People Searching for People: Analysis of a People Search Engine Log (Abstract)

Wouter Weerkamp
University of Amsterdam
w.weerkamp@uva.nl

Richard Berendsen
University of Amsterdam
r.w.berendsen@uva.nl

Bogomil Kovachev
University of Amsterdam
b.k.kovachev@uva.nl

Edgar Meij
University of Amsterdam
e.j.meij@uva.nl

Krisztian Balog
Norwegian University of Science and Technology
krisztian.balog@idi.ntnu.no

Maarten de Rijke
University of Amsterdam
derijke@uva.nl

1. INTRODUCTION

We summarize findings from [4]. Recent years show an increasing interest in vertical search: searching within a particular type of information. Understanding what people search for in these "verticals" gives direction to research and provides pointers for the search engines themselves. In this paper we analyze the search logs of one particular vertical: people search engines. Based on an extensive analysis of the logs of a search engine geared towards finding people, we propose a classification scheme for people search at three levels: (a) queries, (b) sessions, and (c) users. For queries, we identify three types, (i) event-based high-profile queries (people that become "popular" because of an event happening), (ii) regular high-profile queries (celebrities), and (iii) low-profile queries (other, less-known people). We present experiments on automatic classification of queries. On the session level, we observe five types: (i) family sessions (users looking for relatives), (ii) event sessions (querying the main players of an event), (iii) spotting sessions (trying to "spot" different celebrities online), (iv) polymerous sessions (sessions without a clear relation between queries), and (v) repetitive sessions (query refinement and copying). Finally, for users we identify four types: (i) monitors, (ii) spotters, (iii) followers, and (iv) polymers.

Our findings not only offer insight into search behavior in people search engines, but they are also useful to identify future research directions and to provide pointers for search engine improvements.

We seek to answer the following research questions: (A) What are the general usage statistics of a people search engine? (B) Can we identify different types for our information objects (queries, sessions, users)? (C) Can we automatically classify queries into the proposed types? (D) What are interesting findings in people search that indicate future research directions?

2. USAGE STATISTICS

Search system and data. Figure 1 shows the standard interface of the commercial Dutch language people seacrh engine we study. The query log data was collected between September 1, 2010 and December 31, 2010.

Query characteristics. Table 1 lists the characteristics of the individual queries in our log data. We find that a significant amount of queries consists of one term. For these queries the distribution of queries over the number of out clicks has a longer tail than for multiple term queries. This indicates that people are exploring the search results.

Figure 1: Simple search interface: a single search box with a search button.

Table 1: Characteristics of individual queries.

Number of queries	13,331,417	
Number of unique queries	4,221,556	
Number of one term queries	537,365	(4.0%)

Session characteristics. Table 2 lists the characteristics of the sessions. Compared to sessions in web search engines, we find that our people search engine has a much higher percentage of one-query sessions (web search engine logs contain 50–60% one-query sessions [2]). Sessions that do consist of multiple queries, contain on average almost four queries, and these sessions last, on average, just over six minutes. It seems most people use a people search engine to quickly find information on one particular person and leave after the information has been found.

Table 2: Characteristics of sessions (time-out 40 minutes).

Number of sessions	8,125,695
Number of sessions with $>$ 1 query	1,775,880
Average number of sessions per day	67,155

Out click characteristics. Out click statistics are listed in Table 3. When we compare the percentage of queries with at least one out click to out clicks in web search, we notice that the percentages in people search are much lower. Numbers for web search vary greatly (50% in [1], 73% in [3]), but are consistently higher than the 17% for our data. We identify two reasons for the low out click ratio in people search: (i) People search is still a challenging problem, and it is not easy to find relevant results for all person queries, and (ii) the interface already displays information about the person (e.g., related news articles, images, and facts).

The search result page of the people search engine has different parts for different kinds of search results. Their popularity in terms of out clicks is listed in Table 4. Social media results are the most popular and make up 66% of all out clicks, followed by search engine results.

DIR '12 Gent, Belgium
ACM X-XXXXX-XX-X/XX/XX.

Table 3: Characteristics of out clicks.

Number of out clicks	3,965,462	
Number of unique out clicks	2,883,230	
Number of queries followed by out click	2,351,848	17.6%
Number of sessions that include out click	1,625,817	20.0%

Table 4: Interface result categories and number of out clicks.

