A Study of Blog Search

Gilad Mishne Maarten de Rijke

Informatics Institute, University of Amsterdam,
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
gilad,mdr@science.uva.nl

Abstract. We present an analysis of a large blog search engine query log, exploring a number of angles such as query intent, query topics, and user sessions. Our results show that blog searches have different intents than general web searches, suggesting that the primary targets of blog searchers are tracking references to named entities, and locating blogs by theme. In terms of interest areas, blog searchers are, on average, more engaged in technology, entertainment, and politics than web searchers, with a particular interest in current events. The user behavior observed is similar to that in general web search: short sessions with an interest in the first few results only.

1 Introduction

The rise on the Internet of blogging—the publication of journal-like web page logs, or blogs—has created a highly dynamic and tightly interwoven subset of the World Wide Web [10]. The blogspace (the collection of blogs and all their links) is giving rise to a large body of research, both concerning *content* (e.g., Can we process blogs automatically and find consumer complaints and breaking reports about vulnerabilities of products?) and *structure* (e.g., What is the dynamics of the blogspace?). A variety of dedicated workshops bear witness to this burst of research activity around blogs; see e.g., [20].

In this paper we focus on another aspect of the blogspace: searching blogs. The exponential rise in the number of blogs from thousands in the late 1990s to tens of millions in 2005 [3, 18, 19] has created a need for effective access and retrieval services. Today, there is a broad range of search and discovery tools for blogs, offered by a variety of players; some focus exclusively on blog access (e.g., Blogdigger [2], Blogpulse [3], and Technorati [18]), while web search engines such as Google, Yahoo! and AskJeeves offer specialized blog services.

The development of specialized retrieval technology aimed at the distinct features of the blogspace is still in its early stages. We address a question whose answer should help inform these efforts: How does blog search differ from general web search? To this end we present an analysis of a blog search engine query log. We study the intent of blog searches, find out what the user behavior of blog searchers is, and determine the profile of blog searchers in terms of query types.

In the next section we briefly survey related work that guided us in our study. In Section 3 we describe the data used for our analysis and provide basic descriptive statistics about it. In Section 4 we analyze the queries in terms of

user intent. Then, in Section 5 we classsify queries by category, and Section 6 is devoted to an analysis of the sessions in our data. Section 7 wraps up the paper with conclusions, discussions, and future work.

2 Related Work

At the time of writing, no published work exists on blog search engine logs. However, work on search engine log analysis is plentiful: a recent survey paper describes a large body of related work in this area published during the last 10 years [5]. Our work was particularly inspired by some of this work. Most notably, work by Broder [4] on classifying search requests of web users using the (then popular) AltaVista search engine, as well as the follow-up work by Rose and Levinson [13] with Yahoo! data, inspired our attempts at classifying the hidden intents behind blog searches.

In terms of statistical analysis, our work is influenced by one of the first large-scale studies of search logs available to the public, performed by Silverstein et al. [16], and the numerous analyses published by Jansen, Spink et al., which targeted various angles of search engine usage (e.g., [4, 7, 8]), analyzing data not accessible to the majority of the research community.

Finally, our query categorization work was influenced by work done by Pu and Chuang [12], and by Beitzel et al. [1]. Some of the query categorization methods used for the 2005 KDD Cup [9] (which targeted query classification) are similar to our categorization approach, which was developed in parallel.

3 Dataset

Our data consists of the full search log of Blogdigger.com for the month of May 2005. Blogdigger.com is a search engine for blogs and syndicated content feeds (such as RSS and ATOM feeds) that has been active since 2003, being one of the first fully-operational blog search engines. Recently, as major web search engines introduced their capabilities for blog search, it is gradually becoming a second-tier engine. Nevertheless, Blogdigger.com provides some unique services such as local-based search and media search, which attract a relatively large number of users to it. Our log contains both queries sent to Blogdigger's textual search engine and queries sent to its media search engine—a service for searching blog posts (and additional syndicated content) containing multimedia files or links.

Blogdigger.com—like other major blog search engines—serves both ad-hoc queries and filtering queries. Ad-hoc queries originate from visitors to the search engine's web site, typing in search terms and viewing the result pages, in a similar manner to the typical access to web search engines. A user who is interested in continuous updates about the results of a specific query can subscribe to its results: in practice, this means she is adding a request for a machine-readable version of the query results to a syndicated content aggregator (e.g., an RSS reader) she is running. The query results will then be periodically polled; each of these polls is registered as a filtering query in the search log.

