Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

# Melhoria de Desempenho na Preparação de Grandes Bases de Dados para Mineração de Regras de Associação Multiníveis

**Marcelo Ladeira Marques**

JUIZ DE FORA
DEZEMBRO, 2014

# Melhoria de Desempenho na Preparação de Grandes Bases de Dados para Mineração de Regras de Associação Multiníveis

Marcelo Ladeira Marques

Universidade Federal de Juiz de Fora

Instituto de Ciências Exatas

Departamento de Ciência da Computação

Bacharelado em Ciência da Computação

Orientador: Custódio Gouvêa Lopes da Motta

JUIZ DE FORA

DEZEMBRO, 2014

# Melhoria de Desempenho na Preparação de Grandes Bases de Dados para Mineração de Regras de Associação Multiníveis

Marcelo Ladeira Marques

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Custódio Gouvêa Lopes da Motta
Doutor em Engenharia Civil / Sistemas Computacionais pela COPPE / Universidade Federal do Rio de Janeiro (2010)

Raul Fonseca Neto
Doutor em Engenharia de Sistemas e Computação pela Universidade Federal do Rio de Janeiro (1990)

Carlos Cristiano Hasenclever Borges
Doutor em Engenharia Civil pela Universidade Federal do Rio de Janeiro (1999)

JUIZ DE FORA

11 DE DEZEMBRO, 2014

# Resumo

O trabalho tem como objetivo propor uma forma de melhoria de desempenho na preparação de grandes bases de dados, bem como realizar um estudo sobre uma metodologia para mineração de regras de associação multiníveis. Como resultado do estudo, pretende-se obter uma ferramenta para facilitar o pré-processamento dos dados de entrada. A ferramenta contruída também deverá ser capaz de realizar automaticamente a conexão com a base de dados de origem, extrair os dados necessários ao funcionamento dos programas utilizados durante a mineração e integrar as diversas etapas do processo em uma única interface. Para comprovar a eficiência da ferramenta desenvolvida, ao fim do trabalho será realizado um estudo de caso.

**Palavras-chave:** descoberta de conhecimento em bases de dados, mineração de dados, regras de associação.

# Abstract

The work aims to propose a way of improving performance in the preparation of large databases, as well as conduct a study on a methodology for mining multilevel association rules. As a result of the study, it is intended to obtain a tool to facilitate pre-processing of input data. The builded tool also have to be able to automatically perform the connection with the source database, extract the necessary data for the operation of the software used during mining and integrate the various process steps in a single interface. To prove the efficiency of the tool, a case study will be done at the end of the work.

**Keywords:** knowledge discovery in databases, data mining, association rules.

# Agradecimentos

A minha família, em especial ao meus pais Lincoln e Rita e ao meu irmão Lincoln Netto, pelo apoio incondicional oferecido ao longo dos anos e pela compreensão e suporte em todos os momentos em que não pude estar presente, seja no trabalho ou em casa.

Ao professor Custódio, pela orientação, paciência e apoio oferecido desde o princípio.

Aos meus amigos de curso: Igor Russo, Bruno Marques, Thiago Marques e Victor Patrocínio, pessoas que tornaram esses quatro anos na faculdade divertidos e sem os quais não seria possível terminar a faculdade em quatro anos.

A Karen Enes pela amizade, pelo apoio oferecido ao longo do curso e por sempre conseguir me convencer de que no final iria dar tempo, mesmo quando parecia que tudo estava dando errado.

A todos os amigos que, mesmo que de maneira indireta, ajudaram na conclusão deste projeto de faculdade, em especial as amizades construidas durantes os três anos de Ensino Médio.

A todos os meus parentes, pelo encorajamento e apoio.

Aos professores do Departamento de Ciência da Computação pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

*"Não fiz o melhor, mas fiz tudo para que o melhor fosse feito. Não sou o que deveria ser, mas não sou o que era antes".*

*Martin Luther King*

# Sumário

# Lista de Figuras

# Lista de Abreviações

| | |
|---|---|
| KDD | Knowledge Discovery in Databases |
| DM | Data Mining |
| AR | Association Rules |

# 1 Introdução

Devido ao rápido desenvolvimento da ciência da computação e a redução nos custos das tecnologias de armazenamento e processamento de dados, juntamente com o aumento de indivíduos e empresas que registram suas informações em bases de dados, o volume de informação disponível vem crescendo rapidamente. Este crescimento mostra-se benéfico, pois possibilita a extração de conhecimentos que podem ser úteis na solução de diversos problemas. Entretanto, essa tarefa de extração em extensas bases de dados nem sempre constitui uma tarefa simples (Trindade e Costa, 2013).

Em consequência, mostra-se necessário o incentivo ao estudo da Descoberta de Conhecimento em Banco de Dados (KDD - Knowledge Discovery in Databases), a qual atua na detecção de padrões, exceções, riscos e relacionamentos de novas formas de consumo de informação, sendo o emprego correto destas tecnologias capaz de transformar os dados em informações novas e úteis (Fayyad et al, 1996; Stock, 2012).

## 1.1 Motivação

O estudo desta área revela-se pertinente devido à crescente gama de aplicações de KDD. Como exemplos de áreas de aplicações, podem ser citadas (Curotto, 2003; Singh et al, 2013):bancária (aprovação de crédito), ciências e medicina (descoberta de hipóteses, predição, classificação, diagnóstico), comercialização (segmentação, localização de consumidores, identificação de hábitos de consumo), engenharia (simulação e análise, reconhecimento de padrões, processamento de sinais e planejamento), financeira (apoio para investimentos, controle de carteira de ações), gerencial (tomadas de decisão, gerenciamento de documentos), Internet (ferramentas de busca, navegação, extração de dados), manufatura (modelagem e controle de processos, controle de qualidade, alocação de recursos) e segurança (detecção de bombas, icebergs e fraudes).

Como outro indicativo da importância dos estudos que buscam formas de gerir um grande volume de dados, entre eles o KDD, pode-se citar o documento da Sociedade

Brasileira de Computação sobre os desafios da computação até 2016 (SBC, 2006), sendo os desafios relevantes para a área definidos neste relatório: "Gestão da Informação em grandes volumes de dados multimídia distribuídos"; "Modelagem computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem-natureza". Em ambos os casos, os desafios possuem relação com o grande volume de dados envolvido. Sendo esta característica recorrente no "Segundo Seminário Grandes Desafios de Pesquisa em Computação no Brasil" (SBC, 2013), no qual pode-se destacar o tema "Redes Complexas de Colaboração e Gestão da Informação sobre Grandes Volumes de Dados".

Após a realização do processo de KDD, a utilização do conhecimento obtido é realizada através de um sistema inteligente ou de um ser humano como forma de apoio à tomada de decisão.

Especificamente, na mineração de regras de associação, uma das tarefas de KDD mais comumente utilizadas, a maior parte do esforço tem sido voltado para a criação de técnicas mais rápidas e eficientes, ou seja, pouco tem sido feito, relativamente, para investigar, em profundidade, as implicações das análises dessas regras (Nicholas e Zhao, 2009).

Diante deste contexto foi desenvolvida a tese de doutorado, intitulada "Metodologia para mineração de regras de associação multiníveis incluindo pré e pós-processamento" (Motta, 2010). O trabalho mencionado implementou um sistema inteligente que utiliza um método de cálculo das distâncias entre pares de itens selecionados de uma base de dados de transações, a partir da associação entre eles. Esse sistema é capaz de criar estruturas de dados que possibilitam a visualização das associações entre itens no espaço bidimensional, fornecendo assim apoio a pós análise do usuário, seja um profissional em mineração ou um especialista do domínio.

## 1.2 Objetivos

Apesar da metodologia proposta em (Motta, 2010) ter apresentado bons resultados, alguns refinamentos ainda podem ser realizados. Esses aperfeiçoamentos justificam a realização do presente trabalho, o qual terá como principais objetivos:

- Desenvolvimento de uma ferramenta para pré-processamento dos dados;
- Implementação de avanços no método, tendo como foco a sua automação e a integração das diferentes ferramentas utilizadas ao longo do processo;
- Aplicação do método sobre uma base real.

A ferramenta de pré-processamento, juntamente com a automação do processo, contribuirão no sentido de simplificar a aplicação do método, possibilitando assim uma utilização mais ágil e menos trabalhosa em trabalhos futuros.

O estudo de caso tem como objetivo fornecer dados consistentes sobre a utilização do método, de forma a validar as modificações realizadas no decorrer do trabalho. Além disso, serão fornecidas evidências sobre a melhoria de desempenho computacional em comparação com o método original.

## 1.3 Metodologia

O foco principal do trabalho será o desenvolvimento de uma ferramenta com o objetivo de facilitar a utilização do método original proposto em (Motta, 2010), melhorando o seu desempenho. A validade desta implementação será confirmada através de um estudo de caso, que possui como finalidade destacar as melhorias obtidas a partir das modificações realizadas no processo.

Quanto ao tipo de trabalho, optou-se pela pesquisa de laboratório e pela bibliográfica, sendo a primeira responsável por descrever e explicar os resultados obtidos durante o estudo experimental e a segunda por fornecer um referencial teórico para o desenvolvimento do trabalho.

Este referencial será obtido através do estudo de material já publicado na área, como por exemplo, livros, teses, monografias, artigos e periódicos, relacionadas com a metodologia sobre a qual o estudo de caso se aplica.

## 1.4 Organização

Os capítulos dois e três apresentam uma revisão teórica sobre os temas abordados ao longo da monografia. O segundo capítulo é responsável por revisar o conteúdo referente ao processo de Descoberta de Conhecimento em Bases de Dados e a extração de Regras de Associação (Association Rules - AR). Por sua vez, o terceiro capítulo tem como foco a explicação dos métodos e ferramentas utilizados ao longo do trabalho.

O capítulo quatro apresenta explicações sobre o desenvolvimento da aplicação e o quinto capítulo aborda um estudo de caso.

Finalmente, as conclusões, melhorias e trabalhos futuros são descritos no sexto capítulo do presente trabalho.

# 2 Descoberta de Conhecimento em Bancos de Dados, Mineração de Dados e Regras de Associação

O presente capítulo apresenta o embasamento teórico necessário para o desenvolvimento do trabalho, sendo os itens a serem revisados: Descoberta de Conhecimento em Bancos de Dados (Knowledge-Discovery in Databases - KDD), Mineração de Dados (Data Mining - DM) e Regras de Associação (Association Rules - AR).

Em linhas gerais (Han e Kamber, 2006):

- KDD é o processo de extração de novos padrões (conhecimento) embutidos em bases de dados;
- DM é a principal etapa do processo de KDD, sendo composta por um conjunto de técnicas e ferramentas utilizadas para identificar padrões embutidos em grandes massas de dados, ou seja, é a etapa de KDD onde é realizada a extração do conhecimento propriamente dita;
- AR é um dos possíveis resultados produzidos durante a etapa de DM, sendo o algoritmo Apriori uma das opções para a extração destas regras.

## 2.1 Descoberta de Conhecimento em Bancos de Dados

O processo de KDD trata da busca e exploração de informações contidas nos dados através da identificação de padrões, sendo estes, necessariamente: válidos, novos, potencialmente úteis e compreensíveis, podendo a combinação destes fatores ser denominada interessabilidade do padrão descoberto.

KDD é um campo de pesquisa multidisciplinar, ou seja, se apoia em diversas

áreas, como aprendizado de máquinas, sistemas especialistas, banco de dados, estatística, computação de alto desempenho, inteligência artificial, visualização de dados, reconhecimento de padrões, entre outras (Liao et al, 2012).

A descoberta de conhecimento em base de dados é um processo iterativo e envolve diversas etapas, destacando-se a seguinte sequência (Fayyad et al, 1996):

1. Consolidação:

   Etapa na qual os dados são obtidos a partir de diferentes fontes, como por exemplo, arquivos texto, planilhas ou bases de dados, e posteriormente, consolidados numa fonte única. Podendo incluir também um processo de limpeza (data cleaning), a qual consiste na remoção de ruídos e de inconsistências nos dados.

2. Seleção e pré-processamento:

   Nesta etapa os dados relevantes da base são recuperados e preparados para o processo de mineração, buscando garantir que sejam da melhor qualidade possível, ou seja, completos, consistentes e precisos. Diversas transformações podem ser aplicadas sobre os dados, como:

   - Reduzir o número de amostras, atributos ou de intervalos de atributos. Esta redução pode ser feita, por exemplo, através de uma limitação no período de tempo, no grupo de itens ou nos valores envolvidos;
   - Normalização, padronização dos valores e limpeza de dados, caso não tenha sido realizada anteriormente;
   - Reduções e representações alternativas podem ser utilizadas buscando obter uma representação reduzida dos dados originais sem perdas de integridade (Han e Kamber, 2006).

3. Mineração de Dados ou DM:

   É a etapa de extração de padrões propriamente dita, na qual, primeiramente, são feitas as escolhas da atividade e da tarefa de mineração a serem utilizadas, tomando como base para esta decisão o tipo de conhecimento que se espera extrair dos dados. As atividades e tarefas dividem-se da seguinte forma:

- Atividades preditivas (supervisionadas):

  Buscam classificar uma nova amostra, a princípio desconhecida, tomando como base o conhecimento obtido do conjunto de amostras com classes já conhecidas. Desempenham inferência de dados com objetivo de gerar previsões ou tendências.

  As tarefas inseridas no contexto deste tipo de atividade são: classificação, regressão e detecção de desvios.

- Atividades descritivas (não-supervisionadas):

  Trabalham com um conjunto de dados que não possuem uma classe determinada, tentando identificar características e propriedades que possibilitem inferir um padrão de comportamento em comum.

  As tarefas inseridas no contexto deste tipo de atividade são: agrupamento, regras de associação e sumarização.

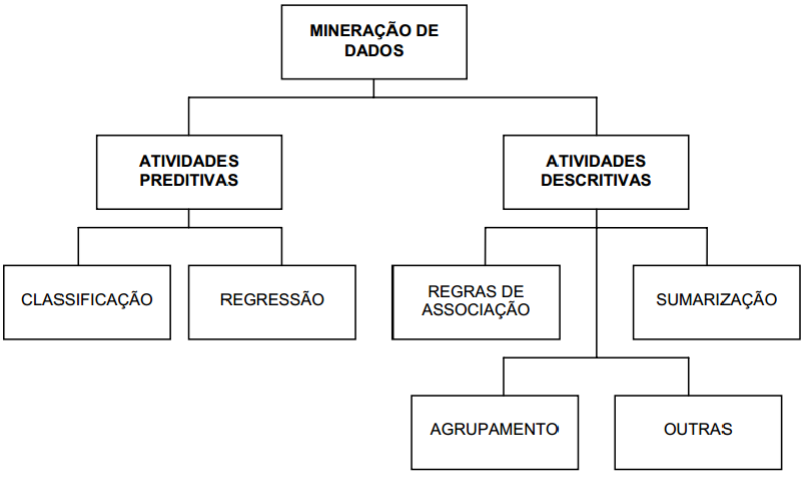



Figura 2.1: Tarefas da Mineração de Dados
(Motta, 2010)

Após a escolha da atividade e, posteriormente, da tarefa a ser utilizada, é necessário optar por um algoritmo condizente e que seja capaz de representar os padrões a serem encontrados. Por se tratar de um tema vasto e, além do escopo deste trabalho, apenas o algoritmo Apriori será detalhado.

