

Sobre Regras de Associação utilizando Mineração de Dados

Gustavo Henrique de Rosa

Faculdade de Ciências (FC) - Campus Bauru
Universidade Estadual Paulista “Júlio de Mesquita Filho”

Orientador: Prof. Dr. João Paulo Papa

Apresentação para a Mostra dos TCCs
04 de fevereiro de 2016

Sumário

- 1 Mineração de Dados
 - Introdução
- 2 Regras de Associação
 - Modelo formal
 - Geração de conjuntos de itens frequentes
 - Princípio Apriori
 - FP-Growth
 - Geração de regras
- 3 Base de Dados
 - Base de dados real
 - Base de dados sintética
- 4 Experimentos
 - Base real
 - Base sintética 1k
 - Base sintética 10k
 - Tempos de execução
 - Regras geradas

Introdução

A forte concorrência do capitalismo tem implicado constantemente em avanços no estudo de técnicas responsáveis por auxiliar em tarefas de **tomada de decisão**.

- Rápido desenvolvimento x *Big Data*

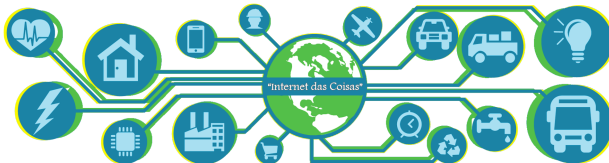


Figura 1: A conexão de dispositivos na futura rede chamada “Internet das Coisas”.

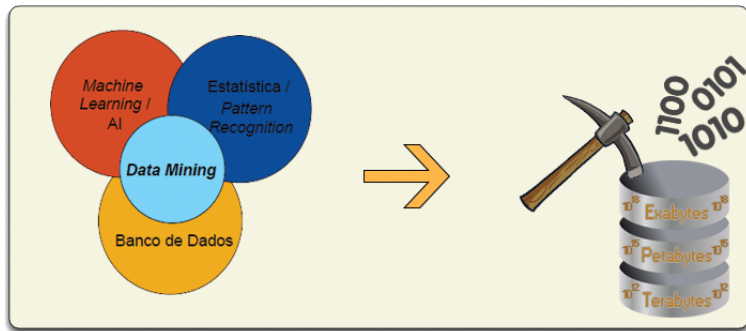
Figura 2: Disposição de produtos em uma prateleira de supermercado.



Exemplo

É de extrema importância a existência de uma lógica entre os dados, por exemplo, a disposição física de produtos dentro de uma prateleira.

Figura 3: Ilustração de um simples processo de mineração em uma base de dados.



A mineração de dados¹ sustenta-se com base em outras áreas da ciência da computação, sendo representada por um aglomerado de diferentes ferramentas capazes de serem utilizadas em conjunto.

- Aprendizado de máquina;
- Reconhecimento de padrões;
- Banco de dados;
- Estatística.

¹R. Agrawal/T. Imielinski/Arun Swami: Database mining: a performance perspective, em: Knowledge and Data Engineering, IEEE Transactions on 5.6 (1993a), pp. 914–925.

A efetiva análise de dados de uma empresa pode resultar em associações e pesquisas de mercado muito qualificadas, tornando-se ótimas opções para enfrentar problemas de tomada de decisão².

²Michael Stonebraker et al.: DBMS Research at a Crossroads: The Vienna Update, em: VLDB '93 Proceedings of the 19th International Conference on Very Large Data Bases, 1993, pp. 688–692.

- A computação demanda um nível de complexidade para cada atividade que efetua;
- O grande volume de dados pode ser visto como uma mina de ouro, embora muitas vezes este ouro esteja escondido e quebradiço.

Qualidade x **Quantidade**

A abordagem utilizada enquadra-se dentro das **regras de associação**³.

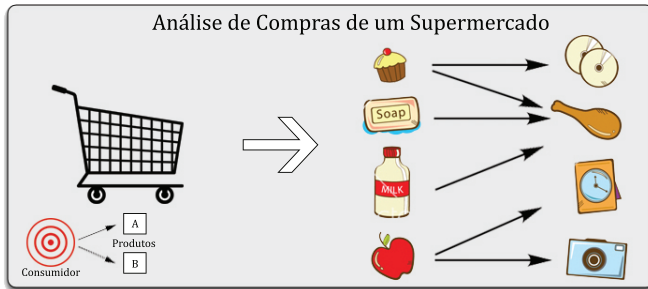


Figura 4: Análise de compras de um supermercado.

³Rakesh Agrawal/Tomasz Imielinski/Arun Swami: Mining Association Rules Between Sets of Items in Large Databases, em: SIGMOD Rec. 22.2 (jun. de 1993b), pp. 207–216.

Exemplo

Informações históricas de consumidores, compras, dentre outros, de uma determinada loja (i.e. supermercados, lojas de departamento, etc).

- As regras de associação utilizam informações armazenadas em um banco de dados;
- A partir destes dados, algoritmos são empregados para sua criação;
- As informações desconexas existentes são inferidas e ordenadas logicamente.

Modelo formal

TID	Conjunto de Itens
1	{Pão, Leite}
2	{Pão, Fralda, Cerveja, Ovos}
3	{Leite, Fralda, Cerveja, Coca-Cola}
4	{Pão, Leite, Fralda, Cerveja}
5	{Pão, Leite, Fralda, Coca-Cola}

Tabela 1: Exemplo de transações de mercado.

Considere os dados apresentados pela Tabela 1 como **transações de mercado**.