Table 1 contains statistics about our log file. Due to the large percentage of duplicates typical of query logs, we provide statistics separately for all queries

	All queries	Unique queries
Number of queries	1,245,903	116,299
Filtering queries	1,011,962 (81%)	34,411 (30%)
Ad-hoc queries	233,941 (19%)	81,888 (70%)
Text queries	1,016,697 (82%)	50,844 (44%)
Media queries	229,206 (18%)	65,455 (56%)
Link queries	2,967 (<1%)	562 (<1%)
Mean terms/filtering query	1.96	1.98
Mean terms/ad-hoc query	2.44	2.71

Table 1: Search log size and breakdown.

and for the set of unique queries in the log (i.e., exact repetitions removed). While filtering queries make up the bulk of all queries, they constitute a relatively small amount of unique terms, and the majority of unique queries originate from ad-hoc sessions. The mean terms/query number for (all) ad-hoc queries is comparable to the mean terms/query numbers reported in the literature for general web search (2.35 [16], 2.21 [6], 2.4–2.6 [17], and 2.4 [7]); while the mean terms/query number for filtering queries appears somewhat smaller (1.96), a closer examination reveals that this difference is caused to a large extent by two specific clients; excluding these outliers, the mean terms/query for filtering queries is 2.5, similar to that of ad-hoc ones.[1]

4 Types of Information Needs

Next, we analyze the information needs in the blogspace, partitioning the queries into two broad classes. Following Broder's influential work [4], queries submitted to web search engines are generally grouped into three classes: *informational* (find information about a topic), *navigational* (find a specific web site), and *transactional* (perform some web-mediated activity). This may not be an appropriate classification for queries submitted to blog search engines—clearly, transactional queries are not a natural category for blog search, and a user searching for a particular site, or even a particular blog (i.e., submitting a navigational query) would not necessarily use a blog search engine, but rather a general-purpose web engine. Our working hypothesis, then, is that the majority of blog queries are *informational* in nature, and a scan of the search log confirms this.

Given this assumption, is it possible to identify different types of informational queries submitted to a blog search service? Ideally, this would be done using a user survey—in a manner similar to the one performed by Broder [4]. Unfortunately, we only have retrospective access to the submitted queries, with no possibility of conducting such a survey. However, Broder's work shows a fairly good correlation between the results of his survey and manual classification of a subset of the queries, leading us to assume that an analysis of the query types based on an examination of the queries in our data is worthwhile.

[1] The two clients issued large amounts of queries in fixed, short intervals; the queries appear to have been taken from a dictionary in alphabetical order and are all single words, pushing down the mean number.

First, we examined a random set of 1000 queries, half of which were ad-hoc queries and half filtering ones, so as to discover likely query types. We observed that the majority of the queries—52% of the ad-hoc ones and 78% of the filtering ones—were named entities: names of people, products, companies, and so on. Of these, most belonged to two types: either very well-known names ("Bush", "Microsoft", "Jon Stewart"), or almost-unheard-of names, mostly names of individuals and companies.[2] An additional popular category of named entities was location names, mostly American cities. Of the non-named-entity queries, most queries—25% of the ad hoc queries and 18% of the filtering ones—consisted of high-level concepts or topics, such as "stock trading", "linguists", "humor", "gay rights", "islam" and so on; the filtering queries of this type were mostly technology-related. The remainder of the queries consisted of adult-oriented queries (almost exclusively ad-hoc queries), URL queries, and other queries for which we could not find specific characteristics.

Next, we examined the 400 most common queries (again, half ad-hoc and half filtering), to find out whether the query types there differ from those found in "the long tail." While the types remained similar, we witnessed a different distribution: 45% of the ad-hoc queries and 66% of the filtering queries were named entities; concepts and technologies consisted of an additional 30% of top ad-hoc queries and 28% of filtering ones. Adult-oriented ad-hoc queries were substantially more common in top ad-hoc queries than in the random set.

