People Searching for People: Analysis of a People Search Engine Log

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

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

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

Edgar Meij  
ISLA, University of Amsterdam  
edgar.meij@uva.nl

Krisztian Balog  
NTNU Trondheim  
krisztian.balog@idi.ntnu.no

Maarten de Rijke  
ISLA, University of Amsterdam  
derijke@uva.nl

ABSTRACT
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.

Categories and Subject Descriptors
H.3.3 [Information Storage and Retrieval]: Search process

General Terms
Theory, Experimentation, Measurement

Keywords
People search, query log analysis, classification

1. INTRODUCTION
As a result of the growth of the amount of online information, search has become one of the most important online activities. Major web search engines are among the most visited web pages.1 with Google, Yahoo!, and Baidu in the global top six. An important aspect of research related to search is understanding how users deploy a search engine: What is it they are looking for? Who is using the search engine? How do they use it? Answering such questions leads to new research directions and, in the end, helps to improve the user experience.

Much of the research in understanding search behavior exploits the log files of search engines. Query (or transaction) logs contain information about the query a user issued, and the subsequent actions (result pages viewed, results clicked, etc.), if any. Early work by Broder [6] shows that there is a fair correlation between findings from query log analysis and user surveys and, in the same paper, he also proposes an influential taxonomy of web queries.

Much of the work on query log analysis was, and still is, focused around web search (see Section 2), despite the increase in so-called vertical search engines. Instead of relying on a single general web search engine to provide information on specific queries, users use a search engine specialized in a single domain or segment of online content. Well-known examples of vertical search engines include scientific literature search [21], medical IR [11], patent retrieval [20], search in cultural heritage [27], and book search [18]. Although previous work on query log analysis has provided us with general insights in users’ search behavior, this behavior might change when searching for a particular type of information. For this reason, research is now also focusing on query log analysis for particular information objects. For example, Jones et al. [17] look at how users search digital libraries, Ke et al. [19] explore search behavior in scientific literature, Mishne and de Rijke [26] analyze blog search, and Huurnink et al. [13] do so for search in an audio-visual archive.

One type of information users frequently look for is people. It is estimated that 11–17% of web queries contain a person name, and, more so, 4% of web queries are person name queries only [1]. No fewer than 57% of adult internet users uses a search engine to search for their own name [23]. In addition to these “vanity searches,” many internet users search for (i) information on people from their past (46%), (ii) their friends (38%), and (iii) business-related persons, like colleagues and competitors (31% of employed internet users). These numbers have increased by 10% in a period of four years, indicating the importance of people search in an online setting.

In this paper, we analyse the query logs of a people search engine. These logs offer us information at three levels: queries, sessions, and users (see Section 3), and we are interested in the structure we can identify within each of these levels. More specifically,
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 each of 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?

The paper makes the following contributions: (1) We describe how a people search engine is being used. (2) We propose a classification scheme for queries, sessions, and users in a people search engine. (3) We identify features to allow for automatic classification of person name queries. (4) We offer recommendations as to future research in, and implementation of, people search technology. To the best of our knowledge, our study is the first to provide a detailed log analysis in the emerging area of entity search.

In Section 2 we discuss previous work on query log analysis and query classification. Section 3 defines the information objects we explore in the paper. In Section 4 we introduce the search system and interface from which our logs originate, and offer insights in the general statistics of our log data. We propose our classification scheme in Section 5 and experiment with automatic classification. Finally, we discuss further observations in Section 6 and conclude in Section 7.

2. RELATED WORK

One of the first large scale query log analysis papers explores search logs of AltaVista [30]. The authors perform a descriptive analysis of the (almost) 1 billion queries in the log, indicating query length (mostly 1–3 term queries), session length (mostly one query sessions), popular query terms (sex related), the number of result pages a user looks at (mostly one page), and how queries are modified within a session. Following several other studies of web search engine logs, Jansen and Spink [15] compare analyses of nine search engine logs between 1997 and 2002. They conclude that most findings are stable over time, but that, e.g., the percentage of users that only looks at the first result page increases. They also show that the percentage of queries related to people, places or things (“entities”) increases from 21% in 2001 to over 41% in 2002, clearly indicating the importance of people search.

When it comes to people search and query log analysis, not much work has been done. Guo et al. [10] propose a method to recognize named entities in queries by learning context for these entities. Although their work shows promise, it focuses on entities like books, movies and music, rather than people. More closely related work is done by Pound et al. [28] and looks at ad-hoc object retrieval; the authors show that over 40% of queries in their dataset is of type “entity” and they specify methods for dealing with such queries in a “web of data” setting.

