Learning about the World through Long-Term Query Logs

MATTHEW RICHARDSON
Microsoft Research

In this article, we demonstrate the value of long-term query logs. Most work on query logs to date considers only short-term (within-session) query information. In contrast, we show that long-term query logs can be used to learn about the world we live in. There are many applications of this that lead not only to improving the search engine for its users, but also potentially to advances in other disciplines such as medicine, sociology, economics, and more. In this article, we will show how long-term query logs can be used for these purposes, and that their potential is severely reduced if the logs are limited to short time horizons. We show that query effects are long-lasting, provide valuable information, and might be used to automatically make medical discoveries, build concept hierarchies, and generally learn about the sociological behavior of users. We believe these applications are only the beginning of what can be done with the information contained in long-term query logs, and see this work as a step toward unlocking their potential.

Categories and Subject Descriptors: H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; J.4 [**Social and Behavioral Sciences**]; J.3 [**Life and Medical Sciences**]

General Terms: Algorithms, Measurement, Experimentation, Human Factors

Additional Key Words and Phrases: query logs, knowledge discovery, user behavior, data mining

ACM Reference Format:

Richardson, M. 2008. Learning about the world through long-term query logs. ACM Trans. Web 2, 4, Article 21 (October 2008), 27 pages. DOI = 10.1145/1409220.1409224 http://doi.acm.org/10.1145/1409220.1409224

1. INTRODUCTION

There are many fields of science devoted to furthering our understanding of humankind, such as anthropology, sociology, psychology, medicine, economics, and political science. In all of these fields, scientific inquiry generally proceeds by making observations about the real world, particularly the complex behavior of the people that live within it. These observations lead to new discoveries about our world and greater understanding about ourselves.

Authors's address: One Microsoft Way, Redmond, WA 98052; email: mattri@microsoft.com

With the advent of Web search engines, a new source of data about people and the world has become available. Every time a person queries a search engine, he provides a small window into his life, his interests, and the world around him. Taken as a whole, across millions of users, these queries constitute a measurement of the world and humanity through time. Viewed another way, it's as if a survey were sent to millions of people asking them to, every day, write down what they were interested in, thinking about, planning, and doing, and mail it back after a year. The purpose of this article is to demonstrate the value of this data when considered as a whole, for computer, medical, and social research.

In many ways, these query logs contain information that would never be available to researchers using conventional data-collection techniques. For example, a medical researcher might discover that people with asthma tend to wear wool, or live in areas with coal power plants; a sociologist could study how ideas spread from one person to an entire community; a political scientist might learn about democracy by studying the evolution of political searches by users in a developing country. The rarer the population to be studied is, the more difficult it is to locate and interview the subjects. The quantity of data in long-term query logs means that a rare situation may still match thousands of users.

Query logs also enable researchers to ask questions that would normally require going backward in time. For example, a medical researcher might study people diagnosed with diabetes today to find out what their primary symptoms were six months ago. Asking them directly, once they learn they have diabetes, may result in subjective bias, and asking thousands of people about their symptoms and waiting to see if any of them develop the condition, would be quite expensive. With long-term query logs, a doctor can look back into the history of people who have developed diabetes to see what their issues were prior to that time. This, in turn, could potentially be used to assist in the development of new, more effective drugs to treat diabetes.

A better understanding of people is not only beneficial for social scientists and medical researchers, but also for computers. For example, by better understanding people's interests, a search engine can do a better job of helping a user find what he is looking for, provide more useful query suggestions, and be smarter about correcting spelling mistakes. There have been many studies demonstrating benefits when applications are provided with information about what items are related to each other (e.g., applications using WordNet [Fellbaum 1998]). Automatically inducing these relationships, even if they are already generally known by people, is an effective technique for providing this knowledge.

Most of the possibilities mentioned above require query logs recording user behavior over a long period of time. To date, however, most research on query logs has focused on either: (1) query sessions: the relations between queries in a short time-frame (usually under 30 minutes), or (2) query frequency: the popularity of a query across all users, over time. Because they either aggregate across all users, or look only at short time slices, they lose crucial information about user interests, history, and so forth. (one notable exception is the work on requerying; see Section 9 for more detail)

In this article, we demonstrate the value of long-term query logs. We study a 12 month corpus of query logs containing millions of users and billions of queries, and make the following contributions:

—*Query effects are long lasting.* We show that knowing that a user has issued a query distinguishes them from the general population of users for days and even weeks after issuing that query. As a corollary, information is lost if only within-session query behavior is considered.

—*The long-lasting effects contain useful and valuable information.* We demonstrate a number of techniques that provide useful information, but lose their value if they consider only within-session data. For example, we show that we can build topic hierarchies, and study the temporal evolution of querying behavior.

—*The information may be used for scientific and research purposes.* We show the potential to discover, for example, the relationship between a medical condition and various potential causes of it.

We begin with a simple motivating example. In Sections 4 and 5, we present our data and the model used for the experiments. Experimentally, we show that query logs can be used to measure term relationships (Section 6.1), and measure how long a query's effects last in distinguishing a user from the general population (Section 6.2). After demonstrating that long-term effects are important, we study how the query logs can be used to observe how a person's interests vary over time (Section 6.3) and deduce topical hierarchies (Section 6.4). In Section 7, we present a novel technique for learning about user behavior by studying how their query distribution varies temporally, relative to a reference query. We then conclude with a discussion about the results, and related and future work.

2. MOTIVATING EXAMPLE

We begin with a simple motivating example. Consider a scientist studying how people's views on traffic and pollution depend on factors such as whether they prefer to live in a house or in a condominium. Typically, this would be done by surveying people in both populations (house vs. condo) and comparing the responses. This is an expensive proposition for many reasons: (1) without knowing ahead of time how the populations are likely to differ, it is difficult to determine what questions should be asked, (2) without knowing how different the responses are likely to be, the scientist must guess how many surveys to conduct in order to accomplish statistically significant results, and (3) manually interviewing people, or mailing out, collecting, and interpreting surveys costs time and reduces the rate at which the scientist can pursue new hypotheses.

An answer to all three of these issues is found in long term query logs. With simple statistics on the collection of logs, we can immediately derive answers to these questions. In our query log sample, we find there are 4.5 million users who have used the word "house" in a query, and 287 thousand who have used the word "condo." By assuming that users query for what they are interested in, and also what describes their lives, we can estimate that there are more homeowners

in the "house" querying population, and similarly for condominium owners.[1] In the housing population, there are 63,000 users who searched for information about hybrid vehicles (query: "hybrid"). Given this, we would expect to find 4,000 condominium owners with the same query. Instead, we find 7,100. This implies that condominium owners (or, at least, people interested in condominiums) are 77% more likely to search for information on hybrid vehicles. If we do the same study for the query term "traffic," we find that the two populations have roughly equal interest in traffic. More generally, we can automatically examine thousands or millions of query terms to find those in which the populations differ. This is a task that is impossible using traditional techniques.

Naturally, this type of study cannot replace a carefully controlled experiment. However, it can indicate what factors should be studied in more depth. Particularly when an issue is not well understood, knowing the top 10 or 100 most likely differences between two populations could be tremendously beneficial in guiding research, from designing experiments (e.g., surveys), to determining how the populations should be sampled, to assessing how many respondents are needed in order to likely find statistical significance.