Consequently, our hypothesis regarding the intents of blog searchers divides the searches into two broad categories:

- **Context Queries**: The purpose of these queries is to locate contexts in which a certain name appears in the blogspace: what bloggers say about it. Most of the named entity queries have this intent; the well-known names might be entities in which the searcher has an ongoing interest (such as politicians) or products she is researching, whereas the lesser-known names are typically vanity searches, or searches for contexts of entities which constitute part of the searcher's closer environment (an employer, organization in which the searcher is a member, etc).
- **Concept Queries**: With these queries the searcher attempts to locate blogs or blog posts which deal with one of the searcher's interest areas, or with a geographic area that is of particular interest to the searcher (such as blogs authored by people from his home town). Typical queries of this type are the various high-level concepts mentioned earlier, as well as location names.[3]

Table 2 shows a breakdown of both the random set and the top-query set according to query type, for ad-hoc and filtering queries separately. For this breakdown, named-entity queries (except location names) were considered as context queries; high-level areas of interest and location names were considered concept queries.

[2] The prevalence of the named entity was confirmed using Google hit counts: well-known names typically had millions of hits; unknown names had few if any.

[3] These queries are somewhat similar to distillation queries as defined by TREC, with target results being blogs rather than websites.

	Top queries		Random queries	
Class	**Ad-hoc**	**Filtering**	**Ad-hoc**	**Filtering**
Context	39%	60%	47%	73%
Concept	36%	34%	30%	23%
Other	25%	6%	23%	4%

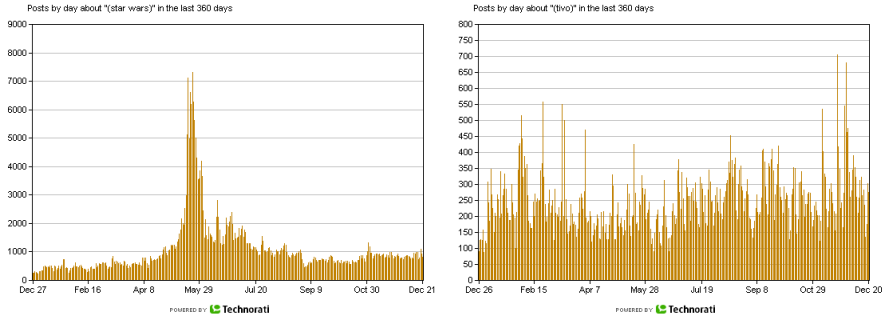
Table 2: Query classes: the top 400 queries vs. a random sample of 1000 queries.

While examining the top queries, we observed an interesting phenomenon which we did not witness in the random set: many of the queries were related to events which were "in the news" at the time of the log. This supports the assumption that blogs are conceived as a source of information and commentry about current events [11]. To quantify the number of news-related queries, we used two independent methods. First, a human decided, for each query, whether it was news-related. This was done by studying the terms in the query, and attempting to locate events related to it that happened during May 2005, the period covered by the log. The second method was an automated one: we obtained daily word frequencies of the terms appearing in the query as reported by Technorati, for the entire year of 2005. Terms which had substantial peaks in the daily frequency counts during May 2005 were considered related to news; sample daily frequencies over the entire year of 2005 are shown in Figure 1. The agreement between our two methods (kappa) was 0.72.

We found that 20% of the top ad-hoc queries and 15% of the top filtering ones are news-related; in the random set, news-related queries were substantially less frequent, amounting to 6–7% of both ad-hoc and filtering queries.

In sum, blog searches have different intents than typical web searches, suggesting that the primary targets of blog searchers are tracking references to named entities and identifying blogs or posts which focus on a certain concept; in addition, searches related to current events are substantially more common in blog searches than in web searches, in particular in the popular queries.

Fig. 1: Sample daily frequency counts in 2005. (Left): a news-related query ("Star Wars"). (Right): a non-news-related query ("Tivo").



Ad-hoc	Filtering	Web
filibuster	Lotus Notes	American Idol
Blagojevich	Daily Show	Google
sex	microcontent	Yahoo
porn	information architecture	eBay
blogdigger	MP3	Star Wars
Madagascar	Streaming	Mapquest
RSS	Google	Hotmail
adult	Wayne Madsen	Valentine's day
Google	Tom Feeney	NASCAR
nude	Clint Curtis	hybrid cars
MP3	digital camera	MP3 players
Los Angeles	DMOZ	NFL
test	desktop search	dictionary
China	manga	Paris Hilton
3G	RSS	Michael Jackson
Star Wars	Abramoff	Hillary Clinton
IBM	knowledge management	heartburn
blog	government	Lohan
music	restaurant	flowers
Bush	information management	Xbox 360

Table 3: Top 20 queries. (Left): Ad-hoc blog queries. (Center): Filtering blog queries. (Right): Web queries.

