ARTICLE IN PRESS

Information Processing and Management xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Real time search on the web: Queries, topics, and economic value

Bernard J. Jansen [a,*], Zhe Liu [a], Courtney Weaver [a], Gerry Campbell [b], Matthew Gregg [b]

[a] College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, United States
[b] Collecta, Santa Monica, CA 90401, United States

ARTICLE INFO

Article history:
Received 21 June 2010
Received in revised form 5 January 2011
Accepted 17 January 2011
Available online xxxx

Keywords:
Real time search
Real time content
Collecta
Twitter
Economic value of search
Search topics

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. Using query log analysis, we investigate searching behavior, categorize search topics, and measure the economic value of this real time search stream. We examine aggregate usage of the search engine, including number of users, queries, and terms. We then classify queries into subject categories using the Google Directory topical hierarchy. We next estimate the economic value of the real time search traffic using the Google AdWords keyword advertising platform. Results shows that 30% of the queries were unique (used only once in the entire dataset), which is low compared to traditional Web searching. Also, 60% of the search traffic comes from the search engine's application program interface, indicating that real time search is heavily leveraged by other applications. There are many repeated queries over time via these application program interfaces, perhaps indicating both long term interest in a topic and the polling nature of real time queries. 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. However, 36% of the queries indicate some geographical affinity, pointing to a location-based aspect to real time search. In terms of economic value, we calculate this real time search stream to be worth approximately US $33,000,000 (US $33 M) on the online advertising market at the time of the study. We discuss the implications for search engines and content providers as real time content increasingly enters the main stream as an information source.

1. Introduction to real time content on the web

Web users of online social media (e.g., Facebook, MySpace, Twitter, StumbleUpon, myYearbook, and other social networking sites and services) create millions of snippets of content in a variety of media (e.g., textual status message, videos, images, links, blog posts, reviews, and more) and share this content with their immediate networks as well as 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 (Brin & Page, 1998); therefore, the Web search engines have typically indexed only a limited amount of real time content, most notably from blogs.

However, search engine companies and others are realizing that real time content from hundreds of thousands or more sources can have significant societal, cultural, and commercial implications. Google and other major search engines are experimenting with methods to archive real time content (Krazit, 2010). Major news organizations routinely rely on the

* Corresponding author.
E-mail addresses: jjansen@acm.org (B.J. Jansen), zul112@ist.psu.edu (Z. Liu), cew5133@psu.edu (C. Weaver), gcampbell@gmail.com (G. Campbell), matthew.gregg@gmail.com (M. Gregg).

0306-4573/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2011.01.007

social Web for insights from frontline observers. People worldwide get near instantaneous observations from major crises, cultural occurrences, and political happenings. Online marketers are working to leverage the social Web for commercial purposes. As such, informational behaviors are changing as people become more familiar with accessing the stream of real time content for a variety of purposes. People are also becoming accustomed to receiving real time content directly from major news events, sometimes within minutes of these events occurring. With this increased market, there is enhanced ecommerce interest in real time searching. Therefore, there are significant opportunities in providing real time search services, with both the major search engines and new firms entering the marketplace offering search technologies for real time content.

However, there has been little to no systemic investigation that we could locate of how real time search is unfolding. How are users engaging with real time search technologies and services? What are the topics of interest for real time search? Is there economic value in this real time search data? How do real time search behaviors compare to traditional Web search? These are some of the motivators for our research.

2. Background of real time content and searching

The theoretical basis for this research is human information processing which is the method of acquiring, interpreting, manipulating, storing, retrieving, and classifying recorded information (Wilson, 2000). Searching is a manifestation of acquiring information, and prior work has examined theoretical aspects of using technology to conduct this search (Jansen & Rieh, 2010; Järvelin & Wilson, 2003) and empirical aspects of Web searching (Jansen, Spink, & Saracevic, 2000; Silverstein, Henzinger, Marais, & Moricz, 1999). Given that different information searching systems provide different affordances (for discussion of affordances, see Gibson, 1977), one could reasonably expect different user behaviors and interactions in different types of searching mediums. Given the unique nature of both real time content and real time searching systems, research on understanding current real time search behaviors and other real time searching aspects could shed light on emerging trends in this novel searching environment.

We define real time content as short status message postings (Jansen, Chowdhury, & Cook, 2010), sometimes with links to longer documents or multimedia content. Real time content is typically generated on social networking and media platforms, such as Twitter tweets, Facebook newsfeed, or LinkedIn shares. Real time content is normally created for the immediate temporal context, to be consumed as soon as produced rather than for archival intentions (Spark, 2009). Real time content addresses the question, "What is going on right now?" (OneRiot, 2009) or "What is going on right now in this location?", when there is a geographical interest. With respect to time, real time content has some resemblance to temporal information streams (c.f., Fenstermacher & Ginsburg, 2003; Kammenhuber, Luxenburger, Feldmann, & Weikum, 2006; Perkio, Buntine, & Perttu, 2004) or geotargeted information (c.f., Gan, Attenberg, Markowetz, & Suel, 2008; Silva, Martinsa, Chavesa, Afonsoa, & Cardosoa, 2006).

Of course, the social aspect of real time content is also a defining characteristic, with the idea being that one can harness the underlying social network of a user to bring more targeted and, therefore, relevant information. In some respects, this social aspect of real time content has parallels with related research in recommender systems (c.f., Resnick & Varian, 1997; Schafer, Konstan, & Riedi, 1999) and collaborative filtering (c.f., Herlocker, Konstan, Terveen, & Riedl, 2004; Kautz, Selman, & Shah, 1997). In the recommender systems research area, one leverages the natural social process of people making recommendations, aggregates these inputs, and then funnels the results to the appropriate individuals. In collaborative filtering, one can cluster people based on selected characteristics, identify information that these people find relevant or desirable, and recommend this information to others with similar characteristics. Like real time content, these areas rely on the social networking aspects of users to improve the quality of the information provided by the technology.

In spite of these similarities, finding relevant real time content can be quite a challenge, given the content's rapid pace of creation, the diminutive content of the individual post, the huge volume of aggregate content, the wide variety of topics, and the lack of standard ranking factors normally used by Web search engines (e.g., anchor text, hypermedia links, term frequency, etc.). Traditionally, web search engines index webpages periodically, return results based on a match to a search query, and rank retrieved results based on a mix of features. These techniques do not work as well with real time content, which typically has a short temporal half life, not a lot of terms within the message, often no hyperlink structure, and few traditional ranking factors. This situation has driven the need for specialized real time search technologies.

Therefore, real time search engines employ different methods than the crawling used for conventional Web content. While the traditional search engines are concerned with relevance, real time search engines factor in relevance, popularity, and temporal immediacy. Real time search employs a variety of implementation techniques for retrieving real time content, but nearly all real time search engines have some similar characteristics: (1) accept a query, (2) poll one or more social media sites, and (3) present an integrated stream of real time content. In this respect, real time search engines are similar to metasearch 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 (i.e., the search engine continuously polls). Therefore, there is no static search engine results page with a set number of links generated.

This new form of search raises questions, currently unanswered, concerning how users will interact with these technologies and what affordances these technologies offer searchers, content providers, and advertisers. Understanding user

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

behavior is not just of academic interest, as there are possible major economic ramifications. Research focusing on the social aspects of real time searching is just beginning to investigate how users will interact with current and possibly new forms of advertising on social media platforms and within social media search. As an example, OneUpWeb (2009) conducted a study using eye-gaze data on 25 participants (60% female) ranging in age from ages 18 to 55 (average = 35) while conducting product searches to investigate how users interacted with advertisements on Facebook and YouTube. Defining areas of interest (AOI) on the webpage, the researchers concluded that rules guiding search behavior on search engines may not apply on social sites. Also, the different searching environment may encourage users to develop different scan habits while engaging with search results (OneUpWeb, 2009).

