

Mineração de padrões sequenciais em voos nacionais

Aluno: Philipp Matthews Rodrigues Mendonça
Orientador: Eduardo Ogasawara

Motivação e Resumo

- Base de dados de voos nacionais(VRA).
- Análise de impacto de atrasos em voos.
- Análise de sequências frequentes.
- Linguagem R e pacote ArulesSequence.

Análise de sequências frequentes

- Busca de sequências que se repetem em diversas transações.
- Geração de regras para as sequências.
- Arranjo de “carrinho de compras”:
 - EventID: Identificador de um evento, evento esse que pode se repetir muitas vezes no tempo.
 - SequenceID: Identificador temporal de quando um evento se repetiu.
 - ItemList: Lista dos itens correspondentes a aquele evento.

Análise de sequências frequentes

SBEG2009/01/01	05	3	1783	1705	1697
SBEG2009/01/01	16	3	3305	3891	1641
SBEG2009/01/02	05	3	1633	1705	1783
SBEG2009/01/02	15	1	1867		
SBEG2009/01/02	17	2	3545	3749	
SBEG2009/01/03	05	2	1705	1783	
SBEG2009/01/03	15	1	1867		
SBEG2009/01/03	16	2	3305	1641	
SBEG2009/01/03	20	1	1938		
SBEG2009/01/04	05	2	1705	1783	
SBEG2009/01/04	15	1	1867		
SBEG2009/01/04	16	1	3305		
SBEG2009/01/04	20	1	1938		
SBEG2009/01/05	05	2	1705	1783	
SBEG2009/01/05	15	1	1867		
SBEG2009/01/05	16	1	3891		
SBEG2009/01/05	17	1	3545		
SBEG2009/01/06	05	3	1705	1783	1697
SBEG2009/01/06	15	1	1867		
SBEG2009/01/06	16	1	3891		
SBEG2009/01/06	18	1	1631		
SBEG2009/01/06	20	1	1938		
SBEG2009/01/07	05	4	1633	1705	1783 1697
SBEG2009/01/07	15	1	1867		

Amostra de um arquivo gerado

Análise de sequências frequentes

- Suporte(X): Probabilidade de uma transação conter X.
- Confiança($X \rightarrow Y$): Probabilidade condicional de uma transação que contém X conter Y também.
- Lift(X, Y): Indica se X e Y tem correlação positiva(>1), negativa(<1) ou independente(=1).

Confiança($X \rightarrow Y$)

Suporte(Y)

Base de Dados

- Diversas informações dos voos nacionais de janeiro de 2009 a fevereiro de 2015:
 - Total de voos: 3067327
 - Voos atrasados: 676756
 - Transações geradas: 416878
 - Aeroportos: 17
- Adequação dos dados para o padrão de entrada da técnica de mineração e do algoritmo usado.

Base de Dados

```
CREATE OR REPLACE VIEW public.v_voos_atrasados AS
SELECT v.dataaero,
       v.horavoo,
       count(v.numerovoo) AS num,
       string_agg(v.numerovoo::text, ' '::text) AS codigos
FROM ( SELECT concat(v1.aeroporto partida, "substring"(to_char(v1.partidaprevista, 'YYYY/MM/DD'::text), 1, 10)) AS dataaero,
                  "substring"(to_char(v1.partidaprevista, 'HH24:MI:SS'::text), 1, 2) AS horavoo,
                  v1.numerovoo,
                  v1.aeroporto partida AS aeroporto,
                  v1.tempoatraso partida AS atraso
        FROM voo v1
       UNION ALL
       SELECT concat(v2.aeroporto chegada, "substring"(to_char(v2.chegadaprevista, 'YYYY/MM/DD'::text), 1, 10)) AS dataaero,
                  "substring"(to_char(v2.chegadaprevista, 'HH24:MI:SS'::text), 1, 2) AS horavoo,
                  v2.numerovoo,
                  v2.aeroporto chegada AS aeroporto,
                  v2.tempoatraso chegada AS atraso
        FROM voo v2) v
WHERE v.aeroporto::text = 'SBGR'::text AND v.atraso > 14::double precision
GROUP BY v.dataaero, v.horavoo
ORDER BY v.dataaero, v.horavoo;
```

