Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

IMPLEMENTAÇÃO DE UM ALGORITMO DE PADRÕES DE SEQUÊNCIA PARA DESCOBERTA DE ASSOCIAÇÕES ENTRE PRODUTOS DE UMA BASE DE DADOS REAL

Aluna: Cecília Henriques Devêza
Matricula: 07.1.4121

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
26 de setembro de 2010

Universidade Federal de Ouro Preto - UFOP

Instituto de Ciências Exatas e Biológicas - ICEB

Departamento de Computação - DECOM

IMPLEMENTAÇÃO DE UM ALGORITMO DE PADRÕES DE SEQUÊNCIA PARA DESCOBERTA DE ASSOCIAÇÕES ENTRE PRODUTOS DE UMA BASE DE DADOS REAL

Proposta de monografia apresentada ao curso de Bacharelado em Ciência da Computação, Universidade Federal de Ouro Preto, como requisito parcial para a conclusão da disciplina Monografia I (BCC390).

Aluna: Cecília Henriques Devêza
Matricula: 07.1.4121

Orientador: Luiz Henrique de Campos Merschmann

Ouro Preto
26 de setembro de 2010

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O constante crescimento de informações decorrentes do desenvolvimento tecnológico tem trazido às organizações um número abundante de dados, aumentando a importância das ferramentas que tem por objetivo extrair informações úteis destes dados. O grande desafio dessas organizações é exatamente transformá-los em conhecimento. Sendo uma organização comercial, este conhecimento oriundo dos dados históricos pode direcionar melhor campanhas de marketing, evidenciar formas mais lucrativas de exibir produtos, bem como mostrar os clientes mais propícios à compra dos mesmos. Sendo uma organização acadêmica, o conhecimento pode trazer respostas de questões impossíveis de serem observadas "a olho nu" por pesquisadores frente a um volume tão absurdo de dados. Além do que, pode encontrar justificativas sobre dúvidas acerca da Medicina, Biologia, e qualquer outra ciência que tem a necessidade de obter respostas ou fazer previsões tendo como base um volume de dados. A técnica de extração de informação mais indicada para este tipo de processo, é a chamada Mineração de Dados.

Mineração de Dados é um ramo da computação que teve início nos anos 80, quando os prossionais das empresas e organizações começaram a se preocupar com os grandes volumes de dados informáticos estocados e inutilizados dentro da empresa. Nesta época, Data Mining consistia essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível. Atualmente, Data Mining consiste sobretudo na análise dos dados após a extração, buscando-se por exemplo levantar as necessidades reais e hipotéticas de cada cliente para realizar campanhas de marketing.[3].

A descoberta de conhecimento entretanto, não pode ser resumida à Mineração de Dados. Esta, é apenas uma etapa de todo o processo, conhecido como KDD (*Knowledge Discovery in Database*). O mesmo pode ser dividido em:

- Pré-processamento de Dados
  Os dados muitas vezes precisam de uma limpeza antes de passar para a etapa de Mineração. Como as bases geralmente são muito grandes, é necessário remover delas tudo que não vai ser utilizado na próxima etapa, como dados redundantes ou inconsistentes. A qualidade dos dados é que determina a eficiência dos algoritmos de Mineração.

- Mineração de Dados
  É a técnica que será abordada neste trabalho, mais precisamente a área de Padrões de Sequência, onde se busca conhecimento em uma base de dados que segue uma ordem temporal, ou seja, onde os dados estão datatos. Essa base possibilita a descoberta de regras do tipo "Se um determinado usuário comprou um produto X, é provável que ele volte e compre o produto Y".

- Pós-Processamento de Resultados
  Após a extração das regras de conhecimento, é preciso verificar o que saiu como "novidade" dessas regras, ou seja, qual conhecimento estava escondido nos dados que alguém não seria capaz de descobrir sem passar pelo processo automatizado. É recomendado que esta etapa seja realizada por alguém que

conhece o processo sob os quais os dados foram extraídos e saiba diferenciar as regras que são ou não são aproveitáveis.

A figura a seguir ilustra as etapas do KDD:

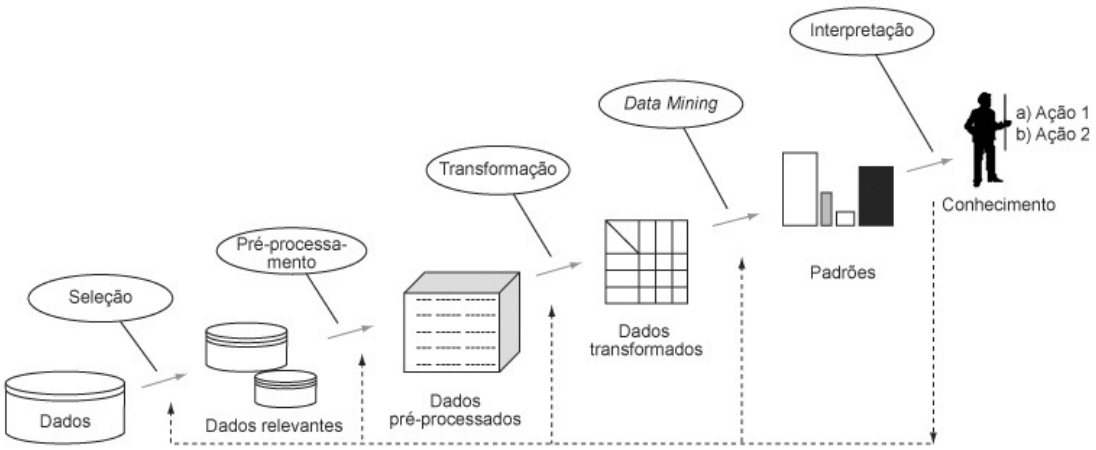

Figura 1: Etapas do KDD

Este projeto irá englobar muitas das etapas exibidas, desde o pré-processamento dos dados até a análise dos mesmos, tendo como ênfase a Extração de Regras de Associação, presente na etapa de *Data Mining*. A descoberta de Padrões de Sequência é uma parte mais específica do processo de descoberta de Regras de Associação, que visa descobrir sequências de ações/acontecimentos presentes em uma base de dados organizada de forma temporal, ou seja, onde os dados estão datados ou seguem um ordem cronológica. Além disso, é preciso que cada transação da base esteja relacionada a um agente para que as ligações entre as transações possa ser realizada de forma coerente.

A tabela a seguir exemplifica uma base de Dados que pode ser utilizada para descobrir padrões de sequência:

ID do Usuário	**Transação**	**Data**
1	3, 6, 7, 1	06/01/2010
2	1, 2	10/01/2010
1	5, 9, 13	03/02/2010
3	2, 5	11/02/2010
3	8	23/03/2010

Tabela 1: Exemplo de Base de Dados.

2 Justificativa

A utilidade deste projeto justifica-se pela crescente necessidade de se encontrar informações úteis de grandes bases de dados. Com a constante automatização de processos das empresas, cresce também o armazenamento de dados enviados e recebidos nestes processos. Entretanto, este dados precisam de um tratamento para que sejam úteis à lucratividade das organizações.

Partindo desde princípio, a empresa GerenciaNet Tecnologia em Pagamentos ofereceu parte de sua base dados que estão relacionados à visualização de produtos em lojas virtuais, sob a condição de que estes dados sejam utilizados única e exclusivamente para fins acadêmicos. A fim de assegurar os direitos dos clientes pela empresa, não foram divulgados: nomes das lojas envolvidas; dados pessoais de clientes que visualizaram produtos nas lojas. Estas informações foram previamente criptografadas.

A base de dados recebida, possui o seguinte formato:

idTransação, idCliente, Produto Visualizado, Loja, Data
idTransação, idCliente, Produto Visualizado, Loja, Data
⋮
idTransação, idCliente, Produto Visualizado, Loja, Data

Onde IdTransação é um valor inteiro que cresce de um a cada tupla, começando por 1. IdCliente é um valor criptografado que representa o cliente que realizou a transação, este valor pode se repetir na base. Produto Visualizado é o nome do produto que foi visto pelo cliente naquela transação (cada tupla mostra a visualização de um único produto). E por fim, Data é o dia, mês e ano que o produto foi visualizado, ou seja, quando a transação ocorreu.

2.1 Proposta

A proposta deste projeto é implementar o algoritmo de Padrões de Sequência GSP, afim de descobrir regras associativas a partir de uma base de dados da empresa GerenciaNet. Detalhes sobre o algoritmo podem ser vistos em [2]. A idéia é construir um algoritmo que poderá ser reutilizado posteriormente para outras bases de dados que segue o padrão mostrado em 1.

3 Objetivos

3.1 Objetivo geral

Este trabalho tem como principal objetivo, a obtenção de padrões de sequência a partir de dados reais de uma loja virtual, para melhor gerenciar campanhas de marketing e promocionais, utilizar de estratégias para fidelizar clientes, e outras ferramentas que visa a lucratividade de uma organização e satisfação de seus clientes.

3.2 Objetivos específicos

- Conhecer os algoritmos de Padrões de Sequência e Mineração de Dados em geral;
- Aprimorar o conhecimento da linguagem de programação utilizando diversos recursos para processamento de dados;
- Descobrir regras de sequência a partir de uma base de dados real.

4 Metodologia

Inicialmente será realizado um trabalho de levantamento e revisão de literatura, com o objetivo de fixar os conceitos de Mineração de Dados acerca da resolução de padrões de sequência. De acordo com o que foi visto até agora, um algoritmo bastante eficiente utilizado para este fim é o GSP (*Generalized Sequential Patterns*). A idéia é implementar este algoritmo, de forma que a entrada dos dados será o arquivo fornecido pela empresa, e a saída será as regras de padrões de sequência contidas neste arquivo.