- Cada linha representa uma transação, a qual contém um identificador único nomeado por **TID** e um **conjunto de itens** comprados por um determinado consumidor.
- A título de exemplificação, a seguinte regra pode ser extraída do conjunto de dados da Tabela 1:

$\{\text{Fralda}\} \longrightarrow \{\text{Cerveja}\}.$

TID	Pão	Leite	Fralda	Cerveja	Ovos	Coca-Cola
1	1	1	0	0	0	0
2	1	0	1	1	1	0
3	0	1	1	1	0	1
4	1	1	1	1	0	0
5	1	1	1	0	0	1

Tabela 2: Representação binária das transações de mercado.

- Seja $I = \{I_1, I_2, \dots, I_N\}$ um conjunto de atributos binários, nomeados de **itens** e T uma base de dados de transações.

- Seja X um conjunto de itens de I . Uma transação t satisfaz X se, para todos os itens $I_k \in X$, $t[k] = 1$.
- A contagem de suporte, $\sigma(X)$, para um conjunto de itens X , pode ser descrita como segue:

$$\sigma(X) = |\{t_i \mid X \subseteq t_i, t_i \in T\}|$$

Uma **regra de associação** implica a forma $X \longrightarrow Y$, onde X e Y são conjuntos de itens disjuntos, ou seja, $X \cap Y = \emptyset$.

- A força de uma regra pode ser mensurada em termos de **suporte** e **confiança**.

O suporte determina o quão frequente a regra é aplicável ao conjunto de dados utilizado, enquanto a confiança estabelece o quão frequente itens de Y aparecem em transações que contenham X .

As definições formais e matemáticas dessas métricas são apresentadas abaixo:

$$\textit{Suporte}; s(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{N},$$

$$\textit{Confiança}; c(X \longrightarrow Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}.$$

Exemplo

Considere a regra $\{\text{Pão, Leite}\} \rightarrow \{\text{Fralda}\}$

- A contagem de suporte para o conjunto de itens $\{\text{Pão, Leite, Fralda}\}$ é **2** e o número total de transações é **5**, portanto o suporte da regra é $2/5 = \mathbf{0.4}$;
- A confiança da regra é obtida através da divisão da contagem de suporte de $\{\text{Pão, Leite, Fralda}\}$ pela contagem de suporte de $\{\text{Pão, Leite}\}$. Dado que são **3** transações que possuem pão e leite, a confiança para essa regra é $2/3 = \mathbf{0.67}$.

O problema de mineração de regras pode ser decomposto em dois subproblemas:

- Gerar todas as combinações de itens que contenham um suporte mínimo determinado. Caso essas combinações tenham satisfeito o limite mínimo serão chamadas de **conjuntos de itens frequentes**.
- Para um dado conjunto de item frequente $Y = \{l_1, l_2, \dots, l_k\}$, com $k \geq 2$, gerar todas as regras que utilizam os itens do conjunto Y . É interessante estabelecer um limite mínimo para a confiança das regras.

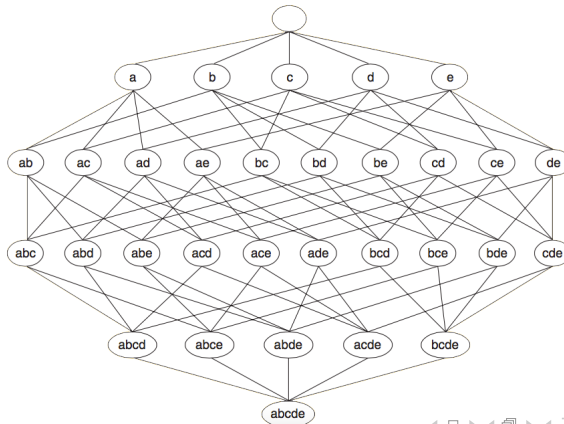
Uma abordagem de força-bruta para minerar regras de associações é computar o suporte e a confiança de toda possível regra. Particularmente, o número total R de possíveis regras extraídas de um conjunto de dados que contém d itens é:

$$R = 3^d - 2^{d+1} + 1.$$

Mesmo para pequenos conjuntos de dados, como o da Tabela 1, esse método requer o cálculo de suporte e confiança para $3^6 - 2^7 + 1 =$ **602** regras.

Geração de conjuntos de itens frequentes

Figura 5: Estrutura em grade de um conjunto de itens.



A Figura 5, composta por uma estrutura em grade, pode ser utilizada para enumerar uma lista de todos os possíveis conjuntos de itens. Neste exemplo, o conjunto de itens da grade é representado por $I = \{a, b, c, d, e\}$.

Usualmente, uma base de dados que contenha k itens pode gerar até no máximo $2^k - 1$ conjuntos de itens frequentes, excluindo o conjunto nulo.

Existem diversas maneiras para se reduzir a complexidade exponencial da geração de conjuntos de itens frequentes, dentre elas:

- **Redução do número de conjuntos de itens candidatos (M):**
Um conjunto candidato é um conjunto a ser analisado pelo algoritmo antes de ser considerado frequente, ou seja, um conjunto gerado prévio à sua contagem de suporte.
- **Redução do número de comparações:** É possível reduzir o número de comparações utilizando estruturas de dados mais avançadas.

Princípio Apriori

O primeiro algoritmo de mineração de regras de associação foi o *Apriori*⁴, tido como o precursor do uso do corte baseado no suporte.

Teorema

Se um conjunto de itens é frequente, então todos os seus subconjuntos também devem ser frequentes.

⁴Rakesh Agrawal/Ramakrishnan Srikant: Fast Algorithms for Mining Association Rules in Large Databases, em: Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), San Francisco, CA, USA 1994, pp. 487–499.



