Análise da Propagação de Atrasos de Voos Brasileiros

Lara Mello e Luana Fragoso

Orientadores: Eduardo Ogasawara e Alice Stenberg

Agenda

- Motivação
- Cenário
- Pré Processamento
- Janela Deslizante
- Padrões Frequentes
- Processo de Mineração
- Resultados
- Considerações finais

Motivação

- Empresas e Aeroportos
  - Minimizar prejuízos
  - Minimizar impacto ambiental
  - Tomadas de decisão

- Passageiros
  - Replanejamento
  - Confiança
  - Satisfação

Cenário

- Em 2013
  - Europa - 36% dos vôos atrasados por mais de 5 min
  - EUA - 31% dos vôos atrasados por mais de 15 min
  - Brasil - 16.3% dos vôos atrasados por mais 30 min
- Causas naturais ou...
  - Compartilhamento de recursos
  - Aeroporto cheio
  - Horário dos funcionários

Cenário

- Propagação de atrasos
  - Assumir um atraso na rede
  - Procurar regras que indicam uma influência nos demais aeroportos

Cenário

- Atrasos em um aeroporto provocam atrasos em outros aeroportos da rede?
- Em quanto tempo os efeitos da propagação aparecem nos diferentes aeroportos da rede?

Pré Processamento

- Remoção de outliers
  - atributos com valores incoerentes
  - atrasos com motivos raros
    - nevasca ou acidente aéreo

- Transformação dos dados
  - Conceito de Hierarquia

Janela Deslizante

- Modelo importante para Mineração
  - tamanho fixo

- Se tamanho for maior que ideal
  - classifico como não-frequente itemsets que só estão frequentes mais recentemente

- Se menor
  - limito minhas transações

Padrões Frequentes

- Relacionamentos recorrentes na base de dados

- Regras de associação
  - suporte
  - confiança
  - lift

Padrões Frequentes

(i) A → B [support, confidence, lift]

(ii) support A → B = P(A U B)

(iii) confidence A→ B = P(B||A) = $\frac{support(AUB)}{support(A)}$

(iv) lift (A, B) = $\frac{P(AUB)}{P(A)P(B)}$

- probabilidade de ocorrência
- probabilidade condicional
- indica a correlação

Apriori

- Algoritmo pra identificar os k-itemsets frequentes
  - k indica tamanho dos grupos

- Percorre a base toda apenas uma vez
  - k-1 usado para definir k-2
  - itemsets candidatos

Regras de Associação

- Resultado do algoritmo Apriori

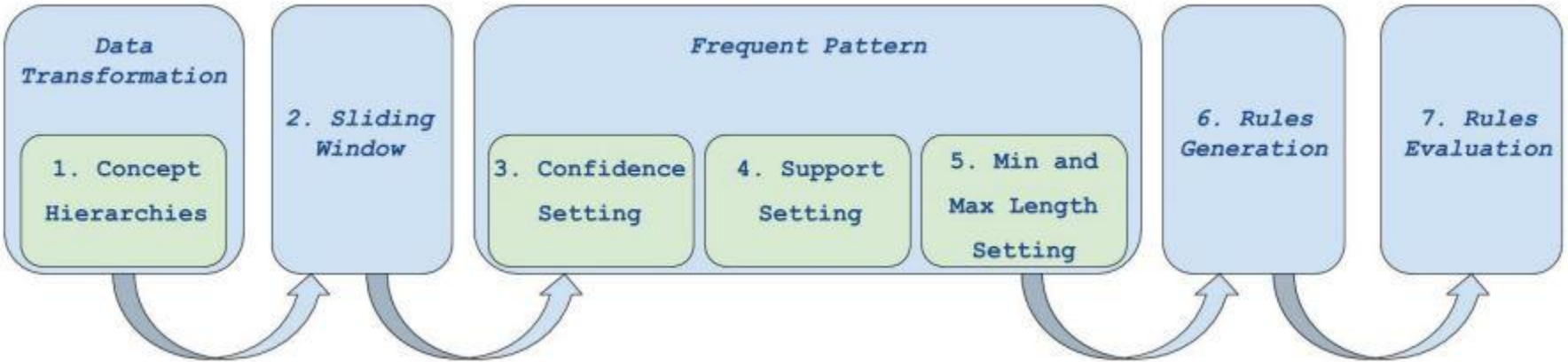
- Ocorrência de um itemset implica na ocorrência de outro
  - correlação
  - A -> B

Processo de Mineração - Base de Dados

- ANAC (Agência Nacional de Aviação Civil)
  - VRA (Base Voo Regular Ativo)
  - 17 aeroportos
- Tabela utilizada: ‘voos’
  - Atributos importantes: hora de partida/chegada prevista, hora de partida/chegada real, aeroporto de chegada, aeroporto de partida.