Queries. What is it users are searching for in a particular search environment? This question is the rationale behind many papers covering queries and query types. Classification of queries is often based on (i) query intent or (ii) query semantics. An influential paper of the former type by Broder [6] looks at queries in a web search engine. An exploration of query log data reveals three types of query: informational, navigational, and transactional. Most queries in a web search engine are informational (40–50%), followed by transactional (30–36%). Later work by Rose and Levinson [29] extends this taxonomy with subclasses. A manual classification of 1,500 web queries shows that the percentage of informational queries is higher than in the original paper (about 60%), at the cost of both other types.

The rise of verticals leads to users interacting with specialized search systems, which in turn might lead to different types of queries and different behavior. Mishne and de Rijke [26] acknowledge this and look at query types in a blog search engine. Since almost all blog queries are informational they propose two new query types: concept and context queries—both of which are informational but quite distinctive in blog search. Another type of vertical search that is explored using query logs are audiovisual archives [13]. Here, the authors do not classify queries, but show general statistics of the logs, indicating that users mainly look for program titles and entities (organizations, people). These two papers show that, by moving towards more specialized search engines, the query typology needs refinement too.

Looking at query classification research based on query semantics, there exists a large body of related work that considers queries that a given query co-occurs with (see “Sessions”). One example is the classification of query refinements, addressed in [12]. A different classification task is proposed by Cao et al. [9], who state that query context (i.e., previous queries in the same session) is needed to classify queries into categories. A similar notion is used by Meij et al. [25], who aim at identifying concepts in queries.

Sessions. Sessions are an important aspect in query log analysis, and various ways of detecting sessions have been proposed. According to Jansen et al. [14], session duration is the interval between the user submitting the first query and the user “leaving” the search engine, resulting in sessions varying from several seconds to a few hours. Most time-based session detection approaches group logged actions by some user id, sort the actions chronologically for each user, and split sessions on intervals longer than a certain cutoff value. The choice of cutoff value is dependent on the goal of the analysis. For example, based on a manual examination Mishne and de Rijke [26] use very small cutoff values between 10 and 30 seconds and show that these values mimic sessions based on query reformulation. Longer sessions (e.g., 30 minutes [16]) allow one to explore the different queries and query types a user issues.

Although the time-based approach is a commonly used definition of sessions, there are alternatives. Huang and Efthimiadis [12] use query reformulations to identify session boundaries. Here, sessions consist of consecutive queries by the same user, where each query is a reformulation of the previous query (e.g., adding or deleting words). The idea is that all reformulated queries address a single underlying information need and should be in one session. Jansen et al. [16] compare query reformulations for session detection to the time-based detection; they conclude that query reformulation results in more detected sessions.

A different approach has been proposed by Lucchese et al. [22], who try to detect sessions based on a user’s task. Since multitasking is very common in web search, they conclude that time-based techniques fail at task-dependent session detection; instead, they propose to cluster queries and use the clusters for session detection.

Users. Research into user behavior from query logs can be challenging, since it can be hard to determine which queries and sessions belong to the same user. White and Drucker [34] counter this issue by using a set of volunteer users. They collect search data from these users over a five month period. From this data, they identify two user types: navigators (users with consistent search behavior) and explorers (variable behavior). A different approach (in the setting of searching literature in CiteSeer) by Manavoglu et al. [24] tries to model user behavior and predicts actions by similar users, based on previous users’ actions.

Where the two studies just mentioned model users based on their actions, Weber and Jaimes [33] describe users’ demographics. For this, they use characteristics per ZIP code, and election results per
county. Combining demographics with what users are searching for and how they do so, allows them to gain insight in the behavior of users with specific characteristics.

3. INFORMATION OBJECTS

In the analysis of our people search query logs, we use four types of information object present in the logs. Here, we detail what we consider these objects to be and how they relate to previous work.

Query A query is a search instance in the query logs. A query consists of a name and possibly a keyword (see Section 4 for a discussion of the interface), and a timestamp. The timestamp is important in that the query type can change over time: a person can be “just anyone” at time $t$, but could become a main player in a news event at time $t+n$, or a celebrity could become “just anyone” after disappearing from television for a while.

Session As mentioned in Section 2, the way to detect sessions is dependent on the type of search system, the goal of the research, and the data available. Since this paper is the first to analyze people search, we take a high-level view of sessions to see how users combine person name queries. For this, we take a long interval (40 minutes) between two actions to signal a session boundary and construct sessions accordingly. Sessions can be characterized by their length (i.e., the number of queries in one session) and their duration (i.e., the time interval between the first and last action within one session). In Section 6 we return to the issue of session detection for people search.

User Identifying users over time can be difficult. We use a persistent cookie to assign a user id to queries, and although different users might use the same computer and browser, it is a fairly accurate way of identifying returning users.

Out click A user clicks on one of the search results; these out clicks are identified by their URL and type (e.g., Facebook, LinkedIn, images, or Blogger).

In the next section we go into details regarding the search system and interface and describe the collected data for each of the objects just mentioned.