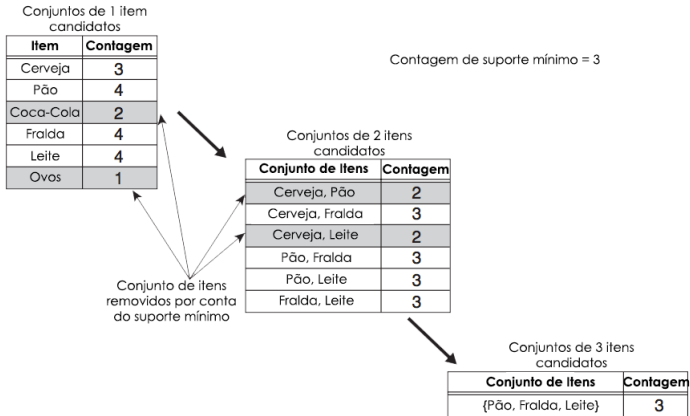
Este diagrama ilustra la estructura de un conjunto de ítems infrecuente y sus superconjuntos. Se muestra una jerarquía de nodos representados por círculos:

- Nivel 1:** Los ítems individuales a, b, c, d, e .
- Nivel 2:** Los pares de ítems $ab, ac, ad, ae, bc, bd, be, cd, ce, de$.
- Nivel 3:** Los triples de ítems $abc, abd, abe, acd, ace, ade, bcd, bce, bde, cde$.
- Nivel 4:** Los cuádruples de ítems $abcd, abce, abde, acde, bcde$.
- Nivel 5:** El conjunto de todos los ítems $abcde$.

Las conexiones entre niveles representan las relaciones de inclusión entre conjuntos de ítems. Se destacan dos grupos con líneas punteadas:

- Conjunto de ítems infrecuente:** Indica el nodo ab y los nodos que lo contienen (abc, abd, abe).
- Superconjuntos cortados:** Indica los nodos $abc, abd, abe, abcd, abce, abde$, que son superconjuntos del conjunto infrecuente.

Figura 8: Ilustração da geração de conjuntos de itens frequentes pelo algoritmo *Apriori*.



FP-Growth

O algoritmo *FP-growth*⁵, contrário ao seu concorrente *Apriori*, utiliza uma metodologia totalmente diferente para a descoberta dos conjuntos de itens frequentes.

A base de dados é codificada a partir do uso de uma estrutura de dados denominada árvore FP, onde os conjuntos de itens frequentes são extraídos diretamente desta estrutura.

⁵Jiawei Han/Jian Pei/Yiwen Yin: Mining Frequent Patterns Without Candidate Generation, em: Proceedings of SIGMOD Rec. 29.2 (maio de 2000), pp. 1–12.

- Representação compacta dos dados de entrada;
- Cada transação é lida e mapeada para um caminho da árvore FP;
- Caso possua itens em comum, seus caminhos são sobrepostos;
- A sobreposição de caminhos reflete em um nível maior de compressão, possibilitando a extração de conjuntos de itens frequentes diretamente da estrutura.

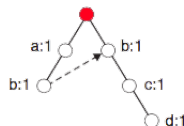
Figura 9: Construção de uma árvore FP.

TID	Itens
1	{a,b}
2	{b,c,d}
3	{a,c,d,e}
4	{a,d,e}
5	{a,b,c}
6	{a,b,c,d}
7	{a}
8	{a,b,c}
9	{a,b,d}
10	{b,c,e}

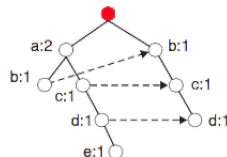
● nulo



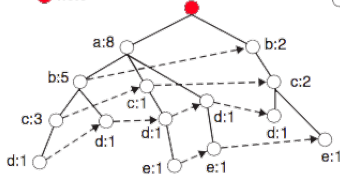
(i) TID=1



(ii) TID=2



(iii) TID=3



(iv) TID=10

The diagram consists of two parts, (a) and (b), each showing a tree structure for the string 'abcd'.

(a) **Árvore de Prefixo** (Prefix Tree): The root node is an empty circle. It has four children: 'a', 'b', 'c', and 'd'. Node 'a' has three children: 'ab', 'ac', and 'ad'. Node 'b' has two children: 'bc' and 'bd'. Node 'c' has one child: 'cd'. Node 'ab' has three children: 'abc', 'abd', and 'acd'. Node 'bc' has one child: 'bcd'. Node 'abc' has one child: 'abcd'. Dashed lines group the nodes into three sets: {'a', 'ab', 'abc', 'abcd'}, {'b', 'bc', 'bcd'}, and {'c', 'cd'}. A dotted line encloses the entire tree structure.

(b) **Árvore de Sufixo** (Suffix Tree): The root node is an empty circle. It has four children: 'a', 'b', 'c', and 'd'. Node 'a' has one child: 'ab'. Node 'b' has two children: 'ac' and 'bc'. Node 'c' has three children: 'ad', 'bd', and 'cd'. Node 'd' has four children: 'abc', 'abd', 'acd', and 'bcd'. Node 'ab' has one child: 'abcd'. Node 'ac' has one child: 'acd'. Node 'bc' has one child: 'bcd'. Node 'ad' has one child: 'abd'. Node 'bd' has one child: 'bcd'. Node 'cd' has one child: 'bcd'. Dashed lines group the nodes into three sets: {'a', 'ab', 'abcd'}, {'b', 'ac', 'bc', 'acd', 'bcd'}, and {'c', 'ad', 'bd', 'cd', 'bcd'}. A dotted line encloses the entire tree structure.

