Temporal Analysis of a Very Large Topically Categorized Web Query Log

Steven M. Beitzel,* Eric C. Jensen, Abdur Chowdhury, Ophir Frieder, and David Grossman
Department of Computer Science, Information Retrieval Laboratory, Illinois Institute of Technology, Chicago, IL 60616. E-mail: {steve, ej, abdur, ophir, dagr}@ir.iit.edu

The authors review a log of billions of Web queries that constituted the total query traffic for a 6-month period of a general-purpose commercial Web search service. Previously, query logs were studied from a single, cumulative view. In contrast, this study builds on the authors' previous work, which showed changes in popularity and uniqueness of topically categorized queries across the hours in a day. To further their analysis, they examine query traffic on a daily, weekly, and monthly basis by matching it against lists of queries that have been topically precategorized by human editors. These lists represent 13% of the query traffic. They show that query traffic from particular topical categories differs both from the query stream as a whole and from other categories. Additionally, they show that certain categories of queries trend differently over varying periods. The authors key contribution is twofold: They outline a method for studying both the static and topical properties of a very large query log over varying periods, and they identify and examine topical trends that may provide valuable insight for improving both retrieval effectiveness and efficiency.

Introduction

Understanding how queries change over time is critical to developing effective, efficient Web search services. The Web is a dynamic, uncooperative environment with several issues that make analysis of a Web search very difficult. These include:

- The Web is a dynamic collection: its data, users, search engines, and popular queries are constantly changing (Beitzel, Jensen, Chowdhury, Grossman, & Frieder, 2004).
- Typical Web search engine traffic consists of many hundreds of millions of queries per day (Sullivan, 2003) and is highly diverse and heterogeneous (Eastman & Jansen, 2003), requiring a large sample of queries to adequately represent a population of even one day's queries.
- It is difficult to accurately capture the user's desired task and information need from Web queries, which are typically very short (Jansen, Spink, & Saracevic, 2000).

To date, studies of Web query logs have been limited to what we term *static analysis*; that is, analysis that examines the log as a closed system, focusing on general, discrete characteristics rather than queries in specific categories or changes and trends over time. One relevant example of this is a recent study by Jansen, Spink, and Pederson (2005) that compares two AltaVista™ logs from single days in 1998 and 2002, presenting mostly static analysis of the two logs. It lacks any detailed analysis of topical trends over a continuous period. Our goal is to develop a method (termed *temporal analysis*) of analyzing very large query logs that is capable of capturing topical trend information over time in addition to traditional static analysis. To aid us in achieving this goal, we introduce metrics that are suitable for examining temporal changes in a query log. It is our hope that these metrics can provide a common field of comparison for studies that may perform temporal analysis on other query logs in the future. Temporal analysis can then be used to foster improvements in both the effectiveness and the efficiency of a Web search. This study builds on our earlier work examining circadian changes in the query stream over the hours in a day (Beitzel et al., 2004). We expand on the conclusions found in that study by examining the behavior of the query stream over longer periods, such as several days, weeks, and months.

We analyzed two query logs from America Online's (AOL) Web search service. The first log contained all the queries (several hundred million) from an entire week in December 2003. We used it to study the changes in both static characteristics and topic-specific behavior in the query stream over the hours in a single day. The second log contained the entire query stream from AOL's Web search service over a continuous 6-month period from September 2004 through February 2005 (several billion queries[1]). This

*Steven M. Beitzel is now with Telcordia Technologies, Piscataway, NJ.

Received August 3, 2005; revised January 29, 2006; accepted January 29, 2006

 ● Published online 22 November 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20464

[1]The exact size of the query log is withheld, as it constitutes AOL™ proprietary data.

was used to examine static and topical changes over longer periods such as days, weeks, and months. These logs represent a population of tens of millions of users searching for a wide variety of topics. With both logs, we focused on analyzing overall query traffic as well as changes in popularity and uniqueness of topical categories. Emphasis on changing query stream characteristics over this temporal aspect of query logs, as well as the sheer volume of data examined, distinguishes this work from prior static log analysis, surveyed in Jansen and Pooch (2001).

We also analyzed the queries representing different topics using a topical categorization of our query stream. These cover approximately 13% of the total query volume. We suspected that traffic behavior for some categories would change over time and that others would remain stable. We examined the traffic characteristics of 16 different topical categories and made the following observations:

- Some topical categories vary substantially more in popularity than others as we move through an average day. Some topics are more popular during particular times of the day, whereas others have a more constant level of interest over time.
- The query sets for different categories have differing similarity over time. The level of similarity between the actual query sets received within topical categories varies according to category.
- Over longer periods, trends in the query sets for particular categories can sometimes demonstrate wildly different behavior. The similarity trends between category query sets can vary substantially depending on the length of time being examined. Categories with potential seasonal changes such as Holidays and Sports can exhibit different behavior over a period of months than they do over an average day.

Although some of these observations may seem obvious in hindsight, the fact that we observed what might be reasonably expected is an indicator that our proposed metrics for temporal analysis have merit. Practical applications made possible by analysis of this type may include improved caching strategies, load-balancing algorithms, query disambiguation and routing, etc. The details of these applications are left to future work, as the main goal of this study is to show that our method of temporal analysis is viable, and gives the researcher a novel way to examine Web search behavior over time.

Prior Work

Examinations of search engine evaluation indicate that performance likely varies over time due to differences in query sets and collections (Hawking, Craswell, & Griffiths, 2001). Although the change in collections over time has been studied (e.g., the growth of the Web; Lawrence & Giles, 1998), analysis of users' queries has been primarily limited to the investigation of a small set of available query logs that provide a snapshot of their query stream over a fixed period.

Existing query log analysis can be partitioned into large-scale log analysis, small-scale log analysis, and some other applications of log analysis such as categorization and query clustering. A survey covering a great deal of relevant prior work in search studies can be found in (Spink & Jansen, 2004). Jansen and Pooch (2001) provide a framework for static log analysis, but do not address analysis of changes in a query stream over time. Given that most search engines receive between tens and hundreds of millions of queries a day (Sullivan, 2003), current and future log analysis efforts should use increasingly larger query sets to ensure that prior assumptions still hold.

Previous studies have measured overall aspects of users' queries from static Web query logs. In the only large-scale study (all others involve only a few million queries), Silverstein, Henzinger, Marais, and Moricz (1999) conclude that users typically view only the top 10 search results, and that they generally enter short queries. This characterization is based on a static analysis of an AltaVista query log taken over 6 weeks in 1998 consisting of 575 million nonempty queries. They also found that only 13.6% of queries appear more than three times, the top 25 queries represent 1.5% of the total query volume, and in 75% of sessions users do not revise their queries. Additionally, co-occurrence analysis of the most frequent 10,000 queries showed that the most correlated terms are often constituents of phrases. No time-based or topic-based analysis of this query load was reported; it does not provide insight into how or when any usage or topical interest changes occur. Other studies examine the effect of advanced query operators on the search service coverage of Google, Microsoft Network (MSN), and AOL, finding that in general they had little effect (Eastman & Jansen, 2003). These overall statistics do not provide any insight into temporal changes in the query log, but do provide some insight into how people use search services.

Jansen and colleagues also provide analysis of query frequency (Jansen & Pooch, 2001; Spink, Wolfram, Jansen, & Saracevic, 2001). Their findings indicate that the majority (57%) of query terms from the Excite log of more than 51,000 queries are used only once, and a large majority (78%) occur three times or less. These studies show that neither queries nor their component terms follow a Zipfian distribution (meaning the frequency of the *nth* most common term in a language is approximately inversely proportional to n), as the number of rare, infrequently repeated queries and terms is disproportionately large. Other studies have focused on user behavior at the query session level and found varying results, with some estimating reformulated queries constituting 40–52% of queries in a log (Spink, Jansen, & Ozmutlu, 2000; Spink, Jansen, Wolfram, & Saracevic, 2002). Wang, Berry, and Yang (2003) examined a log of more than 500,000 queries to a university search engine from 1997–2001. They found trends in the number of queries received by season, month, and day. We extend upon this work by examining the larger community of general Web searchers and analyze trends over continuous periods, such as several days, weeks, and months.