4. SEARCH SYSTEM AND DATA

The main data source for this paper is a large sample of queries, issued to a Dutch language commercial people search engine. This search engine allows users to submit a person name query and offers search results in four different categories:

- social media,
- web search,
- multimedia, and
- miscellaneous.

Social media results consist of profiles from social networking sites like Facebook and LinkedIn, and other social media sites like Twitter, Blogger, Digg, and Last.fm. The web search category returns search results from major web search engines like Google, Yahoo!, and Bing, and vertical search engines for news and blogs. Multimedia results look for images and video about the person, and the miscellaneous category lists related persons (based on last name), facts about the person (e.g., “John Irving is a writer”), tags, and documents (PDF or Word documents).

The people search engine offers two search interfaces. First, the standard (simple) search interface consists of just one search box, in which the user is supposed to type the first and last name of the person she is looking for (Figure 1). The advanced search interface is somewhat hidden and it presents the user with three search boxes: The first box is used for the first name, the second for the last name, and the third can be used to supply the search engine with additional keywords (Figure 2). Besides adding a keyword to the person name query using the advanced search interface, a user can also click on one of the suggested tags after the initial search using the first and last name only. The clicked tag is then added to the query as a keyword. We provide a detailed analysis of the keywords in Section 6.

From the simple interface, the search engine extracts a first and last name, whereas this segmentation is explicitly given by the user in the advanced interface. In cases where a user only enters one name (simple interface) or leaves one of the name fields empty (advanced interface), we end up with a single name query. This happens in 4% of the queries.

4.1 Query logs

The query log data was collected between September 1, 2010 and December 31, 2010. During this period there were no major updates to the search interface, to allow log entries to be comparable. Entries in the query log consist of a number of fields, listed in Table 1. The three query fields (first and last name, and keyword) have been discussed above; Timestamp indicates the date and time when the query was issued, the SearchID can be used to match a query to out clicks, and finally, the UserID is our indication of the user, as explained before. For out clicks, similar fields are available, indicating the URL of the click, the type, and the date and time when the user clicked the result.

<table>
<thead>
<tr>
<th>Queries</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SearchID</td>
<td>unique identifier for the query</td>
</tr>
<tr>
<td>First name</td>
<td>part of the query</td>
</tr>
<tr>
<td>Last name</td>
<td>part of the query</td>
</tr>
<tr>
<td>Keyword</td>
<td>optional; part of the query</td>
</tr>
<tr>
<td>Timestamp</td>
<td>date and time of the query</td>
</tr>
<tr>
<td>UserID</td>
<td>unique identifier using a cookie</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Out clicks</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SearchID</td>
<td>connect out click with query</td>
</tr>
<tr>
<td>Type</td>
<td>name of the result category</td>
</tr>
<tr>
<td>URL</td>
<td>URL of the clicked result</td>
</tr>
<tr>
<td>Timestamp</td>
<td>date and time of the click</td>
</tr>
</tbody>
</table>

In the remainder of this section we give a high-level description of the data in our query logs. Section 4.2 offers insights in individual
queries. Section 4.3 details sessions in the data, Section 4.4 looks at users of the people search engine, and finally, Section 4.5 explores out clicks after a search.

4.2 Query characteristics

Table 2 lists the characteristics of the individual queries in our log data. Our full dataset consists of over 13m person name queries, issued in a four month period, of which over 4m are unique queries. Figure 4 (left) shows the query frequency distribution of the log data, which follows a power law (with slope $\alpha = 2.0$). As we can see, most queries are issued only once. On average, users issued over 110,000 queries per day. In the left plot of Figure 3 we show the number of queries for each day in the dataset. We see a clear cyclic pattern (indicated by the red line), which is due to the popularity of searching on working days compared to weekends. This is clarified in the center plot, which shows the distribution of queries over days of the week. We observe a drop in the number of queries during the weekend; for this plot we looked at the 16 full weeks within our data preventing certain weekdays to occur more often.

Table 2: Characteristics of individual queries.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of queries</td>
<td>13,331,417</td>
</tr>
<tr>
<td>Number of unique queries</td>
<td>4,221,556</td>
</tr>
<tr>
<td>Number of single-term queries</td>
<td>537,365 (4.0%)</td>
</tr>
<tr>
<td>Average number of queries per day</td>
<td>110,177</td>
</tr>
<tr>
<td>Busiest day in number of queries</td>
<td>144,309</td>
</tr>
<tr>
<td>Number of queries with keyword</td>
<td>514,850 (3.9%)</td>
</tr>
</tbody>
</table>

In about 4% of the queries the user submitted only one term (i.e., only a first or last name), and non of these single-term queries is accompanied by a keyword, making it hard to retrieve relevant results for these queries. In Section 6 we get back to single-term queries and their impact on out clicks. In general, keyword usage is low, as only 3.9% of the person name queries contain an additional keyword. The absence of this field in the standard interface is most likely the cause of this. Again, we revisit the issue of keyword usage in Section 6.