Código em R

- Input da base de dados e análise de suporte:

```
require(arules)
require(arulesSequences)
source('C:/Users/Philipp/Aeroportos/Preprocessamento.R')

#Receber Aeroportos atrasados
x <- read_baskets("C:/Users/Philipp/Aeroportos/sbgrAtrasados.txt",
                  info = c("sequenceID", "eventID", "SIZE"));
y <- as(x, "data.frame");

#Minerar para encontrar suportes maiores que 0.05
s1 <- cspade(x, parameter = list(support = 0.05),
             control = list(verbose = TRUE, tidLists = TRUE))
summary(s1)
t = as(s1, "data.frame")
```


Código em R

- Cálculo do valor ideal de suporte:

```
#Curva para valor ideal de suporte  
xp <- c(seq(0.05, 0.50, 0.01))
```

```
yp <- c(length(which(t$support >= 0.05)), length(which(t$support >= 0.06)), length(which(t$support >= 0.07)),  
length(which(t$support >= 0.08)), length(which(t$support >= 0.09)), length(which(t$support >= 0.10)),  
length(which(t$support >= 0.11)), length(which(t$support >= 0.12)), length(which(t$support >= 0.13)),  
length(which(t$support >= 0.14)), length(which(t$support >= 0.15)), length(which(t$support >= 0.16)),  
length(which(t$support >= 0.17)), length(which(t$support >= 0.18)), length(which(t$support >= 0.19)),  
length(which(t$support >= 0.20)), length(which(t$support >= 0.21)), length(which(t$support >= 0.22)),  
length(which(t$support >= 0.23)), length(which(t$support >= 0.24)), length(which(t$support >= 0.25)),  
length(which(t$support >= 0.26)), length(which(t$support >= 0.27)), length(which(t$support >= 0.28)),  
length(which(t$support >= 0.29)), length(which(t$support >= 0.30)), length(which(t$support >= 0.31)),  
length(which(t$support >= 0.32)), length(which(t$support >= 0.33)), length(which(t$support >= 0.34)),  
length(which(t$support >= 0.35)), length(which(t$support >= 0.36)), length(which(t$support >= 0.37)),  
length(which(t$support >= 0.38)), length(which(t$support >= 0.39)), length(which(t$support >= 0.40)),  
length(which(t$support >= 0.41)), length(which(t$support >= 0.42)), length(which(t$support >= 0.43)),  
length(which(t$support >= 0.44)), length(which(t$support >= 0.45)), length(which(t$support >= 0.46)),  
length(which(t$support >= 0.47)), length(which(t$support >= 0.48)), length(which(t$support >= 0.49)),  
length(which(t$support >= 0.50))  
)
```

```
xz = curvature.max(xp, yp)
```

Código em R





- Criação da regra de mineração

```
## Regra para mineração  
r2 <- ruleInduction(s1, confidence = 0, control = list(verbose = TRUE))  
summary(r2)  
dr2 <- as(r2, "data.frame")  
is.redundant(r2, measure = "lift")
```

Regras por aeroporto:

Aeroportos	Regras na faixa de suporte	Número total de regras	Número de transações
SBBE	0	35	7642
SBBR	6	2276	25050
SBCF	0	71	18647
SBCT	0	131	15041
SBEG	0	45	7186
SBFL	0	18	7918
SBFZ	0	24	11609
SBGL	5	1356	22036
SBGO	0	23	7399
SBGR	7	5036	30835
SBKP	0	27	14102
SBPA	0	235	15170
SBRF	0	59	13740
SBRJ	0	68	16415
SBSP	0	1623	23439
SBSV	4	579	19176
SBVT	0	1	8814