5 Popular Queries and Query Categories

Next, we provide a brief overview of the top queries posted, describe a categorization method, and apply this method to the queries in the log, trying to construct the profile of topics blog searchers are interested in.

Simply counting the number of times a query appears in our log yields misleading results regarding the most popular queries. This is due to the fact that the majority of the search requests are automated, and are repeated at regular intervals; agents issuing these queries with high refresh rates will create a bias in the query counts. As a result, we measure the popularity of a query not according to the number of occurrences, but according to the number of different users submitting it. As a key identifying a user we use a combination of the IP address and the user agent string (more details on user identification are given in Section 6).

The most popular queries in the log are shown in Table 3, columns 1 and 2, separately for ad-hoc and filtering queries.

5.1 Comparison to Web Queries To compare the popular queries submitted to blog search engines with those sent to general web search engines, we obtained a set of 3.5M queries submitted to Dogpile/Metacrawler, a second-tier general web search engine,[4] during May 2005—the same timespan as our blog search log. The top 20 queries from this source are listed in Table 3, column 3.

[4] This is a metasearch engine, submitting queries to a number of other engines such as Google and Yahoo! and aggregating the results.

Query: **24** *Yahoo! category*: /Entertainment/Television Shows/Action and Adventure/24 *Froogle category*: /Books, Music and Video/Video/Action and Adventure
Query: **Atkins** *Yahoo! category*: /Business and Economy/Shopping and Services/Health/Weight Loss/Diets and Programs/Low Carbohydrate Diets/Atkins Nutritional Approach *Froogle category*: /Food and Gourmet/Food/Snack Foods
Query: **Evolution debate** *Yahoo! category*: /Society and Culture/Religion and Spirituality/Science and Religion/Creation vs. Evolution/Intelligent Design *Froogle category*: /Books, Music and Video/Books/Social Sciences
Query: **Vioxx** *Yahoo! category*: /Health/Pharmacy/Drugs and Medications/Specific Drugs and Medications/Vioxx, Rofecoxib *Froogle category*: /Health and Personal Care/Over-the-Counter Medicine

Table 4: Example queries and categories.

Some differences between the query lists are clear: the web queries contain many large web sites (Yahoo!, eBay, Hotmail, and so on), perhaps because for some users, the distinction between the search input box and the browser's address bar is unclear. Additionaly, the top blog queries seem to contain a somewhat higher percentage of political and technology-related queries; this strengthens our findings in Section 5.2 regarding the top interests of bloggers.

Other differences between blog queries and web queries require examining more than a small number of top queries. Comparing the most popular 400 queries from both sources, we observed a substantially higher rate of named-entity queries within blog queries than in web queries. As mentioned in Section 4, 45% of ad-hoc blog queries and 66% of the filtering queries were named entities; in comparison, only 33% of the top 400 web queries were named entities, many of which were website names. This suggests that blog searchers—especially those registering filtering queries—are more interested in references to people, products, organizations or locations than web searchers.

As noted earlier, we found a relatively large amount of new-related queries among top blog queries; this type of queries proved to be fairly uncommon in general web search engines, accounting for less than 8% of the top 400 queries, and less than 2% of 400 random ones.

An final difference between the query lists is the presence of very detailed information needs (such as factoid questions) in the web query log: such queries were not found among the blog queries. Finally, as is the case with web searches, adult-oriented queries are an important area of interest for ad-hoc blog searchers; however, these are nearly non-existent in filtering queries.

5.2 Query Categories Current approaches to automatic categorization of queries from a search log are based on pre-defining a list of topically categorized terms, which are then matched against queries from the log; the construction of this list is done manually [1] or semi-automatically [12]. While this approach achieves high accuracy, it tends to achieve very low coverage, e.g., 8% of unique queries for the semi-automatic method, and 13% for the manual one.

We take a different approach to query categorization, substantially increasing the coverage but (in our experience) sustaining high accuracy levels: our approach relies on external "categorizers" with access to large amounts of data.

