DIOGO PEREIRA SANTIAGO

BUSCA DE PADRÕES NAS NEGOCIAÇÕES E COTAÇÕES DA *COMMODITY* CAFÉ NO MERCADO DE FUTUROS

Monografia para a Universidade Federal de Lavras, como parte das exigências do Programa de Graduação em Ciência da Computação, para obtenção do titulo de "Bacharel".

Orientador
Prof. Luiz Gonzaga de Castro Júnior

Co-Orientadora
Profa. Olinda Nogueira Paes Cardoso

LAVRAS
MINAS GERAIS – BRASIL
2002

DIOGO PEREIRA SANTIAGO

BUSCA DE PADRÕES NAS NEGOCIAÇÕES E COTAÇÕES DA *COMMODITY* CAFÉ NO MERCADO DE FUTUROS

Monografia para a Universidade Federal de Lavras, como parte das exigências do Programa de Graduação em Ciência da Computação, para obtenção do titulo de "Bacharel".

APROVADA em 17 de dezembro de 2002.

_____

Profa. Olinda Nogueira Paes Cardoso
DCC/UFLA
(Co-Orientadora)

_____

Prof. Luiz Gonzaga de Castro Júnior.
DAE/UFLA
(Orientador)

LAVRAS
MINAS GERAIS – BRASIL
2002

RESUMO

Este trabalho tem como objetivo o desenvolvimento de um *software* que encontra padrões no Banco de Dados da Tabela de Negociações e Cotações da *commodity* café da BM&F (Bolsa de Mercadorias & Futuros). O programa gera informações sobre o andamento da *commodity*, que são de grande interesse para um profissional, para que ele tenha maior base e facilidade para tomar suas decisões.

Palavras-chave: Mercado Futuro de Café, Busca de Padrões e Aplicações *Web*.

ABSTRACT

The "pattern recognition" consists of the study of repetitions in any set. Such process has larger applicability, among others, in great databases. The purpose of this work is the development of software that finds patterns in the database of Negotiations and Quotations commodity coffee of BM&F (*Bolsa de Mercadorias & Futuros*). The program generates information about tack of the commodity, which are of great interest for an expert, so that it has greater base and easiness to make its decisions.

Keywords: Future Market of Coffee, Pattern Recognition and Web Applications.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

A agricultura, de maneira geral, apresenta fatores variáveis característicos que a classificam como sendo um mercado que se aproxima da concorrência perfeita. Um dos principais problemas que estas características acarretam é em relação a precificação, uma vez que o produtor rural é definido como tomador de preço.

Inserida nesse mercado, a cafeicultura reflete o grave problema da instabilidade de preços, pois com isso, os cafeicultores ficam a mercê das variações de preço do seu produto na hora da venda, tornando o gerenciamento de sua atividade cada vez mais complexo e arriscado. Para [Fontes (2001)] é necessário que o cafeicultor adote uma postura mais empresarial, agindo com racionalidade administrativa e utilizando os diversos instrumentos produtivos, financeiros e comerciais disponíveis, que vão lhe propiciar esta racionalidade.

O surgimento do mercado futuro vem de encontro às necessidades de mecanismos que auxiliem esta racionalidade, pois a principal finalidade do mercado futuro é a fixação de preço da *commodity*, eliminando o risco da variação de preço, pois há uma inter-relação entre os preços futuros e os preços à vista. Quando o cafeicultor busca a comercialização no mercado futuro ele realiza uma operação chamada de *hedge* ou *hedging*, que consiste no ato de defesa contra variações futuras adversas no preço.

Os *hedgers* são agentes de mercado que têm interesse na *commodity* negociada. Podem ser cafeicultores, beneficiadoras, torrefadoras, exportadores, etc. Para a realização do *hedge* é necessário que exista agentes dispostos a correr o risco da variação de preço, pois o que ocorre com o *hedge* é a transferência do risco de variação do preço da *commodity*, dos agentes "hedgeados" para outros que estão dispostos a assumir tal risco. Esses agentes recebem o nome de especuladores, e apesar da imagem negativa que se tem deles, estes são de suma importância para a

operacionalização do mercado derivativo, uma vez que são eles os responsáveis pela liquidez no mercado.

Quando o produtor realiza o *hedge*, ele elimina totalmente o risco de variação do preço, mas passa a sofrer o risco da variação da base. Conforme [Leuthold (1989)] e [Hull (1996)], a base é considerada como sendo a diferença do preço da *commodity* no mercado físico à vista, na praça local de comercialização e o preço futuro para determinado mês de vencimento do contrato.

Quando a variação do preço à vista cresce mais do que a variação do preço futuro, diz-se que houve um fortalecimento de base e o inverso diz-se que houve um enfraquecimento de base.

Portanto, quanto menor for o risco de base, maior será a utilidade dos contratos futuros como mecanismo de transferência de risco e maior garantia de preço para os *hedgers*, propiciando uma maior utilidade do mercado derivativo como instrumento de gerenciamento da comercialização.

Mineração de dados (*Data Mining*) é um nome utilizado para os métodos e técnicas computacionais para a extração de informações úteis em bancos de dados, através de um resumo compacto dos mesmos.

Sendo assim este trabalho tem por objetivo criar um *software* que identifica padrões em um Banco de Dados histórico usando as informações da Bolsa de Mercadorias & Futuros, referentes às negociações e cotações em geral da *commodity* café. Tal *software* poderá ser utilizado como parte do processo de mineração de dados. A identificação dos padrões pode contribuir, por exemplo, na definição de melhores épocas para se negociar nesses mercados.

2. Histórico do Mercado Brasileiro do Café

O tradicional café foi, por muitas décadas, a mola propulsora da economia brasileira. Sua importância pode ser avaliada não só pela responsabilidade direta de grande parcela do desenvolvimento nacional, mas também como gerador de empregos e de considerável renda para todos os agentes envolvidos, além de projetar uma imagem internacional do *agrobusiness* do país.

A importância do produto para o Brasil data da época do Império e sua cultura iniciou-se em nosso país em 1727, no estado do Pará, e isso se deve a Francisco de MELO palheta. Logo de inicio, firmou-se como uma nova fonte de riqueza em decorrência do declínio das produções da cana de açúcar, cacau e algodão. Chegou ao Rio de Janeiro por volta de 1770 e se espalhou por bairros estritamente urbanos, onde a cafeicultura praticada era disseminada por pequenas propriedades, não se registrando, no período, o processo de concentração fundiária [Mario (2002)].