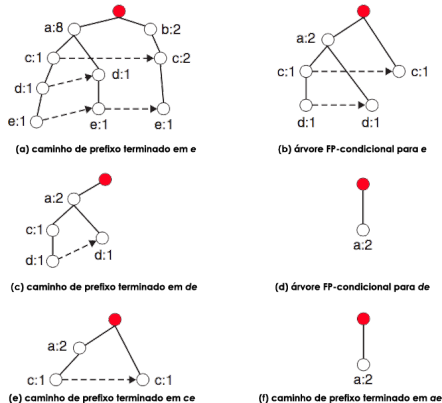
Sufixo	Conjuntos de itens frequentes
e	$\{e\}, \{d, e\}, \{a, d, e\}, \{c, e\}, \{a, e\}$
d	$\{d\}, \{c, d\}, \{b, c, d\}, \{a, c, d\}, \{b, d\}, \{a, b, d\}, \{a, d\}$
c	$\{c\}, \{b, c\}, \{a, b, c\}, \{a, c\}$
b	$\{b\}, \{a, b\}$
a	$\{a\}$

Tabela 3: Lista dos conjuntos de itens frequentes correspondentes aos seus sufixos.

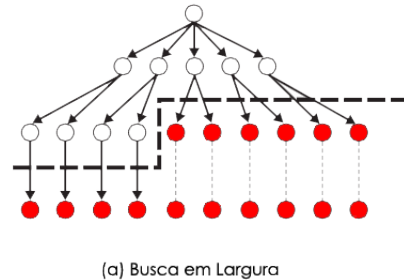
Exemplo

Suporte mínimo = 2

Figura 11: Aplicação do algoritmo *FP-growth* para encontrar os conjuntos de itens frequentes terminados em e.



(a) Busca em Profundidade



Geração de regras

Outra tarefa inerente da mineração de dados é a extração de regras eficientes a partir dos conjuntos de itens frequentes encontrados⁶.

Cada conjunto Y de k itens frequente pode produzir até $2^k - 2$ regras de associação.

⁶Roberto J. Bayardo Jr./Rakesh Agrawal: Mining the Most Interesting Rules, em: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '99), San Diego, California, USA 1999, pp. 145–154.

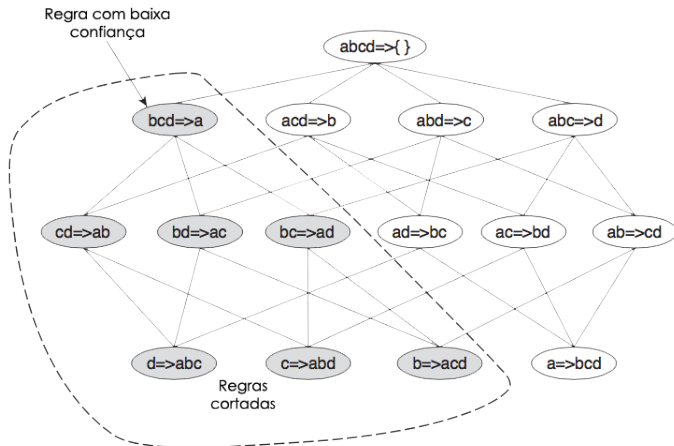
Corte baseado na confiança

Considere as seguintes regras: $X' \rightarrow Y \setminus X'$ e $X \rightarrow Y \setminus X$, onde $X' \subset X$.

A confiança das regras é dada por $\sigma(Y)/\sigma(X')$ e $\sigma(Y)/\sigma(X)$, respectivamente. Como X' é um subconjunto de X , $\sigma(X') \geq \sigma(X)$.

Portanto, a primeira regra não pode ter uma confiança maior que a última regra.

Figura 13: Corte de regras de associação baseado na métrica de confiança.



Base de dados real

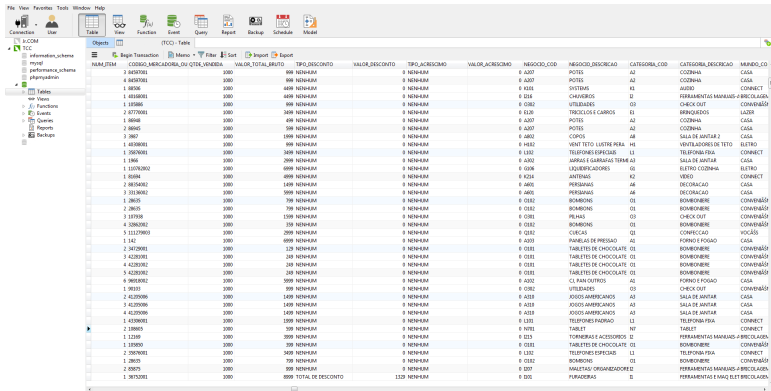
A metodologia para a criação da nossa base de dados real origina a partir de tuplas armazenadas no banco de dados da empresa utilizada.

A tabela principal do banco de dados possui diversos atributos que registram características e informações sobre a venda de seus produtos. Não obstante, alguns desses atributos por não fornecerem a informação central que precisamos, isto é, o **código identificador do produto** e a **transação** a qual pertence, poderão ser descartados.

Atributo	Exemplo
ID	7575
LOJA	C279
DATA_VENDA	2013-04-03
HORA	14
PERIODO	TARDE
NUM_ITEM	2
CODIGO_MERCADORIA	42281001
QNT_VENDIDA	1000
VALOR_TOTAL_BRUTO	249
TIPO_DESCONTO	NENHUM
VALOR_DESCONTO	0
NEGOCIO_COD	O101
NEGOCIO_DESCRICAO	TABLETES DE CHOCOLATE
CATEGORIA_CODIGO	O1
CATEGORIA_DESCRICAO	BOMBONIERE
MUNDO_COD	CONVENIENCIA
MUNDO_DESCRICAO	CONVENIENCIA
DESCRICAO_MERCADORIA	Barra Choc Baton Garoto 76g, Br/Cook

Tabela 4: Exemplo de uma tupla da base de dados original.

Figura 14: Ambiente da base de dados proporcionado pela aplicação Navicat MySQL.