However, limited published information exists on how people actually search for real time content across those social websites. Prior work has examined the informational value of portions of this real time content and methods of people accessing this content. However, we could locate no studies of real time search. Investigating the real time content stream, Jansen, Zhang, Sobel, and Chowdhury (2009) examined the discussion of brands on Twitter, showing that approximately 20% were informational seeking in nature, in addition to the expression of sentiments and opinions. Morris, Teevan, and Panovich (2010) report that the most popular type of information seeking on social networks is recommendations and opinions. Asur and Huberman (2009) relate social media content to the inference of real world events. In terms of searching this content, Mishne and de Rijke (2006) examine a search log from Blogdigger.com, reporting that searching behavior is similar to that of general Web searching, with short sessions in term of length and duration, diminutive queries, and limited examination of results. However, the interface to this blog search engine is similar to that of traditional Web search engines, so it is perhaps not surprising that the user behaviors are similar. When search engines offering diverse affordances, like Aardvark, we see different user behaviors, especially in terms of query formulation (Horowitz & Kamvar, 2010).

Given the limited prior work and uniqueness relative to other forms of Web searching, we do not yet know how users interact with real time search technologies, what these users search for, or whether there is advertising value in this search stream.

In this research, we address these issues, using search data from Collecta, a real time web search engine. Initial results on user behavior were reported in a conference poster (Jansen, Campbell, & Gregg, 2010). In this article, we expand this research with both a topical classification and an economic analysis, presenting a more complete picture of real time search on the web.

3. Research question

With this motivational background, our study of real time searching addresses the following research questions:

- *Research Question 01*: What are the characteristics of queries submitted to real time search services?

For this research question, we are interested in understanding user search behaviors occurring on real time search engines. There have been numerous investigations of user search behaviors with traditional Web search engines (e.g., Cockburn & McKenzie, 2000; Jansen & Spink, 2005; Silverstein et al., 1999); however, given the unique structure and temporal flow of real time content, we are interested in exploring possible differences in behaviors between these two unique searching contexts. Such an understanding of user behavior in the real time search area can lead to better insights for system design and content creation that supports actual web searchers.

- *Research Question 02*: What are the topics of real time content queries?

For this research question, we are interested in the topics searched for by users of real time search engines. There has been significant prior work in query topic investigations (e.g., Beitzel, Jensen, Chowdhury, Grossman, & Frieder, 2004) and investigating techniques for topic identification (e.g., Beitzel, Jensen, Chowdhury, & Frieder, 2007; Shen et al., 2006). However, real time content is perhaps more temporal and more localized than queries in other contexts. Therefore, search topics may differ in the real time search environment. Such an understanding of topics in real time search can advance the future design and implementation of subject specific query reformulation and search algorithms. It would also benefit marketing professionals by enabling them to provide more targeted online advertising.

- *Research Question 03*: What is the economic value of real time search queries?

There has been no online search business model more successful than keyword advertising, financing Web search as we currently know it (e.g., Battelle, 2005; Jansen & Mullen, 2008). The success of sponsored search provides a mechanism for monetizing Web search traffic, spawning a wide variety of search related Web industries, such as search engine marketing and affiliate marketing. Clearly, the search stream from traditional Web search engines has economic value, as evidenced by the success of sponsored search. However, there has been no economic assessment of real time search traffic. Such a monetization provides insights into the potential for future advertising platforms for both real time search engines and related real time content platforms. Information concerning this monetization would also be of great interest and value to marketing professionals.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

We investigate these research questions using a large amount of search log data collected from an operational real time search engine. Given the nature of these research questions, a search log is an appropriate data collection approach, and search log analysis is an appropriate methodological technique for this investigation (Jansen, 2006).

4. Research design

For this research study, we use data from Collecta (http://collecta.com/), one of the most popular real time search services at the time of data collection (http://www.crunchbase.com/company/collecta). However, most real time search engines have similar characteristics of accepting a query, polling real time content sites, and then presenting an integrated content stream. Given that real time search engines provide similar affordances, the results might be applicable to other real time search engines and users of these engines. This has been true in traditional Web searching (Markey, 2007a, 2007b), where similar search engines have similar user behaviors. Naturally, this assumption would have to be empirically tested with data from other real time search engines.

4.1. Collecta

Based on features the time of the study, Collecta provides real time content from Web sources, including 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 queries from searchers, uses XMPP to communicate with social media sites in order to submit the query, and presents a continuous, aggregated, and temporal stream of real time content from these sites. Collecta provides blog posts, comments on blog posts, along with status messages from social media sites, such as Twitter.

Founded in November, 2008, Collecta went live in June, 2009 (Schonfeld, 2009a). Collecta also offers site specific search services for MySpace (http://myspace.collecta.com/). See Fig. 1 for an overview of the Collecta interface and features at the time of the study.

Referring to Fig. 1, the Collecta interface accepts a query via a search bar, with text box and submit button. Once the query is submitted, the platform presents the user with the search status and a stream of search results. By selecting one result in the results stream, Collecta displays that result in the upper right for easier viewing.

Fig. 1. Collecta interface with search features, results, and explanatory notes.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

While the search list in traditional Web searching is static once the query has been submitted, real time search engines continually poll content sites. Therefore, the results are persistently updated if new content matching the query appears. A user can view results that have moved below the fold using the 'Older Results' button. The Collecta interface also has options for hot or trending topics and for searching various type of content. Previous queries submitted by the user during this session are listed, with notification of those previous queries having new results.

5. Data collection and analysis

In a search engine transaction log, we collected trace data of search interactions executed on Collecta from 4 June to 9 December, 2009. We employed standard log analysis methodology (Jansen, 2006) in analyzing the search engine log. The search log contains 1005,296 records, each with five 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*: the terms as entered by the user.
- *Category*: an indication of specific categories as selected by the user. The set of options are: Stories, Comments, Updates, Photos, and Videos, with All being the default (see Fig. 1).

The following are key concepts for Web search (i.e., the process of a searcher interacting with a search engine in order to locate relevant content in response to a query):

- *Term*: a series of characters within a query separated by white space or other separator. We parsed these terms from the query.
- *Query*: a string of terms submitted by a searcher in a given instance of interaction with the search engine. Each query was stored in its own record.
- *Session*: a series of queries submitted by a user and related interactions during an episode of interaction between the user and the Web search engine around a single topic. Due to methods of data collection, we could not adequately determine session boundaries in this dataset. However, we do comment on session related aspects.

5.1. Analysis of querying behavior

The primary unit of analysis for this research is the user query, although we do report some user, session, and term level data. The query level analysis focuses on occurrences of queries both from individual IP addresses and across the entire dataset. We investigate key aspects of user behavior at the query level, namely query length and query occurrences. At the search episode and session levels of analysis, we study searching measures such as average number of queries and single query sessions, in order to compare with Web searching on traditional search engines.

We relied on the user identification code to isolate searchers. Our original intent was to use this as a surrogate for individuals. However, due to the high number of accesses to Collecta via the application program interaction (API) and the nature of logging the queries from the Collecta main Webpage, this was not feasible. Therefore, we had to center this level of analysis at the computer level (i.e., the computer that accesses the Collecta search engine to submit a query).