DOI: 10.1002/asi

Several studies examine query categories in small, static logs. Spink and colleagues analyzed logs totaling more than one million queries submitted to the Excite Web search engine during single days in 1997, 1999, and 2001 (Spink, Jansen, et al., 2002; Spink et al., 2001; Wolfram, Spink, Jansen, & Saracevic, 2001). They classified approximately 2,500 queries from each log into 11 topical categories and found that although search topics have changed over the years, users' behaviors have not. Ross and Wolfram (2000) categorized the top 1,000 term pairs from the one million query Excite log into 30 subject areas to show commonalities of terms in categories (Ross & Wolfram, 2000). Jansen, Goodrum, and Spink (2000) used lists of terms to identify image, audio, and video queries and to measure their presence in the one-million-query Excite log. To examine the differences in queries from users in different countries, Spink, Ozmutlu, Ozmutlu, and Jansen (20020 examined a 500,000 query log from the FAST Web search engine during 2001, believed to be used largely by Europeans at that time, classifying 2,500 queries from it into the same topical categories. They found differences between FAST and Excite in the topics searched. These categories were also used in the comparative static analysis of the 1998 and 2002 AltaVista logs mentioned above and contained in (Jansen et al., 2005).

Other work manually grouped queries by task. Broder (2002) defines queries as informational, navigational, or transactional and presents a study of AltaVista users via a popup survey and manual categorization of 200 queries from a log. Beitzel, Jensen, Chowdhury, and Grossman (2003) implicitly categorized queries from a search log as navigational by matching them to edited titles in Web directories to automatically evaluate navigational Web search. Wolfram et al. (2001) automatically categorized query terms by using results from Web search engines to assign the terms to broad subject categories.

Several studies of query caching examine query frequency distributions from a static log, focusing on the average likelihood of an arbitrary query being repeated over the entire, fixed-length log. Lempel and Moran (2003) evaluated the performance of caching strategies over a log of seven million queries to AltaVista in 2001 and found that the frequencies of queries in their log followed a power law. Eiron and McCurley (2003) compared a query vocabulary from a log of nearly 1.3 million queries posed to a corporate intranet to the vocabulary of Web page anchor text and found that the frequency of queries and query terms follows a tail-heavy power law. Xie and O'Hallaron (2002) studied query logs from the Vivisimo meta-search engine of 110,881 queries over one month in 2001 in comparison to the Excite log of 1.9 million over one day in 1999. They found that although as in other studies over half of the queries are never repeated, the frequencies of queries that are repeated do follow a Zipfian distribution. Saraiva et al. (2001) evaluated a two-level caching scheme on a log of over 100,000 queries to a Brazilian search engine and found that query frequencies follow a Zipf-like distribution. Markatos (2000) simulated the effect of several types of query caches on an Excite query log of approximately one million queries and found that traditional caching methods provide significant improvements in efficiency. Although traditional most recently used (MRU) style caches obviously enhance throughput by exploiting temporal locality at the minute-to-minute level, these studies do not examine changes in the query stream according to the hour of the day that may be leveraged in an enhanced cache design.

It is well known that different users represent the same information need with different query terms, making query clustering attractive when examining groups of related queries. However, as Raghavan and Sever (1995) have shown, traditional similarity measures are unsuitable for finding query-to-query similarity. Wen, Nie, and Zhang (2002) incorporated click-through to cluster users' queries. In evaluating their system, they analyzed a random subset of 20,000 queries from a single month of their approximately 1-million queries-per-week traffic. They found that the most popular 22.5% queries represent only 400 clusters of queries using differing sets of query terms. A more recent study by Chien and Immorlica (2005) used temporal correlation to find sets of similar queries, suggesting that queries with similar frequency patterns are likely to be related (Chien & Immorlica, 2005). They defined a formal metric for temporal similarity between queries and used it to mine sets of related queries from a 6-month MSN™ search log. The presented results are largely anecdotal, but suggest a promising technique if noisy, unrelated queries can be adequately handled.

Many Web search services have begun to offer views of the most popular and/or changing (becoming drastically more or less popular) queries: AOL—Member Trends, Yahoo—Buzz Index, Lycos—The Lycos 50 with Aaron Schatz, Google—Zeitgeist, AltaVista—Top Queries, Ask Jeeves, and Fast (AllTheWeb). These views necessarily incorporate a temporal aspect, often showing popular queries for the current period and those that are consistently popular. Some also break down popularity by topical categories. Systems seeking to display changing queries must address the issue of relative versus absolute change in a query's frequency to find queries whose change is interesting, not simply a query that went from frequency one to two (a 200% jump), or one that went from 10,000 to 11,000 (an absolute change of 1,000).

Overall Query Traffic

We examined two query logs collectively consisting of billions of queries from the AOL search service. Our initial experiments focused on circadian changes in the query stream over hours in a day. These experiments were performed on a log of several hundred million queries representing an entire week of search from December 2003. We expanded on these experiments by also studying temporal changes over longer periods. This was done using the entire query log over a 6-month period from September 2004 through February 2005. These logs represent queries from tens of millions of users. We preprocess the queries in each

DOI: 10.1002/asi

TABLE 1. Aggregate query log statistics.

	One week: December, 2003	Six months: Sept. 2004–Feb. 2005
Number of users	Tens of millions	Tens of millions
Average query length	2.2 terms	2.7 terms
Average popular query length	1.7 terms	1.7 terms
% of users viewing first results page	81	79
% of users viewing second results page	18	15
% of users viewing three or more pages	1	6

log to normalize the query strings by removing any case differences, replacing any punctuation with white space (stripping advanced search operators from the approximately 2% of queries containing them), and compressing white space to single spaces. Some aggregate statistics for each log are shown in Table 1.

We began our analysis by first examining general properties of the query stream as a whole, such as volume progression, frequency distribution, and overlap as time progresses.

The most basic variable to examine is how the volume of query traffic changes over the course of a single day. This analysis was performed using the days in a week, examining changes in volume as we move from peak to nonpeak hours. We show the percentage of the day's total and distinct number of queries for each hour in the day on average over our 7-day period in Figure 1 (all times in our query log are Eastern Standard Time). Only 0.75% of the day's total queries appear from 5–6 a.m., whereas 6.7% of the day's queries appear from 9–10 p.m. Interestingly, the ratio of distinct to total queries in a given hour is nearly constant throughout the day. This shows that the average number of times a query is repeated is virtually constant over the hours in a day, remaining near 2.14 with only a 0.12 standard deviation.

Although the average repetition of queries remained nearly constant, we examined this in greater detail by measuring the frequency distribution of queries at various hours in the day, as seen in Figure 2. From this analysis it is clear that the vast majority of queries in an hour appear only one to five times and that these rare queries consistently account for large portions of the total query volume throughout the course of the day.

We continued this analysis by examining the volume curves and frequency distributions over longer periods of time. For these experiments, we used our larger log containing all queries over 6 months. The average volume for each day of the week averaged across all weeks from our 6-month log is shown in Figure 3, and the corresponding frequency distribution is shown in Figure 4.

From these graphs we can see that query traffic experiences a marked decline on Fridays, and peaks over the weekend. As before, both total and distinct query volume are plotted on the volume curve, and they track each other

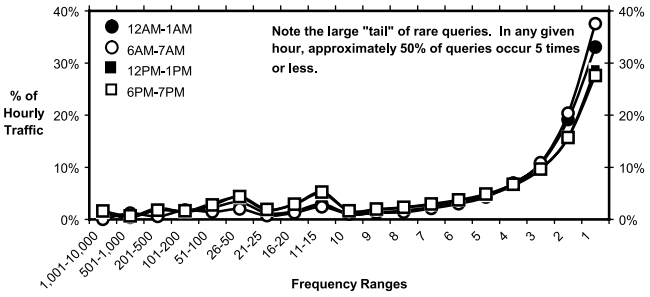

FIG. 2. Frequency distribution for selected hours.

