Real Time Search User Behavior

Bernard J. Jansen
College of Information Sciences and Technology
The Pennsylvania State University
jjansen@acm.org

Gerry Campbell
CEO
Collecta
gcampbell@gmail.com

Matthew Gregg
Collecta
matthew.gregg@gmail.com

Abstract
Real time search is an increasingly important area of information seeking on the Web. In this research, we analyze 1,005,296 user interactions with a real time search engine over a 190 day period. We investigate aggregate usage of the search engine, such as number of users, queries, and terms. We also investigate the structure of queries and terms submitted by these users. The results are compared to Web searching on traditional search engines. Results show that 60% of the traffic comes from the engine's application program interface, indicating that real time search is heavily leveraged by other applications. Of the queries, 30% were unique (used only once in the entire dataset). The most frequent query accounted for 0.003% of the query set. Less than 8% of the terms were unique. The most frequently used terms accounted for only 0.03% of the total terms. Concerning search topics, the most used terms dealt with technology, entertainment, and politics, reflecting both the temporal nature of the queries and, perhaps, an early adopter user-based. Sexual queries were quite low, relative to traditional Web search. Searchers of real time content often repeat queries overtime, perhaps indicating long term interest in a topic. We discuss the implications for search engines and information providers as real time content increasingly enters the main stream.

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.
ACM 978-1-60558-930-5/10/04.

Keywords
Real time search, real time content, Collecta, Twitter

ACM Classification Keywords
J.4 [Computer Applications]: Social and Behavioral Sciences – Sociology

General Terms
Experimentation, Measurement, Human Factors

Introduction
Users of social Websites, such as Facebook, Twitter and FriendFeed, post millions of pieces of content daily. These users generate and share media of all types (i.e.,

Background

- Real time content is many times **short postings** similar to status messages, sometimes with links to other content. Real time content typically **lacks the indexing and ranking factors** use by current Web search engines.

- Real time search engines **gather data by polling** one or more social media services, providing results to the searcher based on a query.

- Real time search engines **do not usually have a static results page**. The search engine continually updates the result listing as additional content enters the real time stream.

- Real time content appears to be **altering** the way people are accustomed to **gathering information**.

- Therefore, **investigations of real time search behaviors** is a productive research area.

status messages, blog postings, video, images, reviews, text snippets, etc.) with both their close social network and the larger Web community. Much of this content is of a short temporal span (a.k.a., real time content) and does not fit into the hypertext ranking structure used by the major Web search engines; therefore, until recently, the traditional Web search engines have typically indexed only a limited amount of this real time content, most notably from blogs.

However, this real time content from hundreds of thousands or more sources can have significant societal, cultural, and commercial implications. Additionally, informational behaviors are changing as people are becoming more accustomed to accessing this real time content for a variety of purposes. Therefore, there are significant opportunities in providing real-time search services, with the major search engines and new firm entering the marketplace to provide technologies for real time content search.

However, there has been little to no investigation, which we could locate, of user behaviors in real time search. How are users engaging with real time search technologies and services? What are the topics of interest for real time search? How does real time search behaviors compare to traditional Web search? These are some of the motivators for our research.

Background

Real time content is often short status message type posting, sometimes with links to longer documents or multimedia content. Real time content is typically generated on the social media platforms, such as Twitter, Facebook's newsfeed, or MySpace status messages. Real time content is typically created for the immediate temporal context, to be consumed as soon as created rather than for archival intentions [5].

Finding relevant real time content can be quite a challenge, given its rapid pace of creation, huge volume of content, information, and lack of standard ranking factors used by Web search engines (i.e., anchor text, hypermedia links, etc.). Traditionally, web search engines index webpages periodically, return results based on a match to a search query, and rank results based on a mix of factors. These techniques do not mess well with real time content, which typically has a short temporal half life, no hyperlink structure, and few ranking factors. This situation has driven the need for specific real time search technologies.