Learning about people is not only useful for scientific endeavors, but for the search engine user as well. By better understanding what a user might be interested in, given their past searches, a search engine can provide many benefits to the user. For example, a search engine that understands that people interested in condominiums are more interested in hybrid vehicles, can do a better job of ranking Web pages when such a user searches for "car" (this is an extension of *personalized search*), or when suggesting additional queries the user might be interested in. Further, with such information, social networking and community sites would be more able to predict a user's interests, leading to improvements such as suggesting a new discussion group or mailing list that the user might be interested in. In all, being able to learn from long-term query logs can prove beneficial to users, and the scientific community at large.

It is important to point out that we are not looking at term cooccurrences just within a search session, but rather across entire query histories. The next section discusses this distinction briefly, before we introduce the data, model, and experimental results.

3. THE ADVANTAGE OF LONG-TERM LOGS

Generally, terms that cooccur within a search session will be terms that are directly related, as the user searches with a particular purpose. Instead, we are asking the question: for users who are interested in condominiums, what else do they find interesting at any time throughout the entire year? Without knowing it, there may be certain activities and interests that users tend to be cointerested in, but that they would not tend to search for nearby in time.

Thus the advantage of using long term query logs is *to* discover correlations that users may not even be aware of. By selecting the pool of users who have

[1]Many queries containing "house" are searches for new housing, and similarly for "condo." A more in-depth study would compare statistics on other terms such as "lawn," "condominium association," and so on, to verify the findings.

expressed an interest in q (by querying for it), and seeing what else they commonly search for, we are looking for terms that are related to each other, not because they have the same meaning, but because they would be of interest to users who care about q. Put another way, in the short term, we would find things that are relevant to the query. In the long term, we would find things that are relevant to the user.

The difference, while subtle, is key. Terms cooccurring in a query session are terms that users already know are related to each other. For example, within a query session, a user may search for "lexus" and then "bmw." This would let us find terms that are directly related to each other, but not terms that are only related because people who are interested in one are interested in the other. For instance, perhaps BMW owners tend to prefer classical music. By looking at long-term correlations, this preference could be detected, even if the users' individual search sessions were always focused only on cars or on music. Note that, the less well-known a relation between two terms is, the more likely it is that the relation will be found only by looking at long-term correlations**.**

In the next sections, we present the data and model used throughout the rest of the article. This is followed by a series of experiments demonstrating the wide variety of results that can be obtained from long-term query logs.

4. DATA AND PRIVACY

We collected a sample of the query logs from the Microsoft Windows Live search engine for one year, beginning in June, 2006. The elements of the logs used in the research were: a user ID, the day and time of the query, and the query itself. Privacy is an important consideration whenever we work with search data. As such, it is important to note that in accordance with Microsoft's Privacy Principles,[2] the user ID in the logs cannot be used to directly or personally identify the user who performed the query.

The Live search services are designed to store search terms separately from account information that personally and directly identifies the user, such as name, email address, or phone numbers. Microsoft has implemented protections to prevent unauthorized correlation of this data, including the use of one-way cryptography to keep the search data deidentified. The user ID comes from a cookie that is set on the user's machine when they log in to a LiveID-enabled service, such as Windows Live Hotmail (this cookie remains in existence after they have logged off, so users do not need to be logged in at the time they issue their query). The cookie is a one-way cryptographic function of their LiveID, so if they delete their cookies or use more than one computer, they will be given the same identifier once they have logged in to a LiveID-enabled service. The advantage of using a one-way cryptographic hash function is that, although this allows us to group the queries by user, it is virtually impossible to link a query back to the LiveID with access to only the query logs. We did not access the LiveID database or the hashing algorithm to do this research. Only queries

[2]http://download.microsoft.com/download/3/7/f/37f14671-ddee-499b-a794-077b3673f186/Microsoft%E2%80%99s%20Privacy%20Principles%20for%20Live%20Search%20and%20Online%20Ad%20Targeting.pdf

from users who have a LiveID account (there are over 300 million people with LiveID accounts) were used in this study.

Our corpus contains only a fraction of the users in the entire query log, and is limited to users who queried the English language search engine. Even so, it contains billions of queries, submitted by millions of users. The corpus is stored on a cluster of servers in a secured environment. Further, experiments were conducted by running programs that compute aggregate statistics over the query log (e.g., the correlation between two terms) and present only these aggregate statistics to the authors. Consequently, the authors were not studying or examining individual search histories.

5. MODEL

In this section, we introduce the model used throughout the rest of the article. Let $U = \{u_1, u_2, \ldots\}$ be a set of users. Let $\boldsymbol{D} = \{\langle u, q, t\rangle\}$ be the set of user-query-timestamp tuples in the data set. Further, let *int(u, q)* mean user u is *interested in* the concept indicated by query q. By interested in, we simply mean the user has something to do with the query, whether it means they own it ("tuba"), do it ("dance"), bought it ("vase"), like it ("monet"), and so on. In this article, we will define *int(u,q)* to be true when user u has queried for q at least once (e.g., $\exists\, t \langle u, q, t\rangle \in \boldsymbol{D}$), but the model is general for any function over $\boldsymbol{D}$. For example, we may require a user to have queried for a concept at least k times, on at least d unique days.

Let $n(q)$ be the number of users interested in q, and $n(q, r)$ be the number of users interested in both q and r:

$$n(q) = \sum_{u \in U} 1_{int(u,q)}$$
$$n(q, r) = \sum_{u \in U} 1_{int(u,q) \wedge int(u,r)}.$$

Finally, let N be the total number of users: $N = |U|$. The probability that a random user is interested in r is given by $p_r = n(r)/N$.

If interest in q and r are independent, we expect $n(q, r) = n(q)p_r$. That is, taking the set of users interested in q, we would find users as being interested in r randomly, with probability p_r. If the two are completely dependent, we would find $n(q, r) = n(q)$. In the experiments below, we will always choose a particular reference query, r, and search for the queries q that are most dependent on r. Since we are simply providing ordered lists, and r is constant, our measure of dependence is simply $dep(q) = n(q, r)/n(q)$. The value ranges from p_r (when independent), to 1 (when dependent).[3]

As expected, smoothing is also an important consideration. Imagine that one of the users who queried for r also happened to search for a random string of characters, q_{unique}. Since $n(q_{unique}, r) = 1$ and $n(q) = 1$, we would get a score of 1. There are many options available for smoothing. For example, we could simply

[3]Note that this measure is the same as the PMI between the user interests in q and r, scaled by a constant (and exponentiated).

require a minimum number of joint counts, compute statistical significance using a t-test [Church et al. 1991], or smooth the counts using Dirichlet or marginal priors. We chose to smooth the counts, and thus our relevance score for query q, with respect to reference query r, is:

$$dep_r(q) = \frac{n(q,r) + mp_r}{n(q) + m}.$$

Intuitively, this behaves as if there are m unobserved users, each of which is interested in q. *A priori*, we believe q and r are independent, which implies there are mp_r unobserved users who are interested in both q and r. Note that as m increases, it requires more and more evidence of dependence before it is believed. In the experiments below, we use $m = 100000$ (in Section 6.4 we show how varying this affects the specificity vs. generality of the terms retrieved).

In some experiments, we additionally constrain the amount of time that passes between query q and query r to be in the range [*minTime*, *maxTime*]. For instance, we may look at queries q that occur between 1 and 7 days before or after r. In this case, we are modifying the definition of $n(q,r)$ to mean the number of users who are interested in r who have also shown interest in q at least *minTime* and at most *maxTime* away from the nearest instance of querying for r.

For most of the results in this article, we operate on a term-level. That is, we split queries into individual terms. This does not affect the results significantly, but eases the exposition of them for the purposes of the article. We give full-query results in Section 6.3.

