Web Usage Mining, Association Rules

Patrick Siehndel(siehndel@L3S.de)

26.05.2014

Web Log Analyse, Association Rules

Given is cleaned Web server log.

1. Indentify the different *Sessions* in the log, assuming that a user is identified by the IP-Address and that a session ends 30 minutes after the last page request.

2. Based on the identified *Sessions* create *Frequent Item Sets* with a *Support* higher than 30%. Use the Apriori Algorithm:

   ```
   Init:  Initial  candidates  are  set  to  be  the  items

   Loop:
     (i)    Combine  n-element  Frequent  Item  Sets  to n+1-element  Frequent Item  Sets
     (ii)  Check  whether  Support  of  the  n+1-element  Frequent  Item Sets  is high  enough
     (iii)  Stop  if  no  new  Frequent  Item Sets  are  found,  else  go  to  (i)
   ```

3. Identify *Association Rules* based on the *Frequent Item Sets* with a *Confidence* higher than 60

4. Which information from the Web Log is not used by the *Association Rules*?

Web Log:

```
10.0.0.127 [06/May/2013:09:03:01] ... "/shop/A.html"
10.0.0.127 [06/May/2013:09:08:11] ... "/shop/C.html"
10.0.0.127 [06/May/2013:09:09:56] ... "/shop/D.html"
10.0.0.143 [06/May/2013:09:10:23] ... "/shop/A.html"
10.0.0.127 [06/May/2013:09:12:34] ... "/shop/B.html"
10.0.0.143 [06/May/2013:09:26:53] ... "/shop/D.html"
10.0.0.156 [06/May/2013:09:39:08] ... "/shop/C.html"
10.0.0.143 [06/May/2013:09:40:12] ... "/shop/C.html"
10.0.0.127 [06/May/2013:09:45:42] ... "/shop/A.html"
10.0.0.127 [06/May/2013:09:48:45] ... "/shop/C.html"
10.0.0.127 [06/May/2013:09:56:12] ... "/shop/E.html"
10.0.0.156 [06/May/2013:09:59:16] ... "/shop/E.html"
```

Solution 1 - Web Log Analyse, Association Rules

1.1 Sessions

A Session is defined as a collection of ressources, which have been visited by a user within a defined time intervall. This time intervall is 30 minutes in this exercise. The resulting sessions are:
$S_1 = \{A, C, D, B\}$ (User 10.0.0.127)
$S_2 = \{A, D, C\}$ (User 10.0.0.143)
$S_3 = \{C, E\}$ (User 10.0.0.156)
$S_4 = \{A, C, E\}$ (User 10.0.0.127)

1.2 Frequent Item Sets

Support of 30% → Item Sets have to appear in min. 2 Sessions together. F_i defines the Frequent Item Set in step i and C_i defines the possible Candidates for F_i in the same step.

Step 1: Init - 1-element Frequent Item Sets:
$C_1 = \{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}\}$
$F_1 = \{\{A\}, \{C\}, \{D\}, \{E\}\}$ (B appears only in one session($\rightarrow$ Support = 25% $<$ 30% = minimum-Support).

Step 2: 2-element Frequent Item Sets:
$C_2 = \{\{A, C\}, \{A, D\}, \{A, E\}, \{C, D\}, \{C, E\}, \{D, E\}\}$
$F_2 = \{\{A, C\}, \{A, D\}, \{C, D\}, \{C, E\}\}$

Step 3: 3-element Frequent Item Sets:
$C_3 = \{\{A, C, D\}, \{A, C, E\}, \{C, D, E\}\}$
$F_3 = \{\{A, C, D\}\}$

Step 4: 4-element Frequent Item Sets:

There are no 4-element candidates from F_3.

The resulting Frequent Item Set are $F = F_1 \cup F_2 \cup F_3$.

1.3 Association Rules

We are looking for implications $X \Rightarrow Y$ With: *"For at least 60% of the cases, containing item-combination X , also item-combination Y is contained.*

$X \Rightarrow Y$	**Confidence**	**bigger 60%?**
$A \Rightarrow C$	$\frac{3}{3}$	yes
$A \Rightarrow D$	$\frac{2}{3}$	yes
$A \Rightarrow C, D$	$\frac{2}{3}$	yes
$C \Rightarrow A$	$\frac{3}{4}$	yes
$C \Rightarrow D$	$\frac{2}{4}$	no
$C \Rightarrow E$	$\frac{2}{4}$	no
$C \Rightarrow A, D$	$\frac{2}{4}$	no
$D \Rightarrow A$	$\frac{2}{2}$	yes
$D \Rightarrow C$	$\frac{2}{2}$	yes
$D \Rightarrow A, C$	$\frac{2}{2}$	yes
$E \Rightarrow C$	$\frac{2}{2}$	yes
$A, C \Rightarrow D$	$\frac{2}{3}$	yes
$A, D \Rightarrow C$	$\frac{2}{2}$	yes
$C, D \Rightarrow A$	$\frac{2}{2}$	yes

1.4 Not used informartion

The order in which pages are visited is not used.