Query Logs Alone are not Enough

Carrie Grimes Diane Tang Daniel M. Russell
Google
1600 Amphitheatre Pkwy
Mountain View, CA 94043
{cgrimes, diane, drussell}@google.com

ABSTRACT
The practice of guiding a search engine based on query logs observed from the engine's user population provides large volumes of data but potentially also sacrifices the privacy of the user. In this paper, we ask the following question: Is it possible, given rich instrumented data from a panel and usability study data, to observe complete information without routinely analyzing query logs? What unique benefits to the user could hypothetically be derived from analyzing query logs? We demonstrate that three different modes of collecting data, the field study, the instrumented user panel, and the raw query log, provide complementary sources of data. The query log is the least rich source of data for individual events, but has irreplaceable information for understanding the scope of resources that a search engine needs to provide for the user.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process*; H.4.m [**Information Systems Applications**]: Miscellaneous.

General Terms
Measurement, Experimentation

Keywords
Web search, information retrieval, user goals, query classification, logs analysis.

1. INTRODUCTION
The goal of a search engine is to have indexed the right pages for any user search, and to effectively retrieve them in response to the query supplied by the user. This goal applies equally to traditional web search results and to other potential corpora such as commercial ads, images, news, etc. In order to accomplish these goals, the search engine must:

- "Understand" the intent of the search query
- Have indexed the correct pages, documents, or materials
- Be able to effectively order and retrieve the right pages with respect to the intent of the user query.

Query logs are one of the largest sources of data potentially available to a search engine. However, the use of these logs poses significant threats to the privacy of the user. Such records can be accidentally or intentionally released to the public or misused in some way internally. Many of these concerns can be mitigated by only examining query logs in aggregate, i.e., how many times a given query occurs, geographic distribution of queries, etc. There also exist other sources of data about the user experience that could potentially provide richer insight into user preferences while providing data only with the explicit consent of the user. The question that remains is: Are query logs critical to the success of a search engine in satisfying its goals?

In this paper, we review three sources of data in the search engine context. The first source is field studies or supervised lab studies where the user is directly observed and interacted with during the course of a search task. The second source of data is a panel of users with instrumentation or other passive observation technology installed on their computer. The third source is aggregated query logs created by user transactions with a search engine. We demonstrate that, although there is significant and necessary information in the first two data sources, there is also unique data in the query logs that can significantly improve performance on the goals of a search engine.

2. RELATED WORK
The first challenge in answering user queries is to understand the distribution of user information needs. Shneiderman, Byrd, and Croft describe this need as "the perceived need for information that leads to someone using an information retrieval system in the first place" [17].

Studies based on web search behavior and labeled intents show that there are a wide variety of task categories that describe user behavior. These tasks range from searching for a specific web site to browsing for general information or looking for broad background information on a topic. Several task taxonomies and descriptions have been proposed as ways to look at log data in terms of human goals and intents [1][12][8][10].

In contrast to task descriptions, some studies seek to identify the moment-by-moment behavior of people as they work. Diary studies and usability field studies provide in-depth information about the intent of a user, as well as the details of behavior that are not visible in any other way.

In addition to understanding the goals of an individual search, studying user behavior on the web provides insight into the overall distribution of resources and materials users are searching for. Trend analysis of World Wide Web searches have shown both an overall popularity of particular resources, and a shift in how frequently people look for online information [15]. It is well known, for instance, that searchers frequently turn to the internet to find health-related information very soon after diagnosis of a serious medical issue [14]. Similarly, the rise of online consumer shopping is evident in the logs data, but a deeper understanding of what people are doing and how they frame their shopping problem is accessible only through field studies and observation [3].

Shifting the emphasis from logs-centered analysis can yield insights that are unique to the human-perspective, and can form

WWW 2007, May 8--12, 2007, Banff, Canada.

the basis for understanding web search behavior as a part of a person's larger set of goals [6][3][19].

3. Definitions & Comparison

For the purposes of comparison, we define the following types of data collection.

