Analysis of Web Search Engine Query Sessions

Workshop on Web Mining and Web Usage Analysis,
WEBKDD 2006, Philadelphia, United States.
August 20-23, 2006

David Nettleton

Web Research Group,
Universitat Pompeu Fabra,
Passeig de Circumval.lació, 8
08003 Barcelona, Spain
david.nettleton@upf.edu

Liliana Calderón-Benavides

Web Research Group,
Universitat Pompeu Fabra,
Passeig de Circumval.lació, 8
08003 Barcelona, Spain
liliana.calderon@upf.edu

Ricardo Baeza-Yates

Yahoo! Research,
Ocata, 1
08003 Barcelona, Spain
ricardo@baeza.cl

Presented by Álvaro R. Pereira Jr
DCC/UFMG/Brazil

1. Introduction

- **Problem to be solved:** identification of user profiles (informational, navigational or transactional) and quality indicators for query sessions from the diversity of queries made by the users and their clicks on the search results.

- **Test data:** extracted from 65,282 queries and 122,184 clicked documents recorded by the "todoCL" search engine [3].

- We use *the Kohonen SOM clustering technique* and the *C4.5 rule induction algorithm* to model web query sessions

1.2 Introduction – Quality of Query Session

We also define four main session quality categories, for the user sessions:

Table 1. Hypothetical user query session quality profile

	Profile (quality of query session			
Descriptive variable	**$high_1$**	**$high_2$**	**low_1**	**low_2**
Average hold time of selected documents		high		low
Ranking of documents chosen	high		low/ medium	
Number of clicks	low		high	high

"high1": user clicks on a few documents which have a high ranking (e.g. in the first five results shown), given that it is reasonable (though not definitive) to assume the ranking of the results is correct with respect to what the user is looking for and has expressed in the corresponding query.

"high2": high hold time, which implies that the user spends a longer time reading/visualizing the clicked document (profile "high2" of Table 1).

"low1": if the user selects many low ranking documents this would also identify that the ordering of the results does not correspond well with the query.

"low2": in the case of an "informational" or "transactional" user type, a lower hold time would indicate that the user has not found the content interesting. If we combine this with a high number of clicks, it would indicate that the user has found it necessary to check many results.

5.1 Clustering in Homogeneous Groups

Table 3. Kohonen clustering of Queries data: averages of input variables for 'level 1' cluster groups

Queries							**Confidence**	
Cluster Group*	Avg. number of terms	Avg. query freq.	Avg. hold time	Avg. ranking	Avg. number of clicks	Number of Queries	Avg. activation	Stdev. activation
11	3.16	**2.63**	30.87	4.94	1.92	191	8.77	2.60
12	2.24	3.53	103.66	6.86	1.92	214	11.07	2.91
21	1.94	**6.84**	126.59	6.97	2.58	205	11.73	**3.40**
22	2.51	4.59	125.01	5.70	3.28	306	11.86	3.21
30	1.93	4.34	**128.78**	**9.86**	**6.88**	449	**14.18**	2.82
40	2.04	2.95	**4.16**	**4.42**	**1.00**	189	**6.44**	2.61
50	**3.45**	2.69	69.24	4.53	1.11	153	7.84	**2.35**
60	**1.53**	4.56	111.03	4.73	2.00	89	9.78	2.97

*8 level 1 clusters, 225 level 2 clusters assigned.

Cluster group 30 has the maximum values for "average ranking of clicked results" (9.86), "average number of clicks" (6.88), and "average hold time" (128.78).

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4	**4.36**	7.64	2.21	**1.29**	**1.21**	14	6.63	1.21
14,6	8.8	5.5	2.5	1.8	1.4	10	7.45	**1.60**
12,2	31.22	**13.22**	2	2.22	2	9	11.54	**1.11**
5,15	**50.11**	5.78	2	**2.56**	2	9	10.08	1.33
Level 2 Query clusters (for level 1 Cluster Group 30)								
7,14	199.27	**8.82**	2	1.36	**6.59**	22	14.44	2.14
6,11	174.25	8.95	2	**3.15**	6.75	20	14.72	1.90
8,2	**90.71**	**16.18**	2	**1**	**10.88**	17	14.3	**1.71**
12,5	143.88	15.5	**9.44**	1.12	7.31	16	17.21	**4.50**
12,6	**343.12**	13.56	3.12	1.38	7	16	16.71	2.62
Level 2 Query clusters (for level 1 Cluster Group 40)								
13,13	0	2	**2**	**2**	1	31	4	**1.1**
14,13	0	2	2	**3**	1	21	4.48	**1.1**
15,13	0	**1.65**	2	**3**	1	20	4.2	1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