At the term level of analysis, in addition to presenting the most frequently submitted terms, we also used the mutual information statistic to investigate term association in the dataset. The mutual information formula measures term association and does not assume mutual independence of the terms within the pair. We calculated the mutual information statistic for all term pairs within the data set. Many times, a relatively low frequency term pair may be strongly associated (i.e., if the two terms always occur together). The mutual information statistic identifies the strength of this association. 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:

$$P(w_1) = \frac{F1}{Q'}; \quad P(w_1) = \frac{F2}{Q'}; \quad P(w1, w2) = \frac{F12}{Q'}$$

The frequency of term pairs is the occurrence of the term pair within the set of queries. However, since a one-term query cannot have a term pair, the set of queries for the frequency base differs. 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:

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

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

where Q_n is the number of queries with n words ($n > 1$), and m is the maximum query length. So, queries of length one have no pairs. Queries of length two have one pair. Queries of length three have three possible pairs. Queries of length four have five possible pairs. This continues up to the queries of maximum length in the data set. The formula for queries of term pairs (Q') accounts for this term pairing.

5.2. Topical classification of real time queries

For the topical classification process, we employed an automatic categorization method by submitting queries to Google Directory (http://directory.google.com/) and retrieving the classifications. The use of Web documents for the topical classification of both queries and user intent has been shown to be a beneficial approach in prior work (Shen et al., 2006; White, Bilenko, & Cucerzan, 2007). Our approach is similar, except that we leverage Google Directory for the labeling of the individual webpages. Google launched its Web directory service by integrating Open Directory Project's (http://www.dmoz.org/) manually annotated data (Chirita, Nejdl, Paiu, & Kohlschütter, 2005) together with PageRank technology to provide a hybrid searching experience which is "uniquely broad and deep" (Sherman, 2000). As a Web index, Google Directory contains a list of hierarchical taxonomies, including 15 top-level parent categories (Arts, Home, Science, Business, News, Shopping, Computers, Recreation, Society, Games, Reference, Sports, Health, Regional, and World) and a number of corresponding sub-categories within each.

Different from the standard search engine results page, Google Directory accepts a query, returns results, and also retrieves a hierarchical list of related categories accompanying each search result. For instance, if we search using the query 'real time Web search' using Google Directory, we will get "Collecta" as the title of our first returned result, together with a list of hierarchical class labels "Computers > Internet > Searching > Search Engines > Specialized" as the related categories of that webpage, as can be seen in Fig. 2.

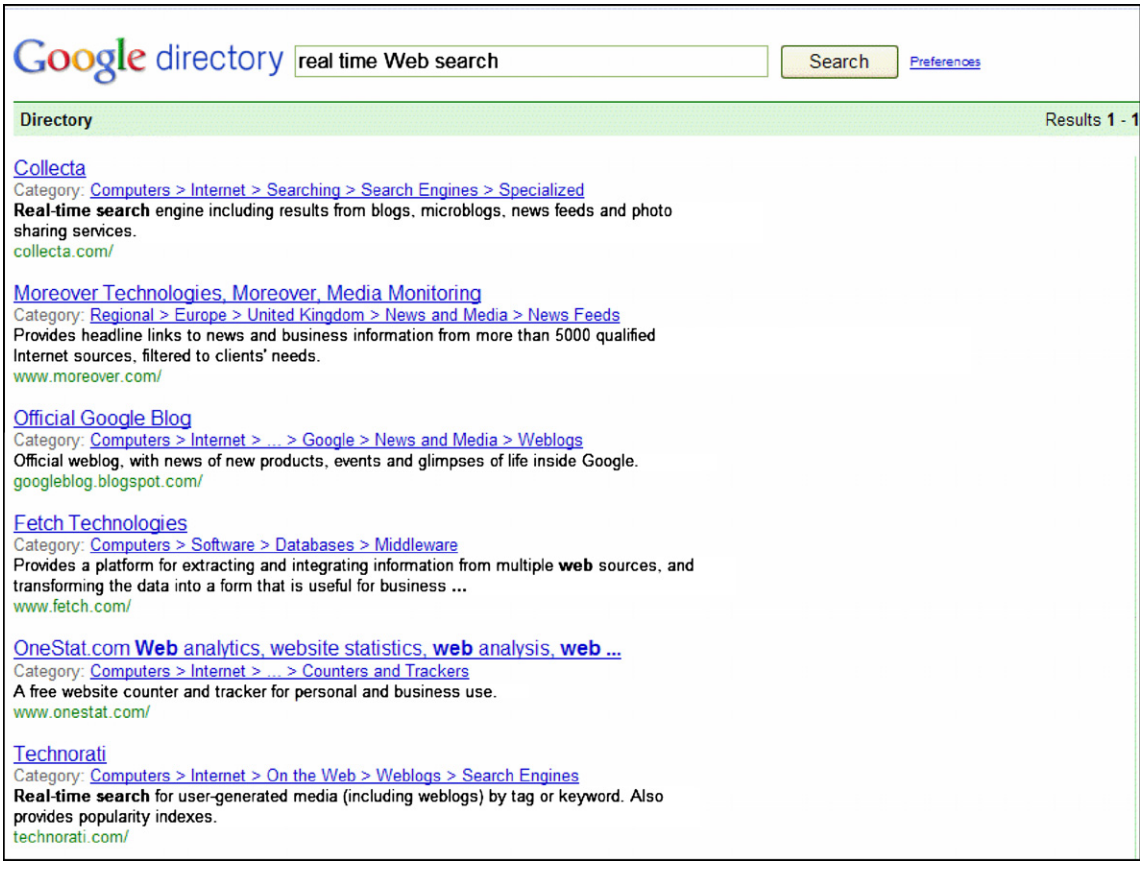

Fig. 2. Illustration of the Google Directory for query of "real time Web search".

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

To obtain the topical categories of the real time content for our research, we used an automated script to submit each unique query from the data log as a request to Google Directory. We then processed these results by parsing out all the top-level categories from each of their corresponding hierarchical class list. By sorting each ten top-level class label according to its frequency of occurrence within that search engine results page, we assigned the most frequently presented class label from the results listing to the query as its primary category classifier.

After the first round of classification, we discovered that a large number of queries were classified into the "Regional" category. In light of the less specific information provided by the "Regional" label compared to the other 14 categories, we decided to further categorize those "Regional" queries in a second round of classification.

By further examining the top ten hierarchical categories of each "Regional" query, we located occurrences of the other 14 top-level parent categories from these queries. In the same way as mentioned in the first round of classification, we assigned the most frequently occurring category labels to those "Regional" queries.

Given the validity of using webpages for topical classification and the robustness of the Google Directory, we believe this methodological approach to be both novel and valid for our purposes.

5.3. Economic value of real time searching

To determine economic value of the real time search traffic, we leveraged Google AdWords (http://adwords.google.com/), which is Google's sponsored search platform (Fain & Pedersen, 2006; Jansen & Mullen, 2008). The Google AdWords platform is unique in that one can leverage the system to determine a monetary value for a term as millions of advertisers can bid on keyphrases that appear in search queries. As such, it is really an unparallel database in this regard for determining the economic value of a term and, by association, a searcher's need.

Specifically, the Google AdWords auction for keywords is an extension of a single-item second price auction, known as a multi-item second price auction or a Generalized Second Price (GSP) auction, to bring more stability to the auction bidding, increase profits, and help reduce (but not eliminate) strategic bidding. The aggregate set of advertisers bids on keywords they are interested in. More bids mean a higher cost per click (CPC) for that keyword. For a walk through of the Google AdWords auction process, see "Introduction to the Google Ad Auction" (http://www.youtube.com/watch?v=K7l0a2PVhPQ). The outcome of these auctions is a market determination of the economic value of the keyword traffic, which we deemed a suitable methodological approach for our purposes of determining the value of the real time search traffic.