- **Field/Lab studies:** In a field or lab study, a researcher or observer watches while a user performs tasks on the search engine. Observational data can be collected of what URLs a user visits, what applications are used, etc. In a lab study, specialized equipment, such as an eye-tracker, may also collect additional data, often in a room with mirrored glass for observation. Field studies, however, are done *in situ*, with the observer watching human behavior in a natural setting. This type of data augments the qualitative data obtained from the researcher interacting with the user.
- **Instrumented panels:** A periodic study where users with some form of instrumentation or browser application agree to take part in observation over a short or long period of time, and have their actions with respect to the search engine and web browsing recorded during that time. Observational data, such as URLs visited or active applications, can be logged. Additionally, more qualitative data regarding what the user's goal, for example, can also be gathered by directly asking the user. However, these questions must be prepared a priori, since they are typically implemented and distributed as part of the browser application and are therefore fixed and standard across all participants.
- **Aggregate query logs:** Records stored by a search engine containing both the query issued by the user and potentially other data about actions such as results clicked on, other queries issued by that user, or additional metadata. Here we consider aggregate analysis of query logs where individual records are discarded and only mass statistics, such as the overall distribution of queries or results clicked on, are considered.

We now discuss these three sources in more detail, and the trade-offs for each source.

3.1 Field or Lab Studies

Field or lab studies are small-scale studies, typically with on the order of tens of participants, conducted in a one-on-one basis between a researcher and a participant. While a less scalable data collection method, it does result in much more detailed information due to the close one-on-one interaction with the participant and the ability to observe unanticipated behaviors.

The studies usually involve asking the user to perform some tasks, either assigned by the researcher or asking the user to come up with his "own" task, optionally followed by some questions (e.g., "did you notice this particular feature"). The studies may also include other techniques, but we focus on user-task-based studies here. Participants are usually compensated for their time.

The data collected usually includes video, audio, URL, and application tracking, the qualitative data obtained from interacting with the researcher, and potentially other data, such as eye-tracking data. Per-participant, the study usually takes up to one to two hours. While it is possible to follow-up with participants over time, that follow-up usually involves quite a few logistic challenges, so these studies are rarely longitudinal. Running a "remote" study, where the researcher calls the participant and then uses screen-sharing techniques to follow the participant's computer screen (e.g., the software packages VNC or similar), makes repeat visits for longitudinal studies easier and allows the researcher's reach to be extended to a wider variety of locations.

These small-scale studies can be conducted either in the lab or in the field. One advantage common to both studies is the ability to revise the study in real-time in response to unexpected findings. The advantages of a lab-based study, relative to a field-based study, are:

- The researcher can use specialized equipment that require a fair amount of set-up, such as an eye-tracker, to collect even more data. This equipment can be used to test how users scan a page, or to understand pre-attentive processing before cognition engages.
- Because the researcher has control over the lab settings, it is often easy to test experimental prototypes and get detailed feedback. This type of experimentation is often harder to do in the field because a non-public product interface may be needed for the test.

In contrast, the main advantage of a field-based study is that it is held in the user's space, which allows the user to have his normal environment that may provide cueing to the user as he completes the tasks. For example, if the participant is asked to come up with his own search task, he might normally look around and see a particular book that he wanted to see if there was a sequel; in a lab, he would not have that type of environmental cue.

Some of the biases with these small-scale studies are due to the nature of the one-on-one interaction. Participants often want to perform or please the researcher, in part because they are usually being compensated for their time. Thus, the nature of the interaction can potentially bias the results, and the researcher needs to be careful to not bias the participant by, for example, pointing out certain features of the page. The order of tasks may also bias participants, if they can notice a trend and make a guess as to what the study is about. Participants may also edit their normal search behavior by doing searches that seem more worthwhile or that they hope to be successful at when the researcher is present.

Due to the small-scale of the study, the participants are often not demographically representative. It is also worth pointing out that the amount of manual labor needed to collect the data is very large, and due to the customization that is possible, there may need to be more manual coding of the data that increases the turnaround time for analysis.

3.2 Instrumented Panels

Instrumented panels are larger studies, typically with hundreds or thousands of participants. Each participant opts-in to the study, either by signing into a web-based application or installing something on their machine that will send back observational data, such as URL's visited, applications that are open, or other data about the computer environment. Participants are usually also compensated in some fashion.