Figure 4: Distribution of (Left:) query frequencies, and (Right:) session length in number of queries. Both follow a power law for slope $\alpha = 2.0$ and $\alpha = 2.6$.

Zooming in on the most popular queries, we list the 10 most frequently queried names, the query counts, the number of unique users searching for these names, and a description of who they are in Table 3. The top 10 shows a mixture of celebrities (persons known to most users), like Geert Wilders and Lieke van Lexmond, and (previously) non-famous people who gained attention through some event. Ranking queries by their frequency or by the number of unique users results in almost the same list, which indicates that, even without user information, we can assume that popular queries are issued by many different users.

Table 3: 10 most popular queries during Sep. 1–Dec. 31, 2010, in terms of query counts and unique users.

<table>
<thead>
<tr>
<th>Name</th>
<th>Count</th>
<th>Users</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suze van Rozelaar</td>
<td>16,929</td>
<td>15,373</td>
<td>mistress of soccer player</td>
</tr>
<tr>
<td>Kelly Huizen</td>
<td>13,005</td>
<td>11,706</td>
<td>teenage girl with sex tape</td>
</tr>
<tr>
<td>Ben Saunders</td>
<td>10,074</td>
<td>9,145</td>
<td>participant of talent show</td>
</tr>
<tr>
<td>Barbara van der Vegte</td>
<td>9,879</td>
<td>8,256</td>
<td>mistress of tv host</td>
</tr>
<tr>
<td>Geert Wilders</td>
<td>8,990</td>
<td>8,483</td>
<td>politician</td>
</tr>
<tr>
<td>Lieke van Lexmond</td>
<td>7,774</td>
<td>6,368</td>
<td>actress</td>
</tr>
<tr>
<td>Quincy Schumans</td>
<td>7,266</td>
<td>6,315</td>
<td>murdered teenage boy</td>
</tr>
<tr>
<td>Joyce Exalto</td>
<td>6,656</td>
<td>5,584</td>
<td>murdered teenage girl</td>
</tr>
<tr>
<td>Aa Aa</td>
<td>6,457</td>
<td>6,442</td>
<td>test query</td>
</tr>
<tr>
<td>Sietske Hoekstra</td>
<td>6,088</td>
<td>5,323</td>
<td>mother, killed her babies</td>
</tr>
</tbody>
</table>

4.3 Session characteristics

As mentioned in Section 3, we detect sessions using a time-out between two subsequent actions by the same user in the log. Applying this detection method to our log data leaves us with over 8m sessions. Characteristics of the sessions are listed in Table 4. We observe that most sessions, over 6m (78.1%), contain only one query, and that the distribution of session length follows a power law (see Figure 4, right plot) with slope $\alpha = 2.6$. 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 [15]). 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 users use a people search engine to quickly find information on one particular person, and leave after the information has been found.

Table 4: Characteristics of sessions.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of sessions</td>
<td>8,125,695</td>
</tr>
<tr>
<td>Number of sessions with &gt; 1 query</td>
<td>1,775,880</td>
</tr>
<tr>
<td>Average number of sessions per day</td>
<td>67,155</td>
</tr>
<tr>
<td>Longest session in hours</td>
<td>08h25m</td>
</tr>
<tr>
<td>Average session duration</td>
<td></td>
</tr>
<tr>
<td>all sessions</td>
<td>1m21s</td>
</tr>
<tr>
<td>sessions with &gt; 1 query</td>
<td>6m9s</td>
</tr>
<tr>
<td>Average session length</td>
<td></td>
</tr>
<tr>
<td>all sessions</td>
<td>1.64</td>
</tr>
<tr>
<td>sessions with &gt; 1 query</td>
<td>3.93</td>
</tr>
</tbody>
</table>

4.4 User characteristics

The log data offers us close to 7m different users (see Table 5) and, similar to sessions, most users only issue one query (and therefore interact in only one session). Still, we have about 500,000 users that use the people search engine in more than one session. These returning users instigate, on average, 3.5 sessions in the four month period: roughly one session each month. Figure 5 shows the distribution of queries over users (on the left), and of sessions over users (on the right). Both distributions follow a power law, with slope $\alpha = 2.5$ for queries and $\alpha = 3.8$ for sessions.

To get a sense of when users deploy the people search engine, we look at the distribution of searches over hours of the day in Figure 3 (right plot). Here, the dashed, red line indicates working days, and the solid, green line weekend days. We see that, for working days,
peaks exist in the afternoon (around 2–3pm) and in the evening (around 9pm), while usage drops during lunch (11am–12pm) and dinner (5–7pm); there is a large drop during the night. When we compare this to weekends, we observe that usage shifts several hours: there are more searches during early night (1–4am) in weekends, but fewer during the morning and afternoon. The highest peak shifts from around 2–3pm for working days to 9–10pm during weekends.