6. EXPERIMENTS

In this section, we provide experimental results showing that query effects are long lasting, that the long-term logs provide information that would be lost when considering only within-session data, and that long term query logs may be used to learn about the world.

6.1 Finding Relations

We first give some basic results, which we then explore further in later sections. One of the motivations of this paper is to show that the logs could be used to assist in medical studies. To this end, we first consider migraines, which were the subject of one of the earliest success stories of mining medical literature for automatic medical discovery [Swanson 1988]. Table I gives $dep_{migraine}(q)$ for a variety of terms. It is well accepted in the medical field that caffeine use is related to migraines. As can be seen, coffee terms score quite high, as compared with unrelated terms such as "dog" and "free." We can find this relation because there are a number of users who, for example, have been looking for information about coffee makers and who, at some other point, also searched for information about migraines. The user may be completely unaware that there is a relationship between these two search activities, but by aggregating over many users, we can discover that there is indeed a relationship here.

Table I. Dependency Score for Selected Terms vs. "Migraine"

Term	$dep_{migraine}(q)$ ($\times 10^{-3}$)
coffee	7.4
tea	8.2
coffee maker	10.1
caffeine	22.3
magnesium	24.7
dog	5.5
free	2.3

We also include the scores for "caffeine" and "magnesium" (magnesium was hypothesized to be related to migraines in Swanson's [1988] work on mining medical literature). The high scores for these terms may be due to migraine sufferers searching online for information, learning that migraines may be caused by caffeine or low magnesium levels, and then searching for more information about these factors. Whether this is considered a shortcoming depends on the application. There are many applications that can benefit from incorporating known relationships. For example, search engines can use the information to find relevant web pages that do not contain the query terms but do contain many relevant terms. Similarly, a news aggregator could benefit from this information to determine which news articles are likely to be interesting to someone it only knows a littlebit about. For such applications, an automatic method for inducing a relationship that says *users who are interested in q are also likely interested in r* can prove useful.

If, on the other hand, the goal is truly novel discovery, such as finding a medical relationship that is unknown to the medical establishment, then additional filtering will need to be done. In Figure 1 we have plotted the daily frequency of each query (total number of times the given term was queried by all users who searched for migraine information) vs. how many days it has been since the first migraine query. Magnesium and caffeine exhibit a much sharper decline in query frequency over time than do tea or coffee. This provides promise that such a feature will do well to discriminate between co-occurrences known to the user, and those that are not. We expect that other features, such as whether the terms frequently cooccur on Web pages (indicating that they are a known relationship) could also help to separate the truly novel discoveries from known relationships.

As another example, we also analyzed sports vs. alcohol to see which sports fans prefer which type of alcohol: wine or beer. For all six sports—football, baseball, basketball, skiing, cricket, and soccer—both beer and wine scored high (ranging from 50-90 in the same scale as Table 1), and beer was always more related than wine. For beer, baseball was the highest-scoring sport, and for wine, skiing was the highest.

6.2 How Long does a Query's Effect Last?

We know that when users are searching for a given interest, their distribution of terms within a session is significantly different from the general population

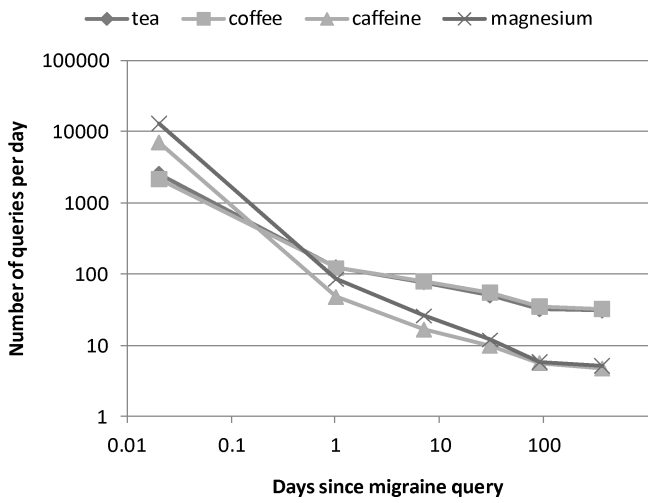

Fig. 1. Though tea, coffee, caffeine, and magnesium are all related to "migraine", they have different behaviors. The terms "tea" and "coffee" still remain popular queries months after the initial migraine query, whereas caffeine and magnesium fall off more rapidly (note the logarithmic scale – the decline in magnesium is 40 times more rapid than coffee). This may indicate that the relationship between caffeine/magnesium and migraines found in the query logs is due to users finding this information on a web page, while the tea (coffee) relationship is due to the fact that the migraine sufferers also happen to drink tea (coffee).

distribution. One natural question is, how long does this different distribution last. In particular, is the distribution of searches different only within the same session? Does it last a day? Does it last months?

In Figure 2, we graph the KL-divergence [Kullback and Leibler 1951] between the query distribution among users who searched for a term, vs. all users. The figure shows the change in divergence over time. For example, we see that if we take users who searched for "mortgage," the queries they searched for 100 days after this query is 0.2 bits different from the overall population.

There are a few interesting conclusions we can derive from this graph. First, we see that queries that identify a more specific population of users ("carabiner") show a higher difference from the general population than generic queries ("restaurant"). Second, we see that, though the divergence is greatest near the time of the query, the decrease is gradual, lasting many days or up to a month before asymptoting. This indicates that users tend to stick to a particular interest in their querying behavior for many days in a row. Finally, and probably most importantly, the divergence asymptotes at a non-zero value.[4] This shows that the behavior of a user even months after submitting a query is somehow related to the query. In this article, we are hypothesizing that this relationship is due to the fact that they've expressed a particular interest, which makes them different from the population in general, and thus allows us to learn from them.

[4]We believe the slight upswing in divergence for > 6 months is due to edge effects from the data spanning only 12 months, though it is possible that there is also a seasonal effect (the endpoints being summer months)

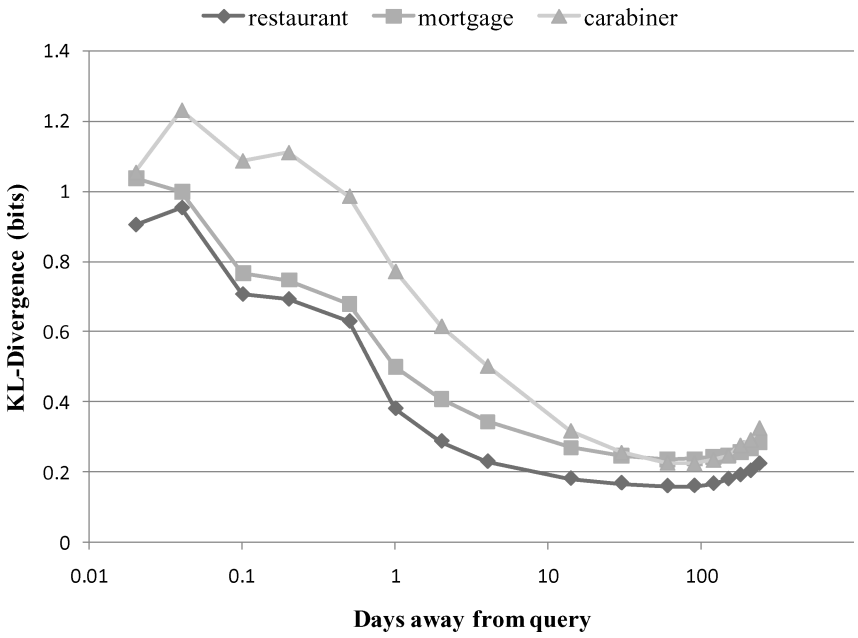

Fig. 2. KL-Divergence between users who issued a given query and the general user population. The divergence decreases over many days, and remains nonzero even months after the query.