Por fim resta a atividade de execução dos algoritmos, sendo esta geralmente a etapa mais custosa em termos de processamento computacional e tempo. É também a

etapa mais interativa, considerando que normalmente os algoritmos empregados necessitam de parâmetros de entrada, os quais são capazes de alterar significativamente os resultados finais (Stock, 2012).

4. Avaliação e Interpretação:

   Nesta etapa é realizada a análise dos resultados e a interpretação dos padrões e regras geradas, identificando assim os padrões de real interessabilidade. Também são avaliados o desempenho do processo e a qualidade dos padrões extraídos, bem como verificada a facilidade de interpretação desses padrões. Estas métricas buscam fornecer um feedback, o qual pode servir como base para a realização de modificações em futuras minerações. Em alguns casos, são utilizados métodos para facilitar a visualização dos resultados obtidas, gerando assim representações alternativas para melhor entendimento, como gráficos, hierarquias e diagramas.

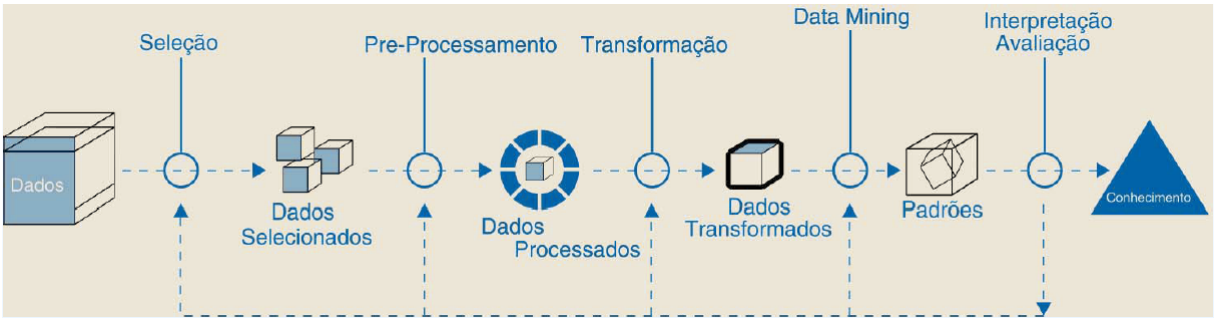



Figura 2.2: Etapas que compõem o Processo de KDD
(Motta, 2013)

## 2.2 Regras de Associação

Mineração de regras de associação é um modelo amplamente estudado pelas comunidades de bancos de dados e de mineração de dados. Seu objetivo é descobrir relacionamentos interessantes escondidos em conjuntos grandes de dados.

Para formalizar o problema de mineração e regras de associação, é apresentado a seguir os conceitos básicos, utilizados na literatura, sobre o assunto.

Seja a base de dados $T = \{t_1, t_2, ..., t_n\}$ contendo um conjunto de $n$ transações.

Seja, agora, o conjunto de $m$ itens $I = \{i_1, i_2, ..., i_m\}$ disponíveis para constituir cada transação $t_i \in T$, tal que $t_i \subseteq I$.

Um conjunto de itens é chamado de um *itemset*. Se um *itemset* possui $k$ itens, ele é um *k-itemset*.

Sejam dois itemsets $A$ e $B$, tais que $A \subseteq I$ e $B \subseteq I$ e não possuem itens em comum, isto é, $A \cap B = \emptyset$. Diz-se que uma transação $t_i$ contem, por exemplo, o *itemset* $A$ se e somente se $A \subseteq t_i$.

Uma regra de associação é uma implicação da forma: $A \rightarrow B$, onde $A \subseteq I$, $B \subseteq I$ e $A \cap B = \emptyset$. Neste caso, lê-se $A$ implica em $B$, onde $A$ é chamado antecedente e $B$ é o consequente da regra.

O objetivo geral da mineração é encontrar, a partir do conjunto de transações $T$, todas as regras que associem a presença de um itemset, $A$, por exemplo, com qualquer outro ($B$, $C$ etc.). Entretanto, como $I$ possui um total de $m$ itens, o espaço de busca para todas as regras é, teoricamente, exponencial $O(2^m)$, pois todos os itens podem constituir *itemsets*.

Na prática, nem todos os itens de $I$ estão presentes nas transações de $T$ e outros, ocorrem em um número muito baixo de transações. Essa esparsialidade é aproveitada pelos métodos de mineração, tornando-os viáveis e eficientes.

Neste sentido, foram criadas duas medidas para as regras, conhecidas como suporte e confiança, que são calculadas a partir da frequência de ocorrência dos *itemsets* envolvidos na regra.

A frequência de um *itemset*, também conhecida como contagem ou contagem de suporte, denotada por $c$, é o número de transações de $T$ que contêm este *itemset*.

O suporte $S$ de uma regra de associação, $A \rightarrow B$, é a porcentagem de transações que contêm $A \cup B$ (ou $A$ e $B$) em relação ao total de transações $n$ de $T$. O suporte, então, pode ser visto como a probabilidade de ocorrência do *itemset* $A \cup B$ em $T$ e é assim calculado:

$$S(A \rightarrow B) = P(A \cup B) = \frac{c(A \cup B)}{n}$$

O suporte indica, portanto, a frequência relativa das regras, podendo ser usado para compará-las, isto é, regras com valores altos para o suporte podem ser interessantes

e merecem a atenção, por se destacarem quantitativamente das demais. Por outro lado, as de suporte muito baixo, podem representar somente uma ocorrência ao acaso. Em última análise, o suporte representa a aplicabilidade da regra.

A confiança $C$ de uma regra de associação, $A \rightarrow B$ , é a porcentagem de transações que contêm $A \cup B$ (ou $A$ e $B$) em relação a todas as transações de $T$ que contêm $A$. Neste caso, $C$ é dado pela probabilidade condicional $P(B|A)$ ou probabilidade de ocorrência de $B$, quando $A$ ocorre. Seu cálculo pode ser feito da seguinte forma:

$C(A \rightarrow B) = P(A|B) = \frac{P(A \cup B)}{P(A)} = \frac{c(A \cup B)}{c(A)}$

A confiança indica a capacidade de predição das regras. As regras com valores altos de confiança se destacam qualitativamente das demais, pelo nível de certeza de ocorrência do consequente da regra, a partir dos casos onde o seu antecedente ocorre. Já as regras com confiança baixa não fornecem segurança de predição e, por isso, são de uso limitado.

É bastante comum nos algoritmos de mineração de regras de associação a adoção de limites pré-estabelecidos pelo usuário para o suporte e para a confiança, conhecidos, respectivamente, como suporte mínimo (sup-min) e confiança mínima (conf-min), reduzindo a amplitude do problema, que, desta forma, passa ser assim enunciado:

Dada uma base de dados $T$ contendo um conjunto de $n$ transações, o problema de mineração consiste em descobrir todas as regras de associação fortes em $T$.

Uma regra de associação forte é aquela que possui suporte e confiança maiores ou iguais, respectivamente, aos limites pré-estabelecidos pelo usuário de suporte mínimo e de confiança mínima.

Após a mineração, é comum apresentar as regras fortes com o seguinte formato:

$A \rightarrow B$[suporte,confiança]

Um *itemset* frequente é aquele que satisfaz ao suporte mínimo, isto é, sua frequência é maior ou igual ao produto do sup-min pelo total de transações de $T$.

Um conjunto de *k-itemsets* é denotado por $C_k$ e um conjunto de *k-itemsets* frequentes, por $L_k$ (Agrawal et al, 1993; Agrawal e Srikant, 1994; Han e Kamber, 2006; Motta, 2010; Liu, 2008).

## 2.3 Algoritmo Apriori

O principal objetivo do algoritmo Apriori é reduzir o número de conjuntos de itens candidatos explorados durante a geração de conjuntos de itens frequentes, ou seja, encontrar todos os *itemsets* frequentes, a partir de um banco de dados de transações $T$ e de um limite de suporte mínimo (*sup-min*), usando a geração de candidatos (Han e Kamber, 2006; Stock, 2012).

A propriedade Apriori determina que, se um conjunto de itens é frequente, então todos os seus subconjuntos também são frequentes. Esta propriedade é verificada facilmente, pois se o *itemset* $B$ é um subconjunto do *itemset* frequente $A$ ($S(A) \geq$*sup-min*), então, todas as transações que contêm $A$ também contém $B$, logo $S(B) \geq S(A) \geq$*sup-min*, isto é, $B$ também é frequente (Han e Kamber, 2006).

A mineração de regras de associação é um processo realizado em dois passos (Han e Kamber, 2006; Agrawal e Srikant, 1994; Liu, 2008):

1. Encontrar todos os itemsets frequentes, etapa na qual o algoritmo Apriori é empregado;

2. Gerar todas as regras de associação fortes a partir dos *itemsets* frequentes.

Deve-se destacar que o desempenho geral da mineração de regras de associação é determinado pelo primeiro passo, sendo o segundo passo bem mais simples. Desta forma, o uso do algoritmo Apriori para encontrar os itemsets freqüentes, torna-se fundamental.

O exemplo a seguir ilustra o funcionamento do processo:

Primeiro Passo:

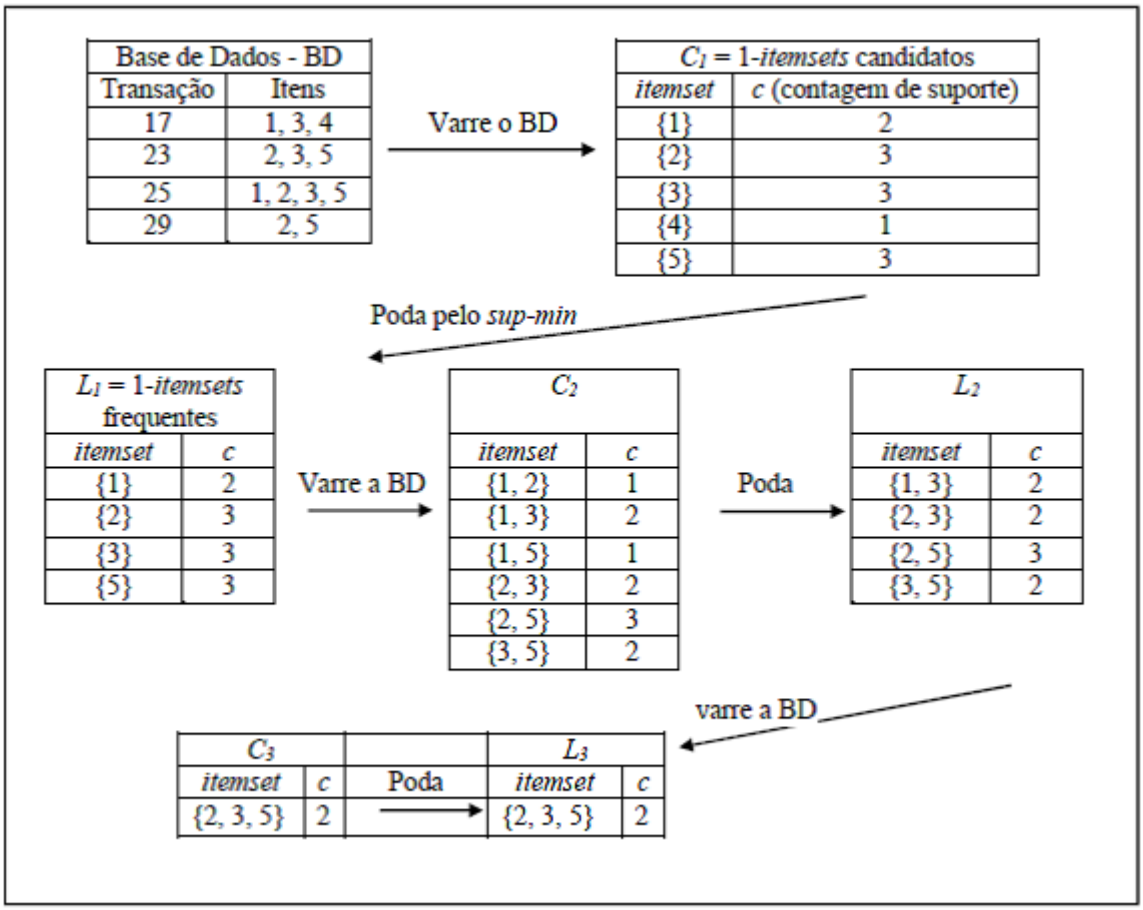



Figura 2.3: Exemplo Algoritmo Apriori
(Motta, 2010)

Segundo Passo:

Para gerar as regras de associação deve-se, inicialmente, distinguir todos os subconjuntos não-vazios *sc* de cada *itemset* frequente *F*. No exemplo, o itemset frequente *F* = {2, 3, 5} possui os seguintes subconjuntos não-vazios: {2, 3}, {2, 5}, {3, 5}, {2}, {3} e {5}.

Em seguida, produzir todas as regras com o formato "$sc \rightarrow (F - sc)$" e que satisfaça a *conf-min*.

No exemplo, produzindo as possíveis regras do itemset frequente $F = 2, 3, 5$ e calculando as suas confianças, tem-se:

$C(2, 3 \rightarrow 5) = 2/2 * 100 = 100\%$

$C(2, 5 \rightarrow 3) = 2/3 * 100 = 67\%$

$C(3, 5 \rightarrow 2) = 2/2 * 100 = 100\%$

$C(2 \rightarrow 3, 5) = 2/3 * 100 = 67\%$

$C(3 \rightarrow 2, 5) = 2/3 * 100 = 67\%$

$C(5 \rightarrow 2, 3) = 2/3 * 100 = 67\%$

Considerando uma conf-min = 80%, são identificadas as seguintes regras fortes:

$2, 3 \rightarrow 5[50\%, 100\%]$

$3, 5 \rightarrow 2[50\%, 100\%]$

A partir da criação do algoritmo Apriori, a grande maioria dos esforços tem se concentrado em duas frentes de pesquisa: (1ª) desenvolver variantes do algoritmo que ofereçam algum tipo de vantagem, especialmente, em relação ao desempenho e à escalabilidade e (2ª) encontrar medidas que possam auxiliar a identificação de regras interessantes.