Já em São Paulo, sua produção concentrou-se, inicialmente, no Vale do Paraíba, atingindo depois outras regiões. Atualmente, há que se destacar a infra-estrutura portuária desse estado, fundamental para o escoamento da produção de outras regiões produtoras e do seu parque industrial de café, o maior do país [Mario (2002)]. No estado de Minas Gerais, no inicio do século XIX, têm-se registros de plantações de café que fizeram desencadear um processo de redistribuição de atividade e áreas economicamente ativas [Mario (2002)]. O Estado do Rio de Janeiro, que era o maior produtor de café do Brasil, cedeu esta posição a São Paulo por volta de 1886, perdendo ainda par Minas Gerais. Em 1928, São Paulo perdeu a posição de terceiro lugar para o estado do Espírito Santo [Mario (2002)].

No período de 1820 a 1850, o Brasil assumiu a liderança da produção mundial de café (40%), o que representava cerca de 70% do valor

das exportações, levando o governo, diante da importância da cafeicultura na economia, a realizar várias intervenções [Mario (2002)].

Até 1989, a cadeia agroindustrial do café era um exemplo de setor, coordenado por agências nacionais e internacionais de regulamentação e o Brasil bancava o controle da oferta do café no mercado mundial. No final dos anos 80, o ambiente institucional do *agrobusiness* do café alterou-se substancialmente. A desregulamentação dos mercados interno e externo deixou para o livre mercado a coordenação desse agronegócio. O Acordo Internacional do Café não foi renovado e houve a extinção, no Brasil, do Instituto Brasileiro do Café. O agronegócio do café, considerado, até então, um setor tradicional dentro do *agrobusiness* brasileiro, se viu diante de uma queda no consumo nos mercados interno e externo, e perda paulatina de liderança o processo [Mario (2002)].

O Brasil em 1906, era responsável por cerca de 80% da exportação e três quartos da produção mundial. Vegro (1994) apresenta alguns dados que confirmam a perda da posição de maior produtor e exportador mundial: a partir de 1950 suas exportações caíram para 40% e na década de 80, representavam cerca de 25%.

Em decorrência de estratégias adotadas, pelo Brasil, como a manutenção de preços artificialmente elevados e perda de qualidade do produto nacional, outros países tornaram-se concorrentes, participando na produção e exportação do produto, colaborando para a perda de mercado externo.

Entre 1985 e 1990, ocorreu uma queda no consumo interno segundo Saes (1995), que indica como principais determinantes: a mudança nos hábitos e costumes alimentares dos consumidores cada vez mais exigentes e seletivos, e além disso, devido ao aumento na diversidade de produtos concorrentes.

Presencia-se, hoje, o esforço das empresas que compõem o setor de bebidas, investindo, continuamente, em novos produtos e estratégias de *marketing* que concorrem diretamente com o café na tentativa de buscar atender aos desejos e necessidades do consumidor, além de criar novas preferências. Tudo isso com base em dados empíricos, opiniões e referências obtidas por meio de propagandas ou experiências [Mario (2002)].

Nos mercados internacionais, principalmente o europeu e o americano, desenha-se uma forte mudança nos padrões de concorrência onde verifica-se um crescimento gradativo da demanda em busca de produtos diferenciados e de maior qualidade – *cafés especiais* -, com a predominância de uma orientação voltada para o mercado. Ressalta-se que tal aumento é decorrente da melhora natural do poder aquisitivo do consumidor, dos programas de *marketing* desenvolvidos e da qualidade da bebida, de acordo com o Agrianual (2002).

O estado de Minas Gerais destaca-se como líder nacional na produção de café (49,68%), no Brasil. Sendo a região sudeste considerada líder no consumo interno, com 54,8% do consumo no mercado brasileiro. Não obstante ainda ser o Brasil o maior produtor mundial de café (27% da produção total), maior exportador e segundo maior mercado interno de café do mundo. Assim, emerge daí a justificativa deste estudo[Mario (2002)].

3. Mineração de Dados

Mineração de dados (*Data Mining*) é um termo genérico utilizado para todos os novos - e os não tão novos assim - métodos e técnicas computacionais para a extração de informações úteis de bilhões de bits de dados através de um resumo compacto dos mesmos.

Na verdade, mineração de dados é um dos passos que compõe um processo maior denominado de Descoberta de Conhecimento em Base de Dados - DCBD (*Knowledge Discovery in Databases* - KDD), que é realizado por ferramentas computacionais em desenvolvimento para crescentes volumes de dados. [Korab (2001)]

> "Mineração de Dados é um passo no processo de KDD que consiste na aplicação de análise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões (ou modelos) particular sobre os dados".

Usama Fayyad, pesquisador sênior na *Microsoft Research* e pioneiro em KDD; trecho extraído de *AI Magazine*, outono de 1996.

3.1. Definição e Conceito Geral

O termo "mineração de dados" é somente um de vários termos, incluindo extração de conhecimento, arqueologia de dados, colheita de informações e *software*. Todos estes termos descrevem na verdade o conceito de descoberta de conhecimento em base de dados. Logo, a idéia de mineração de dados é:

> "O processo não-trivial de identificar padrões válidos, novos, potencialmente úteis e compreensíveis em dados".

O termo Descoberta de Conhecimento em Base de Dados - DCBD (KDD) foi formalizado em 1989, em referência ao amplo conceito de procurar conhecimento em dados. O termo mineração de dados então, é a etapa de aplicação de técnicas/ferramentas para apresentar e analisar dados.

A denominação "mineração de dados" tem sido utilizada por estatísticos, analistas de dados e a comunidade de sistemas de gerenciamento de informação (*Management Information Systems* - MIS), enquanto KDD tem sido utilizada por pesquisadores na área de inteligência artificial.