These results have implications for future research in query log analysis and mining. For instance, they imply that predicting what query a user is likely to issue next can gain information by looking at what he queried for even months ago. Given these results, we were interested to see whether the difference in distribution was due to the same set of terms occurring, or whether the user's interests were also changing over time. The next section explores this question.

6.3 Change in Interests over Time

One of the biggest events to happen in a person's life is buying a house. Not surprisingly, many are turning to the Web for information on mortgages. Among users who searched for mortgage information, what else did they look for within a day, a week, a month, or more?

In Table II we present the results of this experiment. In each column, we give the top 40 novel queries by $dep_{mortgage}(q)$ for a particular time period. As discussed in Section 5, a time period of 7–30 days means queries that occurred at least seven days and no more than 30 days away from the nearest mortgage query (in time). Generally, the top 10–20 queries are the same across time periods (lending, realtor, loan, . . .), so we show only the queries that were not in the top 40 of the previous column. There are a couple of interesting observations we can make from this experiment.

One interesting result is the smooth transition of user interests as we move farther away from the mortgage query itself. Within the same session as the mortgage query, we find queries such as "calculator," "lender," and so on. Looking

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler
servicing	merchants	wells	gas	steakhouse
mortagage	fsbo	wholesale	macy	boat
refinancing	purchase	corporation	dental	labor
option	brokers	calculator	jewelers	water
greenpoint	maryland	arizona	wireless	repair
motgage	subdivision	chase	center	blue
equity	option	florida	gifts	northwest
calculater	deed	court	usps	door
loans	century	income	vacation	ups
fargo	professionals	realestate	jcpenney	clinic
fha	clerk	warranty	systems	recreation
payments	sell	corporate	ranch	mall
phh	finance	home	verizon	sporting
leads	fargo	title	healthcare	pizza
ameriquest	nationwide	management	marriott	corp
everhome	investors	experian	southern	fed
fixed	payment	georgia	gift	supplies
wells	deeds	tile	williams	spa

one to seven days out, we find terms to do with locating properties (realtors, mls, property). From one week to a month, we see the user is interested in official matters (llc, insurance, notary, savings, irs). At one to three months, the user is interested in home furnishings (Kohls, Bed Bath and Beyond, Pottery Barn, Sears). Finally, after three months, users are interested in higher-level housing changes (patio, outdoor, pools, lawn).

The second item to note is that even when we look more than three months before or after a mortgage query, the distribution of terms by users who queried "mortgage" is still significantly different from the background term distribution. Put another way, the population of users who searched for mortgage are still distinguishable from the other users, even months later. These results demonstrate that there is information in long-term query logs that would be lost if they were cut into smaller pieces.

As mentioned earlier, our method works equally well whether queries are cut into individual words or not. If we do not cut the queries into individual words, we get the following queries over the five time periods:

0–30m	mortgage calculator, mortgage rates, countrywide,..
1–7d	capital one, bank America, zillow.com, keller williams,..
7–30	kohls, sams club, sears, best buy, usps, costco, . . .
30–90	radio shack, amazon.com, sears.com, delta airlines,. . .
90+	people magazine, amtrak, enterprise, honda, abc,. . .

Understanding the way people's interests evolve over time is interesting, both from the standpoint of computer science, as well as other fields such as sociology and anthropology. For the former, there are a number of services that could be improved. For example, it could assist search engines in making better query suggestions and spelling corrections. For sociologists, economists, anthropologists, and others, being able to better understand people without having to conduct user interviews could be a huge benefit in productivity. For all concerned, long-term query logs provide an invaluable resource for learning about the way in which people's interests change over time.

6.4 Automatic Topic Generalization

In the previous section, we saw that by selecting users who have an interest in a particular term, we were able to find concepts related to that term (for another example, Table III shows the top related terms to "petunias"). However, note that the users are not only connected by their interest in the particular term, but also in the set of topics the term belongs to. For example, two people who have both queried for "carabiner" (a tool used for mountain climbing) are not only both interested in carabiners, but probably also generally interested in mountain climbing.

Thus if we look for more general words that are still searched for with higher probability within the selected population of users than outside of it, we would expect to find topic clusters of increasing generality. One way to define words that are more general is that they are terms that are searched for more frequently by the population. We can control how many users must have searched for a term by increasing our smoothing parameter, m. Recall that m is the number of unobserved users who we a priori believe query independently for q. Increasing m puts a higher requirement on observing dependence between q and r, which can only be met by observing more users who query for both.

Table III. Top 80 Terms Having the Highest Correlation with the Query "Petunias" for the Full 12 Months of Data (i.e., the terms that have the highest $dep_{petunias}(.)$)

petunias, petunia, impatiens, planting, wave, geraniums, plants, growing, annuals, pansies, flowering, perennials, plant, begonias, perennial, seeds, pruning, grow, shrubs, gardening, geranium, nursery, flower, caring, shade, greenhouse, flowers, seed, hydrangea, hibiscus, nurseries, vine, bushes, trees, tomato, landscaping, grass, marigolds, dwarf, leaves, bulbs, begonia, shrub, hanging, baskets, tomatoes, pots, patio, verbena, roses, landscape, purple, salvia, salad, inca, potatoes, soil, decorating, coleus, lavender, potato, container, lawn, outdoor, baked, corn, lilies, bells, dianthus, roast, clematis, penney, pork, fertilizer, vines, gardens, curtains, vegetable, casserole, planters, . . .

Table IV. Topic Clusters for "Carabiner", Using Queries More than 90 Days Before or After the Original Query. (Shown are the top terms related to carabiner for various values of the smoothing parameter, m. The terms within a given smoothing value tend to belong to the same topic and level of generality.)

m	Terms
1k	mammut, petzl, botach, kydex, clevis, extrication, trijicon, webbing, eotech, aimpoint, sportiva, utilize, boker, surefire, aiming, coolmax, scarpa, 5d11, nomex, armament
10k	surefire, mammut, petzl, webbing, flashlight, trijicon, pouch, hydration, climbing, flashlights, retractable, 5d11, sog, tactical, sportiva, holster, repeater, camelbak, waterproof, botach
100k	climbing, tactical, jacket, waterproof, holder, folding, helmet, rope, flashlight, surefire, vest, garmin, galls, sleeve, racks, container, knife, pouch, belt, promotional
1M	supply, gear, equipment, accessories, light, shirt, custom, fire, supplies, outdoor, plastic, set, review, kit, storage, conversion, systems, bag, safety, mountain
10M	water, center, blue, company, store, american, supply, best, mountain, depot, park, code, fire, products, equipment, state, red, light, office, supplies

In Table IV, we give the results of this experiment. As can be seen, the technique extracts clean topic clusters, and each increasing value for m results in a cluster that is more general than the last. For m of 1k to 10M, we see topics generalize from carabiner manufacturers, to mountain climbing terms, to general outdoor gear, and finally, general outdoor terms. Note that here, we are using only terms that were issued at least 90 days before or after a carabiner query. Hence we find terms that are only related to carabiners in that a person who searched for carabiner more than 3 months ago (or will in the future) would search for the given term.

One question to ask, then, is what happens if we use within-session queries. We give the results of this in Table V. As can be seen, even at the highest level of generality, most of the terms are still immediately related to carabiners (such as keychains, and misspellings of carabiner). This is not surprising, since a query session usually involves hunting for a particular piece of information. In this case, it is information on carabiners. It is not often that carabiner will cooccur within a session with terms such as waterproof, fire supplies, safety, or helmet.