Real time search employs a variety of techniques for retrieving real time content. While the traditional search engines are concerned with relevance, real time search engines factor in relevance, popularity, and temporal immediacy. Therefore, real time search engines (i.e., Collecta, OneRiot, CrowdEye, etc.) employ different methods than the crawling used for conventional Web content. Real time search engines typically use some method of polling (i.e. they accept a query, send it to one or more social media platforms, retrieve and integrate the results). In this respect, real time search engines are similar to meta-search engines. However, given the temporal nature of the content, many real time search engines do not index or store any content themselves. Additionally, as long as the query is active, real time content continues to flow into the engine in response to the query. Therefore, there is no static search engine results page generate with a set number of links.

Mutual Information Statistic

The mutual information statistic (mis) measures term association and does not assume mutual independence of the terms within the pair. The mutual information formula used in this research is:

$$I(w_1, w_2) = \ln \frac{P(w_1, w_2)}{P(w_1)P(w_2)}$$

where $P(w_1)$, $P(w_2)$ are probabilities estimated by relative frequencies of the two words and $P(w_1, w_2)$ is the relative frequency of the word pair (order is not considered). Relative frequencies are observed frequencies (F) normalized by the number of the queries:

$$P(w_1, w_2) = \frac{F_{12}}{Q'}$$

Both the frequency of term occurrence and the frequency of term pairs are the occurrence of the term or term pair within the set of queries. The number of queries for the terms is the number of non-duplicate queries in the data set. The number of queries for term pairs is defined as:

$$Q' = \sum_{n}^{m} (2n - 3) Q_n$$

where Q_n is the number of queries with n words ($n > 1$), and m is the maximum query length.

This new form of search raises questions, currently unanswered, concerning how user will interact with these technologies and what affordances these technologies offer. These are issues we investigate using a log from Collecta, a real time search engine.

Research Question

With this motivation, the research question addressed by this exploratory phase of the study are: *What are the search characteristics of users when looking for real time content, including search engine access, query structure, and use of terms?* These questions are addressed using a large amount of data collected in a search log from an operational real time search engine. In this research, we concentrate on queries and terms as key variables in the interactive search process.

Research Design

We use data from Collecta (http://collecta.com/) for this research, one of the most popular real time search services. However, all real time search engine have some similar characteristics: (1) accept a query, (2) poll one or more social media sites, and (3) present a stream of real time content. So, we believe our results to be applicable to other real time search engines.

Collecta

Collecta provides real time content from the Web, including results from blogs, micro-communication services, blog comments, news feeds, and photo sharing services. Collecta uses Extensible Messaging and Presence Protocol (XMPP), an open XML communications technology. The Collecta engine accepts search queries from users, uses XMPP to communicate with social media sites, and presents a temporal stream of real-time content. Collecta searches blog posts, comments on blog posts, along with social media sites, such as Twitter and Jaiku. Founded in November 2008, Collecta went live in June 2009.[1] Collecta now offers site specific search services for MySpace (http://myspace.collecta.com/). See figure 1 for an overview of the Collecta interface and features.

Data Collection and Analysis

In a log, we collected records of search interactions from 4 June to 9 December 2009 executed on Collecta. The procedure that we used in this research is similar to that used in prior Web search log studies [2]. The log contains 1,005,296 records, each with four fields:

- *User Identification*: a code to identify a particular computer based on the computer's Internet Protocol (IP) address.
- *Date*: the date of the interaction
- *Time of Day*: measured in hours, minutes, and seconds as recorded by the Collecta server on the date of the interaction in Coordinated Universal Time.
- *Query Terms*: the terms as entered by the user.

Key concepts for search (i.e., the process of a searcher interacting with an information system) are:

- *Term*: a series of characters within a query separated by white space or other separator.
- *Query*: a string of terms submitted by a searcher in a given instance of interaction with the search engine.

We used industry standard search log techniques as outlined in [2] for the query and term analysis. At the term level of analysis, we used the mutual information statistic measure, as outlined in the sidebar.