The studies typically include demographic-like questions, such as asking the user how often he uses the internet, preferred websites, male or female, etc. Short-term panels usually have participants perform one or more tasks (such as using a website to complete a task), often followed by user-feedback questions (did you

complete the task successfully, etc.). These user-feedback questions are implemented as part of the application, so per-person customization is non-standard. Thus, the data collected includes observational data, user feedback, and other demographic information. Long-term panels, spanning days, weeks, or months are usually just the demographic-type data combined with observational data and are typically less task-focused and user-feedback focused. Both types of panels have the ability to record much more data about the user's computing environment than would be apparent in the search engine browser window alone. For example, did the user leave the computer entirely during a break? How long did the user spend on a search result, and was he looking at another window/application during that time?

One advantage of both short- and long-term instrumented panels is that due to the size of these panels, participants are often balanced according to a demographic profile, even taking into account the opt-in nature of the panel. Of course, having this balance over several geographic regions (e.g., the US, Europe, and Asia-Pacific) can still be quite a logistical challenge. Long-term panels may be less balanced, since users will be lost over the longer time period. In many cases, up to 50% of the original panel can be lost between the start and end. Also, because long-term panels track users over time, it is worth pointing out that many users do their web search from a variety of machines, but the long-term panel usually only captures behavior on one machine, typically the home machine, rather than home and work.

Additionally, participants are passively observed in the comfort of their own environment, allowing the user to potentially relax and behave more naturally than they might in a laboratory setting. However, panelists might behave differently when they know that they are being watched, even if the watching is only electronically.

Some advantages specific to a short-term user panel, where a representative user population is performing well-formed tasks are:

- The data is well-labeled, with user-feedback about whether they found what they were looking for, and about what their goals are.
- Researchers can test experimental changes or situations that may not be live, such as the effect of branding. Testing experimental changes with a long-term panel is difficult unless the panel is large enough to split into a biased and unbiased population, not to mention the risk of irritating users so much that they remove themselves from the panel.

One difficulty for short-term panels involves choosing the tasks. While it is often easy to get a representative sample from the user population, it is often more difficult to get a representative sample of different task types. As we know, a typical web search system often has many different styles and modes of operation even within just the simplest use. Lam et al. showed that different kinds of tasks (navigational, informational, etc.) can lead to very different user behaviors [9]. At the same time, Russell and Grimes [13] showed that the behavior on "own tasks" (tasks that were the user's own chosen tasks) vs. "assigned tasks" (tasks where the participant is assigned a particular task) also lead to different task types and user behaviors. Thus, even the choice of task can often bias the results of the study, and if the subject performs their own tasks, researchers need to classify the goal into different task types before analyzing the data.

Long-term panels are strong at both giving an accurate portrait of stable web search use, as well as providing a view into slowly changing use patterns in ways that are invisible in short-term panels and difficult to determine from log data. Short-term panels are strong at getting a medium-sized sample of data with basic user labeling and feedback over a relatively short period of time.

3.3 Aggregate Query Logs

Query logs, of the three data collection methods, are the most scalable. Assuming adequate storage and infrastructure exists and user privacy considerations are respected, logging timestamps and queries typed into a search engine is relatively straightforward. This logging happens implicitly and does not disturb the user, leading to unbiased observational data collection.

The main advantage of query logs is that they have a diversity of tasks, queries, user experiences that is difficult, if not impossible, to duplicate in any other data source. However, this diversity is completely unlabeled: the researcher may know that the user did a query, but the researcher does not know what the user meant by it, nor does the researcher know whether the user was happy with the result.

Another advantage of logs is that they measure users in the wild, so assuming that a change is ready for live traffic, it is often possible to test experimental conditions to observe the impact on users (assuming proper experimental design and infrastructure).

While logs are easy to collect in bulk and contain a wealth of observational data, there are many limitations and biases to consider.

First and foremost, logs can only measure the how and the what, rather than the why. For example, if we have a sequence of queries, we only know the sequence of queries, but we have no evidence of why the user is typing in that particular sequence. We might be able to infer the user goal, but it is really only a guess -- we are inherently making up a story. The query sequence does not provide complete information about the goal of a user. While work has been done to connect loggable behavior with task classification, the success rate for such models, even for very general task types, is quite low. Lee, et al [9] showed that it was possible to classify many logged tasks into two general categories: Informational and Navigational using both user behavior and data about the web pages viewed. However, even for their sample of the 50 most popular queries, the classification could only be done for 20 that had sufficient data, leaving the majority of tasks unclassified.