NAME_ITEM	CODIGO_MERCADORIA_ORI_QTDE_VENDIDA	VALOR_TOTAL_LIVRO	TIPO_DESCONTO	VALOR_DESCONTO	TIPO_ACRESCIMO	VALOR_ACRESCIMO	NEGOCIO_COD	NEGOCIO_DESCRICAO	CATEGORIA_COD	CATEGORIA_DESCRICAO	MUNDO_COD
2 84537001	3000	898	NENHUMA	0	NENHUMA	0	A207	POTES	A2	COZINHA	CASA
4 84537001	3000	988	NENHUMA	0	NENHUMA	0	A207	POTES	A2	COZINHA	CASA
1 845356	3000	4898	NENHUMA	0	NENHUMA	0	0204	DISCOS	K1	AUDIO	CONNECT
1 48188001	3000	4898	NENHUMA	0	NENHUMA	0	0204	CHAVEIROS	01	FERRAMENTAS MANUAIS E BRICOLAGEM	CONNECT
1 187586	3000	998	NENHUMA	0	NENHUMA	0	0202	UTILIDADES	03	CHECK OUT	CONVENIÃO
2 87770001	3000	3498	NENHUMA	0	NENHUMA	0	0120	TROCISCO E CARRIS	K1	BRINQUEDOS	LADIN
1 846648	3000	488	NENHUMA	0	NENHUMA	0	A207	POTES	A2	COZINHA	CASA
2 86945	3000	598	NENHUMA	0	NENHUMA	0	A207	POTES	A2	COZINHA	CASA
3 2867	3000	1598	NENHUMA	0	NENHUMA	0	A400	COPOS	A4	SALA DE MANUTENÇÃO	CASA
1 42188001	3000	998	NENHUMA	0	NENHUMA	0	0402	ANT TETO LUSTRE PENA	H1	ILUMINACAO DE TETO	ELETRON
1 25067001	3000	3498	NENHUMA	0	NENHUMA	0	0132	TELEFONES ESPECIAIS	L1	TELEFONIA PRA	CONNECT
1 1866	3000	2099	NENHUMA	0	NENHUMA	0	A202	JARRAS E GUARAFAS TERMAS	A3	SALA DE MANUTENÇÃO	CASA
1 165782902	3000	6899	NENHUMA	0	NENHUMA	0	0206	LICENCIADORES	06	ELETRON	CONNECT
1 6494	3000	4898	NENHUMA	0	NENHUMA	0	K214	ANTENAS	K2	VERSO	CONNECT
2 88154002	3000	1498	NENHUMA	0	NENHUMA	0	A400	PERIGANAS	A4	DECORACAO	CASA
3 33186002	3000	5998	NENHUMA	0	NENHUMA	0	A400	PERIGANAS	A4	DECORACAO	CASA
1 28625	3000	798	NENHUMA	0	NENHUMA	0	0202	BOMBONES	01	BOMBONIERE	CONVENIÃO
2 28635	3000	798	NENHUMA	0	NENHUMA	0	0202	BOMBONES	01	BOMBONIERE	CONVENIÃO
3 187838	3000	1398	NENHUMA	0	NENHUMA	0	0201	PELHAS	03	CHECK OUT	CONVENIÃO
4 22863502	3000	2099	NENHUMA	0	NENHUMA	0	0202	BOMBONES	01	BOMBONIERE	CONVENIÃO
5 11278603	3000	2099	NENHUMA	0	NENHUMA	0	0202	CUBCAS	01	CONEXCAO	VOCASO
1 142	3000	6899	NENHUMA	0	NENHUMA	0	A203	PANELAS DE PRESSAO	A5	FORNO E FOGAO	CASA
3 24729001	3000	125	NENHUMA	0	NENHUMA	0	0201	TABULETOS DE CHOCOLATE	01	BOMBONIERE	CONVENIÃO
3 42282001	3000	248	NENHUMA	0	NENHUMA	0	0201	TABULETOS DE CHOCOLATE	01	BOMBONIERE	CONVENIÃO
4 42282002	3000	248	NENHUMA	0	NENHUMA	0	0201	TABULETOS DE CHOCOLATE	01	BOMBONIERE	CONVENIÃO
5 42282002	3000	248	NENHUMA	0	NENHUMA	0	0201	TABULETOS DE CHOCOLATE	01	BOMBONIERE	CONVENIÃO
9 98058002	3000	5998	NENHUMA	0	NENHUMA	0	A202	CL. PARA OUTROS	A5	FORNO E FOGAO	CASA
1 981031	3000	998	NENHUMA	0	NENHUMA	0	0202	UTILIDADES	03	CHECK OUT	CONVENIÃO
2 42250906	3000	1498	NENHUMA	0	NENHUMA	0	A203	JOSOS AMERICANOS	A3	SALA DE MANUTENÇÃO	CASA
3 42250906	3000	1498	NENHUMA	0	NENHUMA	0	A203	JOSOS AMERICANOS	A3	SALA DE MANUTENÇÃO	CASA
4 42250906	3000	1498	NENHUMA	0	NENHUMA	0	A203	JOSOS AMERICANOS	A3	SALA DE MANUTENÇÃO	CASA
1 43308001	3000	1398	NENHUMA	0	NENHUMA	0	0132	TELEFONES PADRAO	L1	TELEFONIA PRA	CONNECT
2 25067001	3000	3498	NENHUMA	0	NENHUMA	0	0132	TELEFONES ESPECIAIS	L1	TELEFONIA PRA	CONNECT
1 12139	3000	998	NENHUMA	0	NENHUMA	0	0203	TORNABRASIL ACESSORIOS	02	FERRAMENTAS MANUAIS E BRICOLAGEM	CONNECT
1 133850	3000	398	NENHUMA	0	NENHUMA	0	0201	TABULETOS DE CHOCOLATE	01	BOMBONIERE	CONVENIÃO
2 25067001	3000	3498	NENHUMA	0	NENHUMA	0	0132	TELEFONES ESPECIAIS	L1	TELEFONIA PRA	CONNECT
1 28625	3000	798	NENHUMA	0	NENHUMA	0	0202	BOMBONES	01	BOMBONIERE	CONVENIÃO
2 250675	3000	998	NENHUMA	0	NENHUMA	0	0207	MALETAS/ ORGANIZADOR DE	01	FERRAMENTAS MANUAIS E BRICOLAGEM	CONNECT
1 38132001	3000	8998	TOTAL DE DESCONTO	1329	NENHUMA	0	0202	PURIFICADOR	02	FERRAMENTAS MANUAIS E BRICOLAGEM	CONNECT