ANAC ABBREVIATION	COMMERCIAL ABBREVIATION	AIRPORT NAME	CITY - STATE
SBPA	POA	SALGADO FILHO	PORTO ALEGRE - RS
SBFL	FLN	HERCILIO LUZ	FLORIANOPOLIS - SC
SBCT	CWB	AFONSO PENA	CURITIBA - PR
SBGR	GRU	GUARULHOS	SAO PAULO - SP
SBSP	CGH	CONGONHAS	SAO PAULO - SP
SBKP	VCP	VIRACOPOS	SAO PAULO - SP
SBGL	GIG	GALEAO	RIO DE JANEIRO - RJ
SBRJ	SDU	SANTOS DUMONT	RIO DE JANEIRO - RJ
SBCF	CNF	TANCREDO NEVES	BELO HORIZONTE - MG
SBVT	VIX	EURICO DE AGUIAR SALLES	VITORIA - ES
SBGO	GYN	SANTA GENOVEVA	GOIANIA - GO
SBBR	BSB	PRESIDENTE JUSCELINO KUBITSCHEK	BRASILIA - DF
SBEG	MAO	EDUARDO GOMES	MANAUS - AM
SBBE	BEL	VAL DE CANS	BELEM - PA
SBSV	SSA	DEPUTADO LUIS EDUARDO MAGALHAES	SALVADOR - BA
SBRF	REC	GUARARAPES	RECIFE - PE
SBFZ	FOR	PINTO MARTINS	FORTALEZA - CE

Processo de Mineração



2. Janela Deslizante

- Consulta SQL
- Utilizada faixa horária entre 06:00 e 20:00 hrs
- Para cada hora (cada aeroporto) - 4 horas seguintes
  - atributo ‘hora’
- Somente considerados atrasos na partida

2. Janela Deslizante

- Quantidade de atrasos transformada em frequência
  - abaixo de 15% = low
  - entre 15% e 33% = medium
  - acima de 33% = high

	SBPA_0	SBPA_1	SBPA_2	SBPA_3	SBPA_4	SBFL_0	SBFL_1	SBFL_2	SBFL_3	SBFL_4
1	high	high	high	high	low	high	high	high	low	high
2	high	high	high	low	medium	high	high	low	high	high
3	high	high	low	medium	medium	high	low	high	high	high
4	high	low	medium	medium	high	low	high	high	high	low
5	low	medium	medium	high	medium	high	high	high	low	low
6	medium	medium	high	medium	low	high	high	low	low	low
7	medium	high	medium	low	low	high	low	low	low	low
8	high	medium	low	low	high	low	low	low	low	low
9	medium	low	low	high	low	low	low	low	low	low
10	low	low	high	low	low	low	low	low	low	low
11	low	high	low	low	medium	low	low	low	medium	low
12	high	low	low	medium	medium	low	low	medium	low	high
13	low	low	medium	medium	medium	low	medium	low	high	high
14	low	medium	medium	medium	high	medium	low	high	high	high
15	medium	medium	medium	high	high	low	low	low	low	low
16	low	medium	high	low	high	high	low	low	high	medium
17	medium	high	low	high	high	low	low	high	medium	low
18	high	low	high	high	high	low	high	medium	low	high
19	low	high	high	high	high	high	medium	low	high	high
20	high	high	high	high	high	medium	low	high	high	low

3. Ajuste da confiança

- Porcentagem média de atrasos ‘high’ do tempo 0
  - probabilidade condicional de cada aeroporto
  - P(high | delay)
- Aproximadamente 40%

AIRPORT	CONDITIONAL PROBABILITY
SBPA	43,26%
SBFL	30,05%
SBCT	43,92%
SBGR	40,51%
SBSP	44,81%
SBKP	38,40%
SBGL	41,74%
SBRJ	40,90%
SBCF	43,90%
SBVT	37,52%
SBGO	31,54%
SBBR	41,32%
SBEG	18,28%
SBBE	8,57%
SBSV	47,63%
SBRF	41,07%
SBFZ	37,88%
Average	**37,14%**

4. Ajuste do suporte

- Porcentagem entre a quantidade mínima de rotas em um dia e o número total de dias