Table V. Topic Clusters for "Carabiner", Using Within-session Queries. (i.e., queries that occurred within half an hour of the original carabiner query. Shown are the top terms related to carabiner for various values of the smoothing parameter, m. Unlike IV, which used queries from a broader time period, we do not see the same topic generalization as the smoothing parameter increases.)

m	Terms
1k	carabiners, caribiner, carbiner, carabeaner, caribener, carabineer, carabener, carabeener, caribeaner, caribiners, karabiner, carabina, biner, screwgate, caribeener, carabine, carrabiner, carabiener, carabeners, carribeaner
10k	carabiners, caribiner, carbiner, carabineer, carabeaner, caribener, carabener, carabeener, keychains, carabine, keychain, petzl, karabiner, caribeaner, caribiners, keyring, carabina, locking, climbing, biner
100k	carabiners, caribiner, climbing, keychain, keychains, carbiner, locking, carabineer, carabeaner, flashlight, carabine, caribener, carabener, kershaw, carabeener, petzl, chains, promotional, mug, chain
1M	carabiners, climbing, caribiner, key, chain, keychain, keychains, locking, promotional, chains, clip, flashlight, carbiner, carabineer, carabeaner, carabine, kershaw, mug, caribener, pens
10M	key, climbing, carabiners, clip, chain, caribiner, keychain, keychains, locking, promotional, chains, flashlight, carbiner, carabineer, carabeaner, pens, mug, carabine, kershaw, caribener
100M	(same as above)

How long is long enough for finding general topics? Below we give the most general topic cluster (highest m value) achievable for various time spans:

30 min – 1 day	Rope, carabiners, lanyard, keychain
1 – 30 days	mountain, climbing, review, safety
30 – 90 days	Light, center, mountain, gear, store, water

As expected, the level of generality afforded by the data increases as we increase the time separating the query from carabiner. Thus the value of long-term logs for building general topic hierarchies increases with the length of history contained in the logs.

We also show topic generalization for two more queries: "miter" (Table VI(a)) and "salsa dancing" (Table VI(b)). It is interesting to see that we retrieve miter saw models and manufacturers (probably due to people who own miter saws) at the lowest level of generality, then various woodworking tools, then woodworking terms, and finally, general wood oriented terms. Similarly, for salsa dancing, terms generalize from specific salsa dancing terms, to general dancing, to yoga and fitness topics.

7. TEMPORAL QUERYING BEHAVIOR

Besides looking at correlated interests, another set of questions one may want to ask is of the form: given that a user is interested in r at time t_1, how interested would we expect him to be in q at time t_2? Such questions can be used for many purposes. For example, a sociologist may wish to investigate marriage in different parts of the world. He could see how much time typically passes between the queries: "dating," "proposal," "engagement ring", and "wedding

Table VI. Topic Clusters for (a) "Miter", and (b) "Salsa Dancing", Using Queries More Than 90 Days Before or After the Original Query. (Shown are the top terms related to the query for various values of the smoothing parameter, m.)

(a)	m	*Terms*
	1k	dw130v, finishpro, underpinner, outfeed, jointech, 371k, levelite, jessem, lm30, ls1013, incra, ps20,...
	10k	jointer, powermatic, incra, tablesaw, jessem, mlcs, kreg, nailer, festool, senco, woodworker, jointers, nailers,...
	100k	nailer, rockler, woodworking, dewalt, powermatic, sander, makita, senco, woodworkers, woodworker, jointer, ...
	1M	woodworking, dewalt, lumber, saw, craftsman, drill, installing, router, tools, tool, makita, saws, tile, mower,...
	10M	tools, wood, parts, lowes, tool, depot, supply, sears, electric, door, repair, hardware, saw, paint, water, power,...
(b)	m	*Terms*
	1k	dancesport, salsaweb, dancepartner, salsera, salseros, salsero, ndca, orquesta, rueda, nclr, fandangos, plena, jibaro, ...
	10k	ballroom, dancesport, mambo, Rican, sevilla, rumba, capezio, orquesta, tasting, mariachi, Cuban, rhythms, ...
	100k	ballroom, latin, dance, classes, fitness, Puerto, yoga, lounge, Cuban, angeles, Rican, hour, Spanish, rico, ...
	1M	dance, restaurants, fitness, Spanish, California, restaurants, hair, shoes, san, center, los, beach, classes, spa, ...
	10M	City, center, school, new, American, county, state, home, park, college, club, university, map, hotel, restaurant, ...

planning." More simply, do men typically purchase the ring and then figure out how they are going to propose, or vice versa?

As with the correlation experiments just described, the possibilities for learning about people's behavior, their culture, and the world, are nearly limitless. Do people swim before they bike? Do they learn to salsa before they learn to bachata? What concerns people the most in the week leading up to the day they retire? How many people learn how to sail after buying a boat vs. before? And so forth.

Learning the temporal relationship between two queries can also be useful for search engines. Suppose we find that users who search for instructions on playing bridge tend to search for bridge strategies one to seven days later. If such a user comes to the search engine and searches for "bridge," we can rerank results to give bridge strategy results a higher ranking.

7.1 Problem Formulation

Our goal is to measure the *surprise* in query frequency for a given query, q, for each number of days away from of our reference query, r. We also measure the KL-divergence between the expected frequency distribution and the observed distribution for query q.

One way to look at this is the following: over time, there are many users submitting many queries. If we were to plot the number of users querying for, say, "dog", we would see a fairly steady line over time. The number of people interested in dogs remains fairly constant. However, imagine that we line up

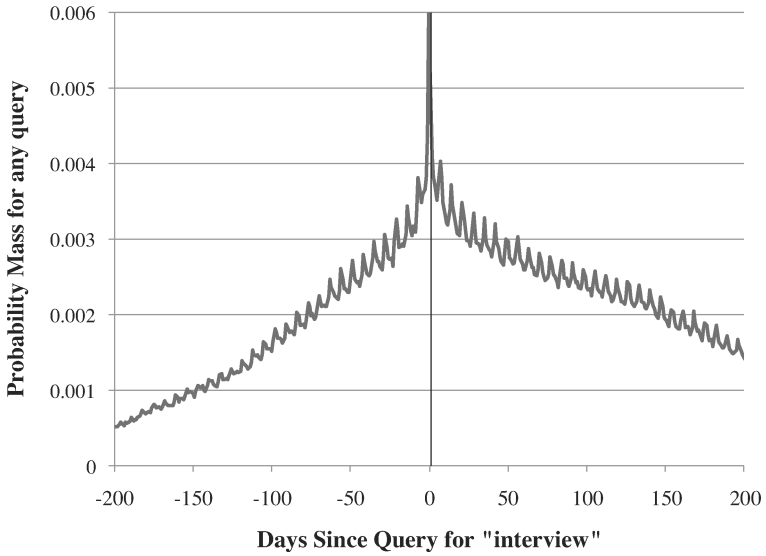

Fig. 3. Distribution of $p(\delta)$—the probability that a user who queried for "interview" queries again δ days later.

all the users so that time 0 is the day they query for "humane society" (an organization from which people can adopt pets), and then plot the frequency of the word dog before and after day 0. We would expect to see a fairly large spike in query activity near day 0, possibly falling off as we look farther in the future or in the past from the day they queried for humane society.