4.5 Out click characteristics

The final information object we explore in our log data are the out clicks: do users click on results after a query? If so, where do they click to? Table 6 shows that about 4m clicks are recorded, of which almost 3m unique ones. About 17% of the queries in the logs are followed by an out click, and for sessions this is 20%. Once again, the distribution of out clicks over both queries and sessions (Figure 6) follows a power law. 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.

4.5.1 Out click categories

More interesting than the overall numbers are the details of the out clicks. We can categorize the out clicks according to the search result interface category they belong to. From this categorization, we obtain the percentages as listed in Table 7. Social media results are the most popular and make up 66% of all out clicks, followed by search engine results. Besides the interface result categories explicitly mentioned in the interface, we identify an additional category that attracts many out clicks: the “alternative sources” area at the bottom of the initial result page. Here, users can click on sponsored links to external sites, mainly dating sites and web shops, to look for this person. The links to dating sites are particularly popular, receiving 154,419 out clicks.

4.5.2 Out click ratio

We zoom in on individual result types, and plot the number of out clicks per site in Figure 7. Social networking site Hyves is by far the most popular result type in number of clicks, and it is followed by fellow networking sites Facebook, Schoolbank (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).
old school friends), and LinkedIn. All of these result types are displayed on the first result page. Web search engines Google, Yahoo!, and Bing are also among the most popular result types, as are dating sites. The first site-specific result type is “related,” which refers to a click on a related person. We see that users prefer to find pages that are directly linked to the person they are looking for (answering the question “Who is this?”), profiles being by far the most popular result type. Multimedia results are not very popular, however, the interface already shows these results without a click necessary and, hence, it is likely that users see many more multimedia results than can be concluded from the log data. Finally, dating sites appear to be a particular popular result type.

5. OBJECT CLASSIFICATIONS

In the previous section we performed a high-level exploration of the logs of a people search engine. In this section we add more context to the contents of these logs. More specifically, for each of the information objects (see Section 3), we propose a classification scheme. This exercise resembles work we discussed in Section 2 but has a specific focus on people search. Section 5.1 introduces the query types we identified for people search; in Section 5.2 we explore session types in people search and in Section 5.3 we propose different types of users of people search engines.

To come to our classification schemes, we sampled random queries from our log data. After assigning the query to one of our query types, we continued to annotate all queries in the same session (in case the session contains more than one query), and annotate the session as a whole. The annotation system that we designed for this purpose then allowed us to annotate all other queries and sessions by the same user, resulting in a user annotation. In total we manually annotated 3,281 queries, 1,005 sessions, and 412 users.

5.1 Queries

Based on an initial exploration of the data, we propose the following query types for people search:

**High-profile queries** These queries involve people that stand out in some way and denote people that are known to a relatively large group of users. We distinguish two types of high-profile people:

- **Event-based** People of this type get a boost in attention based on an event that is either currently happening or took place shortly before the query was submitted. In most cases, these events are news-related and are reported either in traditional media or in social media. This type also includes events not related to world news, like recurring cultural events (e.g., Christmas, Easter).

- **Regular** People that are continuously at the center of attention, like celebrities and public persons. In principle, event-based high-profile people can, in time, turn into regular high-profile people, but our period of data collection is too short to be able to observe this phenomenon.

**Low-profile queries** These queries involve people that are “just anyone”: users can be looking for their own name, names of relatives, friends, or other “unknown” persons. We consider all of these queries low-profile queries.

To further explain the difference between the two high-profile query types, we plot the query volume of three example queries in Figure 8. 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).

During the annotation of queries, we came across instances that could not be classified, mainly because they contained only one query term. After removing these 285 queries, we are left with 2,995 annotated queries. Table 8 lists the counts for each of our query types in our sample. By far most of the queries in our sam-

<table>
<thead>
<tr>
<th>Query type</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low-profile</td>
<td>2,796</td>
<td>93.4%</td>
</tr>
<tr>
<td>High-profile</td>
<td>199</td>
<td>6.6%</td>
</tr>
<tr>
<td>Event-based</td>
<td>144</td>
<td>72.4%</td>
</tr>
<tr>
<td>Regular</td>
<td>55</td>
<td>27.6%</td>
</tr>
</tbody>
</table>

Table 8: Query types and their frequency in a sample.
5.2 Sessions

...mention are ignored. The logs, and result counts for Wikipedia and social media are most... contribution of the individual features shows that search volume in...downsampled to the number of high-profile queries, 396 queries in total), we improve accuracy...high-profile queries with good accuracy.

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. For query classification, we use the following features: (i) search volume in the logs over the previous week; (ii) number of mentions in the Dutch news from September 2010 onwards; (iii) number of mentions in the Dutch news in the previous week; and result counts for the query in (iv) social media (using Topsy) and (v) the Dutch Wikipedia (using Yahoo!). We train a J48 decision tree algorithm on a sample of our annotated set of queries. To counter class distribution skewness, we downsample the more common classes to the size of the least common class, leaving us with 162 annotated queries. We use 10-fold cross-validation, and present results in Table 10.