A listagem abaixo representa o algoritmo Apriori:

```
(1)  L1 = acha_1-itemsets_freqüentes(T);
(2)  para (k = 2; Lk-1 ≠ ø; k++){
(3)      Ck = gera_candidatos(Lk-1);
(4)      para cada transação t ∈ T {   // leitura de T para contagem
(5)          Ct = subconjunto(Ck, t); //recebe os subconjuntos candidatos de t
(6)          para cada candidato cd ∈ Ct
(7)              c(cd)++;
(8)      }
(9)      Lk = {cd ∈ Ct | c(cd) ≥ sup-min};
(10) }
(11) retorna L = ∪k Lk;
```

Figura 2.4: Algoritmo Apriori
(Han e Kamber, 2006)

## 2.4 Regras de Associação Multiníveis

Um ponto que deve ser considerado em um processo de mineração de dados, é o significativo número de regras de associação com baixa importância que podem ser geradas, apresentando baixos valores de Suporte e Confiança. Portanto, classificar e agrupar ob-

jetos em relações mais significativas, a partir de propriedades similares pode reduzir a quantidade de regras de associação, assim como melhorar suas métricas.

Buscando obter regras mais significativas torna-se interessante minerar regras de associação definidas não somente a partir de itens básicos, mas também a partir de itens que representam generalizações destes itens básicos. Um exemplo, é que com regras de associação multinível seria possível minerar a regra genérica ração → medicamentos na base de dados, sem necessidade de minerar regras mais específicas como ração de gato → medicamentos para cães, medicamentos para gatos, medicamentos para aves (Stock, 2012; Motta, 2010).

## 2.5 Mineração com Restrições

A complexidade real da mineração de regras de associação reside na quantidade de dados. O fato de ter que trabalhar com enormes volumes de dados, se reflete no aumento da complexidade da tarefa de mineração propriamente dita. Neste sentido, alguns recursos adicionais podem ser utilizados na busca de redução dessa complexidade. Sendo relevante para o contexto do trabalho apenas as reduções feitas através de restrição.

O uso de restrições permite que seja especificado o conhecimento desejado como resultado da mineração das regras, de acordo com as intenções do minerador, tornando o processo mais efetivo. As principais formas de restrições são (Han e Kamber, 2006; Motta, 2010):

- Restrição de nível de abstração: nas bases de dados onde os itens estão representados em mais de um nível de abstração, é interessante executar a mineração do mais alto para o mais baixo nível;

- Restrição de interessabilidade: essa restrição tem como principal forma os limites de suporte e confiança mínimos;

- Restrição de transações: transações que não contém nenhum *k-itemset* frequente não podem conter nenhum *(k+1)-itemset* frequente. Essas transações são dispensadas de leituras futuras para *j-itemsets* onde $j > k$;

- Restrição de itens: usada quando interessam somente as regras formadas por um conjunto selecionado de itens. Neste caso, pode-se criar novas bases de dados, tanto de itens quanto de transações, contendo somente os itens de interesse;

- Restrição de tempo: como as transações são normalmente colecionadas em ordem cronológica e o padrão de mineração de regras é feito por intervalos fixos de tempo (mês a mês, por exemplo), o uso dessa restrição é bastante comum. Neste caso, as transações realizadas em intervalos de tempo definidos pelo usuário constituem subconjuntos da base de dados de transações;

- Restrição de itemsets ou de regras: especifica a forma das regras a serem mineradas como, por exemplo, o número máximo de itens do antecedente ou do consequente das regras.

# 3 Métodos e Ferramentas

Este capítulo terá como objetivo apresentar a ferramenta "CalcD", proposta na tese de doutorado: "Metodologia para mineração de regras de associação multiníveis incluindo pré e pós-processamento" (Motta, 2010), sendo ela a base para o desenvolvimento do trabalho aqui apresentado.

Segundo Motta (Motta, 2010): "A ferramenta CalcD é um sistema inteligente que utiliza o conhecimento descoberto na mineração das regras de associação mais fortes de todos os 2-itemsets existentes em uma base de dados, para calcular as distâncias entre os itens envolvidos e criar estruturas de dados que possibilitam a visualização das associações entre esses itens no espaço bidimensional".

O sistema CalcD também é complementado por uma ferramenta de projeção gráfica (PEx), que é responsável por acrescentar um módulo de produção de imagens, tornando assim possível a visualização das matrizes geradas durante o processo.

## 3.1 Função de Distância

O resultado obtido durante o processo de mineração de regras de associação é comumente utilizado como fonte de informação para a tomada de decisão baseada em distância.

Um exemplo clássico desta aplicação foi descoberto pela rede de lojas norte-americana Wallmart, a qual optou por expor fraldas e cervejas lado a lado, aumentado assim consideravelmente o volume de venda destes produtos (Bispo, 1998).

Basicamente, o problema da transformação de regras de associação em distância pode ser resumido da seguinte forma: "Dada uma regra de associação entre dois itens A e B, isto é, dada a regra A $\rightarrow$ B, determinar a melhor distância a ser adotada entre A e B". Sendo assim, a melhor distância é calculada a partir de duas estratégias de marketing (Han e Kamber, 2006):

1. Dois itens frequentemente comprados juntos podem ser expostos em locais distantes, buscando assim incentivar a venda de outros que estejam expostos entre eles;

2. Dois itens frequentemente comprados juntos podem ser expostos proximamente, buscando incentivar a sua venda conjunta.

Alguns princípios foram adotados buscando alcançar uma solução para o problema da distância ideal entre itens frequentemente comprados em conjunto, sendo eles (Motta, 2010):

- Como fonte de informação para a tomada de decisão será utilizada a estrutura suporte-confiança, a qual, a partir do algoritmo Apriori, tornou-se comum na grande maioria dos sistemas de mineração de regras de associação;
- Será usada a regra mais forte para obtenção da distância entre os itens A e B, ou seja, aquela que possuir a maior confiança, já que o suporte é simétrico.

Considerando que as frequências de ocorrência de A e B na base de dados de n transações são, respectivamente, $f_a$ e $f_b$, a obtenção da distância entre os itens A e B será obtida através da regra (Motta, 2010):

$f_a \leq f_b, S(A \rightarrow B) = \frac{c(A \cup B)}{n}$ e $C(A \rightarrow B) = \frac{c(A \cup B)}{f_a}$

Tomando como base a equação apresentada anteriormente e o fato da estrutura suporte-confiança garantir que os valores absolutos do suporte são, no máximo, iguais ao da confiança, é possível reescrever as estratégias de marketing da seguinte forma (Motta, 2010):

- Quanto maior o suporte e quanto menor a diferença da confiança para o suporte, maior a distância entre A e B;
- Quanto menor o suporte e quanto maior a diferença da confiança para o suporte, menor a distância entre A e B.

Por fim, pode-se adotar o suporte da regra (S) como distância básica entre A e B e subtrair dessa distância básica um percentual dela própria, correspondente à diferença da confiança (C) menos o suporte, chegando a seguinte função de distância (D) (Motta, 2010):

$D = S - (C - S) * S$

## 3.2 Sistema CalcD

O sistema CalcD é composto por dois programas: GeraD e HierarqD, os quais, trabalhando de forma integrada, geram uma matriz contendo os valores das distâncias calculadas a partir dos pares de itens presentes numa base de dados de transações. O processo tem como finalidade organizar os itens em estruturas hierárquicas, para que, quando essa matriz for utilizada por um sistema de projeção multidimensional, as imagens obtidas tenham uma melhor qualidade de posicionamento dos itens para efeito de análise visual (Motta, 2010).

## 3.3 Sistema GeraD

É responsável por minerar as regras de associação mais fortes de uma base de dados de transações, calculando as distâncias entre os pares de itens e, ao final, gera e grava a matriz de distância original.

O programa recebe dois arquivos de entrada no formato texto, o primeiro contendo a frequência de todos os itens em ordem decrescente e o segundo com o registro de todas as transações a serem mineradas.

Para garantir um melhor desempenho, os itens devem ser identificados por um código gerado a partir do seu posicionamento na ordenação das frequências, por exemplo, o item mais frequente receberá o código um.

Outra informação necessária para o funcionamento do programa é a definição do nível de frequência de cada item, fornecido pelo minerador de forma manual, devendo ser estabelecidos pelo menos dois níveis, sendo eles numerados em ordem decrescente até um (Motta, 2010).

A definição dos níveis de frequência tem dois objetivos importantes (Motta, 2010):

- O sistema CalcD utiliza essa definição para organizar as estruturas hierárquicas, onde cada item do nível de frequência mais alto dá origem a uma estrutura;
- O nível de frequência de um item identifica o tamanho e a cor da sua representação na visualização a ser gerada.

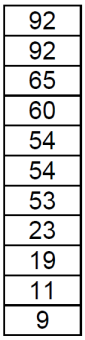



Figura 3.1: Exemplo Vetor de Frequência
(Motta, 2010)

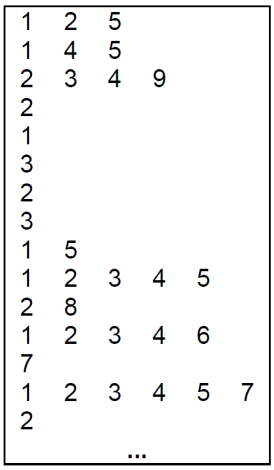



Figura 3.2: Exemplo Arquivo de Transação
(Motta, 2010)

Finalmente, após a entrada dos níveis de frequência, o programa GeraD é efetivamente iniciado, realizando a seguinte sequência de tarefas (Motta, 2010):

- Leitura do arquivo de frequência e criação do vetor FREQ com as frequências dos itens, cujos índices são os seus respectivos códigos;
- Leitura do arquivo de transações e criação da matriz "c" com a contagem de suporte das regras de associação mais fortes de todos os 2-itemsets. Onde a linha representa o lado esquerdo da associação e a coluna o lado direito, formando ao final uma matriz triangular inferior sem a diagonal principal, a qual é representada de forma vetorial, conseguindo assim uma economia de memória de mais de 50%;
- Criação da matriz suporte "S" a partir da divisão de cada item da matriz "c" pelo número de transações;

| Solicitação do Programa | Informação do Usuário |
|---|---|
| | |
| Número de Níveis de Freqüência: | 3 |
| Número de Itens por Nível de Freqüência: | |
| Nível 3: | 2 |
| Nível 2: | 5 |
| Nível 1: | 4 |

Figura 3.3: Exemplo - Número de Itens por Nível de Hierarquia
(Motta, 2010)

- Cálculo da matriz "C" com as confianças das regras, dividindo-se cada linha L da matriz "c" pela frequência do item $L(f_L)$ do vetor FREQ.
- Na sequência, é calculada a matriz D contendo as distâncias entre cada par de itens, fazendo uso da função $D = S - (C - S) * S$;
- Finalizando, o programa grava um arquivo de texto com a matriz de distância original já adaptado o formato de entrada do sistema PEx.

| → | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | | | | |
| 2 | 54 | | | | | | | | | | |
| 3 | 49 | 47 | | | | | | | | | |
| 4 | 45 | 46 | 30 | | | | | | | | |
| 5 | 40 | 36 | 26 | 27 | | | | | | | |
| 6 | 36 | 40 | 33 | 26 | 25 | | | | | | |
| 7 | 38 | 40 | 30 | 30 | 22 | 22 | | | | | |
| 8 | 13 | 12 | 9 | 13 | 5 | 9 | 11 | | | | |
| 9 | 15 | 19 | 17 | 12 | 9 | 9 | 12 | 4 | | | |
| 10 | 8 | 10 | 9 | 7 | 5 | 4 | 11 | 3 | 9 | | |
| 11 | 7 | 9 | 9 | 4 | 5 | 9 | 5 | 2 | 8 | 4 | |

Figura 3.4: Exemplo Matriz Original Gerada
(Motta, 2010)

Neste ponto do processo já é possível extrair um conjunto considerável de informações através da representação gráfica da matriz gerada. Entretanto, alguns problemas no posicionamento das amostras durante a visualização, principalmente no que se refere a itens com frequência baixa, justificaram o desenvolvimento do programa HierarqD, que busca melhorar a representatividade das associações realmente importantes (Motta, 2010).

| [ / ] = | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *c* = | 54 | 49 | 47 | 45 | 46 | 30 | 40 | 36 | 26 | 27 | 36 | 40 | 33 | 26 | 25 | 38 | 40 | 30 | 30 | 22 |

| 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 | 13 | 12 | 9 | 13 | 5 | 9 | 11 | 15 | 19 | 17 | 12 | 9 | 9 | 12 | 4 | 8 | 10 | 9 | 7 |

| 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 4 | 11 | 3 | 9 | 7 | 9 | 9 | 4 | 5 | 9 | 5 | 2 | 8 | 4 |

Figura 3.5: Exemplo Vetor de Representação da Matriz Original (Motta, 2010)

## 3.4 Sistema HierarqD

O HierarqD tem como parâmetros de entrada o arquivo de texto que contém a matriz de distâncias gerada pelo programa GeraD, uma taxa de hierarquização e outra de otimização. Durante o seu funcionamento são executadas as seguintes tarefas: (i) tratamento das distâncias nulas; (ii) hierarquização e (iii) otimização da matriz. Para cada uma destas tarefas é gerado um arquivo de saída (Motta, 2010).

De forma geral, a função do programa é organizar os itens em ordem decrescente de seus níveis de frequência, criando, assim, uma estrutura hierárquica distinta para cada item do nível mais alto e mantendo as distâncias originais entre eles. Os demais itens são alocados nessas estruturas, valorizando e mantendo a menor distância e ajustando as suas outras distâncias, se necessário. Como consequência desta abordagem, os itens de baixa frequência associam-se de forma mais forte com o item de maior frequência possível, configurando assim uma boa solução para o problema do item raro (Motta, 2010).

O tratamento das distâncias nulas mostra-se necessário, pois durante a execução do programa GeraD não são realizadas podas através dos valores de suporte e confiança, podendo então ocorrer situações de distâncias nulas (itens não se associação) na matriz D (Motta, 2010).

Para uma representação eficiente dos itens no $R^2$ é preciso adotar um valor de distância, mesmo nos casos onde os itens não se associam. Para tanto, foi utilizada a distância máxima existente entre todos os itens pertencentes ao maior dos níveis de frequência envolvidos na regra. Ao final desta etapa é gerado um arquivo texto contendo um relatório dos itens que não se associam (Motta, 2010).

A seguir, é iniciado o processo de hierarquização da matriz D, sendo este responsável por valorizar as distâncias mais importantes para efeito da projeção dos itens do $R^n$ no $R^2$. Assim, buscando obter uma redução com uma menor perda de informação, ou seja, mantendo, tanto quanto possível, as distâncias entre os itens no $R^n$ em suas projeções no $R^2$ (Motta, 2010).