3.2. Necessidade de Mineração de Dados

Os algoritmos complexos utilizados na mineração de dados têm existido nas últimas duas décadas. O governo dos EUA tem usado *software* próprio de mineração de dados utilizando redes neurais, lógica *Fuzzy* e reconhecimento de padrões na análise de fraudes em impostos e escuta em comunicações internacionais. Estas ferramentas, até então, têm sido de domínio somente das grandes corporações devido ao alto custo envolvido.

Avanços na coleta de dados científicos (por exemplo, sensores remotos e satélites espaciais), processamento de código de barras e transações governamentais têm aumentado em muito o volume de dados. Aliados aos avanços na área de armazenamento, ao uso extensivo de sistemas de gerenciamento de banco de dados e à tecnologia de D*ata Warehousing* (*vide* próximo tópico), a magnitude dos dados tem evoluído drasticamente. O banco de dados do projeto do código genético humano e

pesquisas astronômicas têm produzido *terabytes*[1] de dados. Imagens remotas de satélites e outros instrumentos espaciais são capazes de produzir 50 *gigabytes* por hora.

Assim, vários fatores têm sido combinados para trazer a mineração de dados no foco das atenções do processo de decisão comercial. Alguns destes são:

- O valor nulo em grandes bancos de dados; (pois um valor nulo em um banco de dados não é um valor 0 ou em branco e sim um valor que pode ser qualquer valor possível para este campo).
- Consolidação dos registros de bancos de dados na direção de clientes individuais;
- Conceito de uma informação ou *Data Warehouse*, através da consolidação de bancos de dados;
- A dramática queda na taxa de custo/desempenho de sistemas de *hardware*, tanto para armazenamento como para processamento. O *Bank of America* gastava US$ 24 por consulta numa base de 800 *Gbytes* de dados em 1995, comparado aos US$ 2,430 por consulta em 15 *Gbytes* em 1985. Um *terabyte* de armazenamento custaria US$ 10 milhões em 1990; agora custa menos de US$ 1 milhão;
- Intensa competição de um mercado em crescente saturação;
- A habilidade de direcionar a manufatura, mercado e propaganda para segmentos pequenos e indivíduos;
- O mercado para produtos de mineração de dados está estimado em cerca de US$ 800 milhões até o ano 2000;

[1] Um *terabyte* equivale a 1.000.000MB

Tecnologias de mineração de dados são caracterizadas pela intensiva computação em grandes volumes de dados. Poder de processamento significante é crítico, e paralelismo é essencial para poder permitir uma mineração de dados significativa. Logicamente, uma arquitetura de sistemas balanceada que suporte I/O, computação e escalabilidade a um custo efetivo é desejável. As duas tecnologias de arquitetura paralela disponíveis atualmente são sistemas de processamento paralelo massivo (*Massively Parallel Processing Systems* - MPP) e sistemas de multiprocessamento simétrico (*Symmetric Multiprocessing Systems* - SMP).

3.3. Data Warehouse

O potencial da mineração de dados pode ser melhorado se os dados apropriados tiverem sido coletados e armazenados em um *data warehouse*. Um *data warehouse* é um sistema de gerenciamento de banco de dados relacional (*Relational Database Management System* - RDMS) desenvolvido especificamente para atender as necessidades de sistemas de processamento de transações.

Superficialmente, pode-se definir *data warehouse* como um repositório centralizado de dados que pode ser consultado para benefícios comerciais, contudo posteriormente, será definido de forma mais clara. O *Data Warehousing* é uma nova e poderosa técnica, tornando possível a extração de dados operacionais e superação de inconsistências entre formatos de dados legados. Assim como é possível a integração de dados através da empresa independente da localização, formato ou requerimentos de comunicação, também é possível a incorporação de informações adicionais.

Em outras palavras, um *data warehouse* fornece dados que já estão transformados e resumidos, tornando apropriado um ambiente para aplicações DSS e EIS mais eficientes [Thearling (2002)].

A primeira etapa importante no processo de mineração de dados é organizar grandes volumes de dados em alguma forma de categoria para facilitar a busca, interpretação e organização por usuários finais. Reunir os dados para a "mineração" pode ser um processo difícil. Normalmente, os dados são armazenados de uma maneira imprópria para a extração.

As ferramentas para *data warehouse* consistem em dois tipos: transformação e limpeza de dados; e ferramentas de acesso para usuários finais. Estas ferramentas asseguram que o *data warehouse* contenha integridade de dados, consistência através do tempo, alta eficiência e baixo custo de operação. O elemento importante de um *data warehouse* é que os dados sejam armazenados em diferentes níveis de detalhamento, permitindo o acesso rápido aos mesmos. Estes tipos de dados podem ser explorados para a mineração de dados

O maior problema em *data warehousing* é a qualidade dos dados. Para evitar o princípio de GIGO (*garbage in garbage out* - literalmente lixo dentro, lixo fora), os dados devem ter valores nulos mínimos, porque isso afeta os resultados da mineração de dados. A chave é continuamente monitorar os dados à medida em que vão sendo adicionados ao *data warehouse* e fazer um exame formal dos dados através de uma mineração preliminar para asseguar a integridade dos mesmos [Thearling (2002)].

3.4. Mineração de Dados *Versus* Consultas Tradicionais

Consultas tradicionais em bancos de dados contrastam com mineração de dados simplesmente, porque estão limitados a questões simples tais como "quais foram as vendas de suco de laranja em janeiro de

1995 em São Paulo?". A análise multidimensional, geralmente chamada de *On-Line Analytical Processing* (OLAP), habilita os usuários a fazerem consultas muito mais complexas, tais como a comparação das vendas programadas e das reais em uma determinada região nos anos de interesse. Novamente, a ênfase em ambas as áreas é que os resultados e valores derivados de uma extração ou agregação de valores existentes.

A mineração de dados, por outro lado, através do uso de algoritmos específicos ou mecanismos de busca, tenta descobrir padrões discerníveis e tendências nos dados e inferir regras para os mesmos. Com estas regras ou funções, o usuário está habilitado a suportar, revisar e examinar decisões em alguma área comercial ou científica relacionada.

Durante o processo de transição de dados comerciais para **informações** comerciais, cada nova etapa incrementou a anterior. Por exemplo, acesso dinâmico a dados é crítico para aplicações de navegação de dados, e a habilidade para armazenar grandes bancos de dados é crítica para a mineração de dados. Do ponto de vista do usuário, as quatro etapas listadas na Tabela 1 foram revolucionárias, porque permitiram que novas questões pudessem ser respondidas de maneira rápida e precisa [Thearling (2002)].