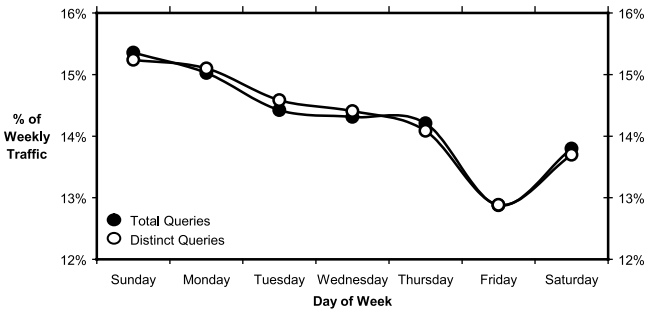

FIG. 3. Average volume of days in the week.

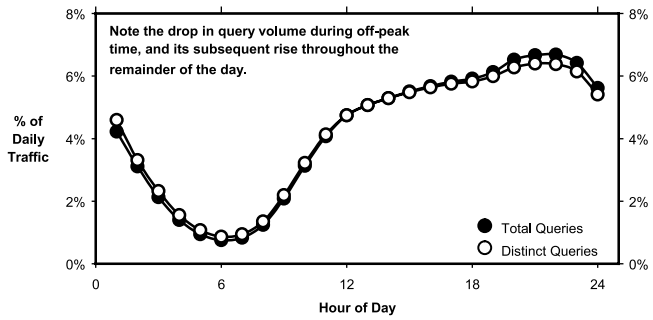

FIG. 1. Query volume over a day.

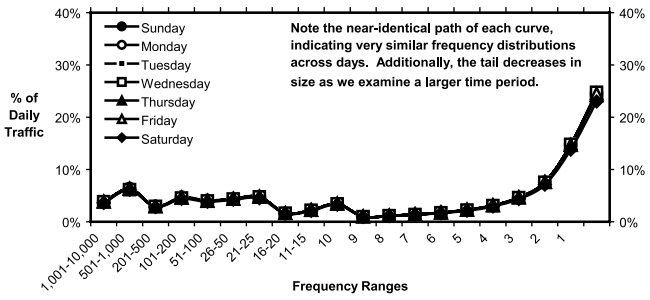

FIG. 4. Average frequency distributions for days in the week.

DOI: 10.1002/asi

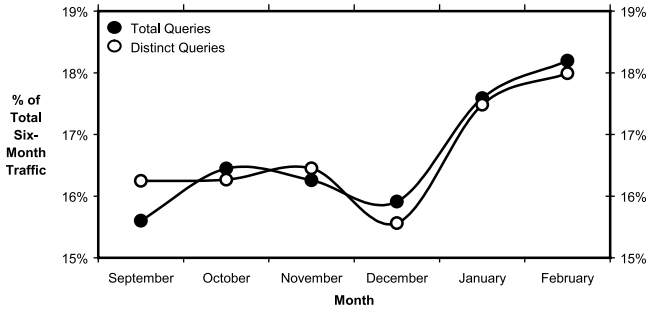

FIG. 5. Query volume by month.

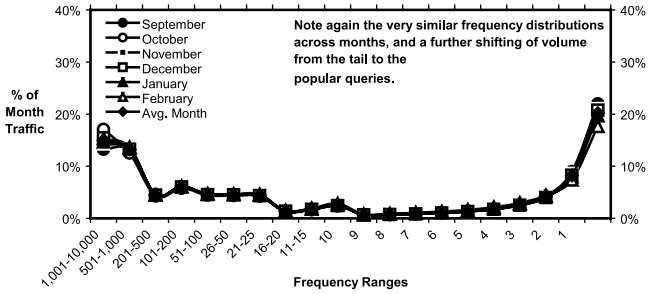

FIG. 6. Frequency distributions by month.

relatively closely. In examining the average frequency distributions for days in a week, we see that individual days in a week have very similar distributions. More importantly, we observe that the tail of the query stream (queries occurring five times or less) is less prevalent over longer periods, and that there is a larger percentage of very popular queries (occurring more than 10,000 times). This trend continues as we examine the volume and frequency distribution data for each month in our 6-month period. These graphs are shown in Figures 5 and 6.

Examining these graphs, we can see that query volume has spiked considerably since the beginning of 2005. This is most likely correlated with the release of an improved search product on the AOL service, generating increased user interest. Additionally, when examining the frequency distributions for each month, we can see that the trend of decreasing tail influence continues, and the increased prevalence of very popular queries.

Although we have shown that the query distribution does not change substantially over the course of a day, this does not provide insight into how the sets of queries vary from one hour to the next. To examine this, we measured the overlap between the sets of queries entered during those hours. We used traditional set and bag overlap measures as given in Equations 1 and 2, respectively. Distinct overlap measures the similarity between the sets of unique queries from each hour, whereas overall (bag) overlap measures the similarity of their frequency distributions by incorporating the number of times each query appears in an hour, $C(q_i; A)$. Although these measures examine the similarity of the sets of queries received in an hour and the number of times they are entered, they do not incorporate the relative popularity or ranking of queries within the query sets. To examine this, we also measured the Pearson correlation of the queries' frequencies. As can be seen from Equation 3 [where $\overline{C(q; A)}$ is the mean number of query repetitions in period A and $S_{C(q;A)}$ is the standard deviation of all the query frequencies in period A], we measured the degree of linear correlation between the frequencies of the queries in each hour. Therefore, 2 hours that had exactly the same queries with exactly the same frequencies would have a correlation of one. Note that this normalizes for the effect of differing query volume, i.e., the correlation of 2 hours with exactly the same underlying query distributions simply scaled by a constant would also have a correlation of one.

$$\textit{Distribution overlap}\ (A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Equation 1: Distinct overlap of query sets from periods A and B

$$\textit{overlap}\ (A, B) = \frac{\sum_{q_i \in A \cup B} \min(C(q_i; A), C(q_i; B))}{\sum_{q_i \in A} C(q_i; A) + \sum_{q_i \in B} C(q_i; B) - \sum_{q_i \in A \cup B} \min(C(q_i; A), C(q_i; B))}$$

Equation 2: Overall overlap of query sets from periods A and B

$$r_{A,B} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(C(q_i;A) - \overline{C(q;A)})(C(q_i;B) - \overline{C(q;B)})}{S_{C(q;A)} S_{C(q;B)}}$$

Equation 3: Pearson correlation of query frequencies from periods A and B

In Figure 7 we show the average level of overlap and correlation between the query sets received during the same hour for each day over our week. As measuring overlap over the set of all queries appearing in our week would be computationally expensive, we used the set of all the tens of millions of queries in the day after our 7-day period as an independent sample and measured overlap at each hour in our week of the queries matching those in that sample.

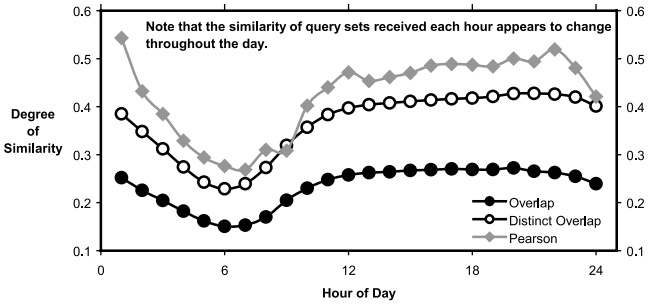

FIG. 7. Average overlap & Pearson correlations of matches from January 2, 2004, over hours in a day.

DOI: 10.1002/asi

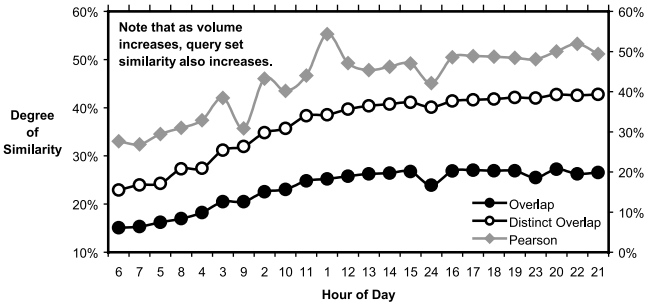

FIG. 8. Figure 7, sorted by increasing volume.