Second, logs are completely unlabeled except for the presence or absence of an event. For example, it is possible to determine whether an ad was clicked, but not necessarily why the ad was clicked or whether the user found the ad useful. Fox, et al [4] inferred session-level satisfaction based on the user behaviors during a search, and were able to do significantly better than baseline: up to 74% accuracy versus a baseline of ~56%. However, this model used explicit judgments of the quality of individual results provided by the user, and without those, the accuracy dropped to 60%. Even in their reduced model, one of the major factors – session exit status – would not normally be observed by a search engine without some browser side instrumentation.

Next, logs can only measure the system being logged: Nielsen's study found that only 20% of a user's time is spent searching, which means that logs from a search engine are missing at least 80% of a user's online activity, more if a user uses multiple search engines [11]. Logs can only capture the specific system: if users

	Depth	"Naturalness"	Flexibility	Scale	Turnaround
Field Studies	Very detailed	Observed, may be artificial tasks	Altered midstream	O(50) users	~ 1 month
Panels	Observes computer environment, multi-tasking	Natural, may be edited by user	Hard to change data collection	O(1000) users	~ 2-4 weeks
Query Logs	Limited; no contextual information	Completely natural	Easy to run experiments on Search Engine side	Everything, millions of users	Real time to ~ 1 week

Table 1. Summary Comparison of Data Sources.

use Google for one type of search and Yahoo for another, then the logs from Yahoo will only contain the second type. Even within the framework of a single search engine, logs record only measurable properties: depending on the system set-up, some things may not be loggable [16]. Similarly, logs cannot capture any social interaction, such as people talking to one another; this consideration is important when testing a user-visible change, for example: the full effect may not be felt until a critical mass has been reached.

Fourth, query logs are noisy, since they include everything, including robots, spam, data outages, recording errors, etc. Using intelligent filters can reduce that noise, but filters are often based on a set of heuristics, with varying degrees of accuracy and stringency (it is often a question of over-filtering vs. under-filtering).

Finally, logs don't necessarily allow long-term studies of a single user. The goals of these studies are often to determine whether a particular change leads to a poorer user experience and users searching less or coming to the site less often, or whether users learn to trigger a particular feature. Even in these studies, it is the still the aggregate behavior that researchers are often interested in, rather than what individual users do (e.g., how many users learned to trigger a particular feature). However, due to concerns regarding respecting users' privacy, sites will usually log a cookie, with no personally identifying information, as a proxy for a user. Cookies, however, are a very imprecise approximation of users: people can clear their cookies, software that people don't even know they have installed can automatically clear cookies, one person may use multiple machines (e.g., one at work, one at home), multiple people may use the same machine (e.g., a shared computer at home), and people may share cookies to try out different experimental conditions [23]. One study from Jupiter cited a weekly cookie churn rate of 30% [21], and another survey found that 40% of users voluntarily clean their cookies weekly [22]. While both numbers are from surveys and have a self-reporting bias, evidence suggests that cookie churn is significant enough to impede these aggregate long-term studies.

Overall, logs are useful for analyses where a large amount of data is needed, and for testing the impact of changes, especially ones that only impact a small proportion of queries. However, it must be emphasized that logs can only ever reveal correlations and not causality, however tempting inferring causality may be.

3.4 Compare & Contrast

While we have discussed several of the trade-offs inherent to each of the three data collection methods, we briefly compare them directly here.

The most obvious trade-off between the three methods is scale and level of detail: field / lab studies are the most detailed but scale the least well, while logs scale the best and have the least detail. This level of information has several implications, since there is simply some information that cannot be obtained at various levels of detail, such as why a user does something (causality vs. association), labeling the data (e.g., user feedback / satisfaction), etc. Also, with the higher scaling comes a lot more noise, but also much more diversity.

Another trade-off is the level of automation and the timeliness / longitudinality of collection. Field / lab studies are very manual, and get a lot of detail, but also take a lot of work to set-up, and are rarely longitudinal. In contrast, query logs and instrumented panels are more automated and can capture longitudinal trends, such as the popularity of baby names. This automation also leads to an immediacy with regard to quickly capturing trends, such as important news events.

Based on these comparisons, the primary advantages of query logs are in the scale and the immediacy of data collection, while the disadvantages are the noise and lack of insight into the user experience.

4. Why is scale critical?

From a statistical standpoint, the sample sizes used in many instrumented panels are perfectly sufficient to estimate fractions of satisfied users, and even to obtain basic feedback on the quality of features within the sample. However, from the point of view of a search engine, these samples provide infinitesimally small coverage of the total space of user queries and potential user tasks.