In fact, Google AdWords may well be the largest auction house in the world, with millions of auctions occurring daily (i.e., whenever a query is submitted for which two or more advertisers have placed a bid on an associated keyphrase). As such, AdWords is a viable, and perhaps the largest, platform for the economic assessment of words. Although we could find no other academic publication using this method, the approach is similar to the use of external databases, such as WordNet (Voorhees, 1993) or tagged search records (Beitzel et al., 2007), for classification of queries.

For this research, we leveraged the overall market assessment from the Traffic Estimator as a gauge of the monetization of the keyword traffic using what advertisers would have paid to market to this search traffic on Google. As a tool for advertisers, Google offers the Traffic Estimator application within AdWords, which provides advertisers a method for determining the cost of adding new keywords to their accounts. The Traffic Estimator offers a range of options for traffic estimation. For this research, we selected US dollars as the currency, English as the language, and the United States as the geographical location. We then submitted each unique query from the data log to the Traffic Estimator, which then generated a variety of measures including CPC, estimated clicks per day, and estimated cost per day, as shown in Fig. 3.

We used only the volume of unique queries given the current vast difference in search traffic between real time search engines and Google. Therefore, our estimate is conservative.

6. Results

6.1. Aggregate results

Overall, the 1005,296 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 API. This is unique compared to the use of traditional Web search engines, in which the majority of the traffic occurs on the website. This access of real time search engines via APIs has helped feed the growth of real time search and has resulted in an intense array of API widgets and development by the real time search engines (Schonfeld, 2009b; Siegler, 2009).

Each IP address for the API accesses submitted an average of 23 queries over the data collection period or 0.002% of the total queries. As shown in Table 1, 37% of the users not on the Collecta website (i.e., those using some form of API) submitted only one query during the data collection period, perhaps indicating some experimentation rather than routine use. Conversely, there are also a small number of users who submitted a large number of queries, indicating the use of some automated tools that one would expect with API utilization. These lengthy searching episodes likely represent a series of topical interests (Jones & Klinkner, 2008).

During the 190 day period, the average number of queries submitted per day was 5291, although there were substantial spikes and dips. It certainly reflects a lower search volume than traditional Web search on the major search engines, which

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

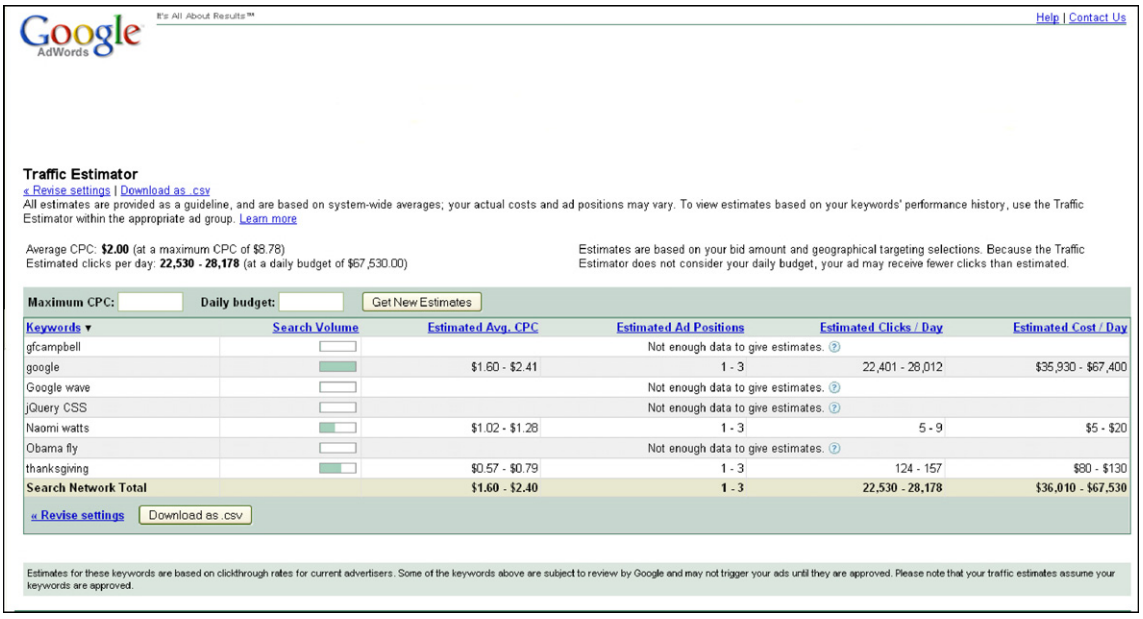

Fig. 3. Search results from Google AdWords traffic estimator.

Table 1
Number and percentage of real time queries by IP address.

Queries by IP address	Occurrences	Percent (%)
1	16,121	37.4
2	8741	20.3
3	5113	11.9
4	3254	7.5
5	2084	4.8
6	1508	3.5
7	1087	2.5
8	841	1.9
9	641	1.5
>=10	3750	8.7
	43,140	100.0

might be caused by the infancy of the real time searching model. Search distribution follows the typically hourly fluctuations (Fig. 4) that one sees from analysis of other search engines, with dips during the early morning hours and typical fluctuations during the week with (Fig. 5) with higher traffic volumes during the workweek and lower traffic on the weekends.

6.2. Querying behaviors

Referring to research question 01, (*What are the characteristics of queries submitted to real time search services?*), at the query level of analysis, of the 1,005,296 queries in the dataset, 297,392 were unique (30%). This is low, with studies of Web search reporting unique queries as high as 63% (c.f., Silverstein et al., 1999), indicating standing queries, persistent information interests by users, or searching on popular topics. Most real time search platforms, like Collecta, provide mechanisms for standing searches, which are typically clickable links that execute a query against the real time search engine. Third party applications also provide this querying service. This continually polling of the real time content stream is a unique aspect that differs from traditional Web search, where the results are relatively more static. Therefore, the correlation of 'query traffic' on traditional search engines and on real time search engines should take this continual pinging into account.

We examined also query length, as shown in Fig. 6. We see that there were a small percent of null queries (i.e., users coming to the search engine and submitting a query with no terms). Again, this is a small percentage, not in line with studies of Web search (c.f., Silverstein et al., 1999), where null queries range from 30% to 40%. 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 (c.f., Beitzel et al., 2004).

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

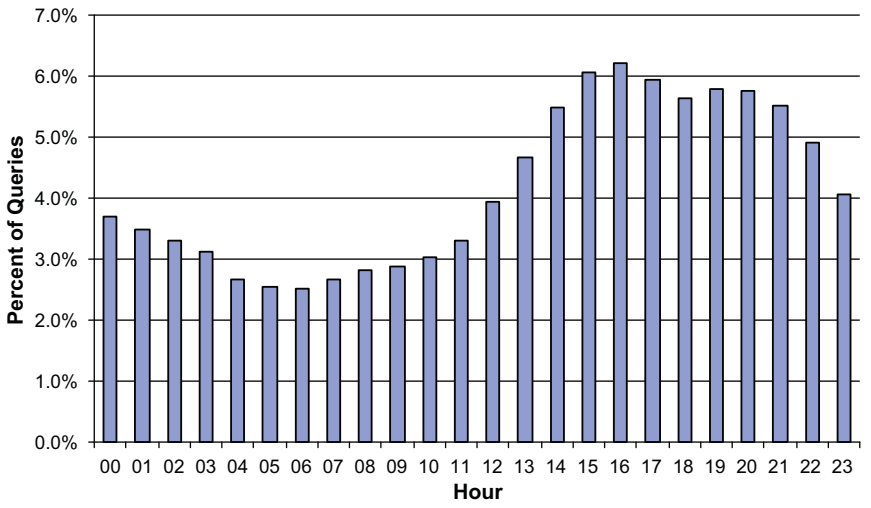

Fig. 4. Distribution of real time queries by hour of the day.