Flight Route Per Day	Flight Route in 6 years	Percentage of Total Routes	
1	2.190	6,68%	
2	4.380	13,35%	
3	6.570	20,03%	
4	8.760	26,70%	TOTAL OF DAYS IN DATASET 32803
5	10.950	33,38%	
6	13.140	40,06%	
7	15.330	46,73%	
8	17.520	53,41%	
9	19.710	60,09%	
10	21.900	66,76%	

5. Ajuste Máximo e Mínimo

- Quantidade que gera regras interessantes
  - mínimo = 2
  - pelo menos um atributo no lado esquerdo
  - máximo = 2

- SBSP_0(high) -> SBRJ_2(high)

6. Geração das regras

- Biblioteca ‘Arules’ do R

- Aplicação do Apriori
  - para cada aeroporto separadamente

7. Avaliação das Regras

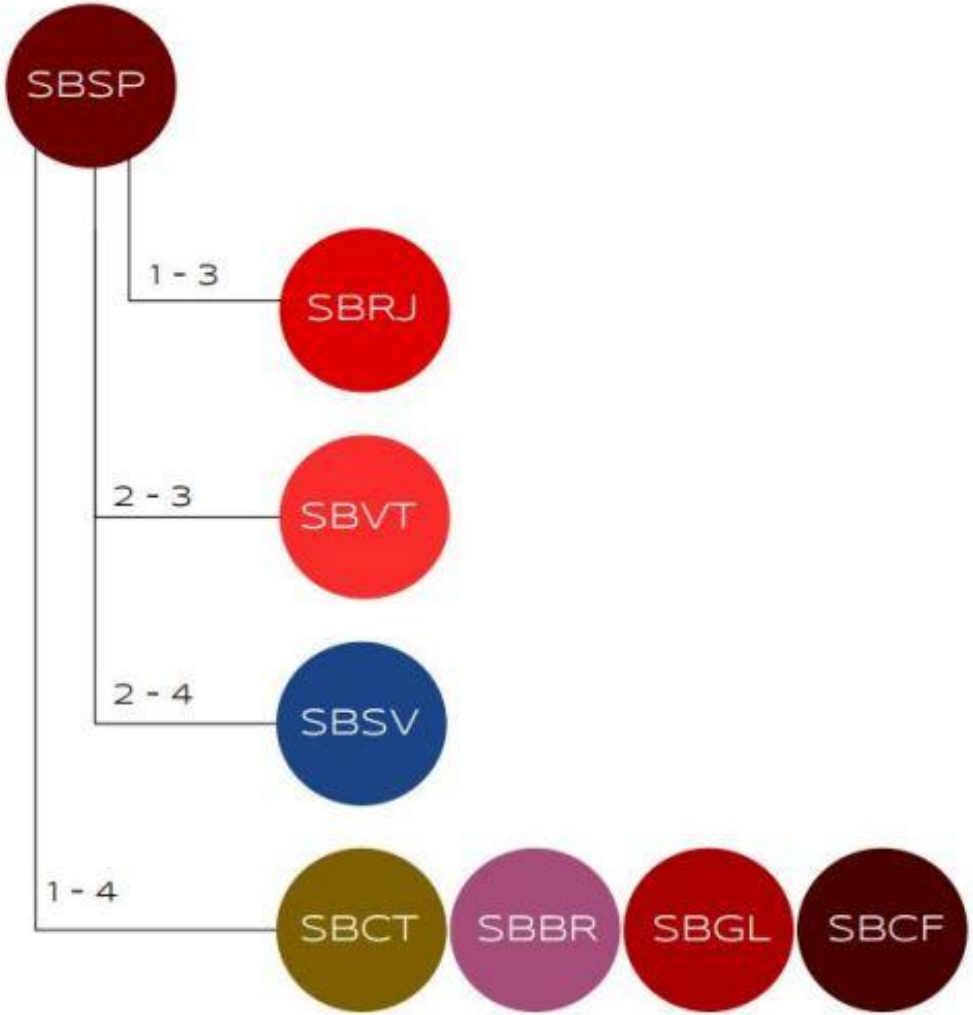
- Regras com lift > 1

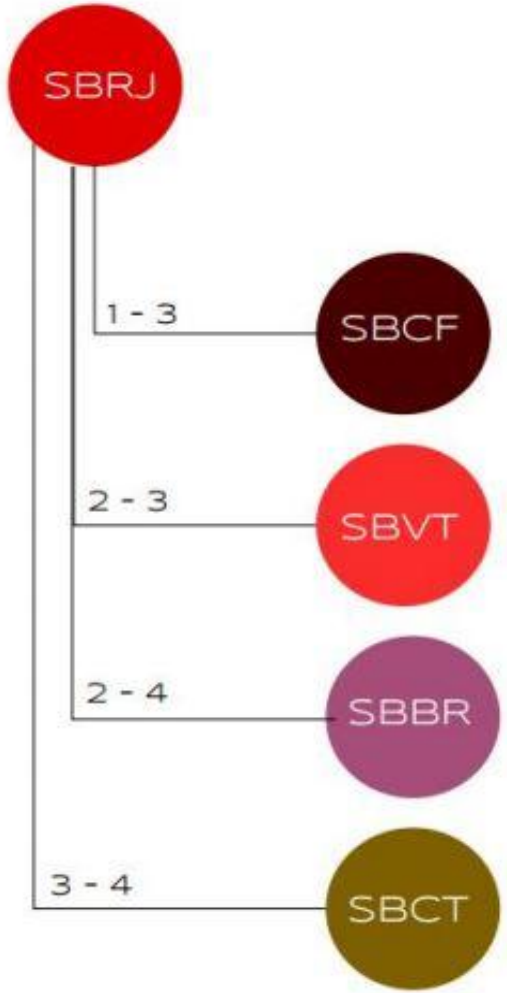
- Restrições do lado esquerdo da regra:
  - somente tempo 0