*in descending order of number of queries assigned

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4	**4.36**	7.64	2.21	**1.29**	**1.21**	14	6.63	1.21
14,6	8.8	5.5	2.5	1.8	1.4	10	7.45	**1.60**
12,2								**1.11**
5,15								1.33
Level 2 Query clusters (for level 1 Cluster Group 30)								
7,14	199.27	**8.82**	2	1.36	**6.59**	22	14.44	2.14
6,11	174.25	8.95	2	**3.15**	6.75	20	14.72	1.90
8,2	**90.71**	**16.18**	2	**1**	**10.88**	17	14.3	**1.71**
12,5	143.88	15.5	**9.44**	1.12	7.31	16	17.21	**4.50**
12,6	**343.12**	13.56	3.12	1.38	7	16	16.71	2.62
Level 2 Query clusters (for level 1 Cluster Group 40)								
13,13	0**	2	**2**	**2**	1	31	4	**1.1**
14,13	0	2	2	**3**	1	21	4.48	**1.1**
15,13	0	**1.65**	2	**3**	1	20	4.2	1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

Navigational: the query-sessions grouped in level 1 **query cluster 40** have a low hold time and a low number of clicks,

*in descending order of number of queries assigned **rounded down to zero if less than 1.0

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4	**4.36**	7.64	2.21	**1.29**	**1.21**	14	6.63	1.21
14,6	8.8	5.5	2.5	1.8	1.4	10	7.45	**1.60**
12,2	31.22	**13.22**	2	2.22	2	9	11.54	**1.11**
5,15	**50.11**	5.78	2	**2.56**	2	9	10.08	1.33
Level 2 Query clusters (for level 1 Cluster Group 30)								
7,14	199.27	**8.82**	2	1.36	**6.59**	22	14.44	2.14
6,11	174.25	8.95	2	**3.15**	6.75	20	14.72	1.90
8,2	**90.71**	**16.18**	2	**1**	**10.88**	17	14.3	**1.71**
12,5	143.88	15.5	**9.44**	1.12	7.31	16	17.21	**4.50**
12,6	**343.12**	13.56	3.12	1.38	7	16	16.71	2.62
Level 2 Query clusters (for level 1 Cluster Group 40)								
13,13								**1.1**
14,13								**1.1**
15,13	0	**1.65**	2	**3**	1	20	4.2	1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

*in descending order of number of queries assigned **rounded down to zero if less than 1.0

Informational: in query **cluster group 30**, clusters were generated which grouped the query-sessions whose number of clicks and hold time is high.

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4	**4.36**	7.64	2.21	**1.29**	**1.21**	14	6.63	1.21
14,6	8.8	5.5	2.5	1.8	1.4	10	7.45	**1.60**
12,2	31.22	**13.22**	2	2.22	2	9	11.54	**1.11**
5,15	**50.11**	5.78	2	**2.56**	2	9	10.08	1.33
Level 2 Query clusters (for level 1 Cluster Group 30)								
7,14	199.27	**8.82**	2	1.36	**6.59**	22	14.44	2.14
6,11	174.25	8.95	2	**3.15**	6.75	20	14.72	1.90
8,2	**90.71**	**16.18**	2	**1**	**10.88**	17	14.3	**1.71**
12,5	143.88	15.5	**9.44**	1.12	7.31	16	17.21	**4.50**
12,6								2.62
Level 2 Query clusters (for level 1 Cluster Group 4)								
13,13	0**	2	**2**	**2**	1	31	4	**1.1**
14,13	0	2	2	**3**	1	21	4.48	**1.1**
15,13	0	**1.65**	2	**3**	1	20	4.2	1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

Transactional: in query **cluster group 12** we can observe medium to high hold times and a low number of clicks

*in descending order of number of queries assigned **rounded down to zero if less than 1.0

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4	**4.36**	7.64	2.21	**1.29**	**1.21**	14	6.63	1.21
14,6								**1.60**
12,2								**1.11**
5,15								1.33
Level 2 Query clusters (								
7,14								2.14
6,11	174.25	8.95	2	**3.15**	6.75	20	14.72	1.90
8,2	**90.71**	**16.18**	2	**1**	**10.88**	17	14.3	**1.71**
12,5	143.88	15.5	**9.44**	1.12	7.31	16	17.21	**4.50**
12,6	**343.12**	13.56	3.12	1.38	7	16	16.71	2.62
Level 2 Query clusters (for level 1 Cluster Group 40)								
13,13	0**	2	**2**	**2**	1	31	4	**1.1**
14,13	0	2	2	**3**	1	21	4.48	**1.1**
15,13	0	**1.65**	2	**3**	1	20	4.2	1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

Quality Profiles: We now interpret the clusters with reference to the session quality profiles presented in Section 1 (Table 1).

***High1*:** in all of **cluster group 40** we can see a high clicked document ranking (low values) and a low number of clicks (all equal to 1), which corresponds to the hypothetical Profile 1 which indicates “high” quality.