4.1 Sampling a Natural Task Distribution

The goal of collecting data about the user experience is to understand and improve the overall web search user. To that end, experimenters often try to collect a "natural" distribution of tasks by observing the user in his or her own environment. Another approach is to choose a stratification of tasks that represent categories of interest. This strategy reflects the user experience to the extent that meaningful categories of experience are understood. Query logs offer the ultimate opportunity to observe a natural, though implied and unlabeled, task distribution.

One obvious example of this distribution is geographic locale: search engines advertise services in over 100 locales [25][27] and many different languages. Internet usage has increased enough that significant portions of the potential search engine market are in countries that are just now developing internet usage, or where a search engine may need to do significant research to understand the local web [26]. The increased diversity of the web searcher population leads to many interesting questions that can only be answered looking at a subgroup of the user population. Each combination of language and location may offer a significantly different search experience to the user. At the simplest level, this is reflected in the difference between users speaking the same language, but coming from different countries, such as Portuguese queries from Brazil and Portuguese queries from Portugal. However, there are less obvious divides: users searching for general topics may be specifically interested in data from their

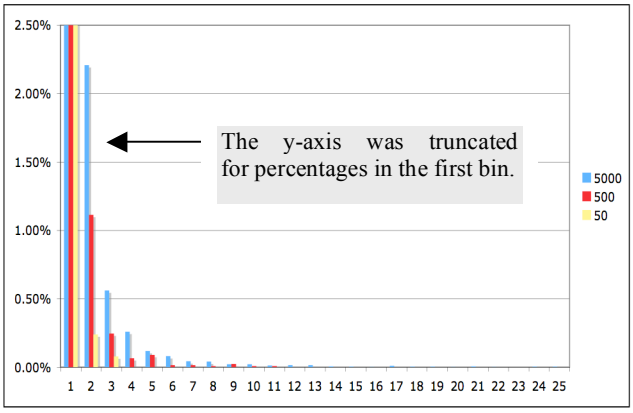

Figure 1. Percent of unique queries in any sample that occurred in one or more samples.

own state, and the available data in the form of yellow page listings or high-pagerank sites may vary greatly by locale.

In addition, researchers often wish to study combinations of behaviors that may be common individually, but extremely rare in combination. By using a complete query log, the researcher can automatically access a sample large enough to isolate even 1% task types or rare behaviors.

A fundamental advantage of having an enormous sample of users is that these complex stratifications can be examined after the fact rather than requiring a study be created to achieve correct stratification of the category.

4.2 Diversity of Queries

Even within a single locale, the diversity of queries is enormous, and determining the necessary stratification (e.g., a priori when creating a study and recruiting participants) can be problematic. The odds of seeing a specific query more than once in a reasonably sized sample, or of seeing a specific query at all, are very small.

With respect to sampling, the query stream is quite diverse. We took a fixed test set of 362M US English queries sampled over one year from traffic on www.google.com. Individual queries may occur more than once in the sample depending on their frequency in the original sample. We then sampled this test set to create 25 independent samples of 50 queries, and repeated the process to create 25 samples of size 500 and size 5000. The intent was to test likely sample sizes that might be observed in a field study or an instrumented panel over a reasonably short period of time such as 2-4 weeks. From this data, we counted in how many of the 25 samples each unique query occurred (queries were all mapped to lowercase before being considered unique, but otherwise were not processed in any way), out of the total pool observed at that sample size. The percentages are remarkably low. For the 50 query samples, 99.7% of queries occurred in only one sample, 0.2% in two samples, and 0.1% in three samples. For samples of size 500, the results were only slightly different: 98.4% in on sample, 1.1% in two samples and 0.3% in three samples. Even at 5000 query samples, 96.5% of queries still only occurred in one sample, 2.2% in two samples, and 0.6% in three.

Figure 1 shows the drop-off in number of sample occurrences (the value for queries occurring in 1 sample has been cropped). While most queries appear only once, there were a handful of very common queries that occur in almost every sample once the sizes get large enough. For example, in the 5000 query samples, there were 5 queries that appeared at least once in all 25 samples, and one query that averaged 8 appearances in each of those samples.

