir esse objetivo foram realizados estudos sobre regras de associação multirelacional apresentados em Ribeiro (2004) e Garcia (2008), para, então, desenvolver recursos instrucionais para sua utilização na ferramenta de mineração de dados Kira.

3. Desenvolvimento

Na primeira parte do projeto foram realizados diversos experimentos, com conjuntos de dados distintos, utilizando dois algoritmos, Connection e ConnectionBlock, desenvolvidos por Ribeiro (2004) e Garcia (2008), respectivamente e que focam a mineração multirelacional. Foram propostos os layouts de telas para a ferramenta de mineração de dados Kira, para dar suporte ao usuário na execução dos algoritmos de Ribeiro (2004) e Garcia (2008). Esses layouts podem ser incorporados na mesma, para assim haver a possibilidade de se trabalhar com a mineração multirelacional na ferramenta.

4. Resultado e Discussão

Neste projeto, a ferramenta Kira foi utilizada com sua base de dados padrão, que acompanha sua instalação, para com isso ter uma visão prática sobre toda metodologia de descoberta de conhecimento em bases de dados e suas fases. Essa base de dados é referente a um conjunto de informações de alunos inscritos para um congresso de tecnologia realizado em 2007 na cidade de Mococa – SP.
Após esse aprendizado, foram então estudadas as abordagens de Ribeiro (2004) e Garcia (2008), as quais possuem seus tratamentos sobre mineração multirelacional e os algoritmos: Connection e ConnectionBlock, respectivamente.
Inicialmente foram feitos estudos teóricos e pequenos testes com tais métodos e algoritmos, para posteriormente, serem propostas algumas especificações das interfaces para o módulo de tratamento de mineração multirelacional na ferramenta Kira. A figura 1 é uma proposta de interface para a escolha do algoritmo que será usado para a realização da mineração multirelacional. A figura 2 é uma proposta de interface para o usuário realizar a seleção dos dados a serem usados na mineração, selecionando as tabelas, atributos e identificadores desejados. Outras protótipos de interfaces foram criados para apoiar as fases do processo de mineração de dados para regras de associação multirelacional usando os algoritmos Connection e ConnectionBlock.

5. Considerações Finais

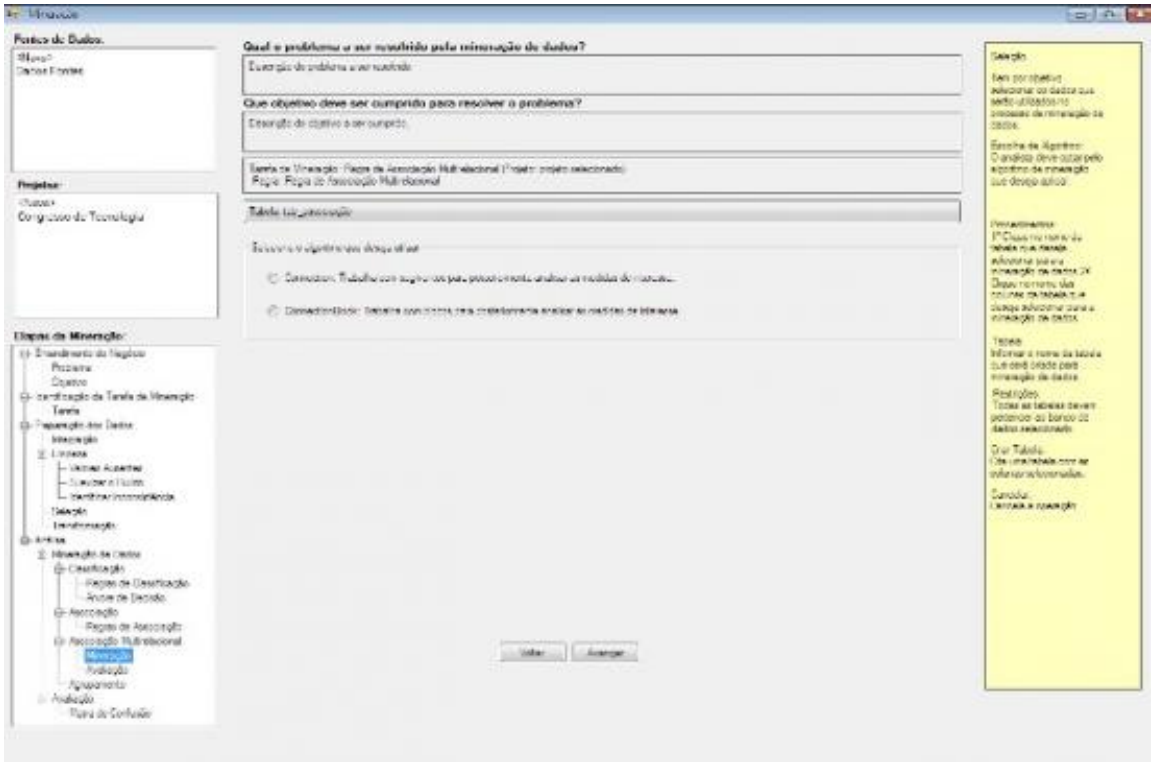
Para este trabalho científico foram estudados conceitos de mineração de dados, focando em mineração multirelacional, seguindo a abordagem de Ribeiro (2004) e Garcia (2008), juntamente com pequenos testes realizados com os algoritmos Connection e ConnectionBlock, respectivamente.
A ferramenta de mineração de dados Kira foi proposta baseada em um conjunto de guias redigidas em Mendes (2009), neste trabalho a mesma proposta de utilizar guias para o desenvolvimento e aplicação da mineração de dados voltada para regras de associação multirelacional.
Juntos com os mesmos foram propostos alguns layouts de telas que foram desenvolvidos para que futuramente seja realizada a implementação desta tarefa na ferramenta de mineração de dados Kira.
Para trabalhos futuros, novos guias poderão ser sugeridos para as outras tarefas de mineração de dados, de forma a facilitar o trabalho do analista de dados.

