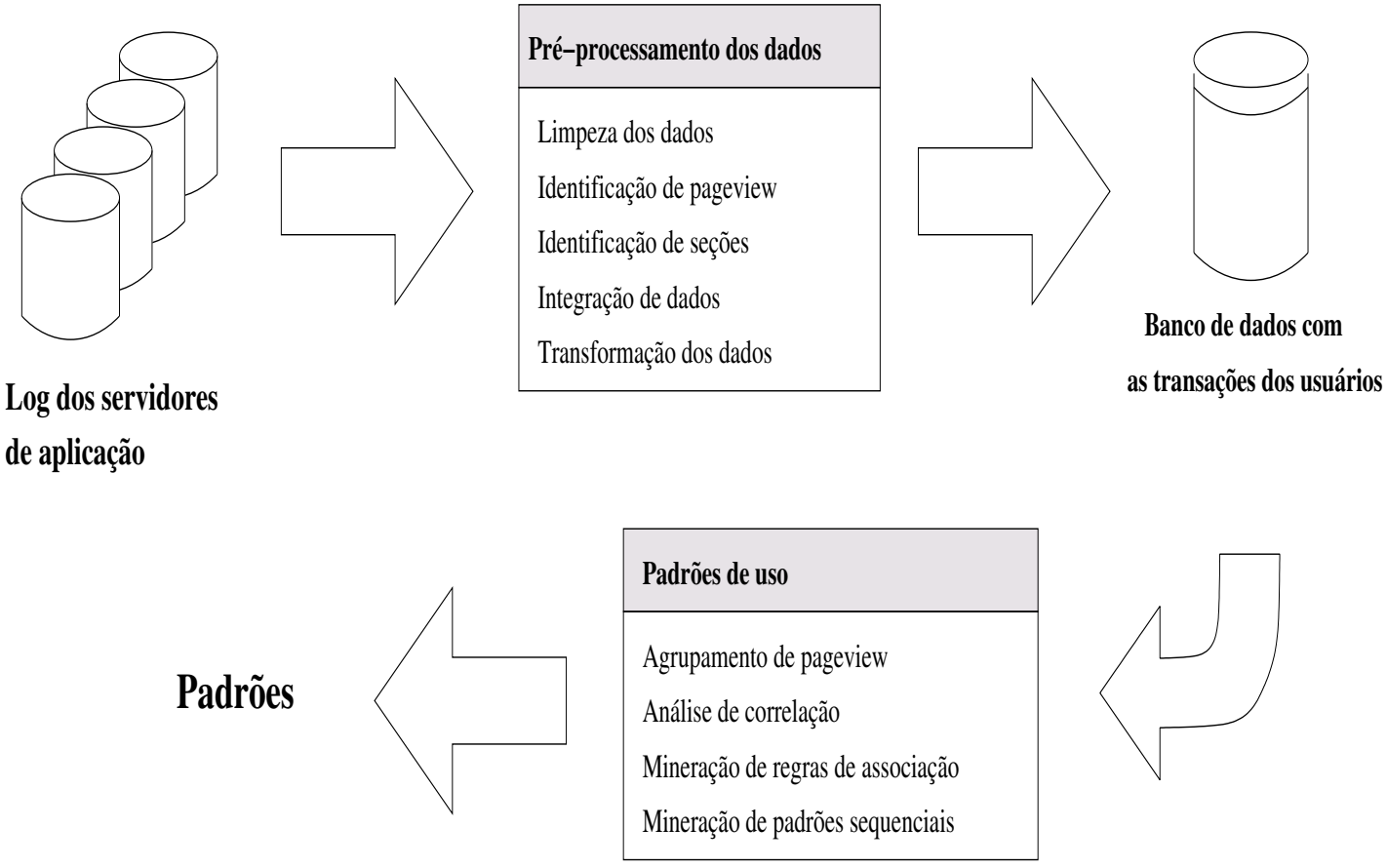
Criação de regras de associação a partir da navegação de usuários em sites Web

Fabrício J. Barth

Faculdades BandTec e VAGAS Tecnologia

Junho de 2013

Processo de mineração de padrões na Web



Exemplo típico de log

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

Pré-processamento do log: identificação de usuários

Time	IP	URL	Ref	Agent
0:01	1.2.3.4	A	-	IE5;Win2k
0:09	1.2.3.4	B	A	IE5;Win2k
0:10	2.3.4.5	C	-	IE6;WinXP;SP1
0:12	2.3.4.5	B	C	IE6;WinXP;SP1
0:15	2.3.4.5	E	C	IE6;WinXP;SP1
0:19	1.2.3.4	C	A	IE5;Win2k
0:22	2.3.4.5	D	B	IE6;WinXP;SP1
0:22	1.2.3.4	A	-	IE6;WinXP;SP2
0:25	1.2.3.4	E	C	IE5;Win2k
0:25	1.2.3.4	C	A	IE6;WinXP;SP2
0:33	1.2.3.4	B	C	IE6;WinXP;SP2
0:58	1.2.3.4	D	B	IE6;WinXP;SP2
1:10	1.2.3.4	E	D	IE6;WinXP;SP2
1:15	1.2.3.4	A	-	IE5;Win2k
1:16	1.2.3.4	C	A	IE5;Win2k
1:17	1.2.3.4	F	C	IE6;WinXP;SP2
1:26	1.2.3.4	F	C	IE5;Win2k
1:30	1.2.3.4	B	A	IE5;Win2k
1:36	1.2.3.4	D	B	IE5;Win2k