O funcionamento do método de hierarquização é baseado no conceito de áreas de influência, o qual pode ser definido, segundo Motta (Motta, 2010), como: "espaço ao redor de um item no $R^n$, definido, no mínimo, pela metade da distância entre esse item e outro com nível de frequência igual ou superior ao dele e que lhe seja o mais próximo".

Tomando como base este conceito, algumas situações de relacionamento entre os itens e as áreas de influência podem ser definidas, sendo elas (Motta, 2010):

- Pontos A e B próximos e de mesmo nível de frequências: área de influência mínima é o espaço em torno de cada um, distante até $\frac{D^n(A,B)}{2}$;
- Ponto C com nível de frequência inferior ao de A e B e que, no $R^n$, está mais próximo de A do que de qualquer outro item com nível de frequência superior, incluindo B: C pertence a área de influência de A, que passa a ser chamado de pai ou pivô de C. Neste caso, se $D^n(A,C) \leq \frac{D^n(A,B)}{2}$, então C pertence a área de influência mínima de A, senão C pertence a área de influência expandida de A;
- Item C pertence a outra área de influência, além da sua própria (no máximo uma outra): a menor distância $D^n(A,C)$ é sempre mantida e é verificado se a maior distância não está invadindo outra área de influência. Caso a invasão seja confirmada, a maior distância é substituída por um percentual dela mesma, o qual é definido pelo usuário, através de uma taxa de entrada denominada ajuste de hierarquização (AH). Desta forma, a menor distância é valorizada e, se necessário, C é empurrado de maneira a passar a pertencer a uma única área de influência.

Por fim, chega-se a seguinte situação: A é pai ou pivô de todos os itens que, como C, forem alocados em sua área de influência e todo item possui uma área de influência onde podem ser alocados outros itens. Vale ressaltar que, o valor atribuído a hierarquização é de fundamental importância para a obtenção de uma boa projeção no $R^2$. Esse valor pode

ser definido como a proporção na qual o item de baixo nível de frequência será deslocado para fora da área de influência de um outro que não seja o seu pivô e possua nível superior ao dele (Motta, 2010).

Terminando o processo de hierarquização é possível obter uma definição clara das estruturas hierárquicas. Porém, os itens com níveis de frequências mais baixos possuem, frequentemente, uma distância muito pequena dos seus pais, acabando por ficar sobrepostos na visualização gerada a partir da matriz hierarquizada, impossibilitando uma análise eficiente do gráfico produzido. Este problema é resolvido pela terceira tarefa realizada pelo programa HierarqD, através de um método complementar de otimização das distâncias em cada área de influência mínima (Motta, 2010).

O processo de otimização consiste na maximização linear das distâncias. A partir do segundo nível de frequência, é determinado se todos os irmãos pertencem a mesma área de influência mínima. Caso isto ocorra, a maior das distâncias entre os irmãos ou entre cada um deles e o pai é maximizada para sua área de influência, sendo então calculado o fator de acréscimo sofrido por esta distância. Em seguida, todas as distâncias entre o pai e os filhos e entre os irmãos são multiplicadas por este fator, mantendo, desta forma, uma proporcionalidade entre essas distâncias e otimizando a ocupação dessa área de influência mínima para uma melhor distribuição de seus itens (Motta, 2010).

Para a execução deste método é utilizada a taxa de otimização, fornecida pelo usuário e funciona como multiplicador do fator de acréscimo obtido durante o processo, possibilitando assim um ajuste mais refinado do resultado final.

Ao fim da execução do programa é gerado um arquivo de texto no formato de entrada do sistema PEx, possibilitando assim uma rápida projeção dos resultados.

## 3.5 PEx

Projection Explorer Graph (PEx-Graph)[1]é uma ferramenta construída em JAVA e que pode ser utilizada para criar e explorar representações visuais de gráficos e de “Social Networks”. Mais especificamente, sua utilização no contexto deste trabalho é voltada à construção de gráficos de distância entre pontos, sendo estes gerados a partir de um arquivo no formato “.dmat”. Esse arquivo contém um matriz com a representação das

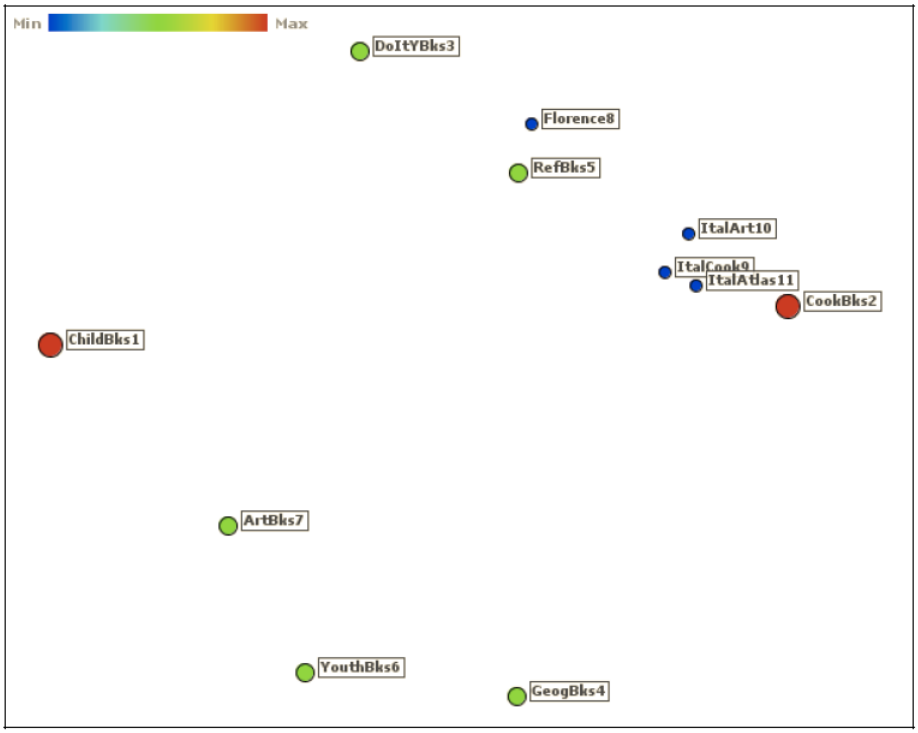



Figura 3.6: Exemplo Gráfico Gerado pelo PEx
(Motta, 2010)

distâncias entre todos os pares de pontos. O método utilizado para a projeção dos pontos é o "Classical Scaling", sendo ele responsável por reduzir as distâncias contidas no $R^n$ para o $R^2$, mantendo, tanto quanto possível, a relação entre os pontos.

## 3.6 CBA

CBA[2]é uma ferramenta para mineração de dados desenvolvida na Universidade Nacional de Singapura, Departamento de Computação. Originalmente tinha como foco a Classificação Baseada em Associações, porém, além desta tarefa, também é capaz de minerar várias formas de regras de associação. Mais especificamente, sua utilização no contexto deste trabalho é baseada em sua capacidade de gerar regras de associação a partir de um arquivo transacional (".tra"), fornecendo os valores de suporte e confiança referentes a cada regra.

[1]Disponível em http://infoserver.lcad.icmc.usp.br/infovis2/InfoVis2
[2]Disponível em http://www.comp.nus.edu.sg/ dm2/

# 4 Desenvolvimento e Implementação

O presente trabalho tem como finalidade facilitar o uso de um método gráfico para auxilio no processo de mineração de regras de associação desenvolvido na tese de doutorado "Metodologia para mineração de regras de associação multiníveis incluindo pré e pós-processamento" (Motta, 2010).

Para atender a esta finalidade, o trabalho terá como foco o refinamento de dois aspectos do trabalho original: automação e integração, buscando assim obter uma melhora no desempenho do processo como um todo.

Para atingir estes objetivos, uma ferramenta contendo uma interface de fácil utilização será desenvolvida. Junto a ela, algumas tarefas serão implementadas, entre elas, conexão direta da interface a base de dados a ser minerada; implementação de meios para realizar o pré-processamento dos dados de forma simples e intuitiva; geração automática das entradas necessárias ao funcionamento dos programas de mineração utilizados (GeraD, HierarqD, PEx e CBA) e subsequente integração destes programas; criação e armazenamento dos arquivos de saída de forma simples e organizada; implementação de métodos e funções com o objetivo de facilitar o processo de mineração de forma geral.

O projeto será desenvolvimento na linguagem JAVA, através da utilização da ferramenta NetBeans IDE e do seu respectivo módulo (Swing) para criação de interfaces.

Todo o processo é divido em três janelas (frames) principais. Cada uma contém os métodos, classes e atributos necessários a sua execução. Contudo, o usuário final tomará conhecimento apenas do funcionamento da interface, ou seja, a complexidade inerente ao processo de programação será tratada em segundo plano, facilitando assim o trabalho do minerador. Assim, as tarefas são realizadas de forma mais transparente, sem a preocupação sobre o que está sendo executado no background do programa.

A seguir, encontra-se um resumo de todo o trabalho realizado, sendo este particionado conforme as três janelas (frames) criadas durante o desenvolvimento do programa. Por não serem particularmente relevantes ao entendimento do projeto, os métodos, classes e atributos referentes ao controle e a criação da interface não serão abordados, assim como

o código propriamente dito, sendo o foco do resumo uma explicação de mais alto nível.

## 4.1 Janela 1 - Frame Seleção

Frame responsável pela apresentação e pré-processamento dos dados para a futura mineração, de forma geral realiza as seguintes tarefas:

- Grava na memória os itens selecionados a cada clique do mouse, evitando assim problemas resultantes do modelo padrão, que consistia em segurar a tecla "CTRL" ou "SHIFT" para selecionar mais de um item ao mesmo tempo, o que causava consequência inesperadas, como perda dos itens selecionados e seleção de itens indesejados;

- Realiza automaticamente a conexão com o banco de dados, povoando as tabelas "Setor", "Seção", "Grupo" e "Produto" presentes na interface com os seus respectivos itens contidos na base de origem;

- Realiza uma limpeza nos dados da base, retirando todos os produtos com frequência nula dentro do prazo estipulado para a mineração, evitando assim que eles sejam utilizados para povoar as tabelas presentes na interface;

- Cria sorters em cada uma das tabelas, possibilitando que elas sejam organizadas com base no atributo selecionado pelo usuário. Ao clicar no cabeçalho, a tabela passa a ser ordenada a partir dos valores dos atributos contidos na respectiva coluna selecionada;

- Cria filtros de busca em cada uma das tabelas, facilitando assim o trabalho do usuário na procura por um item específico. Ativados a partir do botão "Busca" presente em cada tabela;

- Cria filtros entre as seleções realizadas em cada tabela. Como o projeto trabalha com um sistema de seleção multinível, caso um item seja selecionado em um nível superior, um subitem deste não deve ser selecionado em um nível mais baixo, neste caso, o mesmo item seria computado duas vezes. Para tanto, foram criados métodos,

ativados através dos botões “Gravar Seleção” presentes em cada tabela, para desselecionar e, posteriormente, impedir a visualização, dos subitens dos elementos selecionados. Cada vez que o método é chamado as seleções são recomputadas e as visualizações atualizadas. Os níveis e, consequentemente, os filtros, seguem a seguinte ordem de nível de abstração decrescente “Setor”, “Seção”, “Grupo” e “Produto”. Sendo que, ao selecionar um nível, todos os seus subníveis são filtrados;

- Envia a seleção feita para o próximo frame. Ativada através do botão “Exportar Seleção”.

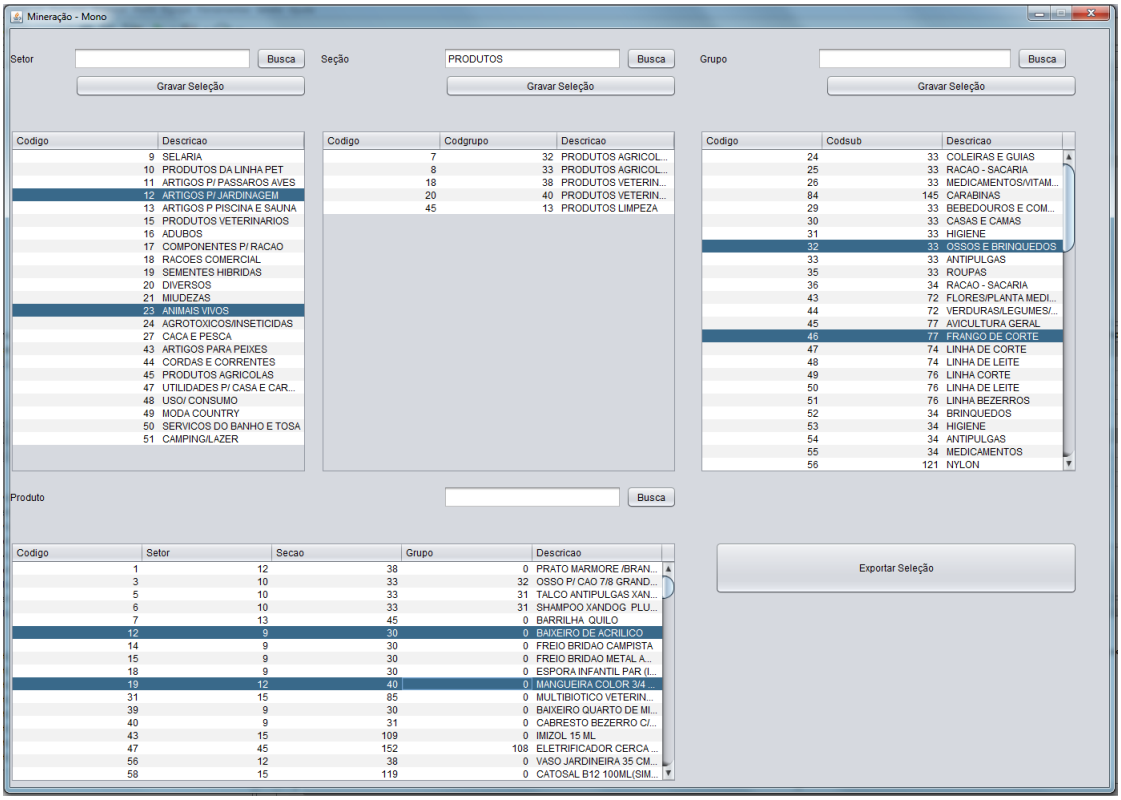



Figura 4.1: Exemplo Frame 1

## 4.2 Janela 2 - Frame Display Seleção

Frame responsável por apresentar para o usuário os elementos selecionados na etapa anterior, assim como por gerar e armazenar apropriadamente os parâmetros de entrada que serão utilizados nas ferramentas de mineração utilizadas. Realiza as seguintes tarefas:

- Apresenta de forma clara os itens selecionados na etapa anterior, separando-os de acordo com as suas tabelas de origem;

- Assim como o frame anterior, cria sorters em cada uma das tabelas, possibilitando que elas sejam organizadas com base no atributo selecionado pelo usuário;

- Conecta-se automaticamente ao banco de dados (processo iniciado ao clicar no botão "Gravar Resultado");

- Gera o arquivo com a frequência de cada item:

  – A partir dos itens selecionados cria automaticamente as querys responsáveis por obter a frequência de cada item nas transações contidas na base de origem. Ainda, garante que o elemento em questão é válido, ou seja, pertence ao período de tempo a ser minerado e não é um produto marcado como inativo na base;

  – Classifica os itens em ordem decrescente de frequência;

  – Cria os novos índices para cada item que, a partir deste momento, passam a ser tratados por sua posição no vetor decrescente de frequências. Para a rápida conversão entre os valores reais dos itens e as chaves obtidas, através da frequência, uma estrutura de mapeamento no modelo "chave $\rightarrow$ valor" foi criada, sendo responsável pelo armazenamento das informações para posterior recuperação;

  – Grava o resultado em um arquivo de texto, garantindo que este esteja no formato e no local corretos para a execução dos programas de mineração. Cada linha do arquivo contém um valor de frequência, sendo estes dispostos em ordem decrescente.