- A base original contém 2, 194, 306 registros de vendas de itens. Dentre todos esses registros, existem 14, 647 diferentes itens que estão mapeados em 7 diferentes categorias;
- Contudo, como os registros são individuais para cada item, é possível que o número total de transações seja menor.

A transação conterá um campo identificador chamado **TID**, seguida pelos **códigos dos itens** que estão presentes nela (i.e. venda de uma barra de chocolate e um saco de amendoim).

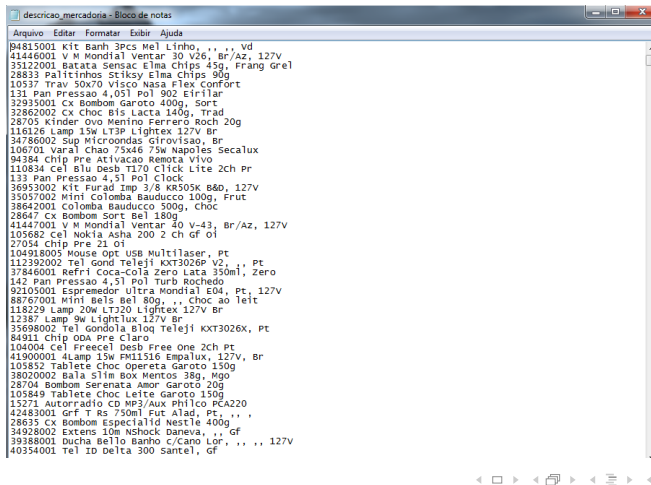
TID	Conjunto de Itens
1	{4228100, 18776002}

Tabela 5: Exemplo de uma transação da base de dados remodelada.

Figura 15: Arquivo da base de dados já padronizado no formato de entrada.

```
base_real.tab 3
53 28635 35057002 34730002 105852
54 37680007
55 37200001
56 107656 87801005 87801005 42203002 108195 83322005 83322005 41897001 39415004
57 135282
58 28934 28934
59 133293
60 135191
61 117845 117845
62 36068002
63 109265
64 82365 42166003 42166003 28704 87068001
65 135063 135063 135063 28736 123157 123157 123157
66 133650 135346
67 94438
68 133774 133774
69 12253
70 135811 132371 35031004
71 91027002 28844
72 37200001
73 105250
74 118765
75 39418002
76 134410
77 135172 91163
78 41162001
79 113371
80 36068002
81 112392002 28705
82 130968
83 8061 2222
84 136156
85 33723001
86 133859 133049
87 11690
88 2642 136210 35122002 98890
89 132135 121919 101693
90 10537 136045
91 131980
92 122695001 135359
93 133650 135054
```

Figura 16: Arquivo texto referente à descrição dos códigos das mercadorias.



Base de dados sintética

Nosso caso de uso baseia-se em uma empresa de telecomunicações que efetua a venda de pacotes para canais de TV e, recentemente, vem sofrendo com um declínio em sua quantidade de clientes.

Duas pesquisas foram realizadas: uma com 1,000 indivíduos a fim de saber quais são seus 50 canais de TV mais assistidos dentro da grade de programação da empresa (**base de dados sintética 1k**), e outra com 10,000 indivíduos contendo os seus 60 canais mais assistidos (**base de dados sintética 10k**).

Atributo	Exemplo
ID	64
NOME	Larissa Lopes
SEXO	F
IDADE	27
PROFISSAO	MUSICO
QNT_FILHOS	0

Tabela 6: Exemplo de uma tupla da tabela 'usuarios' da base sintética.

Atributo	Exemplo
ID	109
NOME	HBO Plus
GENERO	FILME
PACOTE	4

Tabela 7: Exemplo de uma tupla da tabela 'canais' da base sintética.

Atributo	Exemplo
ID	1
USUARIO_ID	4
CANAL_ID	101

Tabela 8: Exemplo de uma tupla da tabela 'canais_usuarios' da base sintética.

A transação final conterá um campo identificador chamado **TID**, seguida pelos **códigos dos canais** que estão presentes nela. Por exemplo, uma transação de um indivíduo que assista mais aos canais BIS (código 31), Fox Life (código 27), Futura (código 54) e TNT (código 96).

TID	Conjunto de Canais
1	{31, 27, 54, 96}

Tabela 9: Exemplo de uma transação da base de dados sintética remodelada.

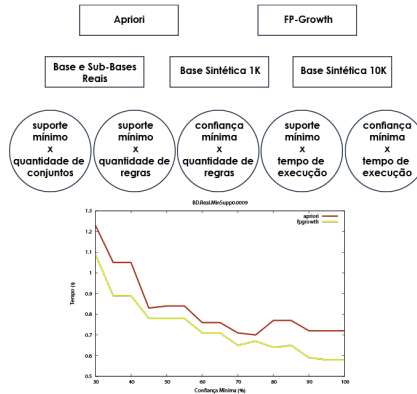
A distribuição dos dados foi feita de forma aleatória a partir de um algoritmo de distribuição probabilística desenvolvido para o próprio projeto.