User 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

User 2

0:10	2.3.4.5	C	-
0:12	2.3.4.5	B	C
0:15	2.3.4.5	E	C
0:22	2.3.4.5	D	B

User 3

0:22	1.2.3.4	A	-
0:25	1.2.3.4	C	A
0:33	1.2.3.4	B	C
0:58	1.2.3.4	D	B
1:10	1.2.3.4	E	D
1:17	1.2.3.4	F	C

Pré-processamento do log: identificação das seções

User 1

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

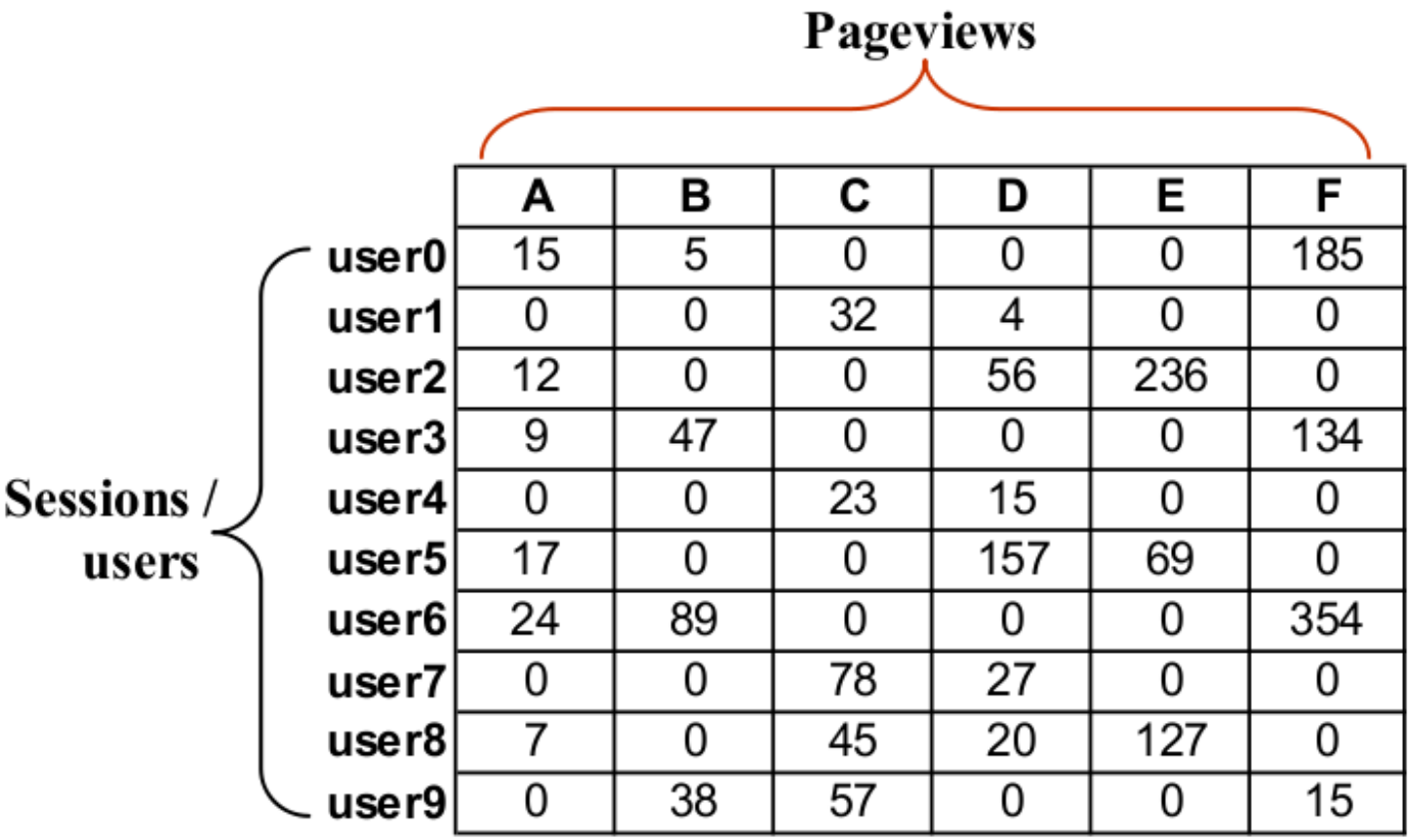
Session 1

0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C

Session 2

1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Matriz de transações



	A	B	C	D	E	F
user0	15	5	0	0	0	185
user1	0	0	32	4	0	0
user2	12	0	0	56	236	0
user3	9	47	0	0	0	134
user4	0	0	23	15	0	0
user5	17	0	0	157	69	0
user6	24	89	0	0	0	354
user7	0	0	78	27	0	0
user8	7	0	45	20	127	0
user9	0	38	57	0	0	15