- Gera o arquivo de transações:

  – Cria automaticamente uma query responsável por obter todas as transações que contenham algum dos itens selecionados na etapa anterior. Realiza o mesmo processo de validação dos itens feito na etapa de obtenção das frequências;

– Percorre a query retornada transformando cada transação para o formato de entrada dos programas de mineração utilizados. Nesta etapa é utilizada a estrutura de mapeamento criada anteriormente, com o objetivo de converter os valores dos itens obtidos pela query em seus respectivos índices relativos a sua frequência. Os arquivos de entrada criados são arquivos de texto contendo, em cada linha, uma das transações da base, sendo a representação de cada item feita pelo índice correspondente a sua posição no vetor de frequência. Os itens em cada transação são classificados em ordem crescente. Porém os arquivos gerados possuem algumas peculiaridades, sendo elas:

  * GeraD: Transações contendo apenas um item são desconsideradas, assim como a repetição de itens dentro de uma mesma transação. Como o programa GeraD trabalha com associações entre pares de itens, não faz sentido armazenar transações com apenas um elemento, sendo elas eliminadas;
  * CBA: Diferente do funcionamento do GeraD, que recebe o arquivo de frequência separadamente, o CBA calcula este valor a partir das transações, sendo portanto necessário incluir também as transações que contém apenas um item. A repetição de itens dentro de uma mesma transação é desconsiderada, sendo apenas um valor registrado.

## 4.3 Janela 3 - Frame Tabela de Frequência

Frame responsável por apresentar os itens selecionados e suas respectivas frequências de maneira resumida, permitindo ao minerador determinar de forma simples o nível de cada item na hierarquia. Também fica a cargo deste frame integrar as diferentes ferramentas de mineração em uma mesma interface, facilitando assim a execução do processo. São realizadas as seguintes tarefas:

- Tabela é povoada com as informações de cada item selecionado anteriormente, sendo essas: seu código (posição no vetor de frequência), seu nome e sua frequência. Além das informações do item, é disponibilizado um campo extra para que o minerador possa preencher o nível que aquele item deve ocupar na hierarquia;

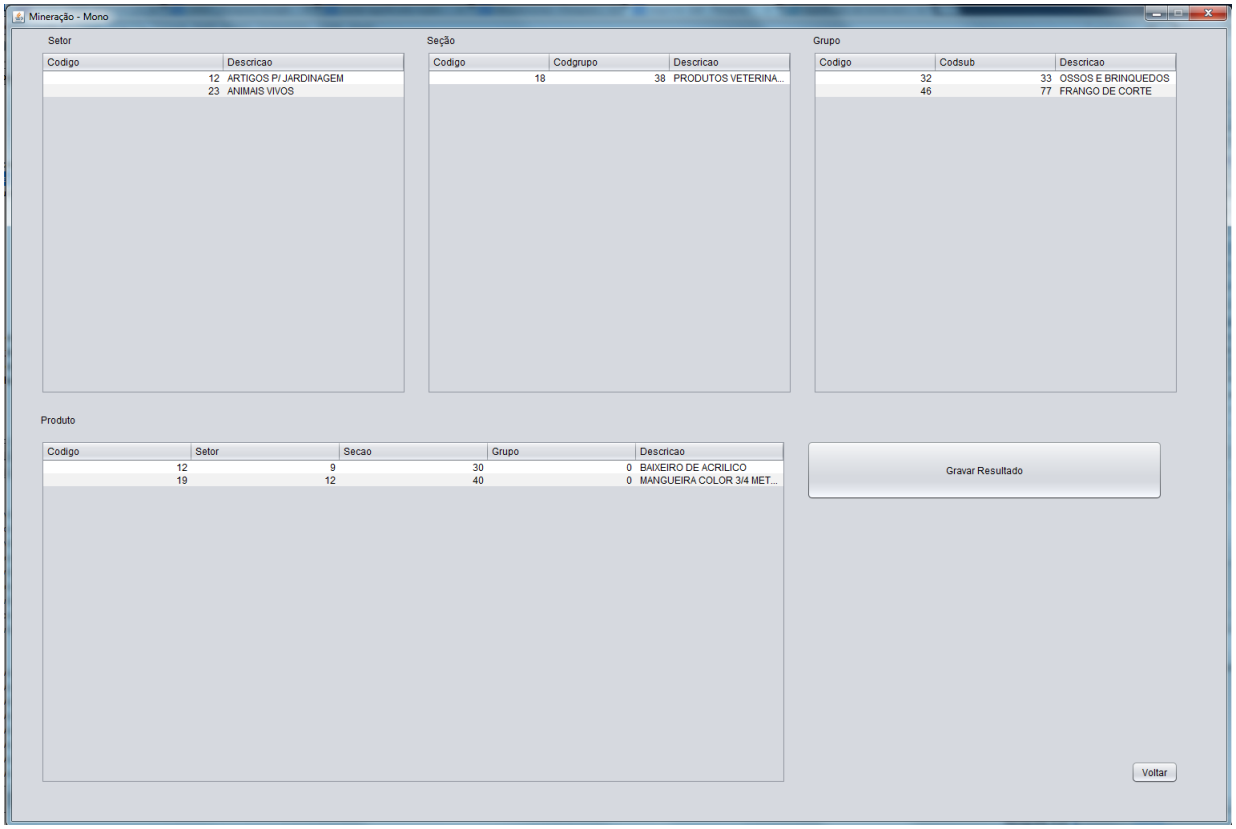



Figura 4.2: Exemplo Frame 2

- Cria um arquivo no formato texto com os nomes de cada item, este será utilizado posteriormente pelo PEx para rotular cada ponto no momento da geração do gráfico;

- Após o preenchimento dos níveis de hierarquia, o sistema faz um cálculo do total de níveis existente e do número de itens por nível. Estas informações são utilizadas como entrada do programa GeraD e seu cálculo automático acelera o processo, evitando que o minerador precise contá-los manualmente. Esta contagem tem início quando o usuário clica no botão “Registrar Nível”;

- Copia sempre que necessário os arquivos entre os diretórios, armazenando-os nas pastas utilizadas para a execução dos programas. O processo de cópia é realizado através de chamadas feitas ao “prompt de comando”, utilizando arquivos de lote (“.bat”);

- Altera a extensão dos arquivos sempre que necessário, deixando-os no formato correto de entrada exigido pelos programas: “.titles” para a rotulação dos pontos no PEx, “.dmat” para a entrada das matrizes na geração dos gráficos do PEx e “.tra”

para a entrada do arquivo de transações no CBA. As alterações são realizadas através de chamadas feitas ao "prompt de comando", utilizando arquivos de lote (".bat");

- Inicia a execução automática do respectivo programa quando um dos seguintes botões é acionado: "GeraD", "HierarqD", "PEx" ou "CBA". Antes da execução, realiza todas as ações necessárias para garantir que os parâmetros de entrada do programa estarão no formato e no local corretos. Inicialização dos programas realizada através de chamadas feitas ao "prompt de comando", utilizando arquivos de lote (".bat").

Durante o estudo de caso foi identificado que, para obter um bom resultado de análise, sempre eram gerados uma grande quantidade de gráficos. Buscando acelerar este processo, mais duas alterações foram realizadas:

- Ao invés do modelo padrão, no qual o HierarqD gerava apenas um resultado por execução, ele foi modificado de maneira a gerar 11 valores automáticos, com otimização igual a zero e hierarquização de zero a um, variando de 0.1 em 0.1, e um valor ajustável;

- Foi também implementado um método de otimização de forma separada, no qual o valor da hierarquização é dado como entrada e são geradas 11 saídas automaticamente, sendo o valor da hierarquização fixo e o de otimização de zero a um, variando de 0.1 em 0.1.

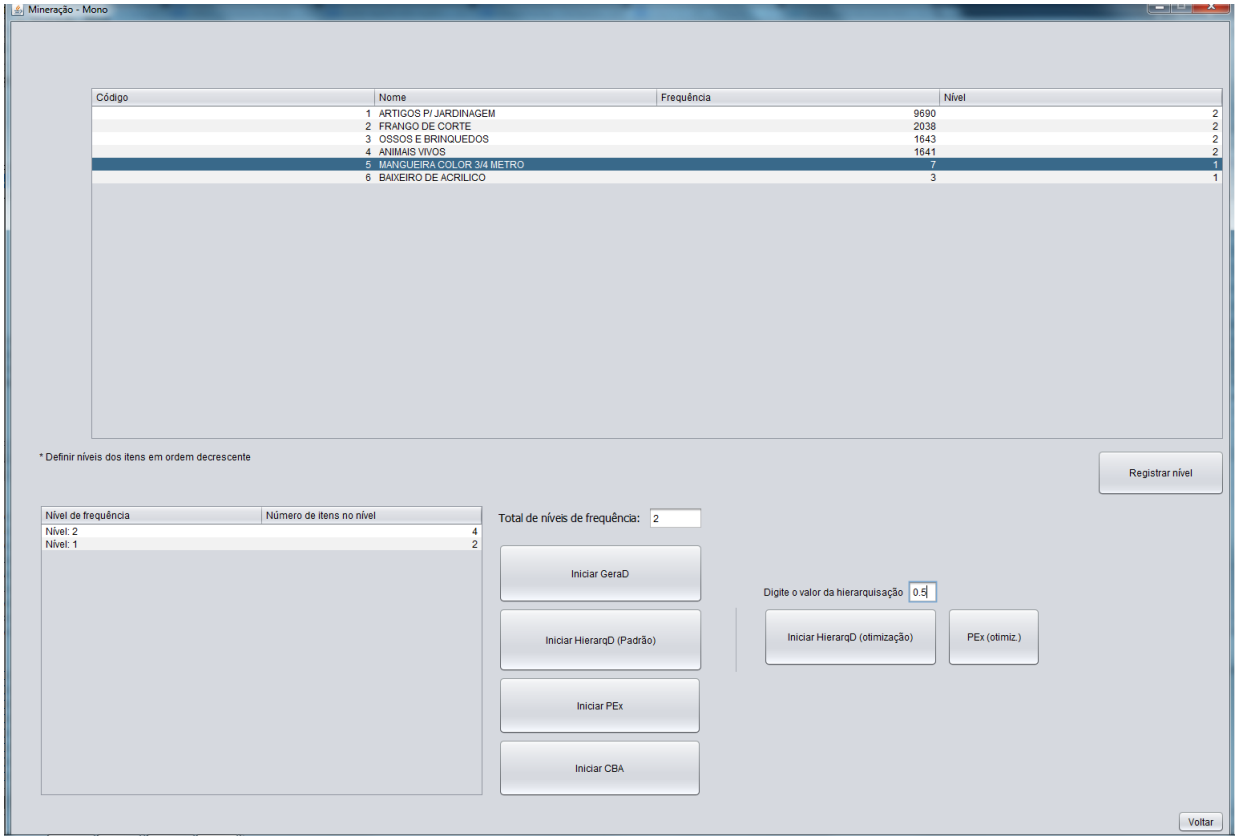



Figura 4.3: Exemplo Frame 3

# 5 Mineração

O presente estudo de caso foi realizado em uma base de dados transacional pertencente a empresa “A Rural Toscana Ltda”, a qual têm como foco o comercio de produtos agrícolas, atuando também na área de vendas de produtos pet. A base utilizada continha quatro níveis de abstração, sendo eles, em ordem decrescente de granularidade: “Setor”, “Seção”, “Grupo” e “Produto”.

## 5.1 Estudo de Caso

Antes de iniciar a mineração, o banco passou por um processo de limpeza dos dados, o qual foi realizado através da implementação de uma função resposável, exclusivamente, por esta limpeza. Como os dados não foram alterados durante a realização do trabalho, a limpeza foi realizada uma única vez, sendo a função chamada apenas durante a primeira execução.

Primeiramente, os itens inativos foram retirados da base, em seguida uma rotina foi criada para encontrar os elementos com frequência nula, excluindo-os também da base. A limpeza dos dados é realizada antes foi implementada como um método a ser acionado

Esse processo foi necessário para evitar que itens indesejados fossem selecionados durante a escolha dos elementos a serem minerados. Também foi realizada uma filtragem por período de tempo, considerando neste estudo as transações entre “01/01/2013” a “01/01/2014”.

Como ponto de partida, foi feita uma análise geral, considerando apenas os itens de mais alto nível (setores) contidos na base. Nesta etapa, alguns destes setores foram descartados pois englobavam subitens de controle interno da empresa, não sendo portanto relevantes no processo de mineração, como “Diversos”, “Miudezas”, “Uso/Consumo” e “Serviços do Banho e Tosa”.

Neste ponto, um conjunto de gráficos foi gerado através da variação da hierarquia e da otimização, porém apesar das tentativas de obtenção de uma boa visualização, os

pontos permaneciam sempre sobrepostos, impossibilitando uma análise satisfatória do resultado.

Através de uma avaliação mais detalhada das frequências correspondentes a cada setor foi possível detectar um fator determinante para a sobreposição dos pontos, nota-se que alguns setores possuíam valores de frequência muito superiores aos demais, o que resultava em uma desbalanceamento no momento da construção da matriz de distâncias, fazendo com que estes setores de alta frequência se mantivessem sempre nas bordas dos gráficos e provocando uma concentração dos demais em seu centro, tornando-o incompreensível.

Visando solucionar o problema da sobreposição dos pontos a seguinte medida foi tomada, os setores que possuíam valores de frequência com uma disparidade muito grande em relação aos demais foram substituídos por seus respectivos subníveis. Para garantir uma mudança mais controlada entre os níveis de hierarquia selecionados, este processo foi dividido em três etapas, sendo cada uma responsável pela subdivisão de um determinado setor.