Social media	2,625,500	66.2%
Search engines	674,079	17.0%
Multimedia	120,874	3.1%
Miscellaneous	337,104	8.5%
"Alternative sources"	187,098	4.7%

3. OBJECT CLASSIFICATIONS

For each of the information objects, queries, sessions, and users, we propose a classification scheme. We summarize the query classification results here, for sessions and users we refer to [4]

We have defined our query types in the introduction. To further explain the difference between the two high-profile query types, we plot the query volume of four example queries in Figure 2. Note that the y-axis has a different scale for each of the plots. We can clearly see a peak in query volume for the two event-based high-profile queries. For both queries we can identify related (news) events that led to this peak: Derck Stabler was the main suspect in the murder of his mother (on October 4); Nathalie Weinreder is a murder victim (on December 12). On the other hand, the query volume for the regular high-profile query is relatively stable, with about 100 queries per day over the whole period. The low-profile query has no peaks, and search volume is very modest (one search on a few days).

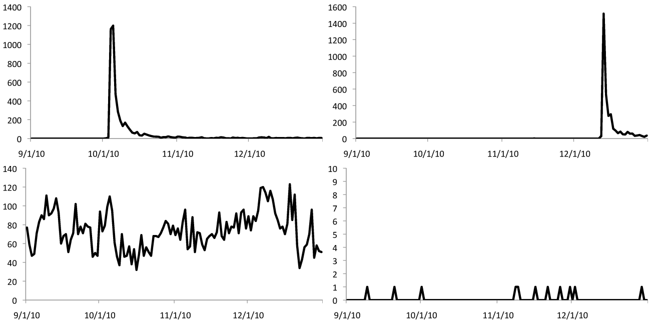

Figure 2: Examples of query volume per day for the two high-profile query types (Top:) event-based queries (Derck Stabler and Nathalie Weinreder, respectively), and (Bottom:) a regular query (Geert Wilders). For comparison, we have included a random low-profile query (Yucel Ugur).

Automatic classification. Being able to automatically classify queries as high-profile or low-profile is useful, both for investigating sessions/users and for a people search system. Based on this classification, the system might prioritize different result types or show additional information sources.

We train a J48 decision tree algorithm on a sample of an annotated set of queries. To counter class distribution skewedness, we downsample the more common classes to the size of the least common class. Table 5 shows our results. Distinguishing low-profile from high-profile queries is possible with good accuracy, but distinguishing event-based from high-profile queries is harder. An analysis of the contribution of the individual features shows that search volume in the logs, and result counts for Wikipedia and social media are most important, while mentions in Dutch news are ignored.

Table 5: Results of automatic query classification using the J48 decision tree algorithm.

Query type	Precision	Recall
Event-based high-profile	0.745	0.759
Regular high-profile	0.739	0.630
Low-profile	0.820	0.926
Low-profile	0.911	0.879
High-profile	0.883	0.914

4. CONCLUSION

We performed an analysis of query log data from a commercial people search engine, consisting of 13m queries submitted over a four month period. It is the first time a query log analysis is performed on a people search engine, in order to investigate search behavior for this particular type of information object.

We focused our analysis on four information objects: queries, sessions, users, and out clicks. The most interesting findings include (i) a significant number of users type just one term (i.e., only a first or last name) and start exploring results; (ii) we observe a much higher percentage of one query sessions in people search as compared to web search; (iii) we observe a low click-through ratio as compared to web search; (iv) social media results are the most popular result type. Furthermore, we have proposed classification schemes for queries, sessions, and users, and shown, through an initial experiment, that automatic classification of queries is doable. Analysis of the features shows the usefulness of social media reports in identifying high-profile queries.

Acknowledgments. This research was partially supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, the PROMISE Network of Excellence co-funded by the 7th Framework Programme of the European Commission, grant agreement no. 258191, the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.061.814, 612.061.815, 640.004.802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, and under COMMIT project Infiniti.

REFERENCES

[1] J. Callan, J. Allan, C. L. Clarke, S. Dumais, D. A. Evans, M. Sanderson, and C. Zhai. Meeting of the minds: An information retrieval research agenda. *ACM SIGIR Forum*, 41 (2):25–34, 2007.

[2] B. J. Jansen and A. Spink. How are we searching the world wide web? a comparison of nine search engine transaction logs. *Inf. Proc. & Management*, 42(1):248–263, 2006.

[3] S. Stamou and E. N. Efthimiadis. Interpreting user inactivity on search results. In *ECIR'10*, 2010.

[4] W. Weerkamp, B. Kovachev, R. Berendsen, E. Meij, K. Balog, and M. de Rijke. People searching for people: Analysis of a people search engine log. In *SIGIR'11*, 2011.