Regras SBGL na faixa de suporte:





	rule 	support 	confidence 	lift 
463	$\langle\{1793\}\rangle \Rightarrow \langle\{1825\}\rangle$	0.16903344	0.5040984	1.757902
669	$\langle\{1793\}\rangle \Rightarrow \langle\{1805\}\rangle$	0.16674301	0.4972678	1.839891
869	$\langle\{1793\}\rangle \Rightarrow \langle\{1735\}\rangle$	0.16262025	0.4849727	1.758630
782	$\langle\{1876\}\rangle \Rightarrow \langle\{1793\}\rangle$	0.15574897	0.5862069	1.748210
1217	$\langle\{1793\}\rangle \Rightarrow \langle\{1348\}\rangle$	0.15208429	0.4535519	1.864602

Regras SBGL na faixa de suporte:

	rule	support	confidence	lift
463	<{1793}> => <{1825}>	0.16903344	0.5040984	1.757902
669	<{1793}> => <{1805}>	0.16674301	0.4972678	1.839891
869	<{1793}> => <{1735}>	0.16262025	0.4849727	1.758630
782	<{1876}> => <{1793}>	0.15574897	0.5862069	1.748210
1217	<{1793}> => <{1348}>	0.15208429	0.4535519	1.864602

Todos os voos com mesma origem, todas as regras intra aeroporto.

Regras SBGR na faixa de suporte:

	rule 	support 	confidence 	lift 
1659	$\langle\{3504\}\rangle \Rightarrow \langle\{3507\}\rangle$	0.1717033	0.5364807	1.8867533
3854	$\langle\{1648\}\rangle \Rightarrow \langle\{1872\}\rangle$	0.1712454	0.7056604	1.7335908
4134	$\langle\{1872\}\rangle \Rightarrow \langle\{1814\}\rangle$	0.1611722	0.3959505	1.1993841
3862	$\langle\{1792\}\rangle \Rightarrow \langle\{1872\}\rangle$	0.1584249	0.7119342	1.7490036
3876	$\langle\{3504\}\rangle \Rightarrow \langle\{1872\}\rangle$	0.1565934	0.4892704	1.2019871
4699	$\langle\{1872\}\rangle \Rightarrow \langle\{1606\}\rangle$	0.1565934	0.3847019	1.6034141
3860	$\langle\{1743\}\rangle \Rightarrow \langle\{1872\}\rangle$	0.1501832	0.7522936	1.8481543

Regras SBGR na faixa de suporte:

	rule ↕	support ▼	confidence ↕	lift ↕
1659	<{3504}> => <{3507}>	0.1717033	0.5364807	1.8867533
3854	<{1648}> => <{1872}>	0.1712454	0.7056604	1.7335908
4134	<{1872}> => <{1814}>	0.1611722	0.3959505	1.1993841
3862	<{1792}> => <{1872}>	0.1584249	0.7119342	1.7490036
3876	<{3504}> => <{1872}>	0.1565934	0.4892704	1.2019871
4699	<{1872}> => <{1606}>	0.1565934	0.3847019	1.6034141
3860	<{1743}> => <{1872}>	0.1501832	0.7522936	1.8481543

Voo 1872 tem como Aeroporto de origem o SBGL, gerando as regras em vermelho como “cross” aeroporto.

Perguntas?



<https://www.flickr.com/photos/colins-airplane-photos/28996035283/in/photostream/>

Referências

- ANAC, 2015a. Agência Nacional de Aviação Civil. Technical Report. <http://www.anac.gov.br/>.
- ANAC, 2015b. Anuário Estatístico do Transporte 2014. Technical Report. <http://www2.anac.gov.br/estatistica/anuarios.asp>.
- DECEA, 2015. Departamento de Controle do Espaço Aéreo. Technical Report. <http://www.decea.gov.br/>.
- Han, J., Kamber, M., Pei, J., 2011. Data Mining: Concepts and Techniques, Third Edition. Morgan Kaufmann, Waltham, Mass.. 3 edition edition.
- Sternberg A., Carvalho D., Murtac L., Soares J., Ogasawara E., 2016, An Analysis of Brazilian Flight Delays Based on Frequent Patterns.