Pacote	Quantidade
1	50%
2	25%
3	12.5%
4	3.125%
5	3.125%
6	3.125%
7	3.125%

Tabela 10: Distribuição da porcentagem de pacotes empregada na geração dos dados.

Experimentos

Figura 17: Detalhamento dos experimentos propostos.



Cada conjunto de experimentos é composto por análises de duas dimensões (**2D**) para uma posterior representação em gráficos **XY**.

Todos os dados foram armazenados em arquivos .dat e grafados utilizando a ferramenta **Gnuplot**⁷ do Linux.

⁷<http://www.gnuplot.info/>

Figura 18: Gráfico do suporte mínimo x quantidade de conjuntos para a base de dados real.

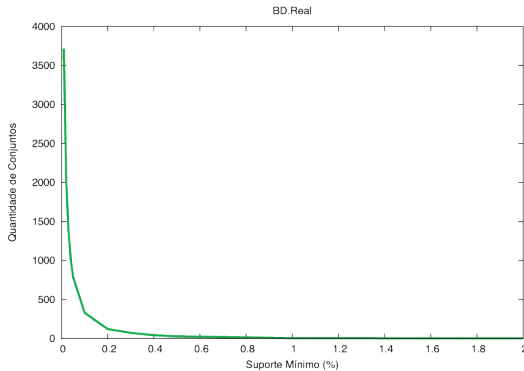




Figura 20: Gráfico da confiança mínima x quantidade de regras para a base de dados real (suporte mínimo 0.001%).

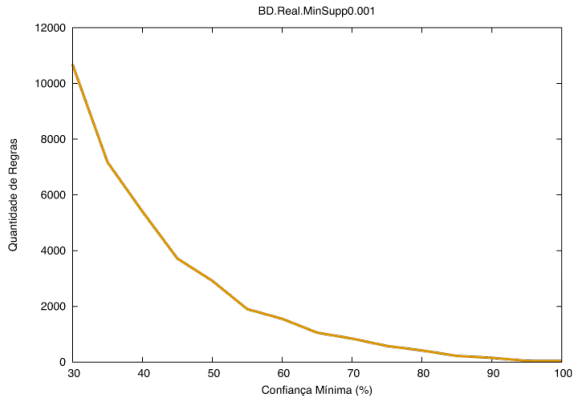
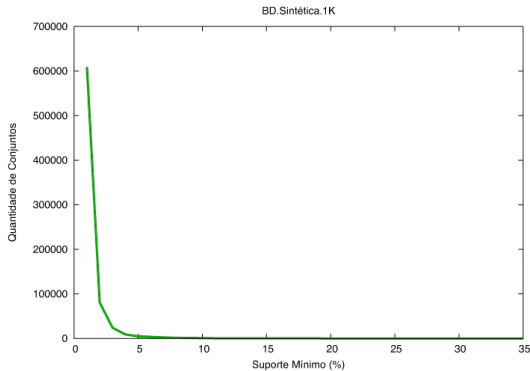


Figura 21: Gráfico do suporte mínimo x quantidade de conjuntos para a base de dados sintética 1K.













Tempos de execução

Figura 27: Gráfico do suporte mínimo x tempo de execução para a base de dados real (confiança mínima 80%).

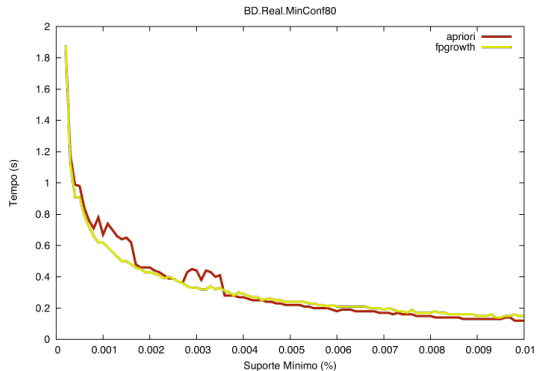


Figura 28: Gráfico da confiança mínima x tempo de execução para a base de dados real (suporte mínimo 0.0005%).

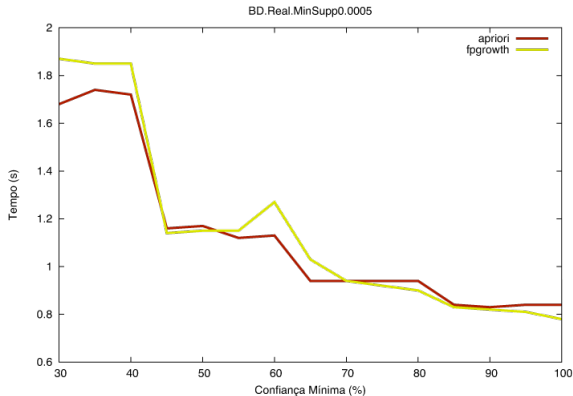


Figura 29: Gráfico do suporte mínimo x tempo de execução para a base de dados sintética 1K (confiança mínima 90%).

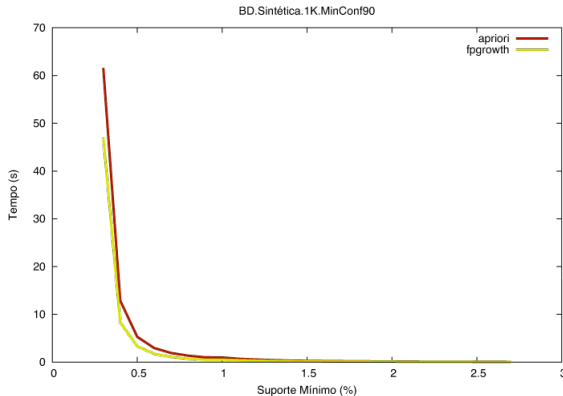


Figura 30: Gráfico da confiança mínima x tempo de execução para a base de dados sintética 1K (suporte mínimo 1%).