Tabela 1 - Etapas na evolução da Mineração de Dados

Etapa evolucionária	**Questão comercial**	**Tecnologias disponíveis**	**Fornecedores de produtos**	**Características**
Coleção de dados (1960s)	"Qual foi minha receita total nos últimos cinco anos”?	Computadores, fitas e discos	IBM, CDC	Retrospectiva, distribuição de dados estática
Acesso a dados (1980s)	"Quais foram as vendas unitárias de São Paulo em março"?	Bancos de dados relacionais (RDBMS), *Structured Query Language* (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospectiva, distribuição de dados dinâmica a nível de registros
Data warehousing & Suporte à decisão (1990s)	"Quais foram as vendas unitárias de São Paulo em março? Avalie também Campinas".	*On-Line Analytical Processing* (OLAP), bancos de dados multidimensionais, *data warehouses*	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospectiva, distribuição dinâmica de dados a múltiplicos níveis
Mineração de dados (atualmente)	"Qual a previsão para as vendas de Campinas no próximo mês? Por quê”?	Algoritmos avançados, computadores multiprocessados, bancos de dados massivos	Pilot, Lockheed, IBM, SGI, e outras (novas empresas)	Prospectiva, distribuição de informação ativa

Fonte: [Thearling (2002)]

As companhias estão começando a perceber que o seu mais precioso patrimônio é a informação que eles possuem do consumidor e dos padrões de compra. O aumento da competitividade depende da qualidade da tomada de decisão e de uma melhora na qualidade da tomada de decisão baseada em transações e decisões anteriores.

A habilidade de melhorar o conhecimento sobre consumidores e mercado permitirá que gerentes direcionem melhor seus produtos e serviços. Por exemplo, vendedores serão capazes de apontar consumidores para promoções e gerenciar bancos de dados de estoque; companhias de telecomunicação serão capazes de estabelecer padrões e características de um dado grupo de usuários, personalizando as contas e análise de lucro; e, instituições financeiras poderão consolidar informações para segmentar produtos financeiros.

Descoberta de Conhecimento em Base de Dados (KDD)

1. **Definição de metas:**
   Definição do problema a ser resolvido através do processo de KDD.

2. **Seleção:**
   Seleção ou segmentação dos dados apropriados para a análise de acordo com algum critério. Por exemplo, todas as pessoas que possuem um carro - desta maneira subconjuntos dos dados podem ser determinados.

3. **Pré-processamento:**
   Eliminação de ruídos e erros, estabelecimento de procedimentos para verificação da falta de dados; estabelecimento de convenções para nomeação e outros passos demorados para a construção de uma base de dados consistente.

4. **Transformação:**
   Redução dos dados através da busca por atributos que representem várias características dos dados.

5. **Mineração de Dados:**
   Aplicação dos algoritmos para descoberta de padrões nos dados; envolve a seleção de métodos/técnicas e modelos que melhor se enquadram no cumprimento das metas estabelecidas.

6. **Interpretação/Avaliação:**
   Pode requerer a repetição de vários passos, mas normalmente é encarada como uma simples visualização dos dados.
   Os padrões identificados pelo sistema são interpretados em conhecimento, que pode então ser utilizado para suportar a tomada de decisão humana.

3.5. Algoritmos de Mineração de Dados

Os objetivos principais da mineração de dados são a previsão e a descrição. A previsão faz uso de variáveis existentes no banco de dados para prever valores desconhecidos ou futuros. A descrição é voltada para a busca de padrões descrevendo os dados e a subseqüente apresentação para a interpretação do usuário.

A relativa ênfase entre previsão e descrição varia de acordo com o sistema de mineração de dados utilizado. Estes objetivos são conseguidos através de vários algoritmos. Tais algoritmos são incorporados em vários métodos de mineração de dados.

Existem vários algoritmos de mineração de dados utilizados para resolver problemas específicos. Estes são categorizados em algoritmos de associação, classificação, padrões seqüenciais e agrupamento. A premissa básica de algoritmos de associação é achar todas as associações em que a presença de um conjunto de itens em uma transação implica em outros itens. Algoritmos de classificação ou geração de perfis desenvolvem perfis de diferentes grupos. Algoritmos de padrões seqüenciais identificam tipos de padrões seqüenciais em restrições mínimas especificadas pelo usuário. Algoritmos de agrupamento segmentam o banco de dados em subconjuntos ou grupos [Yanaga (2002)].

3.5.1. Regras de Associação

Algoritmos de associação têm numerosas aplicações, incluindo supermercados, planejamento de estoque, mala direta para marketing direcionado e planejamento de promoções de vendas. Por exemplo, a regra de associação deriva a partir da mineração de dados de um banco de dados de transações (através do leitor de código de barras de produtos), uma lista contendo o conjunto de itens comprados pelo consumido em uma visita à loja. A regra de associação poderia ser:

```
75% dos consumidores que compram Coca-cola também
compram batata frita.
```

O número "75%" refere-se ao fator de confidência (*confidence factor*), uma medida do poder preditivo da regra. O item do lado esquerdo da regra (*left hand side* - LHS) é Coca-cola, enquanto que batata frita está do lado esquerdo da regra (*right hand side* - RHS). O algoritmo produz uma grande quantidade destas regras e cabe ao usuário selecionar o subconjunto de regras que têm graus de confidência maiores e também porcentagem de listas que seguem a esta regra. Podem existir também múltiplas associações tais como:

```
65% dos consumidores que compram Coca-cola e batata
frita também compram sorvete.
```

É importante para o usuário determinar quando existe algum elemento com chance de correlação (Coca e batata-frita sendo vendidas) ou quando existe alguma correlação desconhecida, mas importante (sorvete também estava sendo comprado). O impacto aqui é como o supermercado pode incrementar as vendas de sorvete? O que acontece se houver uma promoção de Pepsi? Em outras palavras, quais itens devem ser colocados lado a lado na prateleira? Um conjunto de itens relacionados deve estar seguidamente um após o outro.