Matriz de transações com meta-informações sobre as páginas

usuário	$categoria_1$	$categoria_2$	$categoria_3$	$\cdots$	$categoria_m$
$user_1$	0	2	0	$\cdots$	1
$user_2$	1	1	0	$\cdots$	0
$user_3$	2	0	1	$\cdots$	0
$user_4$	0	1	0	$\cdots$	0
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$user_n$	1	1	0	$\cdots$	1

- Cada página pode pertencer a uma categoria (i.e., tipo de livro, tipo de estabelecimento comercial)
- Cada página pode estar associada a uma cidade (i.e., um estabelecimento, uma vaga de emprego)

Regras de Associação

- **Caso do supermercado** (fralda $\rightarrow$ cerveja)
- Quem acessa a página sobre futebol também acessa a página de volei em **90%** dos casos (futebol $\rightarrow$ volei).
- Quem acessa a página de ofertas e a página de material de construção também finaliza a compra em **83%** dos casos (ofertas $\wedge$ material_construção $\rightarrow$ compra)

Algoritmo para criação de regras

Mineração de itens frequentes

- Dado:
  - ⋆ um conjunto $A = \{a_1, \cdots, a_m\}$ de itens,
  - ⋆ uma tabela $T = (t_1, \cdots, t_n)$ de transações sobre A,
  - ⋆ um número β_{min} que $0 < \beta_{min} \leq 1$, o **suporte mínimo**.
- Objetivo 1:
  - ⋆ encontrar o conjunto de **itens frequentes**, tais que o **suporte** de cada conjunto de itens é maior ou igual ao β_{min} definido pelo usuário.

Exemplo de transações

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

0 itens	1 item	2 itens	3 itens
{}: 10	{a}: 7	{a,c}: 4	{a,c,d}: 3
	{b}: 3	{a,d}: 5	{a,c,e}: 3
	{c}: 7	{a,e}: 6	{a,d,e}: 4
	{d}: 6	{b,c}: 3	
	{e}: 7	{c,d}: 4	
		{c,e}: 4	
		{d,e}: 4	

Figure 1: Um banco de dados de transações, com 10 transações, e a enumeração de todos os conjuntos de itens frequentes usando o suporte mínimo = 0,3

Mineração de itens frequentes

- Objetivo 2:
  - ⋆ encontrar o conjunto de regras de associação com confiança maior que um mínimo definido pelo utilizador.

Suporte e Confiança

O suporte de um conjunto de itens Z, $suporte(Z)$, representa a porcentagem de transações na base de dados que contêm os itens de Z.

O suporte de uma regra de associação $A \rightarrow B$, $suporte(A \rightarrow B)$, é dado por $suporte(A \cup B)$.

$$confianca(A \rightarrow B) = \frac{P(A \cup B)}{P(A)} = \frac{suporte(A \cup B)}{suporte(A)} \quad (1)$$

Exemplo de regras geradas

Premises	Conclusion	Support	Confidence
b	c	0.300	1
e, d	a	0.400	1
e	a	0.600	0.857
a	e	0.600	0.857
d	a	0.500	0.833
a, d	e	0.400	0.800

Figure 2: Regras extraídas com confiança maior que 0,8

Exemplo básico de uso

http://rpubs.com/fbarth/regraAssociacao

Medida Lift

Dada uma regra de associação $A \rightarrow B$, esta medida indica o quanto mais freqüente torna-se B quando ocorre A.

- Se $Lift(A \rightarrow B) = 1$, então A e B são independentes.
- Se $Lift(A \rightarrow B) > 1$, então A e B são positivamente independentes.
- Se $Lift(A \rightarrow B) < 1$, A e B são negativamente dependentes.

Esta medida varia entre 0 e ∞ e possui interpretação simples: **quanto maior o valor de $Lift$, mais interessante a regra, pois A aumenta B.**

Dados de click-stream de um site da Hungria

Dados anonimizados fornecidos por Ferenc Bodon -

http://fimi.ua.ac.be/data/kosarak.dat

http://rpubs.com/fbarth/regrasAssociacaoClickStream

Material de **consulta**

- Fabrício Barth. Mineração de regras de associação em servidores Web com RapidMiner[a].
- Iah H. Witteh and Eibe Frank. Data Mining: Practical Machine Learning Tools and Techniques (Third Edition), 2011.
- Gonçalves. Regras de Associação e suas Medidas de Interesse Objetivas e Subjetivas. INFOCOMP Journal of Computer Science, 2005, 4, 26-35.

[a] http://fbarth.net.br/materiais/webMining/webUsageMining.pdf

- Data Mining Algorithms in R - Apriori Algorithm. http://en.wikibooks.org/wiki/Data_Mining_Algorithms_In_R/Frequent_Pattern_Mining/The_Apriori_Algorithm. Acessado em 13 de junho de 2013.
- RDataMining.com: Association Rules. http://www.rdatamining.com/examples/association-rules. Acessado em 13 de junho de 2013.