Web Usage
Lecture as part of the Web Science Course
Leibniz University Hannover, 16 November 2010

Exercises
For these exercises you need to obtain the visit log of one participant of a client-side study from 2005. This log contains the visits of the user's first 14 sessions.
You can obtain the file as either a MySql dump, CSV file or Excel sheet from http://www.l3s.de/~herder/temp/webusage/

1. Calculate the recurrence rate
The recurrence rate is defined as (1 – (distinct pages visited / total pages visited)) x 100%
It is a measure of how repetitive the online behavior of the user is.

Solution
The easiest way to calculate the recurrence rate is via MySql, but the procedure should be the same when using Excel or a Java program. I present just the MySql query and the solution.

MySQL query: select count(distinct uri) as distinct_pages,count(*) as total_pages,1-(count(distinct uri)/count(*)) as recurrence from webusagelog
The log contains 626 distinct pages and 455 total pages. Hence, the recurrence rate is 27.3%.

2. Reconstruct the session time-out heuristic
The heuristic assumes that if there is more than 25 minutes elapsed between two subsequent page visits, a new session is started.

Solution
Iterate each page visit and compare the time of the visit with the previous visit. The time is listed in milliseconds. In Excel you can easily do this by creating a new column M, starting in M3, with the formula “=(E3-E2)/1000” . Expand the column so that it covers all page visits.
Then in column E you can check the session time-out with the boolean “=M3>(60*25)”. You will see a value ' TRUE' where a new session starts.
3. Which percentage of visits is covered by Google?

Further: retrace why this user was searching for 'Thinking about the Unthinkable' in session 8.

Solution

MySQL: select uri,title,session from webusagelog where host like "%google%"

This query yields 76 results. Visits to Google / Total visits = 76 / 626 = 29%.

It is obvious that the user was looking for the book “Thinking About the Unthinkable in the 1980s” by Herman Kahn. He actually bought it at Amazon.

4. List the ten most visited pages and calculate how often these pages have been visited

Bonus exercise: plot the distribution of most visited pages

Solution

Query: select uri,count(*) as freq from webusagelog group by uri order by freq desc limit 10

'http://www.trekanten.org/results/', 11
'http://medusa.aakb.bib.dk/is/www/query-sh.asp', 11
'http://www.trekanten.org/english/news.htm', 10
'http://www.pixar.com/theater/trailers/incredibles/', 8
'http://www.google.de/', 6
'http://www.spiegel.de/', 6
'http://www.aakb.dk/sw2228.asp', 5
'http://www.fit.fraunhofer.de/gebiete/studien_usability_anf1.php3', 5
'http://www.cabal.se/silence/sortenmuld/', 5
The distribution plot can be created by exporting the query results (without the limit 10) to an Excel sheet and creating a bar or line chart. The chart should look like this:

![Distribution Plot](image)

Note that, even with the small amount of data, still the power law distribution can be observed (a few pages visited frequently, other pages only once or twice)

5. How many different pages did the user visit at [www.trekanten.org](http://www.trekanten.org)?

In total, the user visited the site 26 times. How many different pages within this site did he visit and what is his main reason for visiting this site?

*Optional: calculate for each visited site the ratio between the number of distinct pages visited and the total number of visits*

**Solution**

The user visited 7 distinct pages at trekanten.org. At the end of the solution you will find the main reason for visiting the site.
The ratio between the number of distinct pages visited per site can be easily calculated as 
distinct_pages / total_visits x 100%.

In MySql: select host,count(*) as total_visits,count(distinct uri) as distinct_pages,count(distinct 
uri)/count(*)*100 as ratio from webusagelog group by host order by total_visits desc

Below are the first items in the resultset (from left to right: site name, total visits, distinct pages, ratio).

'www.spiegel.de', 80, 67, 83.7
'www.google.dk', 67, 58, 86.5
'medusa.aakb.bib.dk', 33, 20, 60.6
'www.amazon.de', 30, 23, 76.6
'www.trekanten.org', 26, 7, 26.9
'www.pixar.com', 23, 14, 60.8
'www.aakb.dk', 21, 12, 57.1
'www.heise.de', 20, 17, 85.0

For news sites such as Spiegel and Google the ratio is quite high. The user typically visits the portal site 
and clicks on a news item (Spiegel) or issues a query (Google).

For trekanten.org the ratio is very low. When you visit trekanten.org you will learn that this is a Danish 
sports club. This user is apparently a member of the club. Upon closer inspection, you see that he 
regularly checks the results (of the matches) and news items.