Por não serem relevantes aos propósitos deste trabalho, as outras famílias citadas de algoritmos de mineração de dados, regras de classificação, padrões seqüenciais e agrupamento, não serão tratadas.

3.6. Ferramentas de Mineração de Dados

Qualquer algoritmo de mineração de dados é composto de 3 componentes: representação através de modelo, avaliação do modelo e método de busca.

Resumidamente, o modelo deve representar limites flexíveis e suposições adequadas, de uma maneira que os padrões possam ser descobertos; o modelo deve ter validade preditiva - que pode ser baseada em validações cruzadas; então, a busca deveria otimizar os critérios de avaliação do modelo de acordo com os dados observados e a representação do modelo.

As ferramentas de mineração ou mecanismos de busca são usualmente programas ou agentes automatizados inteligentes, incorporando alguma forma de inteligência artificial em bancos de dados relacionais. Os agentes detectam padrões predefinidos e alertam o usuário sobre variações. Vários tipos de ferramentas são utilizadas na mineração de dados: redes neurais, árvores de decisão, indução de regras e visualização de dados. [Yanaga (2002)]

3.6.1. Indução

Um banco de dados é um armazém de informações, mas o mais importante é a informação que pode ser inferida deste. Existem duas técnicas principais de inferência disponíveis: dedução e indução.

- Dedução é uma técnica de inferência de informação que é uma conseqüência lógica da informação no banco de dados, tal como o operador join aplicado em duas tabelas relacionais onde o primeiro diz respeito aos empregados e departamentos e o segundo, departamentos e gerentes - infere em um relação entre empregados e gerentes.

- Indução foi descrita anteriormente como um técnica de inferência de informações que é generalizada através do banco de dados, tal como exemplo mencionado acima para inferir que cada empregado tem um gerente. Este é um nível de informação ou conhecimento alto do ponto de vista de que é uma regra geral sobre objetos no banco de dados. O banco de dados é vasculhado por padrões ou regularidades.

A indução tem sido utilizada das seguintes maneiras na mineração de dados:

3.6.2. Árvore de Decisão

Árvores de decisão são simples representações de conhecimento e classificam exemplos em um número finito de classes. Os nós são rotulados com nomes de atributos, os arcos são rotulados com possíveis valores para este atributo e as folhas são rotuladas com diferentes classes. Objetos são classificados através de um caminho percorrendo a árvore - seguindo os arcos que contêm valores que correspondem a atributos no objeto.

3.6.3. Indução de Regras

Através de um sistema de indução de regras será gerado um conjunto de condições não-hierárquicas, que será utilizado para prever valores em novos itens de dados. Certas aplicações de *software* tendem a avaliar e refinar o conjunto de regras através da seleção das melhores regras e evitam certas regras. As regras utilizadas para a predição são mais gerais e mais poderosas do que as árvores de decisão, utilização florestas de predição (com várias árvores de decisão parciais) com escalas estendidas de valores. Estes modelos preditivos são totalmente transparentes e provêm explanações completas para suas predições.

Uma companhia de cartões de crédito, por exemplo, pode ter registros de consumidores contendo descrições ou atributos. Com histórico de crédito conhecido, os registros pode ser rotulados/classificados como bons, médios ou ruins. Uma técnica de indução pode produzir um modelo de classificação simbólica que gera uma regra estabelecendo "se um portador de cartão ganha >= $25.000, tem entre 45-55 anos de idade, e mora em um determinado CEP, então o portador do cartão tem um bom risco de crédito".

3.6.4. Análise de Grupos

Em um ambiente de aprendizagem não supervisionada, o sistema deve descobrir suas próprias classes e uma maneira de fazê-lo é agrupar os dados em um banco de dados. O primeiro passo consiste na descoberta de subconjuntos de objetos relacionados e então encontrar as descrições, tais como D1, D2, D3, etc - cada uma das quais descrevendo um destes conjuntos.

O agrupamento e a segmentação basicamente particionam o banco de dados de forma que cada partição ou grupo seja similar de acordo com algum critério ou métrica. O agrupamento de acordo com alguma similaridade é um conceito que aparece em muitas disciplinas. Se uma medida de similaridade é disponível, existe um grande número de técnicas para a formação de grupos. A associação aos grupos pode ser baseada em um nível de similaridade entre membros e através disso, as regras de associação podem ser definidas. Outras soluções seriam construir um conjunto de funções que mede alguma propriedade das partições. Esta última solução resulta no que chamamos de taxa de segmentação ótima.

Muitas aplicações de mineração de dados utilizam o agrupamento de acordo com a similaridade para segmentar uma base de clientes/consumidores. O agrupamento de acordo com a otimização de um determinado conjunto de funções é utilizado nas análises de dados, tal como

na determinação de tarifas de seguros os clientes podem ser segmentados de acordo com um número de parâmetros e a segmentação de tarifas ótima pode ser alcançada.

O agrupamento/segmentação em bancos de dados são os processos de separar o conjunto de dados em componentes que refletem um padrões consistente de comportamento. Uma vez que os padrões tenham sido estabelecidos, estes podem ser utilizados para "desmontar" os dados em subconjuntos mais compreensíveis e também podem prover subgrupos de uma população para futuras análises - o que é importante quando lidando com grandes bancos de dados. Por exemplo, um banco de dados poderia ser utilizado para a geração de perfis para marketing direcionado onde a resposta prévia às campanhas de mala direta geraria um perfil das pessoas que responderam; a partir disso, faz-se a previsão de resposta e filtra-se a lista de mala direta para obter o melhor resultado.

3.6.5. Aprendizagem Induzida

Indução é a inferência de informação através de dados e aprendizagem induzida é o processo de construção de modelos onde o ambiente, por exemplo - um banco de dados é analisado em uma visão para a procura de padrões. Objetos similares são agrupados em classes e regras são formuladas onde for possível prever as classes de novos objetos.

Este processo de classificação identifica classes de forma que cada classe tenha um único padrão de valores que forma a descrição da classe. A natureza do ambiente é dinâmica, pois o ambiente deve ser adaptativo de forma que possa aprender.

Geralmente só é possível a utilização de um pequeno número de propriedades para a caracterização de objetos, então fazemos abstrações em que os objetos que satisfazem um mesmo subconjunto de propriedades são mapeados na mesma representação interna.