- Etapa 1 - "Produtos da Linha Pet":

  Este setor possuía frequência quase duas vezes maior se comparado ao segundo mais frequente, sendo portanto o primeiro a ser divido. Os subníveis utilizados em sua substituição são pertencentes a categoria seção, sendo eles: "Hamster", "Cães", "Gatos" e "Outros Pets". Nesta mesma etapa a seção "Outros Pets" foi removida da seleção por possuir um giro muito baixo e por conter itens de pouco interesse para a empresa. Com esta divisão, foi possível obter uma melhor visualização dos dados, mas mesmo através das variações na hierarquização e na otimização ainda não foi possível alcançar um resultado satisfatório.

- Etapa 2 - "Cães":

  Apesar da divisão do setor "Produtos da Linha Pet" uma das seções que o substituíram ainda possuía frequência muito superior aos demais elementos, sendo portanto necessário decompor a seção "Cães" nos grupos "Coleiras e Guias', "Ração - Sacaria", "Medicamentos/Vitaminas", "Bebedouros e Comedouros", "Casas e Ca-

mas", "Higiene", "Ossos e Brinquedos", "Antipulgas", "Roupas", "Ração - Granel" e "Patês e Petiscos". Novamente, foi possível obter uma melhora na apresentação dos dados, porém uma parcela dos pontos ainda encontrava-se sobreposta, sendo portando necessário uma última divisão nos setores.

- Etapa 3 - "Componentes p/ Ração":

  Finalizando o processo de divisão dos setores com frequência muito alta foi necessário decompor o setor "Componentes p/ Ração" em algumas seções, tais como "Farelos", "Sal, Far. De Osso e Melaço" e "Milho e Fubá". Neste ponto, os valores das frequências tornaram-se razoavelmente balanceados, não sendo mais preciso realizar outras divisões nas categorias.

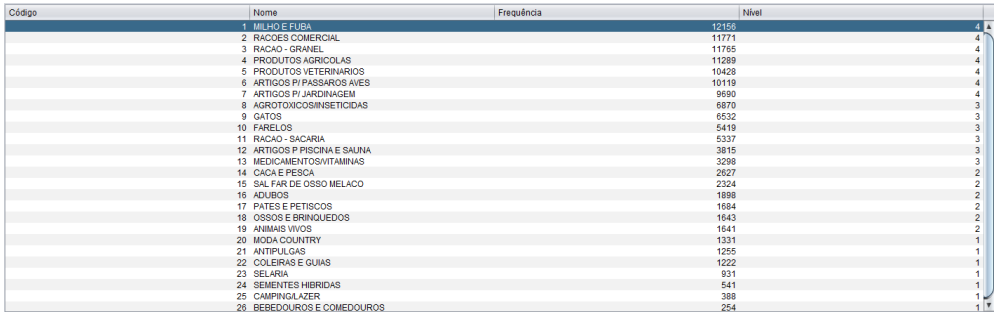
Ao término do processo de decomposição das categorias, uma melhora substancial na visualização do gráfico foi alcançada. No entanto, devido ao grande número de itens agora representados, ainda foram necessárias outras mudanças na seleção dos elementos.

Buscando obter um resultado mais satisfatório na apresentação dos dados foi realizada uma filtragem na seleção feita anteriormente, levando em conta para este procedimento o grau de importância de cada item para a empresa. Portanto, foram retirados apenas os itens considerados de pouco interesse. Os itens removidos nesta fase foram "Roupas", "Casas e Camas", "Artigos p/ Peixes", "Hamster", "Higiene", "Cordas e Correntes" e "Utilidades p/ Casa e Carro".

Neste ponto do processo foi obtida uma representação consideravelmente boa dos dados. Porém, alguns itens permaneciam sobrepostos. Buscando realizar este ajuste final, uma última medida foi tomada, um novo nível de hierarquia foi criado e os itens passaram a ser distribuídos em quatro níveis ao invés dos três anteriores.

Através desta alteração final e da variação nos valores de hierarquização e de otimização foi possível construir um gráfico com uma distribuição mais uniforme dos pontos, facilitando assim sua visualização e consequentemente o processo de análise. Segue a seleção final realizada e o seu respectivo gráfico.

Os Anexos II e III contêm o restante das tabelas que representam a seleção realizada e a distribuição hierarquica dos itens em cada étapa do processo, assim como os gráficos mais relevantes gerados durante a execução da ferramenta.



| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | MILHO E FUBA | 12156 | 4 |
| 2 | RACOES COMERCIAL | 11771 | 4 |
| 3 | RACAO - GRANEL | 11765 | 4 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 4 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 4 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 4 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 4 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 3 |
| 9 | GATOS | 6532 | 3 |
| 10 | FARELOS | 5419 | 3 |
| 11 | RACAO - SACARIA | 5337 | 3 |
| 12 | ARTIGOS P PISCINA E SAUNA | 3815 | 3 |
| 13 | MEDICAMENTOS/VITAMINAS | 3298 | 3 |
| 14 | CACA E PESCA | 2627 | 2 |
| 15 | SAL FAR DE OSSO MELACO | 2324 | 2 |
| 16 | ADUBOS | 1898 | 2 |
| 17 | PATES E PETISCOS | 1684 | 2 |
| 18 | OSSOS E BRINQUEDOS | 1643 | 2 |
| 19 | ANIMAIS VIVOS | 1641 | 2 |
| 20 | MODA COUNTRY | 1331 | 1 |
| 21 | ANTIPULGAS | 1255 | 1 |
| 22 | COLEIRAS E GUIAS | 1222 | 1 |
| 23 | SELARIA | 931 | 1 |
| 24 | SEMENTES HIBRIDAS | 541 | 1 |
| 25 | CAMPING/LAZER | 388 | 1 |
| 26 | BEBEDOUROS E COMEDOUROS | 254 | 1 |

Figura 5.1: Tabela de Distribuição das Hierarquias

## 5.2 Resultados

A partir do gráfico final obtido e da observação dos impactos gerados pela variação na hierarquização foi possível inferir um conjunto de regras, divididas entre as duas estratégias de marketing citadas anteriormente:

1. Produtos que vendem bem separadamente e conjuntamente devem ser expostos em locais distantes, objetivando induzir a venda dos produtos que encontram-se entre eles;

2. Caso um produto venda relativamente pouco e uma quantidade razoável de suas vendas sejam em conjunto com um produto com maior giro, estes devem ser expostos de forma próxima, buscando incentivar a venda do produto menos frequente.

Regras do tipo um:

1. “Ração - Granel” e “Artigos p/ Pássaros Aves”;

2. “Milho e Fubá” e “Ração Comercial”;

3. “Produtos Veterinários” e “Artigos p/ Pássaros Aves”;

4. “Produtos Agrícolas”, “Produtos Veterinários” e “Artigos p/ Jardinagem”.

Regras do tipo dois:

1. “Ração - Granel”, “Adubos”, “Sementes Hibridas” e “Selaria”. Sendo os três últimos os produtos de giro mais baixo que devem ser expostos próximos a “Ração-Granel”;

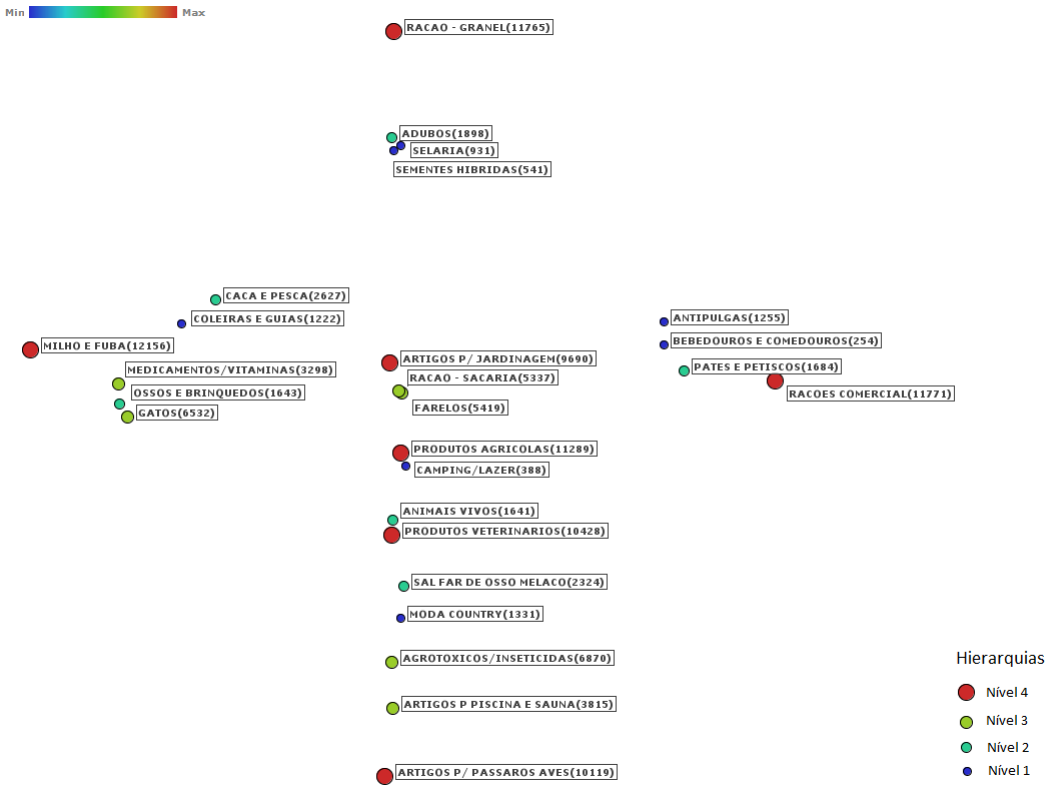



Figura 5.2: Gráfico Resultante do Processo de Mineração

2. “Rações Comercial”, “Patês e Petiscos”, “Anti-pulgas” e “Bebedouros e Comedouros”. Sendo os três últimos os produtos de giro mais baixo que devem ser expostos próximos a “Ração Comercial”;

3. “Artigos p/ Pássaros Aves”, “Agrotóxicos/Inseticidas”, “Artigos p/ Piscina e Sauna”, “Sal Far. Osso Melaço” e “Moda Country”. Sendo os quatro últimos os produtos de giro mais baixo que devem ser expostos próximos a “Artigos p/ Pássaros Aves”;

4. “Milho e Fubá”, “Gatos”, “Medicamentos/Vitaminas”, “Caça e Pesca”, “Ossos e Brinquedos” e “Coleiras e Guias”. Sendo os cinco últimos os produtos de giro mais baixo que devem ser expostos próximos a “Milho e Fubá”;

5. “Produtos Veterinários”, “Animais Vivos” e “Camping/Lazer”. Sendo os dois últimos os produtos de giro mais baixo que devem ser expostos próximos a “Produtos Veterinários”;

6. “Artigos p/ Jardinagem”, “Farelos” e “Ração - Sacaria”. Sendo os dois últimos os produtos de giro mais baixo que devem ser expostos próximos a “Artigos p/

Jardinagem";

7. "Agrotóxicos/Inseticidas" e "Moda Country". Sendo a "Moda Country" o produto de giro mais baixo que deve ser exposto próximo a "Agrotóxicos/Inseticidas".

Buscando validar as regras inferidas a partir da análise do gráfico foi realizado um teste utilizando o programa CBA. Neste teste, as relações encontradas através do gráfico foram utilizadas como entradas no programa. Esse programa realiza um processo de criação de regras de associação, fornecendo suporte e confiança da regra, que foram comparadas com as obtidas inicialmente utilizando os seguintes critérios:

- Regras que contém itens a serem expostos em locais distantes possuem suporte e confiança altos;
- Regras que contém itens a serem expostos em locais próximos possuem suporte baixo e confiança alta, sendo a direção da regra do item de menor para o de maior frequência.

A seguir são apresentadas as regras de associação resultantes. As seleções de itens utilizadas como entrada do CBA encontram-se no anexo IV.

Regras do tipo 1:

1. Artigos p/ Pássaros e Aves → Ração - Granel [62.414%, 54.90%]
2. Milho e Fubá → Ração Comercial [71.458%, 58.08%]
   Ração Comercial → Milho e Fubá [70.044%, 59.25%]
3. Artigos p/ Pássaros e Aves → Produtos Veterinários [65.634%, 50.22%]
4. Produtos Agricolas → Produtos Veterinários [50.560%, 32.29%]
   Produtos Veterinários → Produtos Agricolas [47.237%, 34.56%]
   Artigos p/ Jardinagem → Produtos Agricolas [44.157%, 32.48%]

Regras do tipo 2:

1. Adubos → Ração-Granel [15.317%, 74.56%]
   Selaria → Ração-Granel [7.705%, 71.12%]
   Sementes Hibridas → Ração-Granel [4.304%, 67.86%]

2. Patês e Petiscos → Rações Comercial [13.902%, 73.42%]

   Antipulgas → Rações Comercial [10.389%, 73.21%]

   Bebedouros e Comedouros → Rações Comercial [2.020%, 75.25%]

3. Artigos p/ Piscina e Sauna → Artigos p/ Pássaros e Aves [21.994%, 36.50%]

   Agrotóxicos/Inseticidas → Artigos p/ Pássaros e Aves [39.871%, 39.36%]

   Sal, Far. Osso e Melaço → Artigos p/ Pássaros e Aves [13.205%, 30.03%]

   Moda Country → Artigos p/ Pássaros e Aves [7.896%, 36.10%]

4. Caça e Pesca → Milho e Fubá [13.996%, 34.36%]

   Medicamentos e Vitaminas → Milho e Fubá [17.002%, 34.01%]

   Gatos → Milho e Fubá [33.764%, 38.38%]

   Coleiras e Guias → Milho e Fubá [6.242%, 31.83%]

   Ossos e Brinquedos → Milho e Fubá [8.377%, 30.66%]

5. Animais Vivos → Produtos Veterinários [15.591%, 79.55%]

   Camping/Lazer → Produtos Veterinários [3.623%, 81.07%]

6. Ração-Sacaria → Produtos p/ Jardinagem [35.488%, 44.05%]

7. Moda Country → Agrotóxicos/Inseticidas [19.19%, 79.72%]

# 6 Conclusão

A mineração de dados é, sem dúvida, uma ferramenta fundamental para a obtenção de informações pertinentes em grandes bases de dados, favorecendo fortemente o processo de tomada de decisão. Entretanto, sua execução mostra-se extremamente interativa e iterativa. Em especial no caso das regras de associação, o que demanda um grande esforço por parte do minerador.

Tomando-se estes preceitos como base, mostra-se de suma importância o desenvolvimento de ferramentas que busquem oferecer suporte ao processo de KDD. Assim, objetivando a simplicidade, rapidez e automatização do processo, limitando ao usuário final apenas as tarefas nas quais ele é indispensável, como a análise final e a tomada de decisão.