O arquivo de entrada, entretanto, precisa ser previamente processado. Existem neste arquivo informações desnecessárias à uma mesma pesquisa por padrões de sequência, já que visualizações de produtos de 5 lojas distintas estão misturados. Baseando-se nos produtos (que não encontram-se criptografados), é possível verificar que não existe uma co-relação lógica entre essas lojas, portanto, é recomendável que sejam escolhidos os dados de uma loja por vez. Após a divisão de dados das 5 lojas, a sequência de uma delas será escolhida para os primeiros testes no algoritmo. É também parte deste processamento, a transformação dos nomes dos produtos em valores númericos, primeiramente para garantir a premissa de segurança das informações, e em segundo lugar porque o processamento de valores númericos é bem mais rápido.

Feito o processamento, é necessário realizar a implementação do algoritmo. Na verdade, estas duas fases (pré-processamento e implementação) podem por vezes se misturar, sendo realizadas de forma concorrente já que são etapas independentes. A previsão para realização desta etapa é de 2 meses e meio, sendo considerada a etapa mais importante deste projeto. A implementação será realizada na linguagem C, e deverá aceitar como entrada arquivos do tipo ARFF, que é o mesmo tipo de arquivo utilizado na ferramenta WEKA.

O Weka é uma coleção de algoritmos de aprendizado de máquina para tarefas de Mineração de Dados. O algoritmo pode também ser aplicado diretamente a um conjunto de dados, ou chamado a partir do seu próprio código Java. Weka contém ferramentas para os dados de pré-processamento, classificação, regressão, clusterização, regras de associação e visualização. É também ideal para o desenvolvimento de novos modelos de aprendizagem de máquina.[1]. Esta ferramenta será utilizada como apoio em todo o processo descrito neste projeto, por ser muito conhecida no meio da Mineração de Dados, e oferecer bons recursos de processamento, inclusive o próprio algoritmo de Padrões de Sequência, GSP.

Um mês será gasto apenas na realização de testes e possíveis correções do algoritmo. A ferramenta WEKA será de suma importância neste passo, visto que a comparação de resultados entre o algoritmo criado e o que está acoplado no WEKA ajudará na calibração da implementação aqui defendida. Ainda aqui, podem ser utilizadas bases de dados diferentes para comparar a análise de execução e dados de saída dos dois algoritmos.

Após todas as modificações necessárias, será feita uma análise sobre os dados retornados pelo algoritmo. A regra de padrão de sequência, deve seguir o seguinte formato:

$$<\{5\}, \{2,7\}>$$

Onde os valores que estão entre chaves pertecem à uma mesma transação - ou seja, a um mesmo acesso à loja virtual. Cada transação separada por vírgula indica

que foi realizada em uma data diferente, seguindo uma ordem cronológica da esquerda para a direita. Este exemplo indica que: "Um cliente que visualiza o produto 5, posteriormente retorna à loja e visualiza os produtos 2 e 7". Este conjunto de transações, por seguir uma ordem temporal, não é comutativo, ou seja, não é possível afirmar a partir do exemplo, que a regra $<\{2, 7\}, \{5\}>$ é verdadeira também.

Com as regras retornadas pelo algoritmo, será possível verificar os produtos que, se visualizados, geram uma probabilidade maior de outros determinados produtos serem visualizados também posteriormente, o que pode ser utilizado no marketing da empresa afim de direcionar propagandas a clientes específicos, como criar promoções de produtos Y que provavelmente são adquiridos após produtos X, para aqueles que já visualizaram este último. Enfim, este é o primeiro passo para uma melhor campanha publicitária de uma loja virtual, na qual a propaganda é realmente direcionada a clientes que a esperam, evitando desperdício de gastos da empresa, fidelizando clientes e evitando a imagem de *spammer* daqueles que não desejam receber determinadas promoções por não se interessarem pelo produto divulgado.

5 Cronograma de atividades

A Tabela 2, mostra o cronograma com as atividades a serem realizadas, começando em Outubro de 2010, e terminando em Julho de 2011. A carga horária semanal prevista é de 20 horas.

Atividades	**Out**	**Nov**	**Dez**	**Jan**	**Fev**
- Estudo de Algoritmos de P.S.*	x				
- Levantamento de Referências		x	x		
- Pré-Processamento da Base de Dados			x	x	
- Implementação do Algoritmo GSP				x	x
Atividades	**Mar**	**Abr**	**Mai**	**Jun**	**Jul**
- Implementação do Algoritmo GSP	x				
- Testes do Algoritmo e Correções		x			
- Análise de Resultados			x	x	
- Escrita do Trabalho de Conclusão de Curso				x	x
- Defesa					x

Tabela 2: Cronograma de Atividades.

*Padrões de Sequência.

Referências

[1] Weka - data mining software in java.

[2] Srikant R. Agrawal, R. Mining sequential patterns : Generalizations and performance improvements. pages 3–17, 1996.

[3] Sandra de Amo. Curso introdutório de mineração de dados - compilação de notas de aulas. 1996.