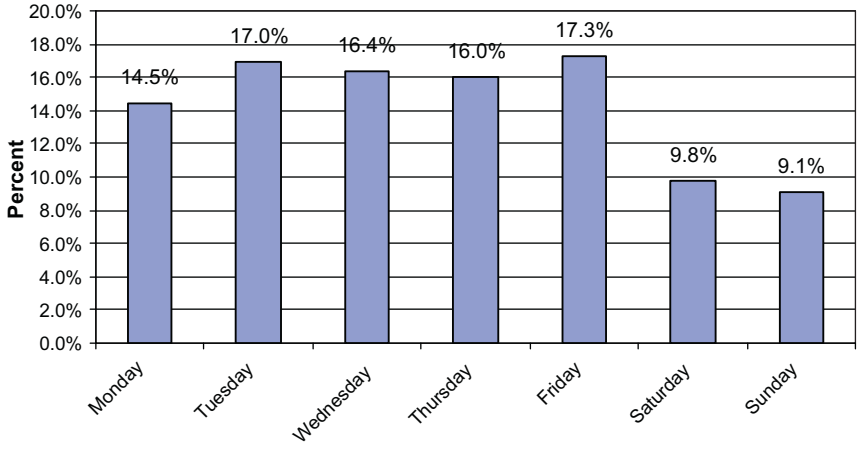

Fig. 5. Distribution of real time queries by day of week.

The most popular queries are shown in Table 2, some of which are consistent with traditional Web searching (Lee, Liu, & Cho, 2005). Most notable is the lack of pornographic queries, which are typical in Web search logs (c.f., Beitzel et al., 2004; Silverstein et al., 1999). 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%), which is in line with term usage in traditional Web searching (c.f., Jansen & Spink, 2005; Silverstein et al., 1999). We see from Table 2, occurrence of some prominent stop words (a.k.a., skip words), indicating use of natural language phrases. There are also high occurrences of technical terms (e.g., *foobar*, *python*, *jquery*, *ajax*), indicating a possible early adopter audience relatively more technical than the general population, which is a trend inline with adoption of other web services. There is also the possible use of the real time service in a navigational role, with use of brand names (i.e., collecta, google).

Examining Fig. 7, we see that the log–log plot of rank-frequency of term distribution adheres to a power law distribution, which is typical for Web query terms. In Table 3, the top most searched terms accounted for 0.03% of term occurrences, and five of the terms (*or*, *de*, *the*, *in*, *of*) are indicative of the use of natural language query phrases.

Examining term pairs occurring more than 100 times using the mutual information statistic (mis) showing strength of term association (Table 4), we list the most strongly associated term pairs. Again, the lack of pornographic phrases is unusual, with most term pairs being associated with people in the news at points during the data collection period. This is consistent with Web searching trends, where popular entertainers are frequent searching topics (see Google Trends, http://www.google.com/trends).

Collecta permits searchers to focus their queries on specific categories of content, with the default being all categories. Concerning searching on specific content, there were a total of 83,351 records with non-empty category field in the data set, as shown in Table 5.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

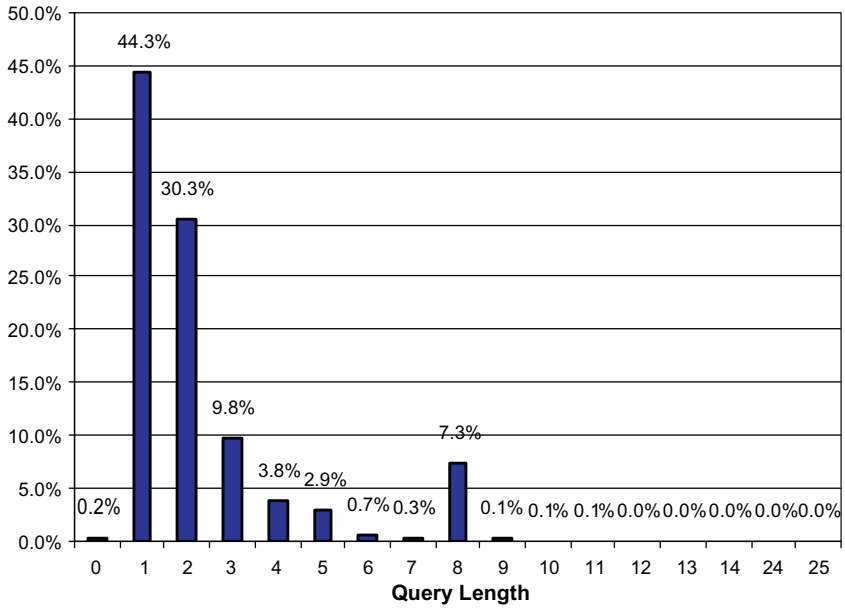

Fig. 6. Distribution of real time queries by query length.

Table 2
Real time queries with more than 1000 occurrences.

Query	Occurrences	Percentage
Maomi watts	3040	0.003
jQuery CSS	2787	0.003
Obama fly	2433	0.002
Thanksgiving	2318	0.002
Google wave	2177	0.002
Google	2136	0.002
Gfcampbell	2022	0.002
	1903	0.002
Sex	1536	0.002
Shark attack comment	1508	0.002
Tiger woods	1386	0.001
Foo	1309	0.001
Apple	1259	0.001
Crazy	1259	0.001
Search obama	1242	0.001
Michael jackson	1236	0.001
Obama or inauguration or inaugural or inaugurate	1203	0.001
Twitter	1193	0.001
Google voice	1184	0.001
Halloween	1167	0.001
Facebook	1155	0.001
Giannoulias	1108	0.001
Ufo	1102	0.001
New moon	1094	0.001
Real time search	1083	0.001
Leweb	1058	0.001
MarcAndreessen OR Netscape OR AOL	1006	0.001

It is apparent that users are following a common path of adhering to the default setting, with nearly 92% of users searching all categories. When specific categories were selected, the distribution was fairly balanced, with perhaps an inclination toward comments (2.68%).

6.3. Topical classification

We now address research questions 02 (*What are the topics of real time content queries?*). Utilizing the aforementioned automatic topical classification method from Google Directory, we categorized 216,963 queries out of the total 297,392

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

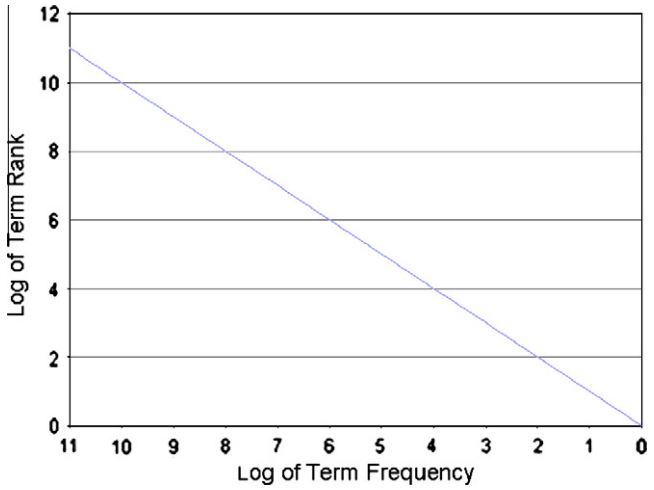

Fig. 7. Log–Log plot of terms from real time queries ranked by frequency.

Table 3
Terms with more than 4000 occurrences from real time queries.