[1] http://www.techcrunch.com/2009/06/18/collecta-enters-the-real-time-search-wars/

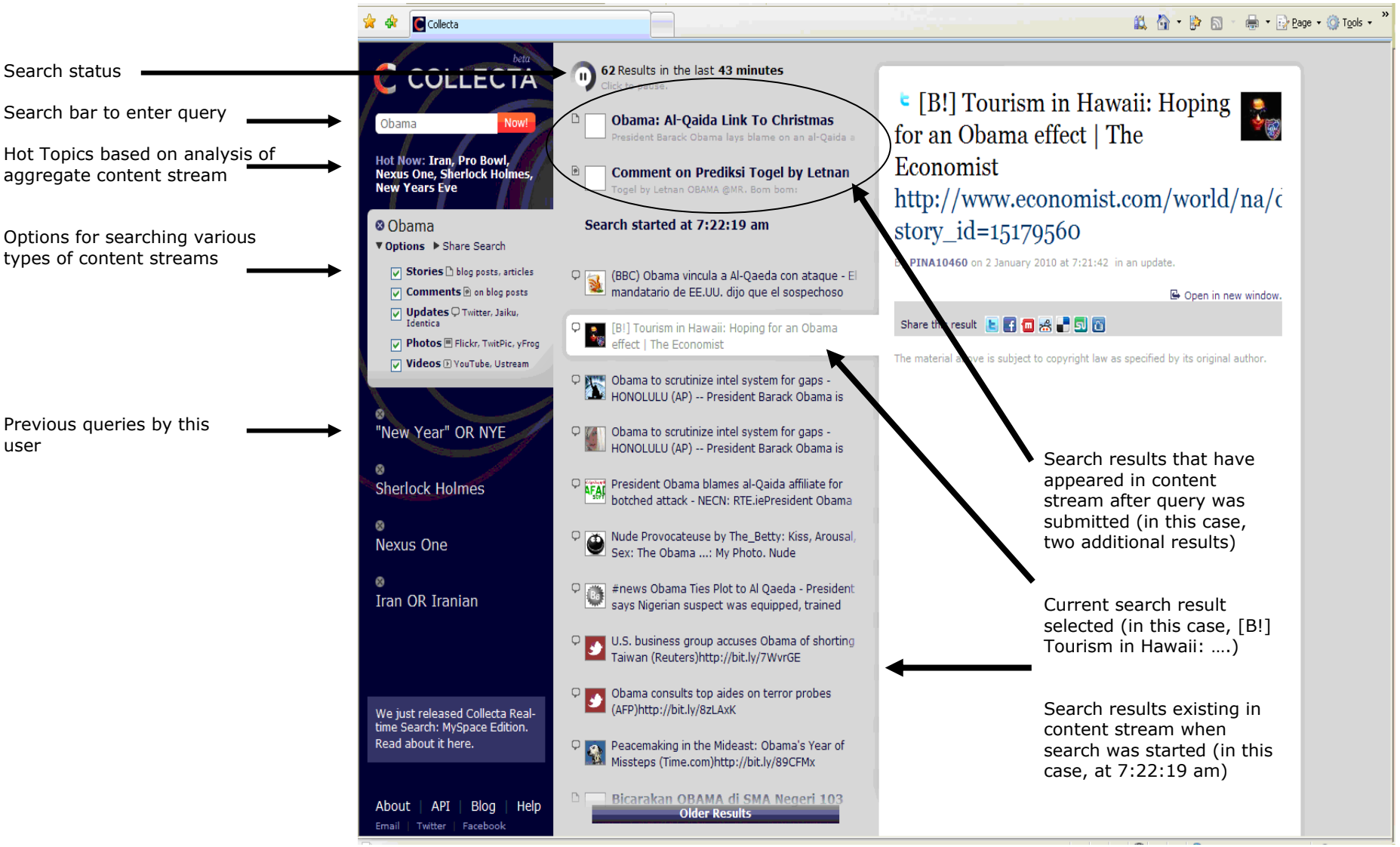

figure 1. Sample of Collecta interface, search features, and results presentation