We thus computed the following. Let $n(\delta, q)$ be the number of users who queried for q, δ days from the first time they queried for the reference query (since everything is being computed relative to a reference query, we drop r from the notation). Also, let $n(\delta)$ be the number of users who queried for anything, δ days from the first time they queried for the reference query. Then $p(\delta|q) = n(\delta, q)/\sum_\delta n(\delta, q)$ is the probability distribution of the users querying for q over time, and $p(\delta) = n(\delta)/\sum_\delta n(\delta)$ is the probability distribution of users querying for anything over time.

It is important to note that $p(\delta)$ is not uniform. In fact, it has the shape shown in Figure 3 (this is $p(\delta)$ for the reference query "interview"). As we would expect, the probability that a user queries for something on the same day as the day he queried for interview, is very high. For many users, they may only use a search engine every few days, so we expect a sudden drop in probability for δ of -1 and 1. The continued decrease with time can be explained by the endpoints of the 12-month corpus. The longer the time period required since the reference query, the less data we have. For example, for a δ of 9 months, only users who queried for interview in the first three months of the corpus can have possibly queried 9 months later. The asymmetry between the positive and negative δ is most likely due to our choice to use the first instance of the reference query. The *spikiness* of the plot is due to weekend-weekday effects (each spike is exactly 7 days apart).

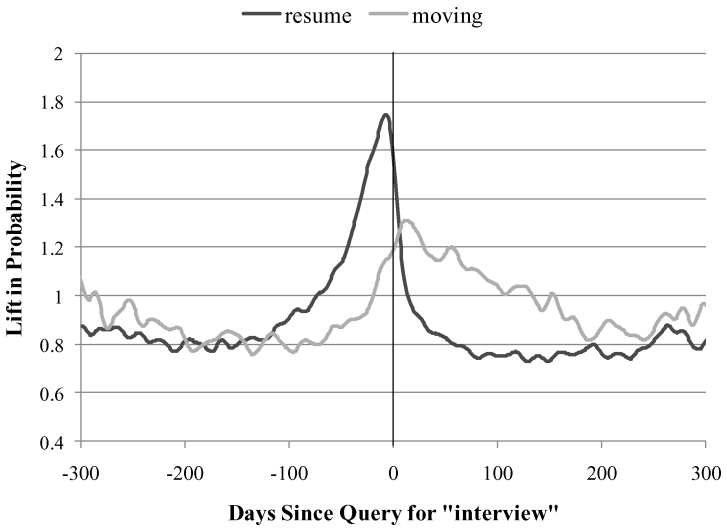

Fig. 4. $S(\delta)$ (lift in probability) for the queries "resume" and "moving" given the reference query "interview". People begin looking for information on resumes up to 100 days before the interview query; most look immediately before. Users become significantly more interested in moving information after the interview query.

Thus we normalize $p(\delta|q)$ by dividing by $p(\delta)$. The resulting quantity: $S(\delta) = p(\delta|q)/p(\delta)$ is similar to the correlation score introduced in Section 5. As with the previous measure, it can also be seen as the exponentiated PMI between δ and q. It is this measure that we plot (vs. time) to observe the temporal effects of a query vs. a reverence query.[5]

We also compute the KL-divergence between the two distributions as a measure of how temporally-related q is to the reference query:

$$D_{KL}(p(\delta|q)||p(\delta)) = \sum_{\delta} p(\delta|q) \log\left(\frac{p\ (\delta|q)}{p\ (\delta)}\right).$$

A query that is unrelated to the reference query, or one that is related but not temporally, will have a probability distribution that looks similar to the baseline distribution p(δ), and hence will have a low KL-divergence.

7.2 Results: Jobs and Interviewing

To demonstrate the potential of this approach, we chose two domains: interviewing for a job, and acquiring a new pet. As will be seen, in both cases we were able to find interesting temporal patterns.

The first domain is that of interviewing for a job. For this, we used the reference query "interview". Two activities associated with interviewing and finding a new job are updating one's resume, and sometimes the need to move after being offered a job. The results are given in Figure 4. As can be seen, users begin

[5]In the following graphs, we smoothed $S(\delta)$ using Gaussian-weighted averaging with a standard deviation of 5 days. This is simply for readability of the graphs and does not affect the results.

Fig. 5. $S(\delta)$ for the queries "tax" and "salary" relative to "interview". People are most interested in taxes one to two months before the interview query, and again seven months later. Interestingly, interest in taxes is lowest when interest in salary is highest.

looking for information on resumes (query: "resume") up to 100 days before their first interview query. Interestingly, resume activity is most intense just before their first interview, and drops dramatically immediately after.

Though the user no longer seems interested in resumes, he has begun searching for information on moving (query: "moving"). This activity tends to begin before the interview, but peaks at day five, and again at day nine, after the first interview query. This may be the typical time between interviews and offers. In the query logs, we also find that the peaks for the term "offer" are 6 and 11 days after the first interview query, further supporting this hypothesis. Of course, we cannot know for sure exactly what users intend by these searches, but the information gleaned from them is indicative, and could inform a researcher conducting surveys, or a machine learning algorithm predicting, for example, when to show an advertisement for a moving company.

For a second set of experiments, we suppose an economist might be interested in better understanding the relationship between people's salaries, changing jobs, and their perceptions on taxes. For instance, he may ask "do salaries affect people's choice of a new job" and "do taxes affect people's choice of a new job." In Figure 5, we give $S(\delta)$ for "tax" and "salary" with reference to the query "interview". Here we see that people are generally interested in salary-related topics, both before and after the interview for approximately two to three months. Thus salary issues do seem to play a part in the job hunting process.

Also interesting is the curve for tax. When the user is most interested in their salary, they are least interested in taxes. As can be seen from the chart, tax queries peak about three months before interviewing, and again about seven months after, and dip to their lowest point right around the day of the

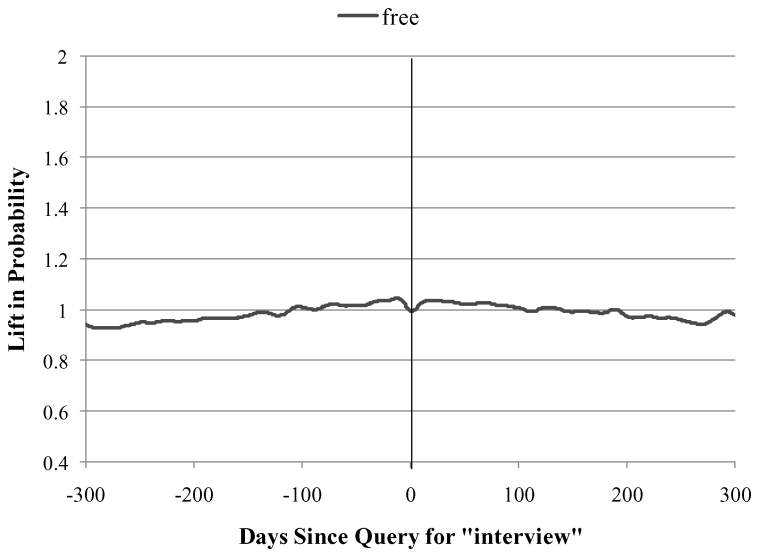

Fig. 6. A temporally unrelated query, ("free"), appears as a flat line.