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	9280	0.004
Story	8188	0.004
Google	7774	0.003
Iran	7064	0.003
De	6909	0.003
The	6793	0.003
Iphone	6622	0.003
In	6588	0.003
Of	5493	0.002
Foobar	5295	0.002
Com	4787	0.002
New	4359	0.002
Search	4123	0.002
And	4073	0.002

unique ones from our Collecta data set. There were 80,429 empty sets (37.1%) among our final topical categorization results (i.e., the query did not match any documents in Google Directory). Ambiguous queries are a standing issue in information retrieval (Sanderson, 2008). Additionally, this may indicate that real time searching on the web is different, at least relative to spread of topics, than traditional web searching. Given the considerable length of time that Google has had to collect and store queries (since 1998), 37% of terms in real time queries being unclassifiable is surprising. However, it has been reported that large percentages of queries from traditional web search engines are new from year-to-year (Beitzel et al., 2007), so this may be just be a phenomenon of web searching.

Table 6 displays the topical classification of the queries that we were able to classify.

As can be seen in Table 6, the topical category that has been retrieved by the users most is "Society," followed by "Arts," and "Computers". This differs from the topical characteristics of the traditional Web search as shown in the studies conducted by Spink, Özmutlu, Özmutlu, and Jansen (2002) and Beitzel et al. (2004). In their study of the Excite search engine,

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

Table 4
Strongly associated term pairs from real time queries with more than 100 occurrences.

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 5
Content categories as specified by searchers of real time queries.

Categories	Occurrences	Percent (not including Default) (%)	Percent (including default) (%)
Comments	26,991	32.38	2.68
Stories	16,759	20.11	1.67
Updates	15,908	19.09	1.58
Videos	12,021	14.42	1.20
Photos	11,672	14.00	1.16
Default	921,945	–	91.71
	1005,296	100.00	100.00

the researchers found that the major topical categories that users retrieved heavily were "Commerce, travel, employment or economy" (24.7%), followed by "People, places or things" (19.7%) (Spink et al., 2002). Beitzel et al. (2004) report that the three largest categories of Web queries were Shopping, Entertainment, and Porn. So, it appears that query topics in real time searching differs from that of traditional Web searching.

Real time searches have some interesting societal aspects, with queries such as *separate but equal*, *divorce rates*, and *abortion*. Art queries included *early irish literature analysis*, *I got a feeling black eyed peas*, and *heidi klum*. There were interesting queries occurring in some of the more narrowly focused categories. Science queries included *climate change*, *eclipse phase*, *shuttle landing*, and *wolfram alpha*. For Heath examples, people searched for *benylin mucus cough*, *laparoscopic surgery*, and *artificial knee replacement*. For the Home category, some example queries are *white bean soup*, *spicy tuna challenge*, and *composting*. As one can see from these example queries, users are searching for an array of topics that one might not at first consider as applicable to real time content subjects of interest. It may be early indications of the impact that real time content and search may have on a variety of human activities.

Table 6
Topical classifications of real time queries based google directories.

Google directory	Occurrence (overall)	Percentage (%) (overall)	Occurrence (unique)	Percentage (%) (unique)
Society	200,810	25.86	47,706	22.00
Arts	136,271	17.55	41,021	18.91
Computers	127,669	16.44	26,462	12.20
Business	106,013	13.65	39,559	18.23
News	38,223	4.93	8571	3.95
Sports	38,037	4.90	9372	4.32
Recreation	33,210	4.28	10,107	4.66
Science	24,788	3.19	8716	4.02
Shopping	21,060	2.71	8402	3.87
Health	19,779	2.55	6216	2.87
Games	12,224	1.57	4329	2.00
Reference	12,164	1.57	4083	1.88
Home	4847	0.62	2080	0.96
Google	1339	0.17	332	0.15
World	18	0.002	7	0.003
Total	776,452	100	216,963	100

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

Table 7
Upper and lower CPC and cost ranges for real time search queries.

Estimated avg. CPC		Estimated clicks/day		Estimated cost/day	
Lower	Upper	Lower	Upper	Lower	Upper
$0.99	$1.29	5807,176	7177,927	$25,180,284	$40,866,356

It is also interesting that, based on our first round of query classification, as explained in the methods section, nearly 36% of the queries were classified as Regional, indicating a localized or geographical inclination to real time search interests. This supports the concept that real time content many times has a geographical attribute or interest. This localized aspect makes real time search especially attractive to small to medium size enterprises that may service a given locality. The high occurrences of Regional topics for queries, also speaks to the immediacy (or 'now' aspect) of real time search, with an apparent interest in an immediate temporal need focused at a particular geographical point.

6.4. Economic value

We now address research question 03 (*What is the economic value of real time search queries?*). Of the 297,392 unique queries, we were able to retrieve economic data on 154,756 (52%). The remaining queries contained foreign characters, were in non-English or for some other reason, were unacceptable to the Google AdWords Traffic Estimator. The 48% of terms with unassigned monetary value was somewhat surprising, given that Google has had since 2000 (the beginning of Google AdWords) to develop monetization values for query terms. However, given the already noted high occurrence of new queries, combined with the foreign characters (perhaps reflecting the technical nature of many of these queries) and the non-English queries (our Google AdWords setting was for the US), this percent of un-monetized queries seems more reasonable.

The overall estimated economic value of the real time traffic was $33,023,320 if the real time content in response to these queries was available on Google search engine results pages, which is an average of the high and low estimate for each query that returned a value. Search results retrieved in response to this real time search traffic would generate an estimated 6,492,552 clicks at an average CPC of $1.14, with specific figures for the traffic shown in Table 7. This would indicate that efforts to design advertising platforms (Lawrence, 2010; OneRiot, 2009) for real time search may be fruitful.

From Table 7, we see the specific cost that an advertiser would have paid to compete with pay per click advertisements utilizing the traffic from the real time queries along with the estimated number of clicks on these advertisements. Although a direct comparison of traditional to real time search traffic has some caveats (such as search volume and advertising expectations, for example), these figures provide an indication of commerce potential of the real time search traffic.

Table 8 shows the top twenty five content keywords based on average cost per day. Naturally, these follow the higher priced keywords in sponsored search, given that we estimated the monetary value using the Google AdWords platform.

From Table 8, see that real time search queries contain terms that are highly competitive in the keyword advertising marketing, including *insurance*, *hotels*, *mortgage*, *travel*, *jobs*, and *flowers*. Although real time content providers, such as Twitter and Facebook, are exploring ways to advertise within the content stream, the interaction on real time search engines is more inline with the adverting platforms of traditional search engines. As these example queries indicate, the keyword advertising model may port to these real time search platforms more directly.

7. Discussion and Implications

As one of the first analyses of real time searching behaviors, these research results highlight interesting aspects of searcher queries, the nature of the searching stream, and the economic value of real time search.

7.1. Implications

First, there appears to be a heavy use of accessing real time search via secondary applications rather than directly from the website. Therefore, it appears that real time search results are being integrated as components of other websites or applications. While the use of the APIs has fueled the growth of real time searching, it also presents additional challenges for marketing and advertising on these platforms, as the advertisements may be easier to strip away when the search results are integrated with other applications.

Second, some of these APIs 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 (Foltz & Dumais, 1992), with fewer unique queries but a sustained information need. Again, this has interesting implications for advertisers, as standing information queries represent the opportunity for greater personalization and increased relevance of online advertisements for these searchers. This is also an interesting counter play to the immediacy of real time searching noted above in that there are real time searching interests that extend over a longer period than the immediate moment. For example, a person may be interested in real time content generated from conference attendees, which may extend over several days before dissipating. From our analysis, this appears to be a common occurrence.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

Table 8
Top 25 real time queries by estimated cost.