Buscando atingir as metas de simplicidade, rapidez e automação, foi desenvolvida no decorrer deste projeto uma ferramenta para facilitar o processo de seleção dos dados. Além disso, foi proposto um método de integração entre os diferentes programas necessários ao modelo de mineração, proposto no método utilizado como base deste trabalho.

A partir dos resultados obtidos durante a execução do estudo de caso, foi possível concluir em relação ao método estudado e a ferramenta proposta:

- Todos as ferramentas necessárias à mineração foram integradas em uma única interface, tornando assim o processo mais rápido e simples;
- Toda a complexidade inerente a modificação e armazenamento de arquivos foi eficientemente tratada em segundo plano, evitando assim que o minerador necessite tomar conhecimento de funções que não possuem relacionamento direto com a sua atividade;
- O processo de KDD, como um todo, foi significativamente acelerado pela utilização da ferramenta, reduzindo o tempo necessário a geração de um gráfico em cerca de 70%. Com a utilização da ferramenta, o pré-processamento dos dados passou a

acontecer de forma simplificada, sendo todos os arquivos necessários ao funcionamento dos programas de mineração gerados automaticamente, possibilitando assim ao minerador realizar suas atividades de forma mais rápida e menos complexa;

- Os resultados obtidos através do estudo de caso indicam o bom funcionamento do método estudado. Esta conclusão pôde ser obtida através da análise do gráfico final e das regras de associação geradas a partir do programa CBA, sendo ambas condizentes com as duas estratégias de marketing citadas no decorrer do trabalho, ou seja:
  - Segundo o gráfico, produtos que deveriam estar expostos próximos apresentaram, durante a execução do CBA, regras de associação com baixo suporte e alta confiança;
  - Ainda em relação ao gráfico, produtos que deveriam estar expostos de forma distante apresentaram, durante a execução do CBA, regras de associação com suporte e confiança altos.

Apesar da ferramenta ter apresentado resultados satisfatórios, muito ainda pode ser feito na tentativa de torná-la mais eficiente. Como trabalhos futuros sugere-se:

- Execução de um estudo de caso nos níveis mais baixos da hierarquia, utilizando como ponto de partida os resultados obtidos no decorrer deste trabalho;
- Tornar o processo de geração dos gráficos mais rápido e iterativo, como por exemplo através da implementação de escalas capazes de variar os valores de hierarquização e otimização em tempo real, possibilitando uma visualização mais intuitiva dos resultados;
- Tornar a ferramenta mais flexível, de forma que ela possa se conectar facilmente a outros modelos de banco de dados;
- Implementar juntamente a ferramenta, um método de clusterização, sendo este responsável por fornecer uma sugestão inicial para a distribuição das hierarquias;

- Propor e avaliar outras funções para o cálculo da distância utilizado pelo sistema CalcD.

- Acrescentar ao processo um método de balanceamento automático das frequências.

# Referências Bibliográficas


Agrawal, R.; Imieliński, T. ; Swami, A. **Mining association rules between sets of items in large databases**. In: ACM SIGMOD Record, volume 22, p. 207–216. ACM, 1993.

Agrawal, R.; Srikant, R. **Fast algorithms for mining association rules**. In: Proc. 20th int. conf. very large data bases, VLDB, volume 1215, p. 487–499, 1994.

Bispo, C. A. F. **Uma análise da nova geração de sistemas de apoio à decisão**. 1998. Dissertação de Mestrado - Escola de Engenharia de São Carlos - Universidade de São Paulo.

Curotto, C. L. **Integração de Recursos de Data Mining com Gerenciadores de Bancos de Dados Relacionais**. 2003. Tese de Doutorado - COPPE/UFRJ.

Fayyad, U. M.; Piatetsky-Shapiro, G. ; Smyth, P. **From data mining to knowledge discovery in databases**. volume Vol. 17, p. 37–54. Al Magazine, 1996.

Han, J.; Kamber, M. **Data mining: Concepts and techniques**. Morgan kaufmann, 2006.

Liao, S.-H.; Chu, P.-H. ; Hsiao, P.-Y. **Data mining techniques and applications–a decade review from 2000 to 2011**. volume 39, p. 11303–11311. Elsevier, 2012.

Liu, B. **Web data mining**. In: Springer-Verlag Berlin Heidelberg, 2008.

Motta, C. G. L. **Metodologia para mineração de regras de associação multiníveis incluindo pré e pós-processamento**. 2010. Tese de Doutorado - COPPE/UFRJ.

Motta, C. G. L. **Mineração de dados: Capítulo 1 - introdução. material de aula, disciplina mineração de dados (dcc 127)**. 2013.

Nicholas, P. D. M.; Zhao, Y. Association rules: an overview. **Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction**, p. 1–10, 2009.

SBC. **Grandes desafios da pesquisa em computação no brasil, 2006/2016, relatório sobre o seminário realizado em 8 e 9 de maio de 2006**. São Paulo, Brasil, 2006. Sociedade Brasileira de Computação.

SBC. **Grandes desafios da pesquisa em computação no brasil, ênfase em grandes desafios do mercado e do governo, relatório sobre o seminário realizado em 15 e 16 de abril de 2013**. São Paulo, Brasil, 2013. Sociedade Brasileira de Computação.

Singh, J.; Ram, H. ; Sodhi, D. J. **Improving efficiency of apriori algorithm using transaction reduction**. volume 3, p. 1–4. Citeseer, 2013.

Stock, M. M. **Pré-processamento de dados para mineração de regras de associação multiníveis**. Monografia - UFJF, 2012. Juiz de Fora, Brasil.

Trindade, A. R.; Costa, M. V. B. Uso de mineração de dados para desoberta de conhecimento: Estudo de caso do vestibular da universidade federal dos vales do jequitinhonha e mucuri (ufvjm). **Revista de Ciência e Tecnologia do Vale do Mucuri**, , n.4, p. 66–78, 2013.

# I Arquivos Gerados

São apresentados a seguir exemplos dos arquivos gerados durante a execução da ferramenta desenvolvida.

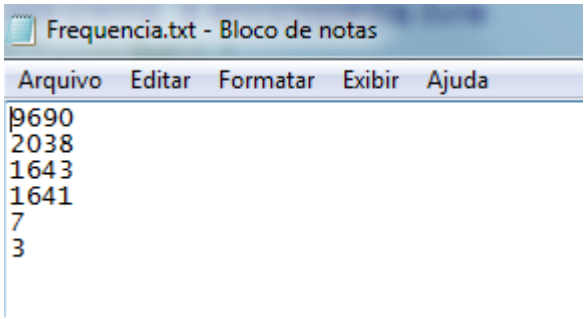



Figura I.1: Exemplo Arquivo Frequência Gerado (GeraD)

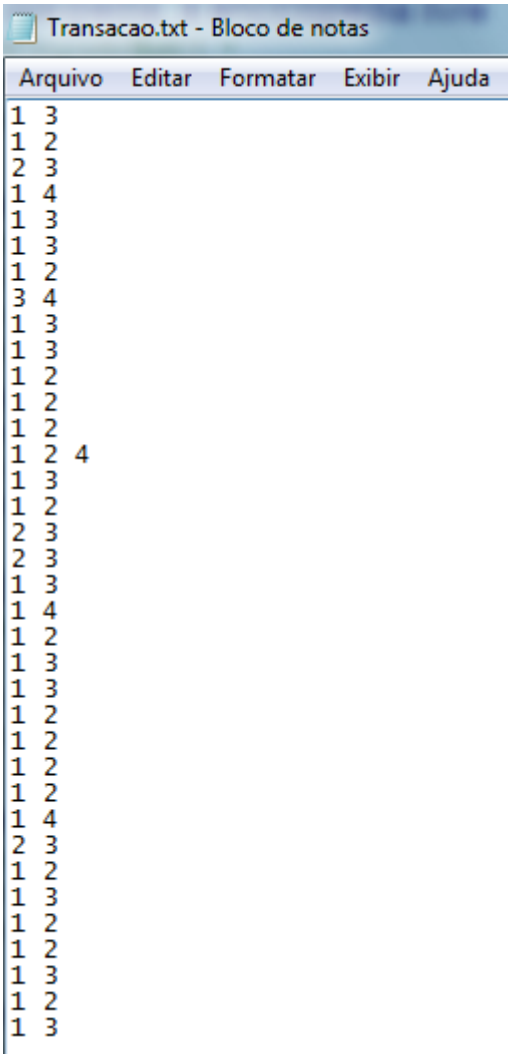



Figura I.2: Exemplo Arquivo Transação Gerado (GeraD)

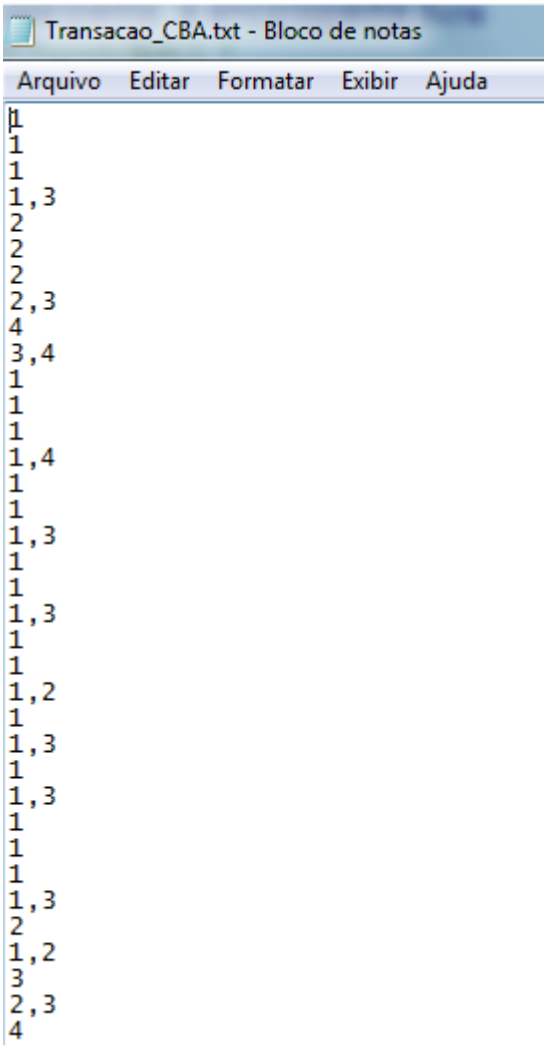



Figura I.3: Exemplo Arquivo Transação Gerado (CBA)

# II Gráficos Gerados

Este anexo contêm os gráficos mais relevantes gerados durante a execução da ferramenta desenvolvida. Devido a grande quantidade de gráficos gerados na mineração, apenas os gráficos que apresentaram melhor visualização dos itens serão apresentados.

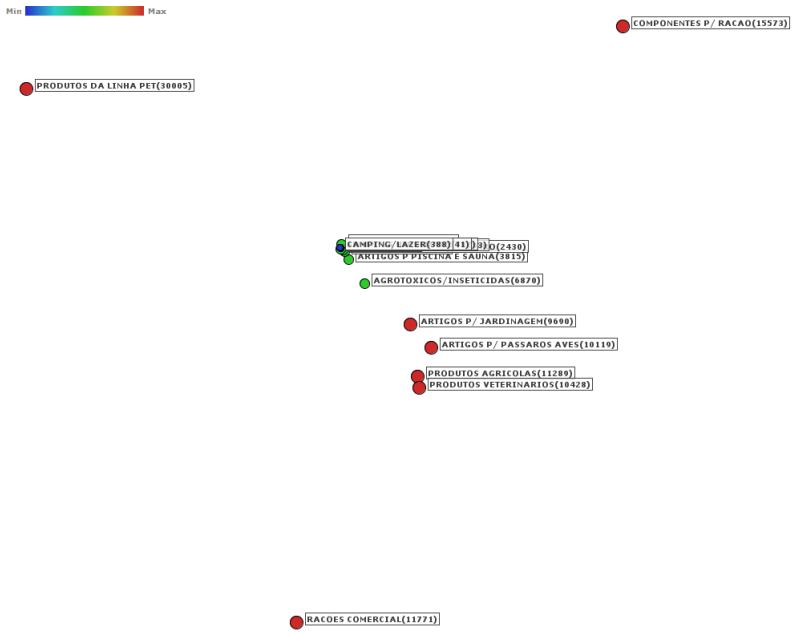



Figura II.1: Gráfico Gerado a Partir de Todos os Setores Relevantes da Empresa - Hierarquização = 0.6

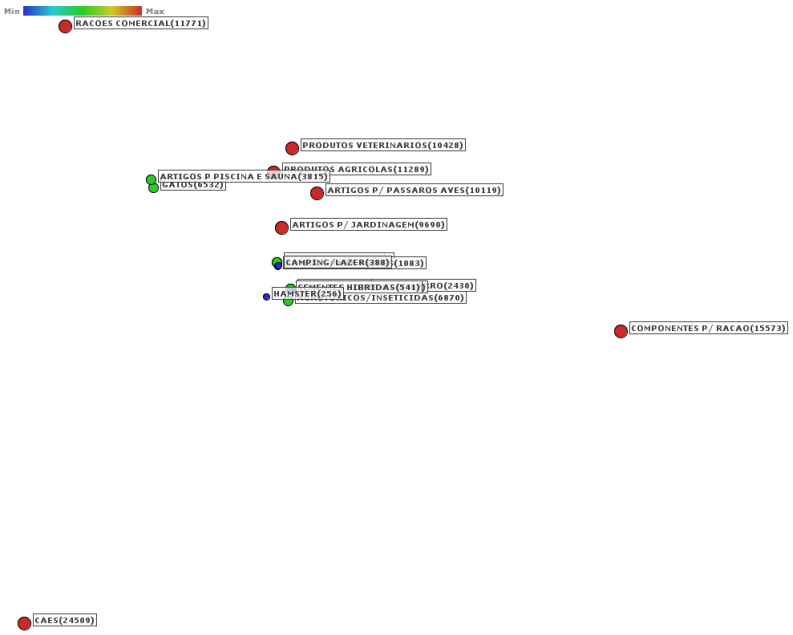



Figura II.2: Gráfico Gerado a Partir da Divisão do Setor Pet - Hierarquização = 0.7