Event	Load Title
Navigate	Soul Asylum - Google Search
Navigate	SoulAsylum.net (http://www.soulasylum.net/)
Back	Soul Asylum - Google Search
Navigate	Soul Asylum: News (http://www/soulasylum.com/…)
URL Entry	(http://www/soulasylum.com)
Navigate	SonyMusicStore: Soul Asylum (http://www.sonymusicstore.com/…)
Back	Soul Asylum: News (http://www/soulasylum.com/…)
Back	Soul Asylum - Google Search
Forward	Soul Asylum: News (http://www/soulasylum.com/…)
Back	Soul Asylum - Google Search
Navigate	Soul Asylum - Google Search
Navigate	Soul Asylum's Pirner waiting for word on home - Katrina hits entertainment (http://www.MSNBC.com/…)

Table 2: Browser side logs record for [soul asylum] searcher in Keynote study.

However, for sample sizes of 500 queries, no single query appeared in more than half the samples. This diversity, even from a relatively small fixed sample of user query traffic, suggests that post-hoc stratification of queries may be very difficult to do. Similarly, methods that rely on repeat instances of a query are virtually impossible to use on smaller samples, except in the case of a few very common queries.

Overall, this data shows that the diversity that we might observe (or be able to create or duplicate) from a small- or medium-scale study would not approach the diversity observable from large-scale query logs.

5. Predicting Intent

Query logs may contain all the queries users issue, and may even record the chains of query sequences, but intent is difficult to determine from a query sequence. While query logs have the material to make aggregate distinctions in meaning, they do not have the depth that other data sources do to understand the goal of a user as expressed by an individual query or query sequence.

5.1 Inferring Intent

To truly understand the user query and the quality of results, a search engine needs to also understand not just topical associations with a query but also the real goal of the user's search. For example, does the user want to buy something or read about something? Is the user searching for a general query in hopes of finding an authoritative source for a more specific question, or because he wants general information on a topic?

In the fall of 2005, Google partnered with Keynote Systems to do a blind study of internet search behaviors. In the study reported here, 401 subjects were recruited and asked to do a web search

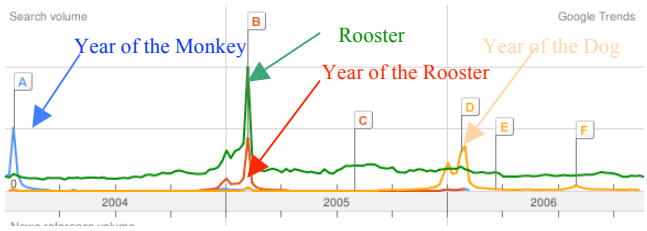

Figure 2. Volume of query trends, 2004-2006 for 4 queries. Letter indicate major news stories containing query terms. Note that spikes for [rooster] and [year of the rooster] coincide at the beginning of 2005.

task in the manner that they "normally would." (Subjects did not know that Google was running the study.) Subjects were asked prior to doing the task what they were searching for, and then the browser events during their searches were recorded by Keynote. After reporting a completed task, the user was then asked again to describe what he or she was searching for, and to categorize that search into one or more topical categories.

One of the most common behaviors that showed a contrast between loggable (on the search engine side) query data and user intent was that the user would state a complex or specific intent, yet do a fairly general query in hopes of reaching the correct information resource to then answer the question. For example, the user in Table 2 stated his or her intent as follows: to find "The band Soul Asylum and any news on the band, including possible tour dates in Minneapolis or country wide." However, examining the browser event log shows only the query [Soul Asylum], repeated several times.

Also significantly, the user again spends a large fraction of their time on this search off the search property – and manually navigates within the band's homepage to locate information. Not surprisingly, this user reported lack of satisfaction with the search due to the fact that "most of them [web results] were pretty outdated and hadn't been updated in a while." Recency is an important attribute for many searches, but in this case, the user had an even more specific information objective (current touring information, potentially in Minnesota or the entire US) that was not at all reflected in the query.

Another interesting phenomenon in this data was that users would often refine their stated intents during the course of a session. For example, a user would start looking for [thyroid problems], but conclude by expressing happiness with the information he or she had learned about "hyperthyroid." Similarly, a user would start by looking for [Italian nougat candy] to look for the brand name "torrone," and then use the brand name to look for recipes. These evolutions in query structure suggest that the queries logged by search engines trace a task trajectory that may be governed by the topical knowledge of the user.