Although we previously saw that the frequency distribution of queries does not substantially change across hours of the day, Figure 7 shows that the similarity between the actual queries that are received during each hour does in fact change. This trend seems to follow query volume, which is apparent if we sort the same overlap data by query volume as is done in Figure 8. Clearly, as query volume increases the queries that compose that traffic are more likely to be similar across samples of those peak time periods.

This finding is consistent with prior analyses of Web query caches showing they significantly improve performance under heavy load. The more redundancy they are able to detect, the more caching algorithms are able to enhance throughput. Although the prior work primarily measures the effect of this redundancy in cache performance, it is obvious that redundancy must exist and be detected for caching to succeed. By examining the overall query stream by hour, we are able to infer the effectiveness of general caching algorithms at those times.

Query Categories

Above we performed a holistic analysis of the entire query log. However, this blanket view of the query traffic does not provide insight into the characteristics of particular categories of queries that might be exploited for enhanced efficiency or effectiveness. For example, a search provider who returns specialized results for entertainment queries cannot determine from general query traffic alone whether a given query is more likely to be referring to entertainment-related content or how to best process and cache that query.

The remainder of our analysis focused on trends relating to the topical category of queries. Our query set was categorized simply by exactly matching queries to one of the lists corresponding to each category. These lists were manually constructed by editors who categorized real users' queries, generated likely queries, and imported lists of phrases likely to be queries in a category (e.g., cities in the United States for the U.S. Sites category). Queries that matched at least one category list comprised 13% of the total query traffic on average, representing millions of queries per day. Although this may seem like an alarmingly small proportion, the categorized queries typically represented popular queries that are repeated often. The remainder of query traffic was comprised of a very large number of rare queries that occur

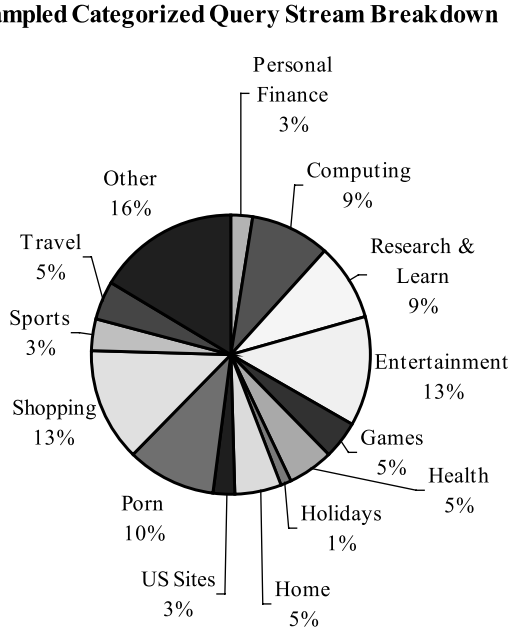

FIG. 9. Breakdown of categorized queries.

only a few times (as shown and discussed in the previous section). This lopsided density in the query stream mitigated the risk of using mostly popular queries to detect trends and changes in user behavior, because most users are, by definition, searching for what is popular. Nevertheless, developing a method of categorizing and analyzing the tail of the query stream is an important avenue for future work.

To verify that our defined category lists sufficiently covered the topics in the query stream, we manually classified a random sample of queries, assigning them to "Other" if they did not intuitively fit into an existing category, as can be seen in Figure 9. To determine the number of queries required to achieve a representative sample, we calculated the necessary sample size in queries (Kupper & Hafner, 1989):

$$SS = \frac{(Z^2\sigma^2)}{\beta^2}$$

Equation 4: Sample size formula

where z is the confidence level value, σ is the sample standard deviation, and β is the error rate. By setting our confidence level to 99% and error rate to 5%, we required a sample of 600 queries. The relative percentages for each category of the approximately 13% of query volume that matched any category list over our week (see Figure 13) were within the error rate of those from our manually categorized sample. This shows that our lists were a reasonable representation of these topical categories.

We focused on a subset of these categories and examined music and movies independent of other entertainment queries. The relative size of each category list we used is given in Figure 10. Obviously, not all queries listed actually match those entered by users, especially when the category contains large imported lists of phrases.

DOI: 10.1002/asi

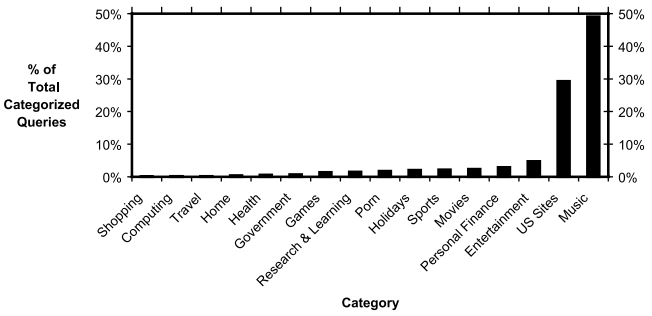

FIG. 10. Relative size of each category list used for matching.

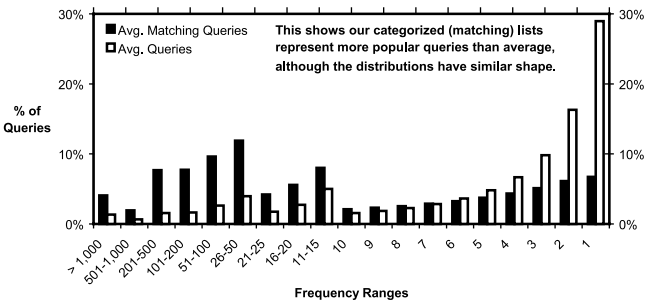

FIG. 11. Compared frequency distributions of queries matching our categorized lists vs. all queries.

Although we have shown that our lists are a fair representation of the topics in the query stream, this does not indicate what portion of the frequency distribution of that stream they represent. To determine this, we measured the average proportion of queries matching any category list that appear at various frequencies each hour and compared them to the average overall hourly frequency distribution of the query stream (see Figure 11). Unsurprisingly, this comparison shows that queries in the category lists represent more popular, repeated queries than average, although the general shape of the distributions is similar.

Trends in Category Popularity

We began our temporal analysis of topical categories by measuring their relative popularity over the hours in a day. First, we examined the percentage of total query volume matching a selected group of category lists, as can be seen in Figure 12. It is clear that different topical categories are more and less popular at different times of the day. Personal finance, for example, becomes more popular from 7–10 a.m., whereas music queries become less popular. Although it is difficult to compare the relative level of popularity shift from one category to another due to the differences in scale of each of their percentages of the query stream, it is clear that some categories' popularity changes more drastically throughout the day than others.

To quantify this, we calculated the KL-divergence (Equation 5) between the likelihood of receiving any query at a particular time and the likelihood of receiving a query in a particular category, as can be seen in Figure 13. This

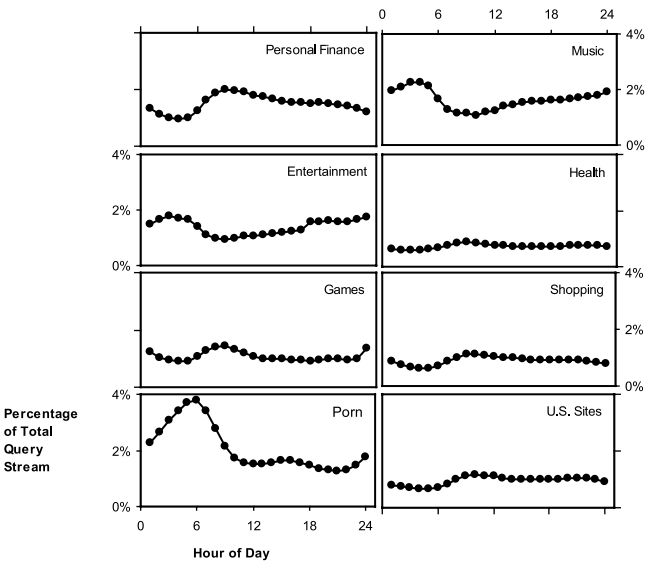

FIG. 12. Percentage of the total query stream covered by selected categories over hours in a day.