Query	Occurrences	%
naomi watts	3,040	0.003
jQuery CSS	2,787	0.003
obama fly	2,433	0.002
thanksgiving	2,318	0.002
google wave	2,177	0.002
google	2,136	0.002
gfcampbell	2,022	0.002
[blank]	1,903	0.002
sex	1,536	0.002
shark attack comment	1,508	0.002
tiger woods	1,386	0.001
foo	1,309	0.001
crazy	1,259	0.001
Apple	1,259	0.001
search obama	1,242	0.001
michael jackson	1,236	0.001
obama or inauguration or inaugural or inaugurate o	1,203	0.001
twitter	1,193	0.001
google voice	1,184	0.001
Halloween	1,167	0.001
facebook	1,155	0.001
giannoulias	1,108	0.001
ufo	1,102	0.001
new moon	1,094	0.001
real time search for	1,083	0.001
leweb	1,058	0.001

table 1. Queries > 1,000 Occurrences.

Results

Overall, the 1,005,296 queries submitted during the 190 days originated from 43,140 unique IP addresses. Of the interactions, 40% came directly from the Collecta Website, and the remaining 60% came from users of the Collecta application program interface (API), which is unique relative to the use of traditional Web search engines. Each IP address for the API submitted an average of 23 queries over the data collection period or 0.002% of the total queries. At the query level of analysis, of the 1,005,296 queries in the dataset, 297.392 were unique (30%). This is very low, with studies of Web search reporting unique queries as high as 59% [3].

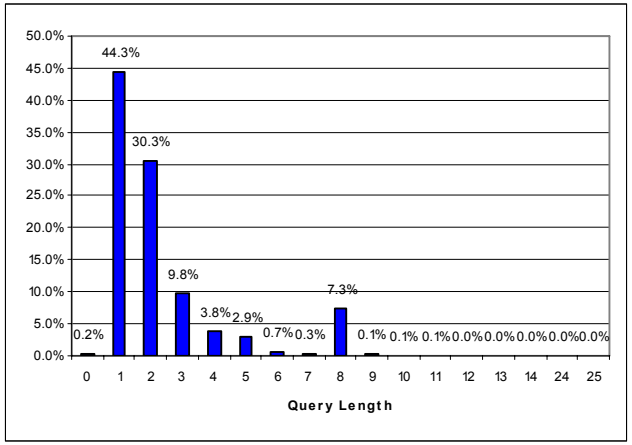

figure 2. Distribution of Queries by Length.

We examined also query length, as shown in figure 2. We see that there were a small percent of null queries (i.e., users coming to the search engine and submit a query with no terms). Again, this is a small percentage, not in line with studies of Web search [3, 4], where null queries are in 30-40% range. More than 44% of the queries contained one term, 30% contained two terms, and nearly 26% contained three terms or more. The average query length was 2.32 terms, which is in line with that of traditional Web search [1, 3]. The most popular queries are shown in table 1. Most notable is the lack of pornographic queries, which are typical in Web search logs [1, 3, 4]. Only one of the top queries was pornographic in nature (e.g., *sex*).

Moving to the term level of analysis, there were 2,331,072 total terms used in all queries in the data set, with 3,477,163 total term pairs. There were 175,403 unique terms (7.5%) and 442,713 unique term pairs (12.7%), inline with Web search [3].

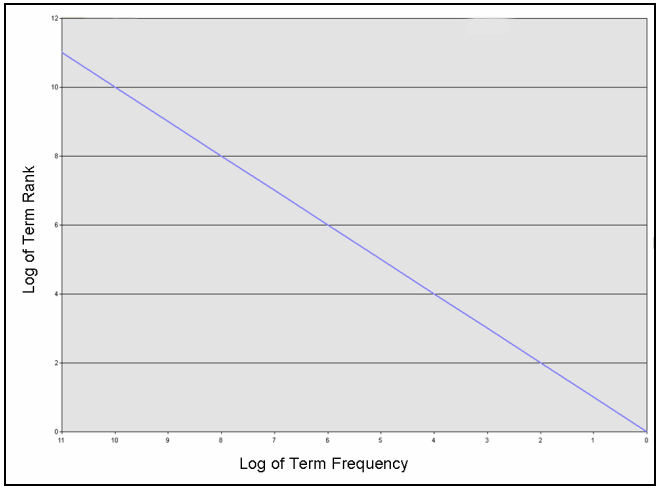

figure 3. Log – Log Plot of Terms Ranked by Frequency.