*in descending order of number of queries assigned **rounded down to zero if less than 1.0

5.2 Analysis of Clusters and Sub-Clustering

Table 4. Average values for key variables used for clustering of query data (corresponding to cluster groups in Table 3), and one comparative variable not used in clustering (Freq of URL 1)

	Average values (for each cluster)						Confidence	Document
Cluster*	Hold time	Ranking	Freq. query	Number of terms	Number of clicks	Number of queries	Avg. activation	Freq. of URL 1
Level 2 Query clusters (for level 1 Cluster Group 12)								
6,7	18	**1.71**	2	2.07	2	14	6,88	1.29
13,4								1.21
14,6								**1.60**
12,2	31.22	**13.22**	2	2.22	2	9	11.54	**1.11**
5,15	**50.11**	5.78	2	**2.56**	2	9	10.08	1.33
Level 2 Query clusters (for level 1 Cluster Group 30)								
7,14	199.27	**8.82**	2	1.36	**6.59**	22	14.44	2.14
6,11	174.25	8.95	2	**3.15**	6.75			
8,2	**90.71**	**16.18**	2	**1**	**10.88**			
12,5	143.88	15.5	**9.44**	1.12	7.31			
12,6	**343.12**	13.56	3.12	1.38	7			
Level 2 Query clusters (for level 1 Cluster Group 40)								
13,13	0**	2	**2**	**2**	1	31	4	**1.1**
14,13								**1.1**
15,13								1.4
13,14	0	**3.53**	2	2	1	19	4.8	1.11
1,5	0	1.82	4.76	**1.41**	1	17	7.95	**3.94**

*in descending order of number of queries assigned **rounded down to zero if less than 1.0

Low1: **cluster group 30** shows a low/medium clicked document ranking and a high number of clicks, which indicates a problem of low quality according to our definition.

This is not necessarily contradictory, given that the queries can show good quality in some aspects, and low quality in other aspects.

High2: **cluster group 30** has the highest average hold time, which is indicative of this quality type.

6 C4.5 Tree and Rule Induction - user type as label

Rule	Used	Errors	Label
1	946	285 (30.1%)	nav
2	228	109 (47.8%)	tra
3	671	292 (43.5%)	inf

We observe that “nav” (navigational) is the easiest user type to predict,

followed by “inf” (informational),

whereas “tra” (transactional) seems to be more ambiguous and difficult to predict.

6 C4.5 Tree and Rule Induction - quality class as label

(a)	(b)	(c)	(d)	<< classified as	%correct by label
367	72	84	41	(a): class high1	65%
59	**73**	39	15	(b): class high2	39%
62	78	**207**	24	(c): class low1	56%
36	21	18	**65**	(d): class low2	37%

We observe that “high1” and “low1” are the easiest quality classes to predict, followed by “high2” and “low2” which gave significantly lower predictive accuracies.

One possible cause of this could be the range assignments which we defined in Section 1.2, or due to ambiguities between the different classes.

7 Conclusions - 1

- The practical objective has been to study the incidence of user profiles in the query session data *and* to confirm the incidence of the quality indicators in the same data set.
  - We have been able to create clusters which show characteristics of *Broder's* classes (**navigational, informational, transactional**), using descriptive variables as input.
  - We have also identified the incidence of the quality profiles (**high1, high2, low1, low2**) in the clusters, which are also related to the user type.

References

1. Baeza-Yates, R., Castillo, C. *Relating web structure and user search behavior* (extended poster). In *Proc. 10th World Wide Web Conference*, Hong Kong, China, May 2001.

2. Baeza-Yates, R., Hurtado, C., Mendoza, M. and Dupret G. *Modeling user search behavior.* In *Proceedings of the Third Latin American Web Congress 2005*, p. 242 – 251. Buenos Aires, Argentina, Oct. 2005.

3. Broder, A.Z. *A taxonomy of web search*. SIGIR Forum, 36(2):3-10, 2002.

4. Hunt, E.B. *Artificial Intelligence*. Academic Press, New York, 1975.

5. Kohonen, T. *Self organization and associative memory*. Berlin, Springer-Verlag, 1984.

6. Lee, U., Liu, Z., Cho, J. *Automatic identification of user goals in web search.* In *Proc. 14th International World Wide Web Conference*, Chiba, Japan, May 2005.

7. Nettleton, D.F. *El uso de tecnologia de minería de datos para la construcción y explotación del datawarehouse*. Novatica, Spain, pp. 52-55, 1999.

8. Nettleton, D.F., Fandiño, V.L., Witty, M., Vilajosana, E. *The use of a data mining workbench for macro and micro economic modelling*. In *Proceedings of Data Mining 2000*, Cambridge University, U.K., July 5-7, pp. 25-34, 2000.

9. Nettleton, D., Baeza-Yates, R. *Web Retrieval: techniques for the aggregation and selection of queries and answers*, (in Spanish), I Spanish Symposium on Fuzzy Logic and Soft Computing, Granada, Spain, Sept. 2005, 183-190.

10. Ntoulas, A., Cho, J., Olston, C. *What's new on the web? The evolution of the web from a search engine perspective*. In *Proc. 13th International World Wide Web Conference*, New York, United States, May 2004.

11. Quinlan, J.R. *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, Calif.. 1993.

12. Sugiyama, K., Hatano, K., Yoshikawa, M. *Adaptive web search based on user profile constructed without any effort from users.* In *Proc. 13th International World Wide Web Conference*, New York, United States, May 2004.