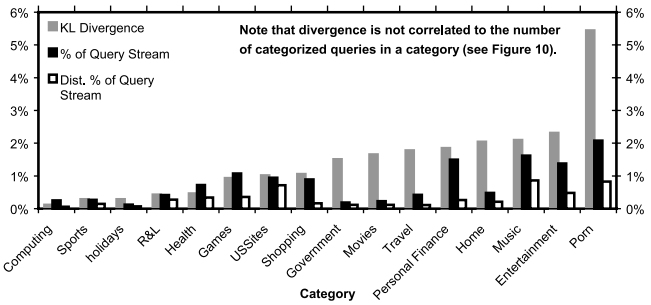

FIG. 13. Average percentage of query stream coverage & KL-divergence for each category over hours in a day.

revealed that the top three categories in terms of popularity are pornography, entertainment, and music.

$$D(p(q|t) \parallel p(q|c,t)) = \sum_q p(q|t)\log\frac{p(q|t)}{p(q|c,t)}$$

Equation 5: KL-Divergence of query occurrence likelihood for category c and total stream at time t

Comparing these divergences to the proportion of categorized queries in each category in Figure 10 quickly illustrates that divergence is not correlated with the number of queries categorized in each category. Also shown in Figure 13 is the average percentage of the entire query volume and distinct queries that match each category. Although the categories that cover the largest portions of the query stream also have the most relative popularity fluctuation, this correlation does not continue throughout all categories.

We also measured average category popularity changes over days throughout the week over our 6-month log. These popularity changes, measured as average percentages of the total query stream, are shown in Figure 14.

DOI: 10.1002/asi

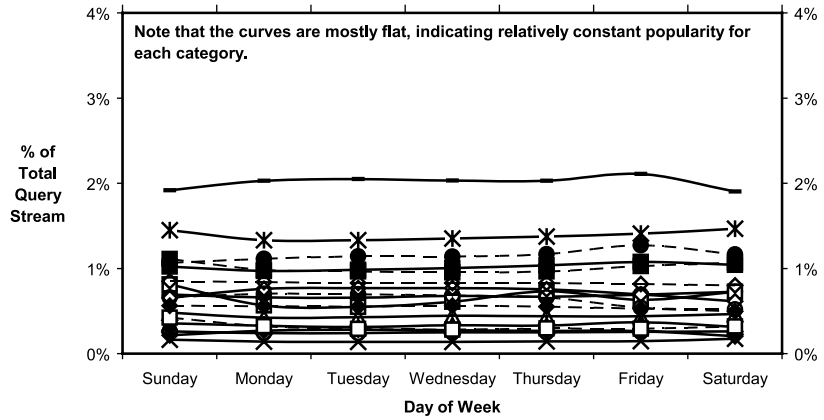

FIG. 14. Average percentage of the total query stream covered by each category over days in a week.

As can be seen in the graph, the behavior of individual categories appears to be less interesting over longer periods. In general, most of the categories exhibit flat curves, taking a relatively constant portion of the total query stream on average for each day in the week. To examine this further, we quantified the changes in popularity for each category using KL-divergence, as above. This is shown in Figure 15.

As with the hourly KL-divergence graph, this one is sorted in increasing order of divergence. One notable difference is that when examining divergences for the larger timescale, they are much smaller in magnitude. The most divergent category over a day (Porn) had a divergence of nearly 6%, whereas the most divergent category over a week (Games) had a divergence of less than 1.5%, which is consistent with the relatively flat popularity-change curves shown in Figure 14. This suggests that while shifts in categorical popularity do occur, they are most evident over shorter periods. In addition, the relative stability of most categories' popularity over a week seems to imply that the changes in popularity observed during a day are cyclical and tend to balance out over time. One caveat to this conclusion is that a week is not a long enough time to allow for any seasonal or other long-term changes in popularity. To examine the behavior across the long term, we also show the matching percentage of the query stream for each category over our entire 6-month period from September 2004 through February 2005 (Figure 16).

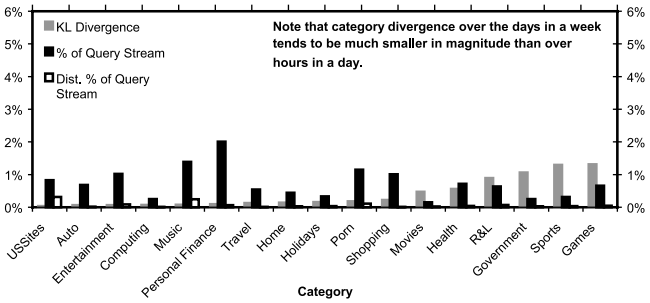

FIG. 15. Average percentage of query stream coverage & KL-divergence for each category across days in a week.

As with the weekly time periods, there are several flat categories illustrated in Figure 16, although some categories do appear to have variance. It is difficult to interpret this data when all the categories are plotted on a single axis, so we returned to KL-divergence to give us an idea of which categories, if any, exhibit interesting trends over a period as long as several months.

Using the KL-Divergence values shown in Figure 17, we can conclude that there are four categories with potentially interesting behavior: Holidays, Shopping, Sports, and

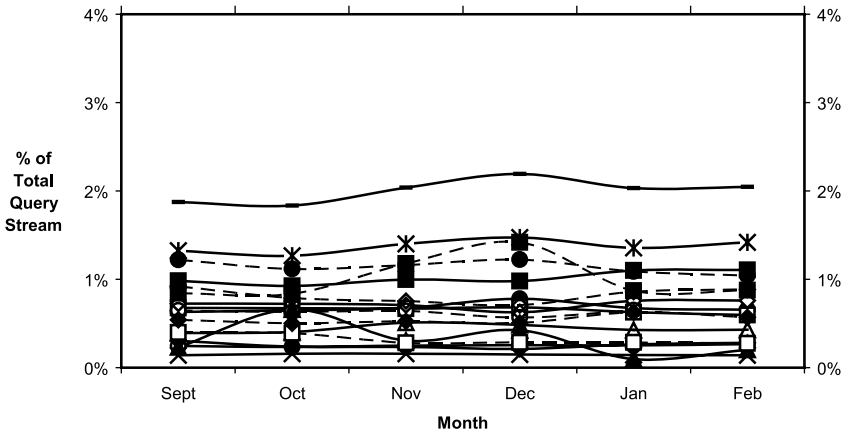

FIG. 16. Percentage of the total query stream covered by each category over 6 months.

DOI: 10.1002/asi

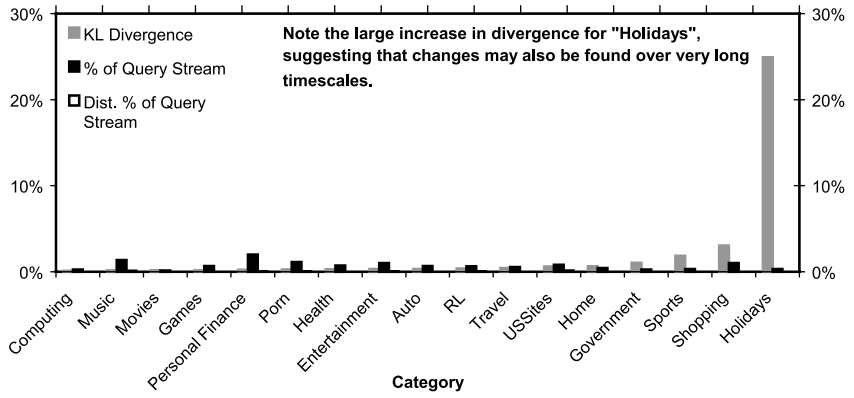

FIG. 17. Percentage of query stream coverage & KL-divergence for each category across 6 months.

Government. To get a clear picture of the change in popularity for these categories, we plotted their matching percentage of the query stream on separate axes.