We submit every unique query in our corpus as a search request to two category-based web search services: Yahoo! Directory (http://dir.yahoo.com) as well as Froogle (http://froogle.google.com). The former is a manually-categorized collection of web pages, including a search service for these web pages; the latter is an online sales search service. We use the category of the top page retrieved by the Yahoo! Directory as the "Yahoo! Category" for that query, and the top shopping category offered by Froogle as its "Froogle Category;" while the Yahoo! Category is a topical category in the traditional sense, the Froogle Category is a consumer-related one, possibly answering the question "if there is potential commercial value in the query, what domain does it belong to?" In spirit, this is similar to the usage of the Open Directory Project to classify web pages by category (e.g., in [15]), except that we classify terms, not URLs. (Similar methods for query classification have been developed in parallel for the KDD 2005 Cup [14].)

The coverage achieved with this method is fairly high: in total, out of 43,601 unique, non-media queries that were sent to Yahoo! and Froogle, 24,113 (55%) were categorized by Yahoo! and 29,727 (68%) by Froogle. Some queries were not categorized due to excessive length, non-standard encodings, and other technical issues, so the coverage over common queries is even higher. An examination of the resulting categories shows high accuracy, even for queries which are very hard to classify with traditional methods, using the query words only. Table 4 lists some examples of queries along with their corresponding categories.

Figure 2(Left) shows a breakdown of the top Yahoo! categories for ad-hoc and filtering queries. Taking into account that "Regional" queries often refer to news-related events, we witness again that current events are a major source of interest for blog searchers. A similar breakdown of the top Froogle categories is given in Figure 2(Center), indicating that most queries which can be related to products deal with intellectual property, such as movies and books.

An added benefit of the Yahoo! and Froogle categories is their hierarchical nature: this enables us to not only examine the most frequent category, but also to evaluate the breakdown of subcategories within a given category. For example, Figure 2(Right) shows the most frequent Yahoo! subcategories within the filtering "Entertainment" queries and the ad-hoc "Business" queries. As is the case for general web searches, adult-oriented searches are the top category for commercial queries, followed by technology-related queries and financial issues. In the entertainment domain, music clearly dominates the scene.

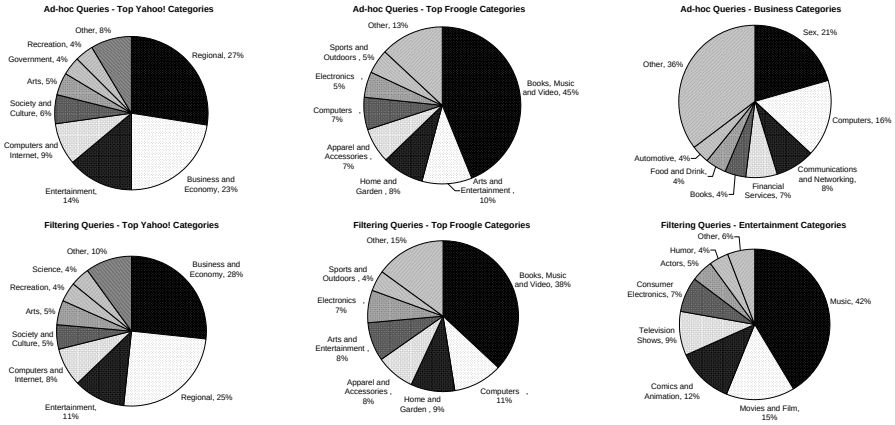
We conclude that in terms of interest areas, blog searchers are more engaged in technology and politics than web searchers, with a noticeable interest in named entities: names of people, brands, companies, and so on.

6 Session Analysis

Next we analyze the query sessions in the log, examining issues such as the amount of queries submitted in a session and the number of viewed results.

Our log does not contain full session information: we do not know how long the user spent examining the results, and which result links she followed. How-

Fig. 2: (Left): Top Yahoo! categories. (Center): Top Froogle categories. (Right): Sample subcategories.



ever, since some identification of the user is given for each query in the log in the form of IP address and user agent string, it is possible to group the queries by sessions and to perform a basic analysis of these.