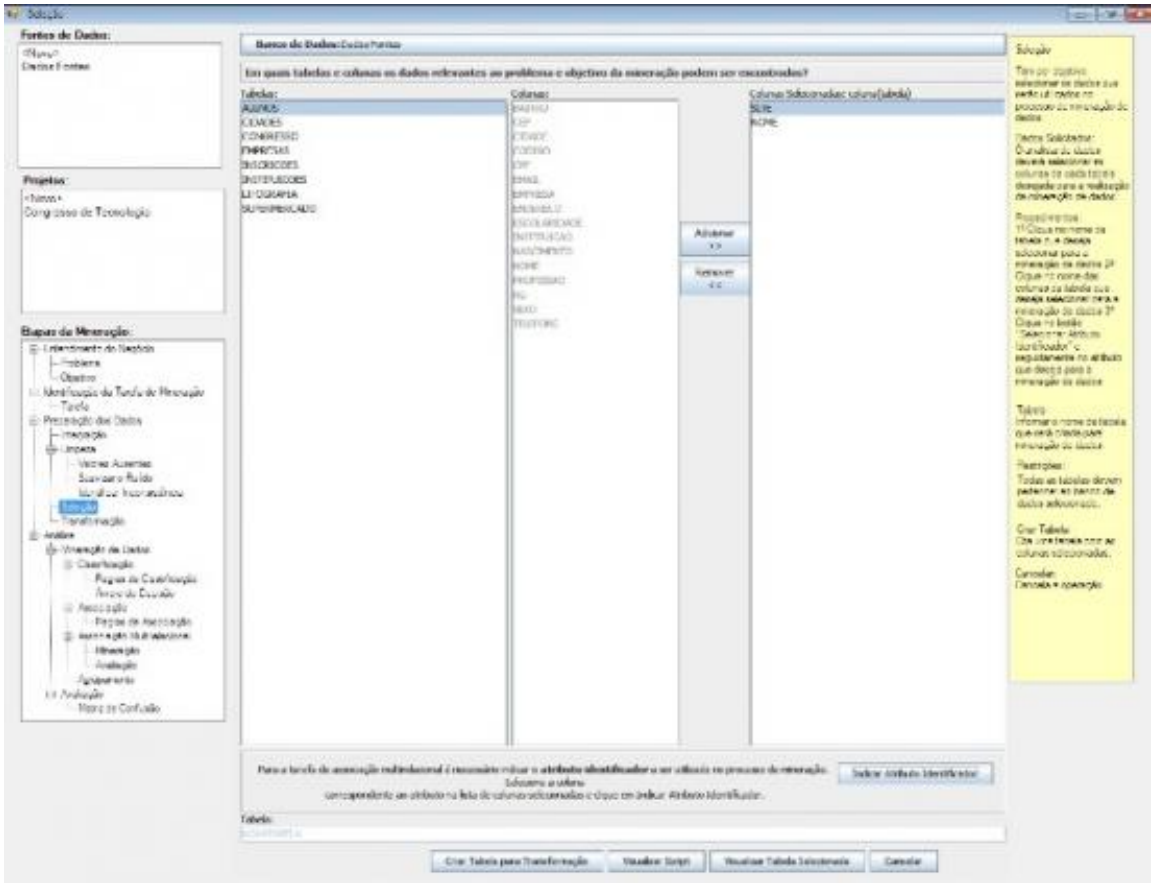
Referências Bibliográficas

ELSMARI, R. & NAVATHE, S.B. **Sistemas de Banco de Dados**. (4ª Edição).Pearson Brasil, 2005.
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.
GARCIA, E. **Mineração de Regras de Associação Multi-Relacional Quantit**ativas. Dissertação (Dissertação de Mestrado) – Faculdade de Ciências Exatas e da Natureza, Universidade Metodista de Piracicaba, Piracicaba, SP, 2008.
HAN, J.; KAMBER, M. **Data Mining - Concepts and Techniques**. 2a edição. Nova York: Morgan Kaufmann, 2006.
HAN, J., PEI, J. e YIN, Y. **Mining frequent patterns without candidate generati**on. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data, 2000, Dallas, Texas, USA. Anais. Dallas, Texas, USA, 2000.
MENDES, Eduardo Fernando. **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.
PIZZI, L. **Mineração de Dados em Múltiplas Ta**belas. 88 f. Dissertação (Dissertação de Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, 2006.
PIZZI, L. C., RIBEIRO, M. X. e VIEIRA, M. T. P. **Analysis of Hepatitis Dataset using Multirelational Association Rule**s. In: Proc. of the ECML/PKDD 2005 Discovery Challenge, 2005, Porto, Portugal. Anais. Porto, Portugal, 2005.
RIBEIRO, M. **Mineração de Dados em Múltiplas Tabelas Fato de Data Wareho**use. 131 f. Dissertação (Dissertação de Mestrado) – Departamento de Computação, Universidade Federal de São Carlos, São Carlos, SP, 2004.

Anexos




Fontes de Dados:
Projetos:
Etapas da Mineração:
Tabelas:
Colunas:
Colunas Selecionadas: coluna(tabela)
Adicionar >>
Remover <<
Tabela:
Criar Tabela para Transformação
Visualizar Script
Visualizar Tabela Selecionada
Cancelar