As illustrated by Figures 17 and 18, when we move out toward examining very large time scales that encompass several seasons (climatic, holiday, sports-related, or otherwise), larger-scale categorical changes begin to reappear in the query stream. The most obvious example of this is the Holidays category, which exhibits the largest divergence by a wide margin. It is followed distantly, but surely, by Shopping, Sports, and Government categories. All of the topics represented by the queries in these categories are prone to seasonal changes, including, but not limited to the following:

- U.S. Federal Labor Day, Halloween, Thanksgiving, Christmas, Chanukah, New Year's Day, Valentine's Day, and other major holidays
- Climatic changes; late summer to autumn, autumn to winter
- The 2004 U.S. presidential and general elections (as well as several debates)
- The 2004–2005 NFL regular season, playoffs, and Super Bowl
- Major League Baseball playoffs and the 2004 World Series
- The start of the 2004–2005 National Basketball Association's (NBA) basketball season
- The proposed start date for the 2004–2005 National Hockey League's (NHL) season (later cancelled)

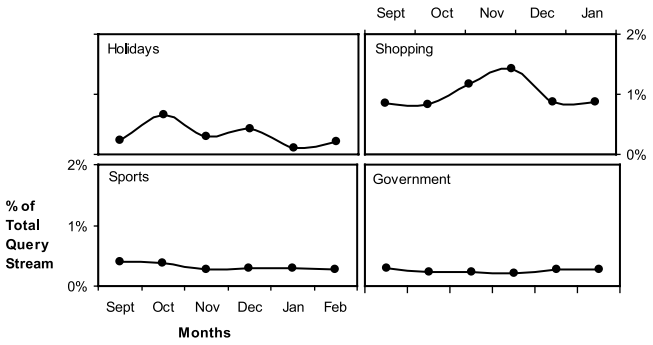

FIG. 18. Percentage of total query stream covered by most divergent categories over 6 months on separate axes.

These findings seem to suggest that there are both short-term and long-term topical trends in the query stream. Short-term trends can be immediately useful to providers of search services by way of intelligent results caching and temporally adaptive load balancing. Long-term trends such as seasonal changes can potentially be very useful for query routing and disambiguation.

Our next goal was to investigate the behavior of the most divergent categories in greater detail. We drilled down into the highly fluctuating categories and examined the behavior of the queries with the most highly fluctuating frequencies in each category. From this, we hoped to gain some insight into the reasons why certain categories fluctuate as well as the effect of terms and queries with very high flux in those categories. We began with our smallest time scale and examined the three most changing queries per day, on average, for the Entertainment category over our weeklong log from December 2003 (Table 2).

All three of these queries are specifically related to events in U.S. popular culture that occurred during or shortly before the time of our logged week. For example, the actress, Gwyneth Paltrow married in secret, and the news of her nuptials broke during the week we analyzed. Hilton Hotel heiress, Paris Hilton had been a popular topic recently, starring in a prime-time reality TV show entitled "The Simple Life." Also popular at the time was Orlando Bloom, the actor who portrayed a popular character in the "Lord of the Rings" movie trilogy. The final installment of the series was released in U.S. theatres shortly before our logged week, so it is no surprise to see his name as a top-changing query for that period.

Drilling down further, we pinpointed some of the specific instances where these popular queries jumped the most. For example, in the afternoon of Friday, December 27, the popularity of the query "gwyneth paltrow" skyrocketed. From 3–4 p.m., it occurred once, from 4–5 p.m. it occurred

TABLE 2. Top three fluctuating entertainment queries over hours in a day.

gwyneth paltrow
paris hilton
orlando bloom

DOI: 10.1002/asi

TABLE 3. Top 25 fluctuating music and entertainment queries over hours in a day.

Music	Entertainment
lyrics	gwyneth paltrow
music	paris hilton
britney spears	orlando bloom
furniture	espn
love	disney
hilary duff	johnny depp
good charlotte	much music
sloppy seconds	disney channel
jessica simpson	hgtv
b2k	disneychannel com
eminem	www disneychanel com
christina aguilera	katie holmes pictures
simple plan	pamela anderson
justin timberlake	cartoon network
free music	hilary duff
linkin park	fake
michael jackson	chad michael murray
beyonce	vivica a fox
jennifer lopez	disneychannel
50 cent	care bears
kinky	sailor moon
napster	www cartoonnetwork com
chic	days of our lives
tupac	charmed
blink 182	tom welling

67 times, and from 5 p.m.–6 p.m. it occurred 11,855 times. The top changing (on average) 25 queries from our week-long log, after normalization, in the Entertainment and Music categories are shown in Table 3.

We also looked at some of the most frequently changing terms to see how they relate to the change of entire queries containing those terms. Some examples of this behavior in the Entertainment category include the terms, *pictures* (the 10th-most changing term) and *duff* (the 17th-most changing term). We looked at the popularity change (i.e., change in frequency) for queries containing these terms and found that several of them also exhibited large changes over time. For example, on the afternoon of December 28 from 12–5 p.m., the query, *hilary duff* changed from an initial frequency of 27 from 12–1 p.m. to a peak of 131 (from 3–4 p.m.), and then stabilized around 70 for the rest of the evening. Similar spikes in frequency for this query occurred at similar times during other days in our period of study.

We also examined top-changing queries over longer time periods using our 6-month log. The KL-divergence and category coverage graphs shown in Figures 17 and 18 seemed to indicate that drastic seasonal changes took place over the course of our 6-month log, particularly in the Holidays category. The top 25 changing queries for the Sports, Holidays, and Government categories are shown in Table 4.

These top changing queries do present plausible justification for the divergence of their respective categories. There is an evident interest in several sporting events, major holidays, and issues of government business (politics, elections, tax information, etc.). Furthermore, the specific targets of interest within each category are highly diverse, suggesting sweeping shifts in popularity throughout the period. These results give more evidence that there are both short- and long-term topical trends in the query stream.

TABLE 4. Top 25 changing queries from September 2004 through February 2005.

Sports	Holidays	Government
baseball	halloween costumes	irs
superbowl	halloween	noaa
fantasy football	costumes	irs gov
super bowl	christmas	john kerry
nfl	mardi gras	internal revenue service
reggie white	christmas songs	election results
mlb	thanksgiving	Fema
bill o reilly	christmas cards	Voting
nfl com	santa	election 2004
red sox	pumpkin carving	george w bush
us open	halloween recipes	Elections
kobe bryant	halloween costume ideas	usps
ryder cup	sexy halloween costumes	norad
serena williams	halloween games	vote 2004
nba	santa claus	hawaii department of revenue
jackie robinson	valentine's day	louisiana dept of wildlife
boston red sox	chinese new year	new hampshire dept of motor vehicles
nba com	adult halloween costumes	larry klayman
world series	christmas crafts	richard ziser
daytona 500	haunted houses	nevada department of health
baseball scores	valentine cards	election returns
disco inferno	halloween crafts	claire mccaskill
yankees	thanksgiving recipes	rock the vote
hendrick motorsports	christmas poems	rhode island prisons
world series tickets	christmas games	ny dept of revenue

DOI: 10.1002/asi

Trends in Uniqueness Within Categories

Although we have shown that different categories have differing trends of popularity over various periods, this does not provide insight into how the sets of queries within those categories change over time. The anecdotal examples of top-changing queries give some measure of identity to what concepts are likely driving the most divergent categories, but there are no hard numbers attached to them. To examine these shifting trends in more detail, we returned to the overlap measures used in the Overall Query Traffic section. First, we examined the overlap characteristics for divergent categories over the hours in a day. Overlap, distinct overlap, and the Pearson correlation of query frequencies for Personal Finance and Music are shown in Figures 19 and 20.