Keywords	Estimated avg. CPC		Estimated clicks/day		Estimated cost/day	
	lower	upper	lower	upper	lower	upper
Insurance	$11.82	$15.73	44,066	55,102	$520,860	$866,680
Quotes	$20.89	$27.01	12,909	16,136	$269,690	$435,860
Home	$3.66	$5.00	65,637	82,085	$239,950	$410,040
Mortgage	$12.29	$16.33	18,272	22,872	$224,660	$373,520
Car	$3.73	$5.17	47,318	59,167	$176,500	$305,800
Hotels	$2.52	$3.17	69,201	86,502	$174,710	$274,060
Free	$1.81	$2.29	91,723	114,657	$166,170	$263,090
Credit	$8.15	$10.52	19,567	24,479	$159,530	$257,470
Auto insurance	$26.21	$34.21	5972	7472	$156,510	$255,590
Hotel	$1.74	$2.58	64,358	80,461	$111,930	$207,690
Travel	$1.94	$2.83	62,691	73,302	$121,670	$207,150
Flights	$1.64	$2.09	67,124	77,842	$110,070	$163,010
City	$2.17	$3.25	44,327	47,716	$96,110	$155,180
Car insurance	$11.86	$16.73	7460	8799	$88,470	$147,240
Debt	$11.02	$14.09	9235	9900	$101,730	$139,470
Jobs	$1.43	$1.79	60,462	75,577	$86,530	$135,200
Flowers	$4.09	$5.18	20,052	25,066	$81,950	$129,740
Card	$4.52	$6.34	16,010	20,023	$72,330	$126,880
Beach	$2.40	$2.99	32,472	40,590	$77,780	$121,520
Auto	$2.84	$3.55	27,082	33,852	$76,840	$120,060
Real estate	$2.42	$3.02	29,071	36,339	$70,260	$109,780
Estate	$2.30	$2.87	29,715	37,143	$68,340	$106,780
Phone	$1.96	$2.62	31,610	39,518	$61,890	$103,480
Mortgage refinancing	$26.06	$38.28	4490	4780	$117,020	$182,940
Refinance	$11.79	$15.83	6304	6308	$74,330	$99,880
Average	$7.25	$9.74	35,485	43,427		
Total					$3505,830	$5698,110

Third, real time search differs in topics relative to Web search. There is a high occurrence of society, entertainment, technology, and politics, with a low occurrence of sexually-related queries. 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 the user experience. There is also a significant occurrence of searches on a focused geographical area, reflecting a strong locality component of real time search. This makes real time search attractive for mobile searching, which already has a significant geo-location attribute.

Finally, it appears that real time has potential as a viable economic platform, with significant monetization of the real time search traffic. With more than 52% of the real time search traffic having current commercial value, much of the real time search traffic can apparently be leveraged for marketing and e-commerce purposes. As real time advertising platforms develop, more refined methods of advertising assessments will surely be developed.

8. Discussion

Real time search is a compelling area of Web interaction with potential as a new channel for information gathering, advertising, and other uses. Given the unique aspect of real time content – naturally social, generally geographic, and inherently temporal – it would seem that users would engage with this information stream with different expectations. These different expectations will impact both searching behaviors and evaluating the search results. As people become more accustomed to using real time content, real time search will become even more important as an information medium. Therefore, understanding how people locate information in this context is critical to the development of future platforms and marketing processes.

Naturally, there are limitations to this research. The data comes from one real time search engine and therefore may not be representative of all users on all real time search engines. However, given the leading role Collecta plays in the real time search market at the time of the study, the variety of traffic, the length of the data collection period, and the similarity among real time search engine technologies, we would expect these results to be representative of users engaging with other real time search engines. Another limitation is that real time search is still in an incipient stage of use, and thus, user behaviors may change over time. However, using traditional Web search as a benchmark, user searching behaviors have remained fairly stable over time (Jansen & Spink, 2005).

There is also a limitation in the methodology of determination of economic value. We used the Google AdWords platform, which is an intention-driven advertising mechanism. That is, the searcher enters a query, expressing an intent or interest. Advertisers bid on these key phrases as an expression of their interest in these potential customers, creating an economic market for the terms. It is a matter of debate whether the same situation holds with real time content. Given its temporal

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

nature in both creation and searcher interest, real time content may not conform to the same advertising mechanisms. However, as the largest economic auction assigning monetary value to words, the Google AdWords platform data can serve as a general benchmark of advertising potential for real time search.

There are several strengths of this study. The data comes from a well known and popular real time search engine, with the data collection period covering more than 6 months. Therefore, we believe that we have a viable sample of the population. We analyzed a wide range of querying behaviors using industry standard methods and leveraged applications from the leading Web search engine at the time of the study for topical classification and monetization of traffic. Therefore, we believe that we present a wide spectrum of insights into the real time search.

9. Conclusion and future research

In this research, we analyzed the queries, search topics, and the economic value of real time search traffic over 190 days, using data from a real time content search engine. Our results show that access to real time search results differ from that of traditional search engines, query topics differ from that of traditional Web search, and there appears to be economic potential in monetizing real time search traffic. There are several areas of future work. On the user side, it would be interesting to investigate the contexts in which users desire real time search results, along with what situations real time search results would be of benefit. On the system side, the current integration of real time search results into the aggregate stream with other search results is rather straightforward. Clearly, more sophisticated methods of integration might improve precision of these results when combined with other content. On the business side, research into effective methods of leveraging the economic value of real time search results will be needed for the development of workable advertising and marketing platforms. The results reported from this research provide foundational understanding of real time content search, an increasingly important aspect of Web search.

Acknowledgements

We thank the three anonymous reviewers for their efforts during the processing of this manuscript. We especially thank *Reviewer #2* for the many content and editorial recommendations that substantially improved the presentation of this research.

References

Asur, S., Huberman, B. A. (2009). *Predicting the future with social media.* <http://www.hpl.hp.com/research/scl/papers/socialmedia/socialmedia.pdf>.

Battelle, J. (2005). *The search: How Google and its rivals rewrote the rules of business and transformed our culture*. New York: Penguin Group.

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D., Frieder, O. (2004). In M. Sanderson, K. Järvelin, J. Allan, P. Bruza (Eds.), *Hourly analysis of a very large topically categorized web query log* (pp. 321–328). Paper presented at the 27th annual international conference on research and development in information retrieval, Sheffield, UK (25–29 July).

Beitzel, S. M., Jensen, E. C., Chowdhury, A., Frieder, O. (2007). In C. L. A. Clarke, N. Fuhr, N. Kando (Eds.), *Varying approaches to topical web query classification* (pp. 783–784). Paper presented at the 30th annual international ACM SIGIR conference on research and development in information retrieval, Amsterdam, The Netherlands (23–27 July).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems, 30*(1), 107–117.

Chirita, P. A., Nejdl, W., Paiu, R., Kohlschütter, C. (2005). In R. Baeza-Yates, N. Ziviani (Eds.), *Using ODP metadata to personalize search* (pp. 178–185). Paper presented at the 8th annual international ACM SIGIR conference on research and development in information retrieval, Salvador, Brazil.

Cockburn, A., & McKenzie, B. (2000). What do web users do? An empirical analysis of web use. *International Journal of Human–Computer Studies, 54*(6), 903–922.

Fain, D. C., & Pedersen, J. O. (2006). Sponsored search: A brief history. *Bulletin of the American Society for Information Science and Technology, 32*(2), 12–13.

Fenstermacher, K. D., & Ginsburg, M. (2003). Client-side monitoring for web mining. *Journal of the American Society for Information Science and Technology, 54*(7), 625–637.