Table VII. KL-Divergence for Temporal Distribution of Various Terms Relative to the Query "Interview"

Term	KL-Divergence ($\times 10^{-2}$)
resume	5.67
tax	3.12
moving	2.21
salary	2.03
free	0.14

interview query. There are two possible explanations for this. One possibility for this is that people are interested in taxes approximately six months after they interview, because of something to do with the interview (e.g., they are receiving their paycheck or making plans for tax-free retirement accounts). Another possibility is that most interviews just happen to occur about six months before April 15th (the day taxes are due in the United States). By examining the distribution of interview queries to see whether they peak in September, we could determine which is correct. Note that the peaking of interview queries would have to be quite sharp, since the peak for tax is very sharp. Also, the time between the two "tax" peaks is only 291 days. If both peaks were due to tax being correlated with time, the peaks would occur 365 days apart.

For an entirely unrelated query, we expect $S(\delta) = 1$. In Figure 6, we plot the query "free" using the reference query "interview"—two terms that we expect to be unrelated to each other. As anticipated, we see a roughly flat line. In Table VII, we give the KL-divergences for each term in these plots. As expected, "tax" and "moving" have the most divergent probability distributions from the expected distribution. Also note that "free" has a significantly lower KL-divergence than that of the related terms. Again, this is expected, but it

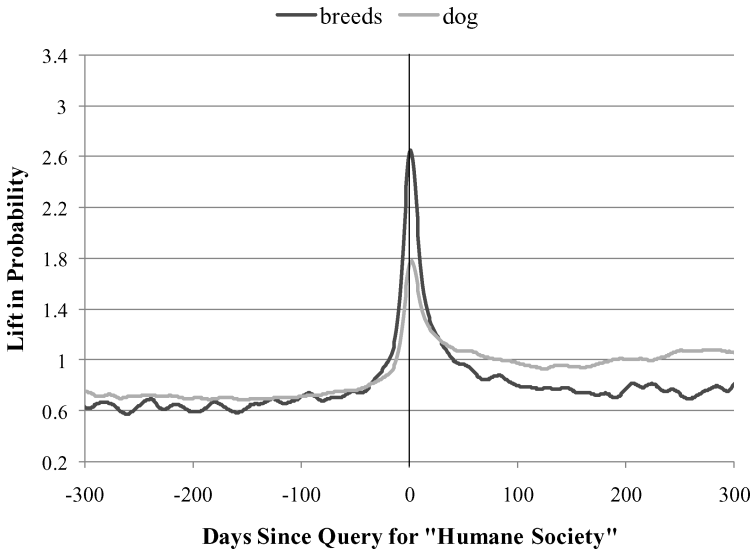

Fig. 7. $S(\delta)$ for "breeds" and "dog" given the reference query "humane society." Interest in dogs remains higher after the reference query than before, indicating the people have probably adopted a pet.

also demonstrates the use of KL-divergence as a possible metric for identifying interesting temporally-related terms.

7.3 Results: Pet Adoption

We will briefly describe one more experiment, involving pet adoption. Again, we can imagine a veterinarian who is interested in the issues that affect pet adoption. By plotting "breeds" and "dog" vs. "humane society", a few interesting patterns emerge (see Figure 7).

First, we see that dog breed searches are highly temporally correlated with queries for the humane society. This is, of course, not surprising, since a person who is planning to visit the humane society (a pet adoption agency) will want to know what type of dog or cat he wants to adopt. Also, not surprisingly, we see that queries for dog are also very highly correlated with humane society.

It is interesting to note that long after the humane society query, users' interest in general dog information remains higher than before the humane society query. From this, we can reasonably conclude that many of the users who visit the humane society (or at least, query for it—they may end up visiting other pet adoption agencies) end up adopting a pet (we see the same pattern for the query "cat"). The difference in levels is quite significant: there are 693 users who searched for dog fifty days before humane society, and 1176 users fifty days after.

In contrast, queries for breeds drops almost to the level it was before. Perhaps this means users have adopted a pet—once they own a pet, their interest in breeds is likely to drop back to the level it was before they started looking for one. We recognize that these results are not as reliable as a true study that

surveys humane society visitors, but they do demonstrate the wide variety of experiments and research that could be informed or supported by the techniques in this article.

8. DISCUSSION

This article is a first attempt to: (1) measure whether query effects extend beyond the immediate short-term around the query, (2) determine whether those effects can provide valuable information, and (3) investigate the possible scientific and research value of the information contained in long-term query logs.

As the previous results have conclusively shown, there are relationships in long-term query logs that extend weeks and even many months beyond the original query. When conditioning on a given query term, the distribution of other terms can vary from the background distribution as far as 12 months away, and we expect the term distributions would not converge for at least many years, if at all. The implications are quite large: chopping long-term logs into shorter term sessions will lose information.

We believe we have demonstrated that this long-term information is interesting and valuable, both for users of online services, as well as for researchers in other fields. Search engines, for example, can use the information to improve query suggestion and ranking. Query suggestions could be based on more generally-related topics, rather than terms only directly related to their query. A search for "carabiner" could produce suggestions such as "mammut" and "petzl" (two manufacturers of carabiners) as well as "waterproof flashlights" and "rope". Similarly, understanding how users' interests change over time (such as an increased interest in dogs after querying for humane society) could lead to improvements in search engine ranking. Whenever a search engine can better estimate what a user cares about, it can do a better job of tailoring its ranking to that user.

To us, however, the most exciting opportunity provided by long-term query logs is the way in which they may be able to contribute to scientific endeavors in other disciplines, such as medicine, sociology, epidemiology, and so on. For a medical researcher investigating a rare condition, finding enough people to conduct a sizeable study is a huge obstacle. Even worse is investigating the correlation between two rare populations. In query logs, a rare condition may affect thousands of people, each of whom has effectively provided a list of their interests, hobbies, and activities. By using these techniques to increase their understanding about medical conditions, researchers may be able to more quickly bring treatments to those affected by the condition.

Also particularly interesting is the potential for *reverse importance sampling* of people. That is, answering question such as: for all people who signed up for a "stop smoking class" today, what was their opinion 100 days ago on whether they could ever stop smoking. The only way to answer this question at the moment is to survey thousands of people on their thoughts about quitting smoking, in the hopes that some of them will end up joining a class in 100 days. Not only is this difficult, but it also introduces a bias into the study. The other alternative, choosing people who joined the class and asking them what they were thinking

100 days prior, has problems with subjective bias. With query logs, a medical researcher could answer the question by using the previous query logs for people who joined a class.

We believe we have demonstrated some of the potential that long-term query logs have for scientific research. While we recognize that there is still significant work to be done (primarily in ensuring that query activity can be correlated to real world activity—see Section 10 for further discussion), we believe this approach to scientific research will eventually lead to novel discoveries in a variety of fields.

8.1 Making the Logs Available

One question is, though there is information contained in the logs that can be used by researchers, how can they access that information unless they work for a search engine company? We realize this is an important question, but it is one that requires primarily legal, rather than technical, solutions. The primary goal of this work was to demonstrate the logs' potential value, leaving it up to the organizations that hold the data to determine the best course of action. That said, one technical solution may be for the query log owner to provide a service (Web or otherwise) that returns the highest-scoring terms for a given query, or provides the score for a given pair of terms. Measures to protect privacy, such as only returning queries that have been issued by at least k unique users, would need to be incorporated as well. Recent work on privacy-preserving data mining [Blum et al. 2005] has shown that there are methods for mining sensitive data that are both powerful and provide strong privacy guarantees; these may be able to be incorporated into such a service.

9. RELATED WORK

Most existing work on mining query logs either ignores individual users (i.e., uses overall query volume), or only considers individual users' short-term activities (e.g., search sessions).