A aprendizagem induzida em que o sistema infere conhecimento por si só através da observação de seu ambiente tem duas estratégias principais:

- aprendizagem supervisionada - é a aprendizagem através de exemplo onde o professor auxilia o sistema a construir um modelo através da definição de classes e fornecimentos de exemplos para cada uma. O sistema deve achar uma descrição de cada classe, tal como as propriedades comuns dos exemplos. Uma vez que a descrição tenha sido formulada, a descrição e a classe formam uma regra de classificação que pode então ser utilizada para a previsão de classes de objetos ainda não vistos. Esta técnica é similar a análise discriminativa em estatística.

- aprendizagem não-supervisionada - é a aprendizagem através de observação e descoberta. O sistema de mineração de dados é suprido com objetos mas nenhuma classe é definida de forma que este deve observar os exemplos e reconhecer padrões (descrição das classes) por si mesmo. Este sistema resulta em um conjunto de descrições de classes, cada classe descoberta possui uma descrição no ambiente. Novamente, isto é similar à análise de grupos em estatística.

Logo, indução pode ser entendida como a extração de padrões. A qualidade do modelo produzido pelos métodos de aprendizagem induzida é tal que o modelo pode ser utilizado para prever o desenvolvimento de situações futuras. O problema é que a maioria dos ambientes tem diferentes estados e conseqüentemente diferentes mudanças entre eles, de modo que não é possível sempre verificar um modelo através de todas as suas situações possíveis.

Dado um conjunto de exemplos, o sistema pode construir múltiplos modelos - alguns dos quais podem ser mais simples que os outros. Os modelos mais simples têm maior probabilidade de estarem corretos se nós aderirmos ao Ockhams razor, que especifica que se existirem múltiplas

explicações sobre um fenômeno particular, fará sentido a escolha do mais simples, porque é mais provável que este capture a natureza do fenômeno.

3.6.6. Estatística

A estatística tem uma sólida fundamentação teórica, mas os resultados da estatística podem ser grandes demais e difíceis de interpretar, pois necessitam do usuário para verificar onde e como analisar os dados. A mineração de dados, entretanto, permite que os conhecimentos do especialista sobre os dados e técnicas de análise avançada do computador trabalhem de maneira conjunta.

Sistemas de análise estatística tais como SAS e SPSS têm sido utilizados por analistas para a detecção de padrões incomuns e explicação de padrões utilizando modelos estatísticos tais como modelos lineares. A análise estatística tem um campo enorme de utilização e a mineração de dados não irá substituir tais análises, e sim utilizar análises mais diretas baseadas nos resultados da mineração de dados. Por exemplo, a técnica de indução estatística é algo como a taxa média de falha nas máquinas.

4. Importância das Tecnologias Computacionais

Historicamente, observa-se que as organizações, na tentativa de melhorarem o processo de tomada de decisão, são cada vez mais pressionadas a obterem informações de forma mais rápida e confiável. Nas últimas décadas, a Tecnologia da Informação evoluiu consideravelmente, dos primeiros computadores centrais até os atuais sistemas distribuídos.

Diante dessa realidade, indiscutivelmente, a área computacional tornou-se uma ferramenta vital no processo de tomada de decisão. Assim, para o desenvolvimento deste projeto, algumas tecnologias foram amplamente úteis, uma vez que supriram perfeitamente as necessidades surgidas durante o desenvolvimento.

4.1. Apache

O Apache é um programa para gerenciar o acesso de páginas em programas, via Internet, no computador onde este está instalado. Este computador passa a ter o nome de servidor de páginas. Então, pode-se acessar as páginas e os programas deste computador de qualquer lugar pela Internet.

4.2. PHP

É uma linguagem de programação para a Internet, mas só funciona em servidores de páginas. Para se poder usar este tipo de linguagem de programação é preciso primeiro instalar um servidor de páginas.

É muito segura pois, as informações que trafegam na Internet são só as informações de visual, as informações do programa ficam todas no

servidor dificultando assim o roubo de informações confidenciais sobre o programa ou a ação do usuário.

4.3. MySQL

É um servidor de banco de dados de fácil utilização, a linguagem PHP já possui todas as funções de acesso ao MySQL e este também é gratuito.

5. Metodologia

5.1. Dados do Trabalho

Os dados utilizados neste trabalho foram obtidos na BM&F e compreendem o período de 29 de outubro de 1996 a 18 de outubro de 2002.

5.2. Operacionalização

Para criar o programa que faria a busca de padrões foi utilizado um outro programa como molde, o Sistema de Simulação em Mercados de Derivativos (SimHedge). Este sistema, que propicia o treinamento em Mercados Futuros de café (BM&F) através da simulação, foi desenvolvido pelo autor, durante um projeto de pesquisa no Departamento de Administração e Economia (DAE), no período de agosto de 2001 a julho de 2002. Tal programa usa informações como cotações de preço, dólar, número de negociações, entre outros.

O processo inicial de construção do sistema foi definido a partir do servidor Apache de Internet configurado rodando em conjunto com a linguagem PHP, um banco de dados MySQL e uma aplicação *Web* com uma base de dados da BM&F, que já havia sido preparado para o desenvolvimento do SimHedge.

Então usando os novos materiais deste Referencial Teórico pôde-se desenvolver um programa que superou as expectativas como pode ser conferido no capitulo seguinte.

6. Resultados & Discussão

6.1. Descrição

Busca Completa

A forma encontrada para realizar uma busca completa, automaticamente, em menos de 5 minutos, foi através do uso de um contador. Esse contador fornece a indicação de qual busca está sendo processada.

Busca por Colunas

Para descobrir quais colunas deveriam ser usadas, os conceitos da analise combinatória foram essenciais para a obtenção dos resultados práticos. As técnicas dessa área matemática, encaixaram-se perfeitamente às necessidades circunstanciais dessa parte do desenvolvimento do projeto. Buscas usando somente três colunas exigem combinações, conforme descrito abaixo:

C(3, 1) + C(3, 2) + C(3, 3) = 3 + 3 + 1 = 7 buscas

Diante dos resultados, evidências de que a quantidade de grupos e a quantidade de elementos em cada grupo correspondiam ao Triangulo de Pascal foram verificadas. Assim, pode-se constatar que tratava-se de um Triângulo de Pascal sem as C(n, 0), ou seja, sem a primeira coluna:

0 => 1
1 => 1 1
2 => 1 2 1
3 => 1 3 3 1
4 => 1 4 6 4 1
...