Foltz, P. W., & Dumais, S. T. (1992). Personalized information delivery: An analysis of information filtering methods. *Communications of the ACM, 35*(12), 51–60.

Gan, Q., Attenberg, J., Markowetz, A., Suel, T. (2008). Analysis of geographic queries in a search engine log. In S. Boll, C. Jones, E. Kansa, P. Kishor, M. Naaman, R. Purves, A. Scharl, E. Wilde (Eds.), 1st international workshop on Location and the web (pp. 49–56). Beijing, China.

Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing*. Hillsdale, NJ: Lawrence Erlbaum.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5–53.

Horowitz, D., Kamvar, S. D. (2010). In M. Rappa, P. Jones (Eds.), *The anatomy of a large-scale social search engine* (pp. 431–440). Paper presented at the 19th international conference on World Wide Web (WWW10), Raleigh, NC.

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

Jansen, B. J., Campbell, G., Gregg, M. (2010). In E. Mynatt D. Schoner (Eds.), *Real time search user behavior* (pp. 3961–3966). Paper presented at the ACM conference on human factors in computing systems (CHI2010), Atlanta, GA (10–15 April).

Jansen, B. J., Chowdhury, A., Cook, G. (2010). The ubiquitous and increasingly significant status message. *Interactions* May/June, 15–17.

Jansen, B. J., & Mullen, T. (2008). Sponsored search: An overview of the concept, history, and technology. *International Journal of Electronic Business, 6*(2), 114–131.

Jansen, B. J., & Rieh, S. (2010). The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Sciences and Technology, 61*(8), 1517–1534.

Jansen, B. J., & Spink, A. (2005). How are we searching the World Wide Web? A comparison of nine search engine transaction logs. *Information Processing Management, 42*(1), 248–263.

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

Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Sciences and Technology, 60*(11), 2169–2188.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007

Järvelin, K., Wilson, T. D. (2003). On conceptual models for information seeking and retrieval research. *Information Research, 9* (1) <http://InformationR.net/ir/9-1/paper163.html>.

Jones, R., Klinkner, K. L. (2008). In J.G. Shanahan (Ed.), *Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs* (pp. 699–708). Paper presented at the 7th ACM conference on information and knowledge management, Napa Valley, California, USA (26–30 October).

Kammenhuber, N., Luxenburger, J., Feldmann, A., Weikum, G. (2006). In *Web search clickstreams* (pp. 245–250). Paper presented at the 6th ACM SIGCOMM conference on internet measurement, Rio de Janeriro, Brazil.

Kautz, H., Selman, B., & Shah, M. (1997). Referral web: Combining social networks and collaborative filtering. *Communications of the ACM, 40*(3), 63–65.

Krazit, T. (2010). *Google launches Twitter timeline search.* <http://news.cnet.com/8301-30684_3-20002453-265.html> (18 May).

Lawrence, D. (2010). *How companies should approach the new Twitter advertising model.* <http://mashable.com/2010/04/14/twitter-advertising-strategies/> Retrieved 05.01.11 (14 April).

Lee, U., Liu, Z., Cho, J. (2005). In A. Ellis, T. Hagino (Eds.), *Automatic identification of user goals in web search* (pp. 391–400). Paper presented at the 4th international conference on World Wide Web, Chiba, Japan (10–14 May).

Markey, K. (2007a). Twenty-five years of end-user searching, part 1: Research findings. *Journal of the American Society for Information Science and Technology, 58*(8), 1071–1081.

Markey, K. (2007b). Twenty-five years of end-user searching, part 2: Future research directions. *Journal of the American Society for Information Science and Technology, 58*(8), 1123–1130.

Mishne, G., & Rijke, M. D. (2006). A study of blog search. In M. Lalmas, A. MacFarlane, S. Rüger, A. Tombros, T. Tsikrika, & A. Yavlinsky (Eds.), *Advances in information retrieval* (pp. 289–301). Berlin: Springer.

Morris, M. R., Teevan, J., Panovich, K. (2010). In E. Mynatt, D. Schoner (Eds.), *What do people ask their social networks, and why? A survey study of status message behavior* (pp. 1739–1748). Paper presented at the 28th international conference on Human factors in computing systems (CHI 2010), Atlanta, Georgia, USA (10–15 April).

OneRiot (2009). *The inner workings of a reatime search engine: Thoughts on realtime search.* <http://www.docstoc.com/docs/16947406/OneRiot-Inner-Workings-of-a-Realtime-Search-Engine/>.

OneUpWeb, (2009). *Seeing search go social: An eye tracking analysis on social networking sites.* Traverse City, MI: OneUpWeb.

Perkio, J., Buntine, W., Perttu, S. (2004). In *Exploring independent trends in a topic-based search engine* (pp. 664–668). Paper presented at the 2004 IEEE/WIC/ACM international conference on web intelligence (WI '04), Beijing, China (20–24 September).

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56–58.

Sanderson, M. (2008). In T.-S. Chua, M.-K. Leong (Eds.), *Ambiguous queries: Test collections need more sense* (pp. 499–506). Paper presented at the 31st annual international ACM SIGIR conference on research and development in information retrieval, Singapore, Singapore (20–24 July).

Schafer, J. B., Konstan, J., Riedi, J. (1999). Recommender systems in e-commerce. In S. Feldman, M. Wellman (Eds.), *1st ACM conference on electronic commerce* (pp. 158–166). Denver, Colorado.

Schonfeld, E. (2009a). *Collecta enters the real time search wars.* <http://techcrunch.com/2009/06/18/collecta-enters-the-real-time-search-wars/> Retrieved 05.01.11 (18 June).

Schonfeld, E. (2009b). *Collecta releases its real-time search api; oneriot responds with a challenge.* <http://techcrunch.com/2009/09/10/collecta-releases-its-real-time-search-api-oneriot-responds-with-a-challenge/> Retrieved 05.01.01 (10 September).

Shen, D., Pan, R., Sun, J.-T., Pan, J. J., Wu, K., Yin, J., et al (2006). Query enrichment for web-query classification. *Transactions on Information Systems, 24*(3), 320–352.

Sherman, C. (2000). Google introduces web directory using Netscape's open directory project data. *Information Today, 17* (5), *14*, *17*(5), 14.

Siegler, M. (2009). *Oneriot real-time search api now open to all.* <http://techcrunch.com/2009/07/09/oneriot-real-time-search-api-now-open-to-all/> Retrieved 05.01.11 (9 July).

Silva, M. J., Martinsa, B., Chavesa, M., Afonsoa, A. P., & Cardosoa, N. (2006). Adding geographic scopes to web resources Computers. *Environment and Urban Systems, 30*(6), 378–399.

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

Spark, D. (2009). *Real-time search and discovery of the social web.* San Francisco, CA: Spark Media Solutions.

Spink, A., Özmutlu, S., Özmutlu, H. C., & Jansen, B. J. (2002). US versus European web searching trends. *SIGIR Forum, 32*(1), 30–37.

Voorhees, E. M. (1993). In R. Korfhage, E. Rasmussen, P. Willett (Eds.), *Using wordnet to disambiguate word senses for text retrieval* (pp. 173–180). Paper presented at the 16th annual international ACM SIGIR conference on research and development in information retrieval Pittsburgh, PA (27 June–01 July).

White, R. W., Bilenko, M., Cucerzan, S. 2007. Studying the use of popular destinations to enhance web search interaction. In W. Kraaij & A.P.d. Vries (Eds.), *The 30th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 159–166). Amsterdam, The Netherlands

Wilson, T. D. (2000). Human information behavior. *Informing Science, 3*(2), 49–55.

Please cite this article in press as: Jansen, B. J., et al. Real time search on the web: Queries, topics, and economic value. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2011.01.007