Table 9: Subclasses of the event-based high-profile queries and their percentage.

<table>
<thead>
<tr>
<th>Event-based subclass</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deaths</td>
<td>33.3%</td>
</tr>
<tr>
<td>Criminals</td>
<td>22.9%</td>
</tr>
<tr>
<td>Related to celebrities</td>
<td>9.7%</td>
</tr>
<tr>
<td>Related to other high-profiles</td>
<td>9.7%</td>
</tr>
<tr>
<td>Television</td>
<td>9.0%</td>
</tr>
<tr>
<td>Sex related</td>
<td>6.3%</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>9.0%</td>
</tr>
</tbody>
</table>

Users mostly deploy the people search engine to search for, e.g., relatives, co-workers, neighbors, friends, the guy from the pub last night, or themselves: low-profile people. Occasionally they search for information on high-profile people, and here we notice that event-based queries are about three times as common as “celebrity” queries. One of the reasons for this could be that general search engines already allow us to get easy access to information about celebrities, but this might be harder for people that were low-profile up to the point they became part of an event. An in-depth analysis shows that users are mainly attracted by “sensational” events, related to murders, child abuse, and fatal crashes.

We introduced a fifth session type during annotations: the repetitive session. Sessions of this type consist of either a sequence of identical queries or queries with small corrections in one of the names (which is similar to query refinement in web search). About 35% of the sessions in our sample are of this type, and this high percentage could indicate the need for “person name suggestion” techniques. The system suggests a person name either when no relevant result is found or when the queried name is very similar to another popular person name.

We are interested in the type of results users click on for the various session types. For the spotting and event session, there is not enough data available to perform this analysis. For the remaining three session types we plot the percentage of out clicks per result type in Figure 9. We observe some interesting differences: In family sessions, users are more likely to click a “related” result, and focus less on Hyves results. In repetitive sessions, users click more often on search engine results. Polymorous sessions follow roughly the same distribution as all queries (Figure 7).

Table 10: Results of automatic query classification using the J48 decision tree algorithm.

<table>
<thead>
<tr>
<th>Query type</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event-based high-profile</td>
<td>0.745</td>
<td>0.759</td>
</tr>
<tr>
<td>Regular high-profile</td>
<td>0.739</td>
<td>0.630</td>
</tr>
<tr>
<td>Low-profile</td>
<td>0.820</td>
<td>0.926</td>
</tr>
<tr>
<td>Low-profile</td>
<td>0.911</td>
<td>0.879</td>
</tr>
<tr>
<td>High-profile</td>
<td>0.883</td>
<td>0.914</td>
</tr>
</tbody>
</table>

The results of the automatic query classification show our features are sufficient to classify low-profile queries with good accuracy. Distinguishing between the two high-profile query types proves to be challenging. Taking one step back, and trying to classify high-profile vs. low-profile queries (downsampled to the number of high-profile queries; 396 queries in total), we improve accuracy on both types: see the bottom half of Table 10. 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 the Dutch news mentions are ignored.

5.2 Sessions

http://www.topsy.com

Based on our query types and initial data observations, we propose four different session types:

Family session In a family session, a user issues several queries trying to find information about relatives. This session type will mainly consist of low-profile queries, with repetitive use of the same last name(s).

Event session Events (e.g., in the news) usually have several main players involved. The event session is centered around an event, and its queries relate to this event. Most of the queries in this session will be of the event-based high-profile type.

Spotting session Users try to “spot” celebrities in the real world, and do the same in an online environment. When trying to spot several celebrities in one session, we have a spotting session. Here, most queries in the session are of the regular high-profile type.

Polymorous session For sessions that show a mixture of the three above mentioned types, or that contain various low-profile queries without clear relation between them, we have a polymorous session type.

We manually annotated 1,005 sessions. Since we are unable to determine a session type for one query sessions, we remove the 540 sessions that contain just one query, leaving us with 465 annotated multiple query sessions. The counts and percentages of the session types in our sample are listed in Table 11. Most users engage...
5.3 Users

We select a random sample of 412 users and manually look at their characteristics and typology. We discern the following types.

**Monitor** To track their own (or someone else’s) web presence, monitors regularly return to the people search engine with the same query (associated characteristics: recurring queries, large interval between queries, few out clicks).

**Spotter** Based on the physical activity of spotting celebrities in cities, spotters use people search engines to spot celebrities online (characteristics: high frequency queries, high-profile queries, out clicks to social media and multimedia).