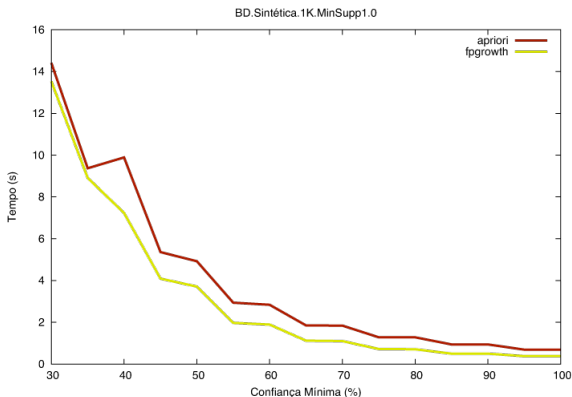


Figura 31: Gráfico do suporte mínimo x tempo de execução para a base de dados sintética 10K (confiança mínima 80%).

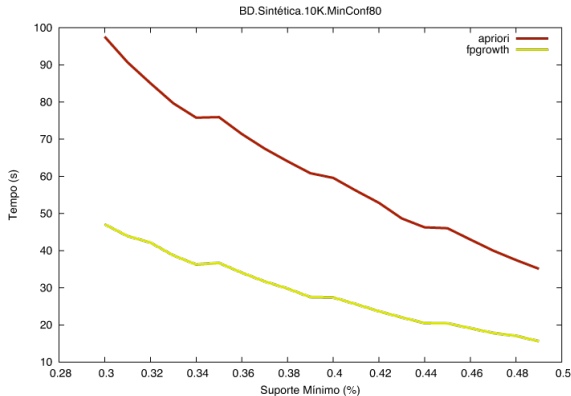
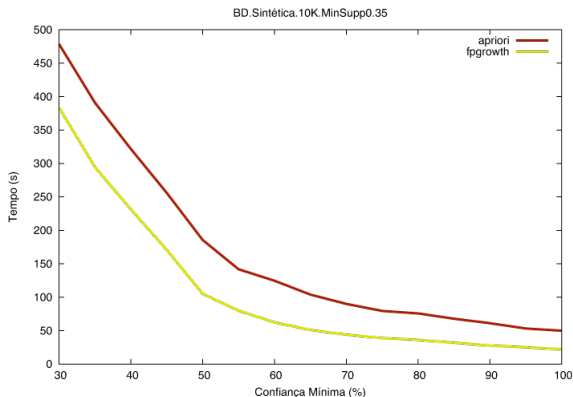


Figura 32: Gráfico da confiança mínima x tempo de execução para a base de dados sintética 10K (suporte mínimo 0.35%).



Regras geradas

Regra	Suporte (%)	Confiança (%)
TRosto Senegal Azul, TRosto Senegal Vermelha, TBanho Senegal Azul → TBanho Senegal Vermelha	0.0022	100
Samsung Chat 222 Pt Vivo → Chip Pre Vivo	0.0044	100
Alcatel OT255 2Ch Pt Vivo → Chip Pre Vivo	0.0238	97.85
Modem USB 3G Huawei Tim → Chip Pre Infinity 21 Tim	0.0308	92.72
Samsung E1182 Duos Basic Pr Vivo → Chip Pre Vivo	0.0295	92.39
Cel Alcatel Desb OT217 2Ch Pt → Chip ODA Pre Claro	0.1451	86.72
Cel Freecel Desb Free One 2Ch Pt → Chip ODA Pre Claro	0.1718	81.25

Tabela 11: Regras com maiores valores de suporte e confiança para a base de dados real.

Regra	Suporte (%)	Confiança (%)
Nick Jr. (2), Disney Channel (2), Ideal TV (1) → FOX (2)	1.1	100
Sony (2), Curta! (1), Canal Rural (1), Ideal TV (1) → NBR (1)	1.1	100
Glitz* (2), TV Brasil (1), TV Justiça (1) → Off (1)	1.9	94.74
FX (2), Rede Vida (1), TV Câmara (1) → NBR (1)	2.1	90.48
Space (2), GNT (1), SHOPBUY (1) → Multishow (1)	2.2	81.82
Warner Channel (2), Off (1), TV Justiça (1) → MTV (2)	2.3	82.61

Tabela 12: Regras com maiores valores de suporte e confiança para a base de dados sintética 1K.

Regra	Suporte (%)	Confiança (%)
TV Aparecida (1), Curta! (1), Band (1), Rede do Bem (1), Rede Mundial (1), Shoptime (1) → Canal Universitário (1)	0.3	90
Play TV (1), Mix TV (1), TV Brasil (1), RIT TV (1), TV Senado (1), Shoptime (1) → Canal Legislativo (1)	0.3	90
Rede 21 (1), Rede do Bem (1), TV Justiça (1), Record (1), RIT TV (1), Canal Legislativo (1) → TV Aberta (1)	0.33	81.82
Viva (1), NBR (1), Rede 21 (1), RDBTV (1), Record (1), Ideal TV(1) → TV Escola (1)	0.38	86.84
Discovery Turbo (3), Off (1), Band News (1), Rede do Bem (1), Leomax Shop (1) → TV Câmara (1)	0.41	85.37
Viva (1), TV Canção Nova (1), CNT (1), Multishow (1), Canal Universitário (1), Leomax Shop (1) → Globo (1)	0.45	80

Tabela 13: Regras com maiores valores de suporte e confiança para a base de dados sintética 10K.

Obrigado pela atenção!

Perguntas ou dúvidas?