Before describing our approach to session recovery and discussing characteristics of the extracted sessions it is important to note the difference between sessions that contain ad-hoc searches and sessions that contain filtering searches. The former are similar to standard web search sessions, and consist of different queries that a user submitted to the search engine during her visit. These different queries include, in many cases, reformulations of a query, or highly-related terms which indicate the user is trying to collect more information regarding her interest. In contrast, "sessions" containing filtering searches are actually sets of queries registered by the same user: in practice, they are not queries submitted during a single visit to the search engine, but a list of queries the same user expressed ongoing interest in, possibly added over a long period of time.

6.1 Recovering Sessions and Subscription Sets We assume two queries to belong to the same session if the following conditions hold: (1) The queries originate from the same IP address; (2) The user agent string of the two queries is identical, and (3) The elapsed time between the queries is less than k seconds, where k is a predefined parameter.

The main drawback of this method is its incompatibility with proxy servers: queries originating from the same IP address do not necessarily come from the same user: they can also be sent by different users using the same proxy server; this is a common scenario in certain environments, such as companies with a single internet gateway. While the usage of the user agent string reduces the chance of mistaking different users for the same one, it does not eliminate it completely. Having said that, anecdotal evidence suggests that the recovered sessions are in fact "real" sessions: the conceptual and lexical similarity between queries in the same session is high for the vast majority of sessions we examined. Additional evidence for the relative robustness of this method can be seen in

Type	Queries
Session	autoantibodies ; autoantibodies histamine ; histamine
Session	firmware dwl 2000 ap+ ; dwl 2000 ap+ ; dwl-2000 ap+
Subscription set	"XML Tag Monitor Report" ; "XML Search Selector"
Subscription set	imap ; imap gmail ; Thunderbird IMAP ; imap labels ; rss email ; thunderbird label ; imap soap

Table 5: Example sessions and subscription sets; queries belonging to the same session or subscription set are separated by semicolons.

		Blog queries Sessions	Blog queries Subscriptions	Web queries Sessions [16]
Length	Mean	1.45	1.53	2.02
	Length 1	70.2%	75.8%	77.6%
	Length 2	20.9%	13.7%	13.5%
	Length $\geq$3	8.8%	10.4%	9.9%
Page views	Mean	1.09	N/A	1.39
	1 result page	94.9%	N/A	85.2%
	2 result pages	3.4%	N/A	7.5%
	3 or more pages	1.7%	N/A	7.3%

Table 6: (Top): Session and subscription set lengths (number of unique queries). (Bottom): Result page views for ad-hoc queries, per session.

the fact that, when used on the set of all queries, it produces less than 0.5% "mixed sessions" – sessions containing both ad-hoc and filtering queries, which are unlikely to be a real session.

We performed our analyses independently for ad-hoc and filtering queries; to avoid confusion, we use the term "sessions" only for ad-hoc sessions—which are indeed sessions in the traditional sense; for filtering sessions, we use the term "subscription sets" (which denotes lists of filtering queries done by the same user within a short timeframe).

6.2 Sessions and Subscription Sets We experimented with various values of k; manual examination of the recovered sessions suggests that values between 10 and 30 seconds yield the most reliable sessions for ad-hoc queries. For filtering queries, the session time is much shorter, in-line with intuition (since the queries are automated): reliable sessions are found with k values of 2–5 seconds. The thresholds were set to 20 seconds for sessions and 5 seconds for subscription sets; this produces 148,361 sessions and 650,657 subscription sets.

Many sessions and subscription sets contain simple reformulations such as different uses of query operators; others are composed of related terms, and yet others consist of seemingly unrelated queries, matching different interests of the same user. Table 5 provides example sessions and subscription sets, and Table 6(Top) details statistics about the session length (the number of unique queries per session), comparing our findings to those for general web searches [16].

The short session length is similar to the one observed in web search engines, e.g., in [16]. While subscription sets also exhibit a short length on average, the actual lengths of the sets vary much more than those of sessions—as can be seen from the much higher variance (5.10 for subscriptions vs. 0.87 for sessions). Users

may subscribe to any amount of queries, and, in our data, some users registered as much as 20 queries.

For ad-hoc queries, an additional interesting aspect is the number of result pages the user chooses to view (each containing up to 10 matches). As with web searches, we find that the vast majority of users view only the first result page: see the detailed breakdown in Table 6(Bottom), again comparing our findings to those presented for general web searches in [16]. While there is a statistically significant difference between the two samples (blog sessions vs web sessions), the bottom line is similar: most users do not look beyond the first set of results.[5]

In sum, while we found that query types in the blogspace differ from the types of queries submitted to general web search engines, we discovered a very similar user behavior for issuing queries and viewing their results.