**Polymer** Has no clear-cut behavior; combines various session and query types.

In our annotated sample, we observe that for 320 users we cannot determine their type. As indicated in Table 5, we can only ascertain more than one query for 21.7% percent of the users. So, for the bulk of the users we observe a single query, making the classification of these difficult if not impossible. For the remainder we find that 69 users are polymers, 22 are monitors, and 1 is a follower.

6. DISCUSSION AND IMPLICATIONS

In this section we take the results of our people search log analysis, and discuss observations with regard to people search aspects, and pointers to interesting research directions.

**Keywords.** As mentioned in Section 4, the search engine offers users the opportunity to add keywords to their search. Since this field is not part of the standard search interface, its usage is limited: about 4% of all person queries contain keywords, the bulk of which are single terms. Table 12 shows the ten most popular keywords; a quick look reveals that many of the keywords are Dutch cities or keywords indicating the type of result the searches want to see. To investigate the use of the keyword field in more detail, we take a sample of 250 keywords and manually classify these. Table 13 lists the classes we identified from this sample. We see that most key-

---

Table 12: 10 most popular keywords.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Count</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amsterdam</td>
<td>4,733</td>
<td>Dutch city</td>
</tr>
<tr>
<td>Com</td>
<td>3,451</td>
<td>top level domain</td>
</tr>
<tr>
<td>Jan</td>
<td>3,009</td>
<td>January</td>
</tr>
<tr>
<td>Rotterdam</td>
<td>2,782</td>
<td>Dutch city</td>
</tr>
<tr>
<td>Foto</td>
<td>2,519</td>
<td>photo</td>
</tr>
<tr>
<td>Facebook</td>
<td>2,411</td>
<td>social networking site</td>
</tr>
<tr>
<td>Anonymous</td>
<td>2,377</td>
<td>name of the search engine</td>
</tr>
<tr>
<td>WWW</td>
<td>2,265</td>
<td></td>
</tr>
<tr>
<td>Profiel</td>
<td>2,135</td>
<td>profile</td>
</tr>
<tr>
<td>Groningen</td>
<td>2,069</td>
<td>Dutch city</td>
</tr>
</tbody>
</table>

**Person name disambiguation.** The task of person name disambiguation is an interesting and active research topic (see, e.g., [1–3]), and it is an important and very challenging aspect of people search. The same name can refer to many different persons: data from 1990 suggests that in the U.S., only 90,000 different names are shared by 100 million persons [3]. Clearly, returning relevant results for person name queries is the challenging.

Our analysis so far revealed several aspects to person name disambiguation: First, as we saw in the previous paragraph, users use the keyword field to give pointers on how to disambiguate people sharing the same name. To this end they mainly enter a location or activity (job, hobby); these two types of keywords combined cover 33% of all keywords. Second, we find evidence of person name disambiguation in the out clicks. Consider the number of different profiles users go to after searching for the same name; Table 14 shows the person names with the largest number of different profiles clicked (Facebook profiles left, LinkedIn profiles right). Except for “Joran van der Sloot” (a high-profile person with many fake profiles and hate groups), all names are very common Dutch names. To support this claim, Table 15 lists the most common Dutch last names:

---

Table 13: Keyword classes for people search, their frequency, and examples.

<table>
<thead>
<tr>
<th>Keyword class</th>
<th>Percentage</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locations</td>
<td>22.8%</td>
<td><em>Amsterdam, Rotterdam, ...</em></td>
</tr>
<tr>
<td>Person names</td>
<td>15.6%</td>
<td><em>Maaike, Peter, Snelders, ...</em></td>
</tr>
<tr>
<td>Result types</td>
<td>13.6%</td>
<td><em>Facebook, pictures, website, ...</em></td>
</tr>
<tr>
<td>Activities</td>
<td>10.4%</td>
<td><em>gardener, swindler, soccer, ...</em></td>
</tr>
<tr>
<td>Date</td>
<td>3.2%</td>
<td><em>November, Monday, jan, ...</em></td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>34.4%</td>
<td></td>
</tr>
</tbody>
</table>

---

3The name of the search engine in Table 12 has been hidden to preserve anonymity.

Table 14: Person names with most unique Facebook (left) and LinkedIn (right) results clicked.

<table>
<thead>
<tr>
<th>Name</th>
<th>Count</th>
<th>Name</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joran van der Sloop</td>
<td>18</td>
<td>Herman de Vries</td>
<td>11</td>
</tr>
<tr>
<td>Jeroen de Vries</td>
<td>14</td>
<td>Michiel Bakker</td>
<td>11</td>
</tr>
<tr>
<td>Rob van Dijk</td>
<td>14</td>
<td>Nicole Bakker</td>
<td>11</td>
</tr>
<tr>
<td>Marieke de Jong</td>
<td>14</td>
<td>Nynke de Vries</td>
<td>10</td>
</tr>
<tr>
<td>Peter de Vries</td>
<td>13</td>
<td>Mirjam de Vries</td>
<td>10</td>
</tr>
<tr>
<td>Peter van Dijk</td>
<td>13</td>
<td>Marjan de Jong</td>
<td>10</td>
</tr>
<tr>
<td>Peter Visser</td>
<td>12</td>
<td>Annemieke de Vries</td>
<td>10</td>
</tr>
<tr>
<td>Saskia de Vries</td>
<td>12</td>
<td>Arjan Visser</td>
<td>10</td>
</tr>
<tr>
<td>Karin de Jong</td>
<td>12</td>
<td>Bas Alberts</td>
<td>10</td>
</tr>
<tr>
<td>Marieke de Vries</td>
<td>12</td>
<td>Frank Driessen</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 15: Ten most common last names in the Netherlands.