Although the uniqueness of queries in categories in general appears to be correlated with that of the entire query stream (Figure 7), that of particular categories appears to be substantially different from one to the next. For example, if we compare the overlap characteristics of personal finance with those of music, we see they are quite different. Not only does personal finance have generally higher overlap, but it also has a much higher overall overlap than distinct overlap. This indicates that personal finance is likely to be a category that is dominated by a small number of very popular queries. In contrast, overall overlap and distinct overlap are nearly equal for music, suggesting that music is a more diverse category. In addition, the Pearson correlation of frequencies for personal finance queries is very high all day, indicating searchers are entering the same queries roughly the same relative number of times; this is clearly not the case for music queries, as the Pearson correlation for the music

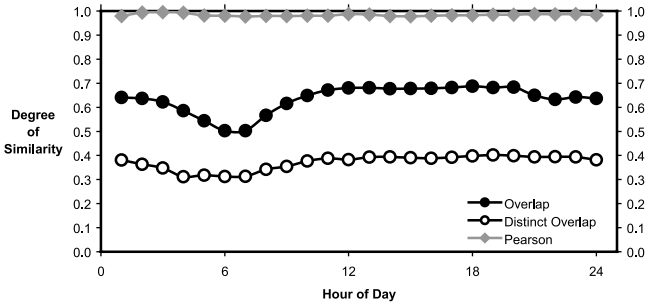

FIG. 19. Overlap and Pearson correlation for the personal finance category over hours in a day.

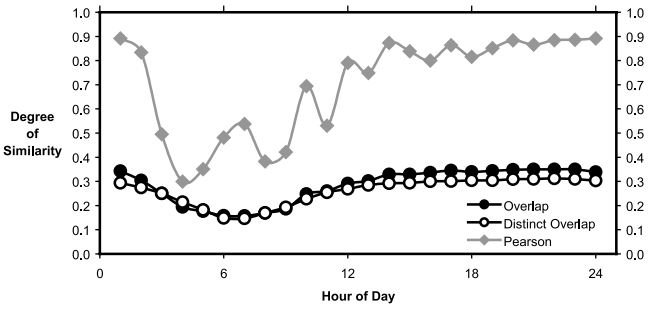

FIG. 20. Overlap and Pearson correlation for the Music category over hours in a day.

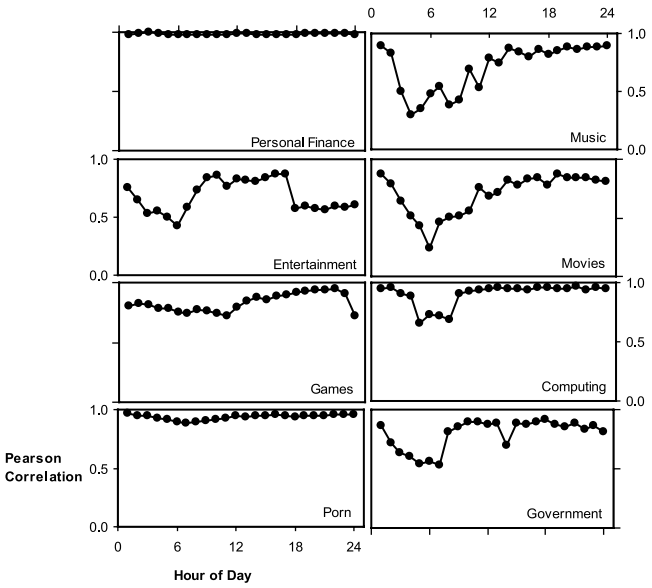

FIG. 21. Pearson correlations of matching query frequencies for selected categories averaged over hours in a day.

category varies widely throughout the day. To get a more complete sense of this behavior, we graph the Pearson correlations for several selected categories in Figure 21.

It is clear that some categories have very similarly ranked queries by frequency throughout the day (Personal Finance, Porn), suggesting a certain stability whereas others exhibit drastic variance in the frequencies of the queries received at each hour (Music, Movies, Entertainment). Referring back to Figures 10 and 13, uniqueness of queries in particular categories does not appear to be correlated with the number of queries in their respective category lists, the proportion of the query stream they represent, or the number of distinct queries they match.

We also measured the overlap and Pearson correlations for categories over time periods in our 6-month log, to get an idea of the behavior of categorical trends over longer periods of time. The Pearson correlations from day to day over the course of a week for all categories are shown on a single axis in Figure 22.

As we show in Figure 22, categorical trends tend to be more stable as we move out to examining the average behavior for each day in the week. Most categories exhibit much more stable behavior over the course of an entire week as opposed to a single day, with minor fluctuation in the Entertainment and Sports categories. This behavior is generally in keeping with the divergence for days in a week shown in Figure 15. We then moved on to examining the categorical trends in terms of pairwise Pearson correlations across our 6-month period. This is illustrated in Figure 23.

Again, we see that over longer periods, most categories become even more stable, with notable exceptions in Holidays, Government, Movies, and Sports. This also concurs with the per-category divergence results shown across our 6-month period in Figure 17, with the added information of exactly how the Pearson correlation between each pair of months as we move forward in time through the period. To get a closer look at the behavior of the four divergent categories, we graphed them on separate axes, as shown in Figure 24.

DOI: 10.1002/asi

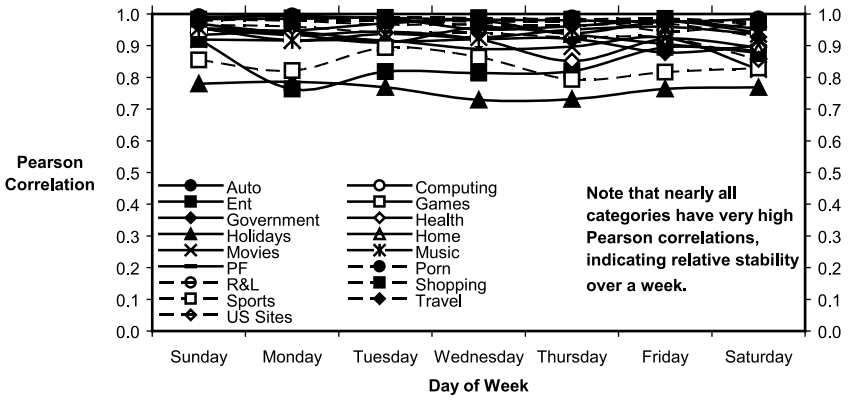

FIG. 22. Pearson correlations of matching query frequencies for each category averaged over days in a week.

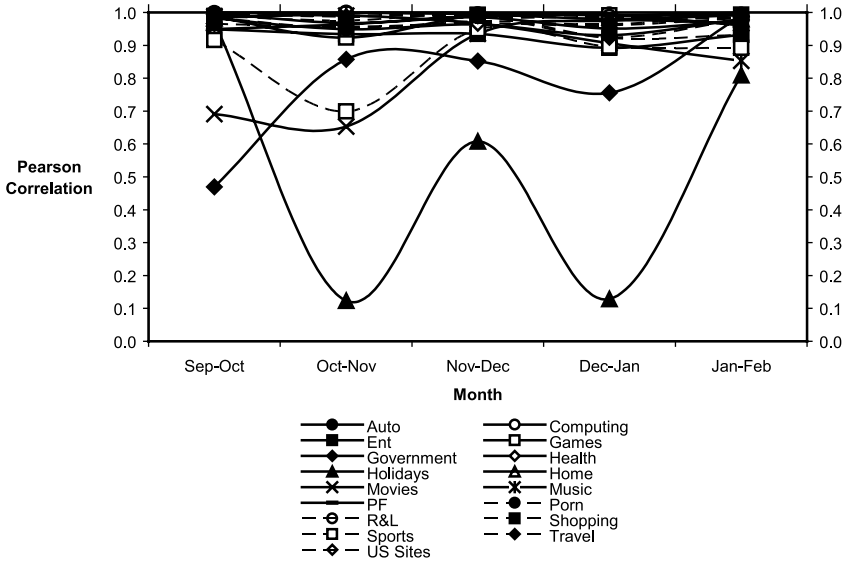

FIG. 23. Pearson correlations of matching query frequencies for each category over 6 months.

The Holidays category exhibits the most sweeping shifts in correlation, which is exactly in keeping with its disproportionately large KL-divergence across the 6-month period. In turn, the Government and Sports categories, which have highly variable Pearson correlations over the 6-month period, also are shown to fluctuate in popularity (Figure 16) and have relatively high KL-divergence (Figure 17), which strongly indicates that some of the possible trends discussed above are likely to be occurring.