  - somente as frequências ‘medium’ e ‘high’

Resultados

- SBSP é o aeroporto mais influente
  - influenciou atrasos nas regiões norte, nordeste, sul e sudeste do país
  - mantém sua influência até 4 hrs depois

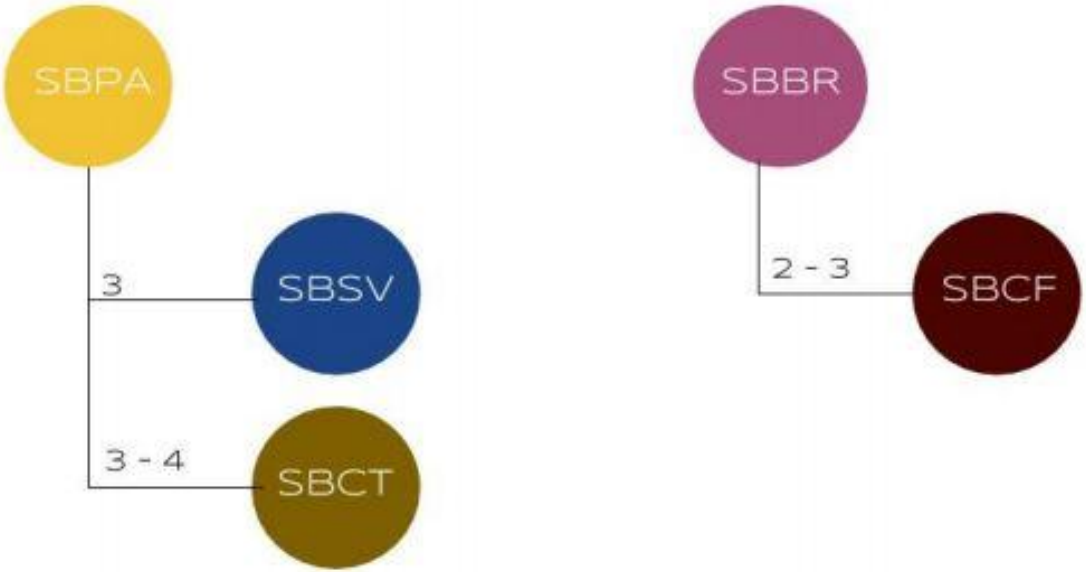
- Comparação dos lifts
  - lift da auto influência < lift do aeroporto influenciador


SBSP
1 - 3
SBRJ
2 - 3
SBVT
2 - 4
SBSV
1 - 4
SBCT
SBBR
SBGL
SBCF



- Segundo aeroporto mais influente

- Grandes Aeroportos porém com pouca influência



Considerações Finais

- Leitura do cenário aéreo brasileiro
  - conhecimento da relação entre os aeroportos
  - identificação dos aeroportos mais influentes
- Possibilidades para continuação da pesquisa
  - automatização dos processos
  - exploração dos parâmetros