7 Conclusions

We presented a study of a large blog search engine log, aimed at analyzing the type of queries issued by users in this domain, the user behavior in terms of amount of queries and page views, and the categories of the queries. The query log covers an entire month, and contains both ad-hoc and filtering queries.

Our main finding in terms of query types is that blog searches fall into two broad categories—context queries, attempting to track the references to various named entities within the blogspace, and concept queries, aimed at locating blogs and blog posts which focus on a given concept or topic. The distribution of these types differs between ad-hoc and filtering queries, with the filtering ones being more context-oriented. In addition, we found that blog searches tend to focus on current events more than web searches.

As to user behavior, the behavior observed is similar to that in general web search engines: users are typically interested only in the first few results returned, and usually issue a very small number of queries in every session.

Finally, using external resources to categorize the queries, we uncovered a blog searcher profile which is substantially more concentrated on news (particularly politics), entertainment, and technology than the average web searcher. Hence, it may be useful for blog search engines to identify and exploit named entities (and parts of them), especially in the domains mentioned above.

Acknowledgments We thank Blogdigger.com, and especially Greg Gershman and Michael Miller for generously providing the data set without which this research could not have been conducted. Many thanks also to our anonymous reviewers for useful suggestions. This work was supported by the Netherlands Organization for Scientific Research (NWO) under project number 220-80-001.

[5] Also, the number of page views for web searches is constantly decreasing, as search engine technology is improving and more relevant documents appear in the first few results.

8 References

[1] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In *Proceedings SIGIR '04*, pages 321–328, New York, NY, USA, 2004. ACM Press.

[2] Blogdigger, 2005. Search engine for RSS and blogs. URL: http://blogdigger/, accessed January 2006.

[3] Blogpulse, 2005. Automated trend discovery system for blogs. URL: http://blogpulse.com/, accessed January 2006.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] F. M. Facca and P. L. Lanzi. Mining interesting knowledge from weblogs: a survey. *Data Knowl. Eng.*, 53(3):225–241, 2005.

[6] B. Jansen and U. Pooch. Web user studies: a review and framework for future work. *J. American Society of Science and Technology*, 52(3):235–246, 2001.

[7] B. Jansen and A. Spink. An analysis of Web searching by European AlltheWeb.com users. *Inf. Process. Manag.*, 41(2):361–381, 2005.

[8] B. J. Jansen, A. Spink, and T. Saracevic. Real life, real users, and real needs: a study and analysis of user queries on the web. *Inf. Process. Manag.*, 36(2): 207–227, 2000.

[9] KDD Cup, 2005. URL: http://kdd05.lac.uic.edu/kddcup.html, accessed January 2006.

[10] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 568–576, New York, NY, USA, 2003. ACM Press.

[11] M. Ludtke, editor. *NIEMAN REPORTS: Journalist's Trade - Weblogs and Journalism*, volume 57,3. Bob Giles, 2003.

[12] H. T. Pu and S. L. Chuang. Auto-categorization of search terms toward understanding web users' information needs. In *ICADL 2000: Intern. Conference on Asian Digital Libraries*, 2000.

[13] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings WWW '04*, pages 13–19, New York, NY, USA, 2004. ACM Press.

[14] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Q2c@ust: Our winning solution to query classification in kdd cup 2005. *SIGKDD Exploration*, 2006.

[15] X. Shen, S. Dumais, and E. Horvitz. Analysis of topic dynamics in web search. In *WWW '05: Proceedings of the 14th intern. conf. on World Wide Web*, 2005.

[16] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[17] A. Spink, B. Jansen, D. Wolfram, and T. Saracevic. From e-sex to e-commerce: Web search changes. *IEEE Computer*, 35(3):107–111, 2002.

[18] Technorati, 2005. Blog tracking service. URL: http://technorati.com/, accessed January 2006.

[19] Technorati, 2005. State of the Blogosphere according to Technorati. URL: http://www.sifry.com/alerts/archives/000298.html/, accessed January 2006.

[20] Weblogging Ecosystem, 2005. WWW 2005 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics. URL: http://www.blogpulse.com/www2005-workshop.html, accessed January 2006.