5.2 Disambiguating Queries

While inferring intent from query logs is difficult, there are areas in which aggregate logs have great benefits. One of the biggest challenges in "understanding" the query for a search engine is to detect meaning shifts or associate meaning to disambiguate a query. Figure 2 shows an example created using Google Trends [23] where a broad query, [rooster], experiences a sudden upsurge in response to the beginning of the Year of the Rooster in the Chinese zodiac. Although the query [rooster] is much more popular during the rest of 2005 than [year of the rooster], there is

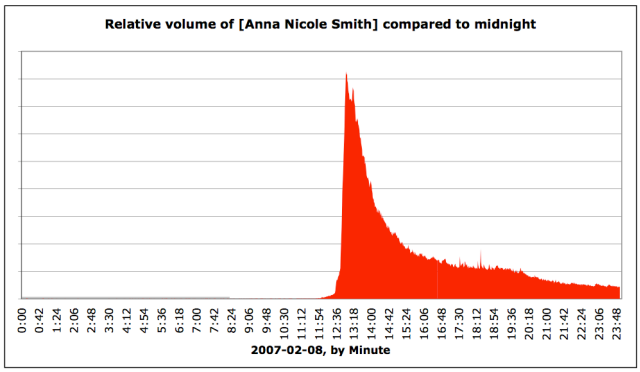

Figure 3. Volume of queries containing [anna nicole smith] *relative to* 12:00 AM (PST), 2007-02-08. News of Anna Nicole Smith's death was first reported around 11 AM (PST).

a strong association between the rise of [rooster] and [year of the rooster] at the beginning of 2005. Given that there are a number of possible meanings for the more general query, this type of timely information can be key to ordering possible user intent of the query – at this specific time, information on "year of the rooster" might be more relevant than general information about roosters or rooster products.

However, this type of topical association is only one form of query "disambiguation" in the broader sense. The ability to understand what type of resources the user is looking for: shopping, reviews, products or other type of information, and how satisfied the user is with the content of each of those resources is far beyond the scope of the logs. Field studies can get at this type of labeling since the researcher is talking directly with the participant and can therefore label the data directly. Panels, especially task-oriented ones, are often directive and therefore directly labeled. In a panel setting, the researcher can ask the user to complete a task with a labeled goal. Therefore, in depth data collected is more easily (and correctly) labeled than logs.

6. Immediacy of Data

Another critical area of differentiation between data sources is the speed at which data can be collected. Search engines provide access not only to the static vast content of the World Wide Web, but also to up-to-the minute news and additions to the web. Fast feedback from the user is an important part of understanding how interest in a topic may have shifted, or how a previously used query term assumes new meaning.

Field studies generally require prearrangement and can be slow to collect. Similarly, diary studies may require a personal interview. Collecting data using either mechanism is human intensive. Instrumented panels could potentially be run continuously with an equally prompt data feed, but traditionally this has been non-standard. Query logs are the most "real-time" of the options, in that they can be collected in high volume very quickly. Ideally, a search engine would not only be able to detect the increased volume in a query, but also evaluate its performance in a direct feedback loop.

Figure 3 above illustrates the speed at which a news topic can appear in the user query stream. The graph represents the prevalence, compared to a baseline of 12:00 AM (PST), of queries related to Anna Nicole Smith at the granularity of minutes. Smith's death was first reported in the news around 11:00 AM, and the query stream starts to reflect this change almost immediately. This data is consistent with similar hourly

observations on fast-rising topics from other logs sources [1], and serves to emphasize speed at which news transmission impacts user queries on a search engine.

In conjunction with traditional methods to search a news corpus for rising events, query logs offer the most immediate access to the interests of users, and are able to provide feedback on a minute-by-minute basis at a level that other data sources are unable to match. Differences in the query stream at a 15-minute level are critical to satisfying the information needs of users, especially for rapidly changing news stories.

7. Conclusions

There are several key benefits to using query logs. First, the diversity and distribution of queries that a search engine receives can only be captured at the scale of query logs. The overall distribution of queries in any small sample of user queries (even on the order of 10,000 queries) has only a small overlap with the entire population. Thus, building good "intent" coverage for the population of user queries, even if unlabeled, requires the diversity of large-scale query logs. Next, past queries are critical for determining the intent, or possible set of intents of a query. Finally, using query logs in aggregate is to improve recency of search results. The interpretation of a query may shift or a new term be introduced very quickly in the presence of current events.