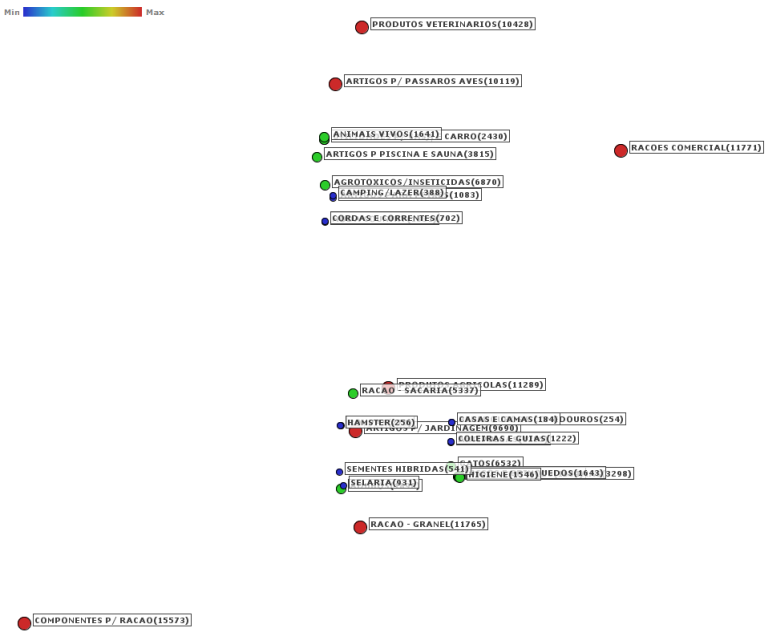



Figura II.3: Gráfico Gerado a Partir da Divisão da Seção Cães - Hierarquização = 0.6


Figura II.4: Gráfico Gerado a Partir da Divisão do Setor Componentes para Ração - Hierarquização = 0.7


Figura II.5: Gráfico Gerado a Partir da Limpeza Final Feita nos Itens Selecionados - Hierarquização = 0.6

# III Tabelas Geradas

Este anexo contêm as tabelas que representam a seleção realizada e a distribuição hierarquica dos itens em cada étapa do processo de mineração.

| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | PRODUTOS DA LINHA PET | 30005 | 3 |
| 2 | COMPONENTES P/ RACAO | 15573 | 3 |
| 3 | RACOES COMERCIAL | 11771 | 3 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 3 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 3 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 3 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 3 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 2 |
| 9 | ARTIGOS P PISCINA E SAUNA | 3815 | 2 |
| 10 | CACA E PESCA | 2627 | 2 |
| 11 | UTILIDADES P/ CASA E CARRO | 2430 | 2 |
| 12 | ADUBOS | 1898 | 2 |
| 13 | ANIMAIS VIVOS | 1641 | 2 |
| 14 | MODA COUNTRY | 1331 | 2 |
| 15 | ARTIGOS PARA PEIXES | 1083 | 1 |
| 16 | SELARIA | 931 | 1 |
| 17 | CORDAS E CORRENTES | 702 | 1 |
| 18 | SEMENTES HIBRIDAS | 541 | 1 |
| 19 | CAMPING/LAZER | 388 | 1 |

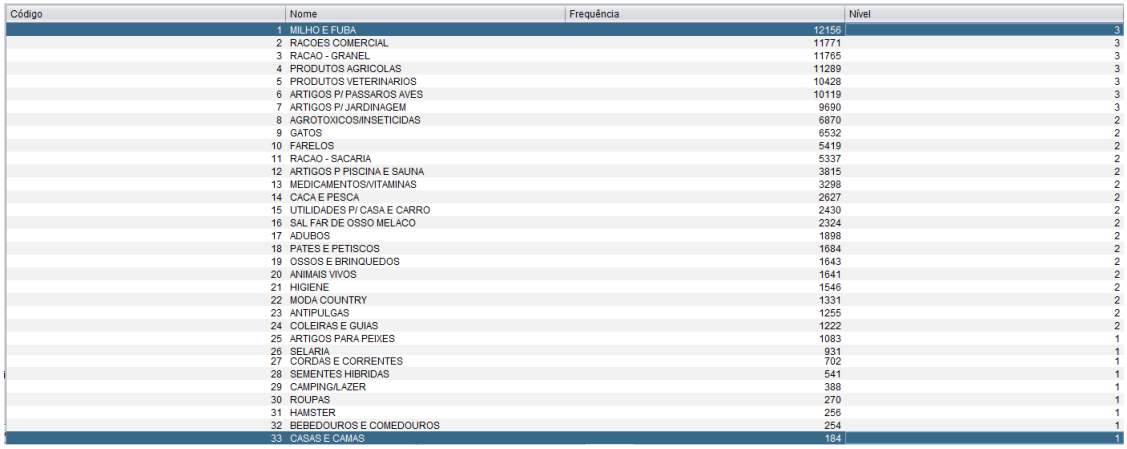
Figura III.1: Seleção de Todos os Setores

| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | CAES | 24509 | 3 |
| 2 | COMPONENTES P/ RACAO | 15573 | 3 |
| 3 | RACOES COMERCIAL | 11771 | 3 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 3 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 3 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 3 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 3 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 2 |
| 9 | GATOS | 6532 | 2 |
| 10 | ARTIGOS P PISCINA E SAUNA | 3815 | 2 |
| 11 | CACA E PESCA | 2627 | 2 |
| 12 | UTILIDADES P/ CASA E CARRO | 2430 | 2 |
| 13 | ADUBOS | 1898 | 2 |
| 14 | ANIMAIS VIVOS | 1641 | 2 |
| 15 | MODA COUNTRY | 1331 | 2 |
| 16 | ARTIGOS PARA PEIXES | 1083 | 1 |
| 17 | SELARIA | 931 | 1 |
| 18 | CORDAS E CORRENTES | 702 | 1 |
| 19 | SEMENTES HIBRIDAS | 541 | 1 |
| 20 | CAMPING/LAZER | 388 | 1 |
| 21 | HAMSTER | 256 | 1 |

Figura III.2: Seleção com Setor Pet Dividido

| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | COMPONENTES P/ RACAO | 15573 | 3 |
| 2 | RACOES COMERCIAL | 11771 | 3 |
| 3 | RACAO - GRANEL | 11765 | 3 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 3 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 3 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 3 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 3 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 2 |
| 9 | GATOS | 6532 | 2 |
| 10 | RACAO - SACARIA | 5337 | 2 |
| 11 | ARTIGOS P PISCINA E SAUNA | 3815 | 2 |
| 12 | MEDICAMENTOS/VITAMINAS | 3298 | 2 |
| 13 | CACA E PESCA | 2627 | 2 |
| 14 | UTILIDADES P/ CASA E CARRO | 2430 | 2 |
| 15 | ADUBOS | 1898 | 2 |
| 16 | PATES E PETISCOS | 1684 | 2 |
| 17 | OSSOS E BRINQUEDOS | 1643 | 2 |
| 18 | ANIMAIS VIVOS | 1641 | 2 |
| 19 | HIGIENE | 1546 | 2 |
| 20 | MODA COUNTRY | 1331 | 2 |
| 21 | ANTIPULGAS | 1255 | 2 |
| 22 | COLEIRAS E GUIAS | 1222 | 2 |
| 23 | ARTIGOS PARA PEIXES | 1083 | 1 |
| 24 | SELARIA | 931 | 1 |
| 25 | CORDAS E CORRENTES | 702 | 1 |
| 26 | SEMENTES HIBRIDAS | 541 | 1 |
| 27 | CAMPING/LAZER | 388 | 1 |
| 28 | ROUPAS | 270 | 1 |
| 29 | HAMSTER | 256 | 1 |
| 30 | BEBEDOUROS E COMEDOUROS | 254 | 1 |
| 31 | CASAS E CAMAS | 184 | 1 |

Figura III.3: Seleção com Seção Cães Dividida



| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | MILHO E FUBA | 12156 | 3 |
| 2 | RACOES COMERCIAL | 11771 | 3 |
| 3 | RACAO - GRANEL | 11765 | 3 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 3 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 3 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 3 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 3 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 2 |
| 9 | GATOS | 6532 | 2 |
| 10 | FARELOS | 5419 | 2 |
| 11 | RACAO - SACARIA | 5337 | 2 |
| 12 | ARTIGOS P PISCINA E SAUNA | 3815 | 2 |
| 13 | MEDICAMENTOS/VITAMINAS | 3298 | 2 |
| 14 | CACA E PESCA | 2627 | 2 |
| 15 | UTILIDADES P/ CASA E CARRO | 2430 | 2 |
| 16 | SAL FAR DE OSSO MELACO | 2324 | 2 |
| 17 | ADUBOS | 1898 | 2 |
| 18 | PATES E PETISCOS | 1684 | 2 |
| 19 | OSSOS E BRINQUEDOS | 1643 | 2 |
| 20 | ANIMAIS VIVOS | 1641 | 2 |
| 21 | HIGIENE | 1546 | 2 |
| 22 | MODA COUNTRY | 1331 | 2 |
| 23 | ANTIPULGAS | 1255 | 2 |
| 24 | COLEIRAS E GUIAS | 1222 | 2 |
| 25 | ARTIGOS PARA PEIXES | 1083 | 1 |
| 26 | SELARIA | 931 | 1 |
| 27 | CORDAS E CORRENTES | 702 | 1 |
| 28 | SEMENTES HIBRIDAS | 541 | 1 |
| 29 | CAMPING/LAZER | 388 | 1 |
| 30 | ROUPAS | 270 | 1 |
| 31 | HAMSTER | 256 | 1 |
| 32 | BEBEDOUROS E COMEDOUROS | 254 | 1 |
| 33 | CASAS E CAMAS | 184 | 1 |

Figura III.4: Seleção com Seção Componentes para Ração Dividida

| Código | Nome | Frequência | Nível |
|---|---|---|---|
| 1 | MILHO E FUBA | 12156 | 3 |
| 2 | RACOES COMERCIAL | 11771 | 3 |
| 3 | RACAO - GRANEL | 11765 | 3 |
| 4 | PRODUTOS AGRICOLAS | 11289 | 3 |
| 5 | PRODUTOS VETERINARIOS | 10428 | 3 |
| 6 | ARTIGOS P/ PASSAROS AVES | 10119 | 3 |
| 7 | ARTIGOS P/ JARDINAGEM | 9690 | 3 |
| 8 | AGROTOXICOS/INSETICIDAS | 6870 | 2 |
| 9 | GATOS | 6532 | 2 |
| 10 | FARELOS | 5419 | 2 |
| 11 | RACAO - SACARIA | 5337 | 2 |
| 12 | ARTIGOS P PISCINA E SAUNA | 3815 | 2 |
| 13 | MEDICAMENTOS/VITAMINAS | 3298 | 2 |
| 14 | CACA E PESCA | 2627 | 2 |
| 15 | SAL FAR DE OSSO MELACO | 2324 | 2 |
| 16 | ADUBOS | 1898 | 2 |
| 17 | PATES E PETISCOS | 1684 | 2 |
| 18 | OSSOS E BRINQUEDOS | 1643 | 2 |
| 19 | ANIMAIS VIVOS | 1641 | 2 |
| 20 | MODA COUNTRY | 1331 | 2 |
| 21 | ANTIPULGAS | 1255 | 1 |
| 22 | COLEIRAS E GUIAS | 1222 | 1 |
| 23 | SELARIA | 931 | 1 |
| 24 | SEMENTES HIBRIDAS | 541 | 1 |
| 25 | CAMPING/LAZER | 388 | 1 |
| 26 | BEBEDOUROS E COMEDOUROS | 254 | 1 |

Figura III.5: Seleção com Limpeza de Itens Realizada

# IV Seleções de Entrada CBA

Este anexo contêm as tabelas que representam as seleções utilizadas como entradas do CBA

| Código | Nome | Frequência |
|---|---|---|
| 1 | RACAO - GRANEL | 11765 |
| 2 | ARTIGOS P/ PASSAROS AVES | 10119 |

Figura IV.1: Seleção Referente a Regra Tipo 1 Número 1

| Código | Nome | Frequência |
|---|---|---|
| 1 | MILHO E FUBA | 12156 |
| 2 | RACOES COMERCIAL | 11771 |

Figura IV.2: Seleção Referente a Regra Tipo 1 Número 2

| Código | Nome | Frequência |
|---|---|---|
| 1 | PRODUTOS VETERINARIOS | 10428 |
| 2 | ARTIGOS P/ PASSAROS AVES | 10119 |

Figura IV.3: Seleção Referente a Regra Tipo 1 Número 3

| Código | Nome | Frequência |
|---|---|---|
| 1 | PRODUTOS AGRICOLAS | 11289 |
| 2 | PRODUTOS VETERINARIOS | 10428 |
| 3 | ARTIGOS P/ JARDINAGEM | 9690 |

Figura IV.4: Seleção Referente a Regra Tipo 1 Número 4

| Código | Nome | Frequência |
|---|---|---|
| 1 | RACAO - GRANEL | 11765 |
| 2 | ADUBOS | 1898 |
| 3 | SELARIA | 931 |
| 4 | SEMENTES HIBRIDAS | 541 |

Figura IV.5: Seleção Referente a Regra Tipo 2 Número 1

| Código | Nome | Frequência |
|---|---|---|
| 1 | RACOES COMERCIAL | 11771 |
| 2 | PATES E PETISCOS | 1684 |
| 3 | ANTIPULGAS | 1255 |
| 4 | BEBEDOUROS E COMEDOUROS | 254 |

Figura IV.6: Seleção Referente a Regra Tipo 2 Número 2

| Código | Nome | Frequência |
|---|---|---|
| 1 | ARTIGOS P/ PASSAROS AVES | 10119 |
| 2 | AGROTOXICOS/INSETICIDAS | 6870 |
| 3 | ARTIGOS P PISCINA E SAUNA | 3815 |
| 4 | SAL FAR DE OSSO MELACO | 2324 |
| 5 | MODA COUNTRY | 1331 |

Figura IV.7: Seleção Referente a Regra Tipo 2 Número 3

| Código | Nome | Frequência |
|---|---|---|
| 1 | MILHO E FUBA | 12156 |
| 2 | GATOS | 6532 |
| 3 | MEDICAMENTOS/VITAMINAS | 3298 |
| 4 | CACA E PESCA | 2627 |
| 5 | OSSOS E BRINQUEDOS | 1643 |
| 6 | COLEIRAS E GUIAS | 1222 |

Figura IV.8: Seleção Referente a Regra Tipo 2 Número 4

| Código | Nome | Frequência |
|---|---|---|
| 1 | PRODUTOS VETERINARIOS | 10428 |
| 2 | ANIMAIS VIVOS | 1641 |
| 3 | CAMPING/LAZER | 388 |

Figura IV.9: Seleção Referente a Regra Tipo 2 Número 5

| Código | Nome | Frequência |
|---|---|---|
| 1 | ARTIGOS P/ JARDINAGEM | 9690 |
| 2 | FARELOS | 5419 |
| 3 | RACAO - SACARIA | 5337 |

Figura IV.10: Seleção Referente a Regra Tipo 2 Número 6

| Código | Nome | Frequência |
|---|---|---|
| 1 | AGROTOXICOS/INSETICIDAS | 6870 |
| 2 | MODA COUNTRY | 1331 |

Figura IV.11: Seleção Referente a Regra Tipo 2 Número 7