Um programa onde o usuário pode configurar cada parâmetro, como:

Qual o banco de dados a ser usado?

Qual a tabela a ser usada?

Qual a medida a ser usada? Porcentagem ou quantidade?

Qual o valor da medida anterior será usado como restrição?

Quantas colunas da tabela serão utilizadas?

Quais as colunas da tabela serão utilizadas?

Tais questões são apresentadas na Figura 1.

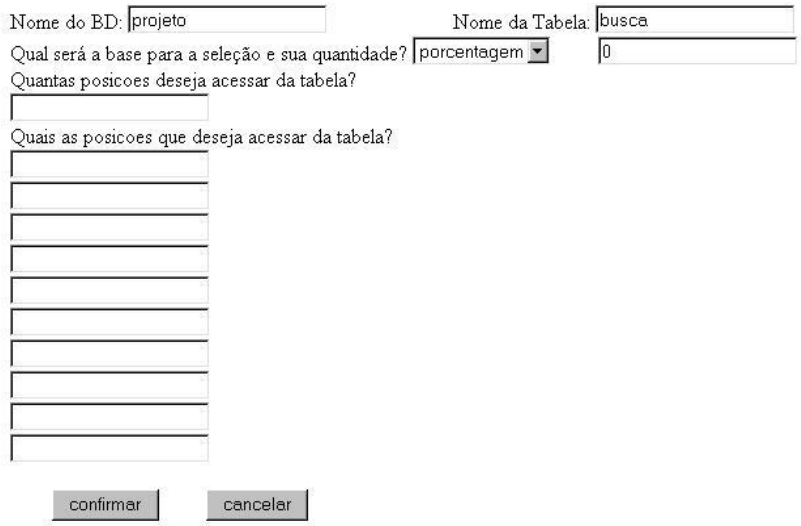

Figura 1 - Tela de Configuração

O programa fará todas as buscas necessárias para a configuração dos parâmetros e ainda mostrará os padrões resultantes da busca realizada. Além disso, o programa também será capaz de responder às perguntas feitas pelo usuário e também ficará aguardando por elas.

A proposta de busca de padrões foi atingida e estendida à possibilidade de permitir ao usuário a realização de perguntas sobre os padrões descobertos e, como se não bastasse, ainda à apresentação da

relevância do resultado obtido. Relevância esta, referente à porcentagem de ocorrências da resposta em cada padrão.

A diferença entre o programa desenvolvido e um *data mining* é que neste último, o usuário realiza os questionamentos em linguagem natural e as respostas também são em linguagem natural, construindo assim, uma relação amigável entre homem x máquina . Por outro lado, o programa desenvolvido requer um especialista que formule as perguntas de maneira clara e, ao mesmo tempo, técnica (ou seja, para o computador entender) para que diante dos resultados estatisticamente corretos tire conclusões para tomar decisões coerentes. A exigência da existência do especialista deve-se ao fato de as respostas serem retornadas sob a forma estatística, sendo portanto, de difícil domínio para pessoas não conhecedoras da área.

6.2. Algoritmo

Primeiramente, é preciso realizar a configuração dos campos citados no tópico anterior, para que o programa possa fazer a busca de padrões.

1. Descobrir quantas iterações faltam;
2. Testar se todas as possibilidades ainda não foram testadas, caso tenham, avançara para passo 8;
3. Descobrir quantas colunas combinar;
4. Descobrir quais as colunas combinar e seus respectivos nomes;
5. Descobrir todas as classes das colunas combinadas;
6. Testar e admitir as classes que suportarem a restrição;
7. Voltar ao passo 1;
8. Testar se há uma pergunta, se não houver ir para o passo 11;
9. Testar se a pergunta é valida, caso não seja avançar para o passo 12;
10. Mostrar os padrões, as respostas da pergunta e avançar para o passo 12;
11. Mostra os padrões;
12. Espera uma pergunta e vai para o passo 8;

6.3. Exemplo

O programa foi configurado para pesquisar no Banco de Dados projeto, na Tabela busca, a restrição para ser padrão é ser mais de 0% do banco de dados (lembrar que a porcentagem é inteira), 3 colunas da tabela e as colunas são a 1, a 2 e a 3. (DIA, MÊS e ANO, respectivamente).

1. No início faltam 7, que são C(3, 1) + C(3, 2) + C(3, 3) = 3 + 3 + 1;
2. As possibilidades ainda não acabaram;
3. Primeiro combina-se as colunas uma em uma;
4. Combinar-se a primeira coluna e seu nome é DIA;
5. Dias = 1-31;
6. Todos os dias são padrões de acordo com as restrições;
7. Voltar ao passo 1;
1. Agora faltam 6;
2. As possibilidades ainda não acabaram;
3. Ainda combina-se de uma em uma coluna;
4. Combina-se a segunda coluna e seu nome é MES;
5. Meses = jan-dez;
6. Todos os meses são padrões de acordo com as restrições;
7. Voltar ao passo 1;
1. Agora faltam 5;
2. As possibilidades ainda não acabaram;
3. Ainda combina-se de uma em uma coluna;
4. Combina-se a terceira coluna e seu nome é ANO;
5. Anos = 96-02;
6. Todos os anos são padrões de acordo com as restrições;
7. Voltar ao passo 1;
1. Agora faltam 4;
2. As possibilidades ainda não acabaram;
3. Agora se combina de duas em duas colunas;