<table>
<thead>
<tr>
<th>Name</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>De Jong</td>
<td>0.53%</td>
</tr>
<tr>
<td>Jansen</td>
<td>0.46%</td>
</tr>
<tr>
<td>De Vries</td>
<td>0.45%</td>
</tr>
<tr>
<td>Van der Berg</td>
<td>0.37%</td>
</tr>
<tr>
<td>Van Dijk</td>
<td>0.35%</td>
</tr>
<tr>
<td>Bakker</td>
<td>0.35%</td>
</tr>
<tr>
<td>Janssen</td>
<td>0.34%</td>
</tr>
<tr>
<td>Visser</td>
<td>0.31%</td>
</tr>
<tr>
<td>Smit</td>
<td>0.27%</td>
</tr>
<tr>
<td>Meijer</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

Relationship finding. Current research in entity retrieval focuses, among other things, on finding relationships between entities, or finding related entities [4, 5, 7]. Our analysis of people search logs show that users are indeed interested in finding combinations of people or finding the relationship between people. As observed in the “keyword” paragraph, users of the people search engine currently use the keyword field to achieve this goal. An interesting example is the female first name “Maaïke,” which is frequently used as a keyword. Table 16 shows person name queries with which this keyword is being used, and explains the relation between the two people. Note that, although we are looking at the same name (Maaïke), searchers seem to be referring to different people. Improvements in the interface and in search algorithms should, in the future, facilitate searching for combinations of people or for relationships between persons.

Table 16: Queries issued with person (first) name “Maaïke” as keyword, and the relation between query and keyword.

<table>
<thead>
<tr>
<th>Queried person</th>
<th>Relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ben Saunders</td>
<td>Maaïke is ex-girlfriend of talent show participant Ben</td>
</tr>
<tr>
<td>Sietske Hoekstra</td>
<td>Maaïke and Sietske are relatives</td>
</tr>
<tr>
<td>Jaap Siewertsz van Reesema</td>
<td>Jaap and Maaïke were both finalists of a talent show</td>
</tr>
</tbody>
</table>

Single-term queries. As mentioned in Section 4.2, we encountered many log entries with only one term in the query (4% of all queries). These single-term queries are likely to be used in two ways: (i) last name search, where the goal is to explore people that share the same last name, and (ii) first name search, aimed at finding the right person and thus that person’s full name. About 16.6% of the single-term queries have at least one out click, which is one percent lower than for all queries (17.6%). However, when we look at the top 10 queries with most out clicks, six of these queries are single-term queries. To explore this finding in more detail, we plot the percentage of queries with their number of out clicks (Figure 10): we binned the out clicks to make the difference apparent, and split the data over two plots for the same reason: The left plot shows bins for 2, 3–5, and 6–10 out clicks, and the right plot those for 11–20, 21–30, and > 30. As we can see, the tail of the single-term queries (gray columns) is “fatter” than for multiple term queries, indicating that users are more likely to try various results for single-term query than for multiple term queries. Users seem to use just one term, to start an exploration of the results. Future work on interfaces and algorithms should account for the fact that users use exploratory search for people search too, and again, person name disambiguation is an important aspect here.

Figure 10: Percentage of queries (y-axis) with their number of out clicks (binned, x-axis) for single-term queries (gray) and multiple term queries (white).

Session detection. In our current setup, we used a (rather long) time-out between actions to detect sessions. From our analysis at the session level (Section 5.2) we observe that we have many polymorphous, and a significant portion of these sessions contain “subsessions” (e.g., a sequences of (almost) identical person names, or some event-related queries, followed by searches for relatives). It would be interesting to apply more advanced session detection methods, based on, for example, query types or overlap in content, to the log data. Offering smarter session detection also allows research into session prediction (i.e., given an initial observation of two or more queries, can predict the session type and suggest follow-up queries).

7. CONCLUSION

In this paper 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. Our results provide hints for future research in terms of both algorithms and interfaces for people search (or entity search in general).

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.

Our analysis of search behavior in people search has revealed many directions for future work, including (i) improved session detection methods for people search, (ii) person name disambiguation, (iii) query prediction within sessions, and (iv) a longitudinal study of users.
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 nr 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