Examining figure 3, we see that the log-log plot of rank-frequency of term distribution adheres to a power law distribution. A power law distribution is typical for

Term	Occurrences	Probability
python	70,205	0.030
ball	70,030	0.030
co	69,957	0.030
pet	69,933	0.030
skin	69,896	0.030
monty	69,869	0.030
burmese	69,848	0.030
snake	69,633	0.030
or	56,214	0.024
obama	53,397	0.023
jquery	19,352	0.008
css	17,881	0.008
ajax	16,093	0.007
collecta	12,391	0.005
category	9,280	0.004
story	8,188	0.004
google	7,774	0.003
iran	7,064	0.003
de	6,909	0.003
the	6,793	0.003
iphone	6,622	0.003
in	6,588	0.003
of	5,493	0.002
foobar	5,295	0.002
com	4,787	0.002
new	4,359	0.002
search	4,123	0.002
and	4,073	0.002

table 2. Terms > 4,000 Occurrences.

Web query terms. In table 2, the top most searched terms accounted for 0.03% of term occurrences. Examining term pairs occurring more than 100 times using the mutual information static showing strength of term association (table 3), we list the most strongly associated terms. Again, the lack of pornographic terms is unusual, with most term pairs being associated with people in the news during the data collect period.

Term	Term	mis	Occurrence
dickie	peterson	12.20	132
khalid	mohammed	11.65	122
Yo	yo	11.61	127
minimum	wage	11.60	107
tacos	yummy	11.59	106
sheikh	mohammed	11.55	102
gerrit	zalm	11.52	106
Hong	kong	11.48	104
lembrancinhas	casamento	11.36	126
Sts	128	11.33	124
fausto	nilo	11.33	139
Hip	hop	11.31	103
captain	albano	11.26	195

table 3. Strongly Associated Term Pairs >100 Occurrence.

Discussion and Implications

As one of the first analyses of real time searching behaviors, these exploratory results highlight interesting aspects. First, there appears to be a heavy use of accessing real time search via secondary applications rather than from the Website. Second, these API are submitting the same query multiple times a day and repeating the query over multiple days. In this respect, it is similar to information filtering, resulting in fewer unique queries. Third, real time search differs in topics from current Web search. There is a high occurrence of technology, entertainment, and politics, with a low occurrence of sexual querying. Implications are that real time search engine technologies can leverage these behaviors, such as providing features to save searches and switch media verticals to improve user experience.

Real time search is a compelling new area of Web interaction with potential as a new channel for information gathering, advertising, and other uses. As people become more accustomed to using real time content, real time search will become still more important. Therefore, understanding how people locate information in this context is critical. Future research will examine specific searching aspects shedding light on needed technologies and societal impact.

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. In *Proceedings of the 27th annual international conference on Research and development in information retrieval* (Sheffield, U.K., 25-29 July, 2004), 321-328.

[2] Jansen, B. J. Search log analysis: What is it; what's been done; how to do it. *Library and Information Science Research*, 28, 3 (2006), 407-432.

[3] Jansen, B. J., Spink, A., Blakely, C. and Koshman, S. Web searcher interactions with the Dogpile.com meta-search engine. *Journal of the American Society for Information Science and Technology*, 58, 4 (2006), 1875-1887.

[4] Silverstein, C., Henzinger, M., Marais, H. and Moricz, M. Analysis of a Very Large Web Search Engine Query Log. *SIGIR Forum*, 33, 1 (1999), 6-12.

[5] Spark, D. *Real-Time Search and Discovery of the Social Web*. Spark Media Solutions, 2009, from http://www.sparkminute.com/?p=1261