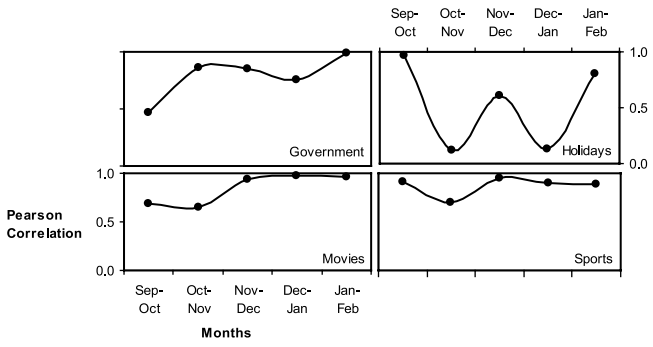

FIG. 24. Pearson correlations of divergent categories over 6 months on separate axes.

These findings imply that select topical categories experience multiple, overlapping trends over long periods of time, and that the behavior of a category in a single day can often be quite different from its behavior over a week, a month, or several months. This type of data is potentially of great use to query-caching algorithms. For example, if it is known a priori that queries for certain categories are similarly ranked throughout the day, they can be given higher priority in a query-caching scheme. Similarly, queries in categories whose rankings change vastly over time might be given low caching priority. Additionally, search services can fine-tune both their search results and their advertising strategies to maximize effectiveness when trends for certain seasonal periods can be anticipated (such as sports playoffs, holidays, etc.).

Conclusions and Future Work

This study focused on investigating the nature of changes in the query stream of a very large search service over time. Understanding how users' queries change over time is critical

DOI: 10.1002/asi

to developing effective, efficient search systems and to engineering representative test sets and evaluations that drive this development. We have proposed a method of temporal log analysis that can be used to study changing topical trends over time in addition to the traditional static analysis found in prior studies. For this, we used a set of topical categories created by human editors that represented approximately 13% of the total query traffic. Using this analysis, we have found trends over time that are stable despite continuing fluctuation in query volume. We have found that certain topical categories can exhibit both short-term (over hours in a day) and long-term (over several weeks or months) query trends, and that these trends and their behavior may vary wildly depending on the category and the length of time being studied. Many of the trends made observable by our analysis seem to be intuitively expected in hindsight, which suggests that our metrics and methods of analysis are viable for use in further research.

Future work in log analysis points in the direction of investigating the tail of the query stream, i.e., the rare queries (low in individual frequency but of collectively high volume) that are not often matched by our categorized topical lists. Tracking changes in the query stream tail would provide insight into whether rare queries are changing similarly to popular queries. One method for approaching this might be to incorporate automatic query classification methods to extend our basic lists.

Additionally, providers of search services can use the trends detected by our analysis to enhance both the efficiency and effectiveness of Web searching. Intelligent load-balancing and results-caching algorithms that take advantage of shifting query trends can clearly provide users with more timely and efficient access to query results, perhaps using machine learning on trend data to determine optimal cache-refresh times. Topical trend information can also be used to assist in query disambiguation and query routing (consider the query, *eagles*, as a prime example), allowing for queries to be identified and sent off to specialized back-end databases for highly relevant information on a topic. Such an approach has clear potential for increasing the effectiveness of a search service.

References

Beitzel, S.M., Jensen, E.C., Chowdhury, A., & Grossman, D. (2003). Using titles and category names from editor-driven taxonomies for automatic evaluation. In the Proceedings of the 12th ACM International Conference on Information and Knowledge Management (pp. 17–23). New York: ACM.

Beitzel, S.M., Jensen, E.C., Chowdhury, A., Grossman, D., & Frieder, O. (2004). Hourly analysis of a very large topically categorized web query log. In K. Järvelin, J. Allan, & P. Bruza (Eds.), the Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 321–328). New York: ACM.

Broder, A. (2002). A taxonomy of web search. SIGIR Forum, 36(2), 3–10.

Chien, S., & Immorlica, N. (2005, May). Semantic similarity between search engine queries using temporal correlation. Paper presented at the 14th International Conference on the World Wide Web (WWW), Chiba, Japan.

Eastman, C.M., & Jansen, B.J. (2003). Coverage, relevance, and ranking: The impact of query operators on web search engine results. ACM Transactions on Information Systems, 21(4), 383–411.

Eiron, N., & McCurley, K.S. (2003). Analysis of anchor text for web search. In J. Callan, D. Hawking, & A. Smeaton (Eds.), the Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 459–460). New York: ACM.

Hawking, D., Craswell, N., & Griffiths, K. (2001, May). Which search engine is best at finding online services? Paper presented at the 10th International Conference on the World Wide Web, Hong Kong, P. R. China.

Jansen, B.J., Goodrum, A., & Spink, A. (2000). Searching for multimedia: Video, audio, and image web queries. World Wide Web, 3(4), 249–254.

Jansen, B.J., & Pooch, U. (2001). A review of web searching studies and a framework for future research. Journal of the American Society for Information Science and Technology, 52(3), 235–246.

Jansen, B.J., Spink, A., & Pederson, J. (2005). A temporal comparison of AltaVista web searching. Journal of the American Society for Information Science and Technology, 56(6), 559–570.

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.

Kupper, L.L., & Hafner, K.B. (1989). How appropriate are popular sample size formulas? The American Statistician, 43, 101–105.

Lawrence, S., & Giles, C.L. (1998, April 3). Searching the world wide web. Science, 280, 98–100.

Lempel, R., & Moran, S. (2003, May). Predictive caching and prefetching of query results in search engines. Paper presented at the 12th International World Wide Web Conference, Budapest, Hungary.

Markatos, E.P. (2000, May). On caching search engine query results. Paper presented at the 5th International Web Caching and Content Delivery Workshop, Lisbon, Portugal.

Raghavan, V.V., & Sever, H. (1995). On the reuse of past optimal queries. In E.A. Fox, P. Ingwersen, & R. Fidel (Eds.), the Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 344–350). New York: ACM.

Ross, N.C.M., & Wolfram, D. (2000). End user searching on the internet: An analysis of term pair topics submitted to the excite search engine. Journal of the American Society for Information Science, 51(10), 949–958.

Saraiva, P.C., de Moura, E.S., Ziviani, N., Meira, W., Fonseca, R., & Riberio-Neto, B. (2001). Rank-preserving two-level caching for scalable search engines. In the Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 51–58). New York: ACM.

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.

Spink, A., & Jansen, B.J. (2004). Web search: Public searching of the web (1st ed.). New York: Springer.

Spink, A., Jansen, B.J., & Ozmutlu, H.C. (2000). Use of query reformulation and relevance feedback by excite users. Internet Research: Electronic Networking Applications and Policy, 10(4), 317–328.

Spink, A., Jansen, B.J., Wolfram, D., & Saracevic, T. (2002). From e-sex to e-commerce: Web search changes. IEEE Computer, 35(3), 107–109.

Spink, A., Ozmutlu, S., Ozmutlu, H.C., & Jansen, B.J. (2002). U.S. versus European web searching trends. SIGIR Forum, 36(2), 32–38.

Spink, A., Wolfram, D., Jansen, B.J., & Saracevic, T. (2001). Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology, 52(3), 226–234.

Sullivan, D. (2003). Searches per day. Retrieved February, 2003, from http://searchenginewatch.com/reports/article.php/2156461

Wang, P., Berry, M.W., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. Journal of the American Society for Information Science and Technology, 54(8), 743–758.

Wen, J.-R., Nie, J.-Y., & Zhang, H.-J. (2002). Query clustering using user logs. ACM Transactions on Information Systems, 20(1), 59–81.

Wolfram, D., Spink, A., Jansen, B.J., & Saracevic, T. (2001). Vox populi: The public searching of the web. Journal of the American Society for Information Science, 52(12), 1073–1074.

Xie, Y., & O'Hallaron, D. (2002). Locality in search engine queries and its implications for caching. In the Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM) (Vol. 3, pp. 1238–1247). Piscataway, NJ: IEEE.

DOI: 10.1002/asi