4. Combina-se a primeira e a segunda coluna e seus nomes são DIA e MES;
5. Dias = 1-31 combinando com os Meses = jan-dez;
6. Nenhum destes foi considerado padrão;
7. Voltar ao passo 1;
1. Agora faltam 3;
2. As possibilidades ainda não acabaram;
3. Ainda combina-se de duas em duas colunas;
4. Combina-se a primeira e a terceira coluna e seus nomes são DIA e ANO;
5. Dias = 1-31 combinando com os Anos = 96-02;
6. Nenhum destes foi considerado padrão;
7. Voltar ao passo 1;
1. Agora faltam 2;
2. As possibilidades ainda não acabaram;
3. Ainda combina-se de duas em duas colunas;
4. Combina-se a segunda e a terceira coluna e seus nomes são MES e ANO;
5. Meses = jan-dez combinando com os Anos = 96-02;
6. A maioria das combinações dos meses com os anos são padrões de acordo com as restrições;
7. Voltar ao passo 1;
1. Agora falta 1;
2. As possibilidades ainda não acabaram;
3. Agora combina-se de três em três colunas;
4. Combina-se a primeira, a segunda e a terceira coluna e seus nomes são DIA, MÊS e ANO;
5. Dia = 1-31 combinando com os Meses = jan-dez e os Anos = 96-02;
6. Nenhum destes foi considerado padrão;
7. Voltar ao passo 1;
1. Agora não falta nenhuma combinação, pular para o passo 8;
8. Como não há perguntas pular para o passo 11;

11. Mostrar os padrões;
12. Esperar uma pergunta e pular para o passo 8;

Agora o sistema fica esperando por perguntas sobre os padrões encontrados. Estas perguntas precisam ser feitar de maneira formal para que o programa possa entender e dar respostas satisfatórias.

7. Conclusões

Ao analisarmos este trabalho, observamos que o programa desenvolvido facilita do trabalho do especialista, pois lhe dá informações de algumas oportunidades para se ganhar ou não perder (dinheiro, alguma disputa, etc.) e este ainda pode fazer muitas "perguntas para o programa" para ajudá-lo em suas decisões. Pois o programa encontra todos os padrões seguindo uma métrica de padrão especificada pelo usuário, com isso o programa encontra padrões que o usuário dificilmente iria encontrar ou até mesmo procurar.

Este programa que inicialmente tinha propósito de procurar padrões na Tabela de Negociações e Cotações da *commodity* café, agora pode realizar esta função em qualquer tabela de qualquer banco de dados, bastando somente este estar em um servidor de banco de dados MySQL e o usuário na tela inicial do programa configurá-lo corretamente. Além disso, para o programa as colunas da tabela que será usada não precisam estar dispostas na mesma ordem da tabela original, para maior comodidade tanto nas buscas quanto nas perguntas.

Como este programa foi feito para rodar em um servidor de Internet este dá aos usuários imensa comodidade, pois estes podem estar em qualquer lugar como: na empresa, em casa ou até mesmo em um cliente para fazer suas consultas ao programa, desde que estes tenham os devidos recursos (*Notebooks*, celulares, *Palms*, etc.).

Existem vários trabalhos futuros possíveis como:

1. A continuação do desenvolvimento de um programa de mineração de dados.
2. Melhorias na interface do programa.
3. Testar o programa em vários outros bancos de dados, entre outros

8. Referências Bibliográficas

[BM&F (2002)] Bolsa de Mercadorias & Futuros. 14 de julho de 2002.

http://www.bmf.com.br/

[Teixeira (1992)] TEIXEIRA, Marco Aurélio. **Mercados Futuros: Fundamentos e Características Operacionais,** Bolsa de Mercadorias & Futuros, 1992

[Shouchana (2000)] SHOUCHANA, Félix. **Introdução aos mercados Futuros e de Opções Agropecuários no Brasil** 2. ed. rev. e atual. São Paulo: Bolsa de Mercadorias & Futuros, 2000.

[Fontes (2001)] FONTES, R. E. **Estudo Econômico da Cafeicultura no Sul de Minas Gerais**. Lavras: UFLA, 2001. 94p. (Dissertação – Mestrado em Administração Rural).

[Hull (1996)] HULL, J. **Introdução aos mercados futuros e de opções**. 2. Ed. São Paulo: Bolsa de Mercadorias e Futuros e Cultura Editores Associados, 1996. 448p.

[Mario (2002)] MARIO, Talestre Maria do Carmo. **Dinâmica Comportamental dos Consumidores de Café: Um Fator Gerador de Ações Mercadológicas.** Lavras: UFLA, 2002.

[Saes (1995)] SAES, M. S. M. **A racionalidade econômica da regulamentação no mercado brasileiro de café.** 1995. 166p. Tese (Doutorado em Economia) – FAE–USP, São Paulo.

[Vegro (1994)] VEGRO, C.L.R et al. **O prazer e a excelência de uma xícara de café expresso:** um estudo de mercado. São Paulo: Ceres, 2002. 111p.

[Exterior (2000)] EXTERIOR, **Comércio (Informe BB).** ed. 31 novembro 2000.

[Iezzi (1981)] IEZZI, Gelson ... [et al.]. **Tópicos de Matemática v.2,** 2.ed. São Paulo Editora ATUAL EDITORA LTDA, 1981.

[Leuthold (1989)] LEUTHOLD, R. M.; JUNKS, J. C.; CORDIER, J. E. **The theory and pratice of future markets**. Massachusetts: Lexington Books, 1989. 410p.

[Yanaga (2002)] YANAGA, Edson; CALVO, Robson Aparecido. **Mineração de Dados**. 14 de julho de 2002. http://www.din.uem.br/ia/mineracao/

[Carvalho (1999)] CARVALHO; SAMPÁIO; MONIOVI. **Utilização de Técnicas de "Data Mining" para o Reconhecimento de Caracteres Manuscritos**, Anais do XIV Simpósio Brasileiro de Banco de Dados, outubro 1999.

[Korab (2001)] KORAB, Holly. **Striking gold in mountains of data**. 2001.
http://access.ncsa.uiuc.edu/Archive/AOarchive/Welge_index.html

[Thearling (2002)] THEARLING, Kurt **An Introduction to Data Mining** - Discovering hidden value in your data warehouse. 2002.
http://www.santafe.edu/~kurt/text/dmwhite/dmwhite.shtml

[Kira (2001)] KIRA, Taraponoff (organizadora). **Inteligência Organizacional e Competitiva.** Brasília: Editora Univercidade de Brasília, 2001.

[Bakken (2001)] BAKKEN, Stig Seather, SCHMID, Egon. **Manual do PHP.** 09 de maio de 2001.
http://snaps.php.net/manual/