However, to fully understand user satisfaction and user intent requires a depth of data unavailable in query logs but possible to acquire from other sources of data, such as one-on-one studies or instrumented panels.

8. REFERENCES

[1] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., and Frieder, O. Hourly Analysis of a Very Large Topically Categorized Web Query Log. SIGIR '04, (July 2004).

[2] Broder, A. A taxonomy of web search, in SIGIR Forum, 36, 2 (2002), 3-10.

[3] Brown, B. and Sellen, A. Exploring Users' Experiences of the Web. First Monday, 6, 9, (June 2004). http://www.firstmonday.org/issues/issue6_9/brown/

[4] Clark, L., Ting, I., Kimble, C., Wright, P., and Kudenko, D. Combining ethnographic and clickstream data to identify user Web browsing strategies, Information Research, Vol. 11 No. 2 (Jan 2006)

[5] Fox, S., Karanwat, K., Mydland, M., Dumais, S., and White, T. Evaluating implicit measures to improve web search. ACM Transactions on Information Systems (TOIS), 23, 2, (April 2005), 147-168.

[6] Hargittai. E. Beyond logs and surveys: in-depth measures of people's web use skills, Journal of the American Society for Information Science and Technology, 53,14 (2002), 1239-1244.

[7] Jansen, B.J., Spink, A., and Saracevic. T. Real life, real users, and real needs: A study and analysis of user queries on the Web. Information Processing and Management, 36,2 (2000), 207 – 227.

[8] Kellar, M. "An Examination of User Behaviour During Web Information Tasks," PhD Thesis, Dalhousie University, (2007).

[9] Lam, H, Russell, D, Tang, D., Munzner, T. Session Viewer: Supporting Visual Exploratory Analysis of Web Session Logs. submitted to Infovis 2007.

[10] Lee, U., Liu, Z., and Cho, J. Automatic identification of user goals in web search. Technical report, UCLA Computer Science, (2004).

[11] Nielsen, J. Users interleave sites and genres. Use-it website article. (February 2006). http://www.useit.com/alertbox/cross_site_behavior.html

[12] Rose, D.E. and Levinson, D. Understanding user goals in web search in WWW '04: Proceedings of the 13th international conference on World Wide Web. (2004), 13-19.

[13] Russell, D.M., and Grimes, C. Self-chosen tasks are not the same as assigned web search tasks. HICSS, Kona, HI, (Jan. 2007).

[14] Schwartz, K.L., Roe, T., Northrup, J., Meza, J., Seifeldin, R., and Neale, A.V. Family Medicine Patients' Use of the Internet for Health Information: A MetroNet Study. The Journal of the American Board of Family Medicine, 19 (2006), 39-45.

[15] Sellen, A.J. and Murphy, R. The future of the mobile internet: Lessons from looking at web use. Technical Report HPL-2002-230, Information Infrastructure Laboratory, HP Laboratories, Bristol, (August 2002).

[16] Sen, A., Dacin, P., and Pattichis, C. Current Trends in Web Data Analysis. Communications of the ACM 49, 11 (Nov 2006), 85-91.

[17] Shneiderman, B., Byrd, D., and Croft, B., Sorting out searching: A user-interface framework for text searches, Communications of the ACM 41, 4 (April 1998), 95-98.

[18] Silverstein, C., Henzinger, M., Marais, H., and Moricz, M. Analysis of a Very Large AltaVista Query Log. http://citeseer.ist.psu.edu/silverstein98analysis.html (1998).

[19] Spink, A., and Jansen, B. *Web Search: Public Searching of the Web.* Springer (2005).

[20] Wolfram, D., Spink, A., Jansen, B. J., and Saracevic, T. Vox populi: The public searching of the web. Journal of the American Society of Information Science 52, 12. (2002), 1073-1074.

[21] http://weblogs.jupiterresearch.com/analysts/peterson/archives/007970.html.

[22] http://www.marketingsherpa.com/!newsletters/bestofweekly-4-22-04.htm.

[23] http://www.google.com/trends

[24] http://galide.jazar.co.uk/2006/03/new-google-ui.html.

[25] http://www.google.com/intl/en/corporate/facts.html

[26] http://www.internetworldstats.com/stats.htm

[27] http://www.msn.com/worldwide.ashx