One of the most well known examples of observing query volume over time, but aggregating across all users (i.e., ignoring who issued the query), is Google Trends (http://trends.google.com). This application allows people to chart the overall query popularity of various keywords. Beitzel et al. [2004] studied trends such as this over short periods of time to learn about the behavior of the querying population at large. Work by Chien and Immorlica [2005] showed that terms can be clustered using their correlation in temporal query frequency. For example, queries for online news sources tend to peak around noon on each weekday, have low volume on weekends, and have particularly high volumes on Friday evenings. By finding all queries that match this pattern, a collection of online news agency queries can be built. Vlachos et al. [2004] also used temporal similarity to find semantically similar queries by decomposing each query into its set of Fourier coefficients. They were able to compute a lower-bound on the distance between two queries given this compressed representation, which allowed them to rapidly find a query's nearest neighbors (in temporal similarity). Additionally, they demonstrated a method to detect bursts of query activity, and

considered a similarity metric based on these bursts. Another interesting paper on temporal mining of query logs is that by Adar et al. [2007], in which they compared the query frequency in query logs to that of blog posts and news articles to view how concepts *travel* through the different audiences. Note, these all differ significantly from our work on temporal querying behavior in that they consider overall query popularity instead of individual user query behavior, and thus are limited to patterns and trends in the entire population. By lining users up at a particular reference query, we have shown that more information can be deduced about people's interests and behaviors.

There are a multitude of papers on mining information from short-term search sessions; we can only provide a brief overview of a few of them. For example, Cucerzan and Brill [2005] used within-session information for query suggestion. They computed which queries tended to immediately follow a given query (more often than would be expected *a priori*) and considered those to be related terms (for example, for the query "Jaguar", they retrieved "Ford", "BMW" and "Mercedes" as uncommonly popular next queries). As we have shown, by looking at correlation on a longer time scale, we find queries that are not just semantically similar, but also may be related more generally (because the same user is interested in both). They also showed that by looking at the extensions of a set of semantically similar queries, they could find concepts related to the queries. For example, for the set of car-related terms given previously, they retrieved the concepts "performance", "club", "forums", and "auto parts". Their work on common query extensions complements our work on using longer-term logs for finding broader concepts, and we would like to explore the use of both. Similar is the work on query substitutions by Jones et al. [2006] and Rey and Jhala [2006]. In these works, a distinction is drawn between *substitutable queries* (those which have very similar meaning) and *associated queries* (those which the user may also be interested in, and are on the same topic, but not the same meaning). The substitutable queries are mined by considering queries that immediately follow each other, and associated queries are considered to be those that cooccur in the same session, but are not typically substitutable. In our work, by looking at longer time periods than a session, we tend to find associated relationships rather than substitutable.

There have been a number of papers on using query log clicks or search results to cluster terms (see, e.g., Baeza-Yates and Tiberi [2007], Beeferman and Berger [2000], Bollegala et al. [2007]). The primary difference between these works and our own is that they are searching for relationships that are already known (e.g., terms that cooccur on Web pages), while we hope to find novel relations due to the same user being interested in both. Session information has also been used for query reranking [Shen et al. 2005]. See also Grimes et al. [2007] for a discussion of the advantages (diversity) and disadvantages (sometimes difficult to determine user intent) of using query logs. Query reformulations were also used by Jones et al. [2007] on a variety of studies involving geographical queries. Some of their techniques resulted in interesting observations about human behavior. For example, "map" queries tend to be about locations far from the user's current location, whereas "pizza" queries were about locations close to the user.

Recently, there has been some work studying user querying behavior over longer periods of time, but to the best of our knowledge, our work (with many millions of users having 12 months of history) is the largest study of long-term user behavior to date. Teevan et al. [2007] examined user requerying behavior, using long-term logs and individual user identifiers. Though they used a corpus of only 114 users from Yahoo's query logs, this study is the largest long-term query log study we are aware of that maintains statistics on individual users. Another body of work that uses query log behavior is that of personalized search. Dou et al. [2007] used a 12-day query log history of 10,000 users to evaluate various personalized search methods. Tan et al.'s [2006] work on long-term query logs covers a longer period of time (70 days), but for only four users. In these works, the goal was to improve search retrieval accuracy by biasing the search results toward pages that were also related to earlier queries (and clicks) by the same user. For this task, they found that using the entire history proved useful, but primarily only for queries that were repeats of an earlier query. Wedig and Madani [2006] have done an excellent analysis of the user statistics in a subset of the Yahoo! query logs, with a goal of measuring how plausible it is to perform search personalization. For instance, they show how much history a user is likely to have when they issue a query, and they also find that users tend to remain interested in the same query topics over long periods of time.

Anonymizing user query logs is also an active area of research (see, e.g., the *Workshop on query log analysis* at WWW2007). The issues here are complex, and involve social, technological, and ethical discussion. We would like to incorporate the work by Adar [2007] on dividing user histories by topic to see what effects that would have on the discoveries that can be made. As mentioned in Section 8.1, another approach is to enlist the use of privacy-preserving data mining [Blum et al. 2005; Agrawal and Srikant 2000]. Such techniques may allow external researchers to perform queries on the data with guaranteed privacy safeguards. We feel this is a very promising direction of future work; particularly, investigating how the specific task of finding terms that are related can be conducted with guarantees that may be stronger than those provided for general data mining.

Previous studies have demonstrated the promise of data mining as a technique to make scientific discoveries. Perhaps the most well known is the work by Swanson [1986] discovering a connection between Raynaud's syndrome and fish-oil, and later linking magnesium deficits to migraines [Swanson 1988], by mining medical literature for undiscovered connections. In these works, Swanson used a combination of mining and domain expertise to find the links; our hope is that similar benefits could be obtained using a combination of our technique and domain experts (e.g., a medical researcher who knows what factors are already known to be related, and which are plausible given the way a syndrome works). The better we are able to filter the known relations, the more useful the system will be to a domain expert, and so this will be a significant direction for future work. More recently, there has been a flurry of work on mining the MEDLINE public database of medical abstracts for automatic (or guided) medical research. See Hirschman et al. [2002] for a nice overview. We

believe we have demonstrated the potential of mining long-term query logs for similar purposes.

10. FUTURE WORK

The primary direction for future work, now that we have demonstrated that we can detect long-term patterns in query logs, is to show that these patterns correspond with actual user interests or activities in the real world. Though we have not solved this problem in general, it is evident from the results in Sections 6.1 and 6.4 that the most-correlated terms do correspond to real-world interests. If that were not the case, we would not have found generalities of "carabiner" or "miter" by looking at other terms queried by those users more than 3 months later. However, for scientific value, we will not be interested in only the most correlated terms, so the question remains: can we determine the probability that a user who queries for q is actually interested in (or owns, has as a condition, etc.), q?

To make this concrete, suppose we are attempting to identify whether users who climb mountains tend to get knee injuries. First, we can require that they have issued at least some number of queries for climbing equipment, on at least some number of different days. We can also look at query patterns during the day vs. the night, on weekdays vs. weekends, and so on. Also, we can filter out users who issue too many queries for outdoor equipment (i.e., these may be outdoor equipment manufacturers or marketing firms). Terms near the query can also be used to deduce the level of interest, such as words like "buy" or "repair", as opposed to "gift" or "sell". We plan to study these and other techniques using a real world data set. For instance, suppose we knew the true correlation between people who climb mountains and people who have knee injuries. Given that data, we can fit a model that predicts who actually climbs mountains and has knee injuries, given their query behavior. In fact, any study that can be done using long-term query logs is also an opportunity for the reverse: to learn more about how to use query logs by knowing the real-world answer to the study. This is currently being explored and we hope to publish the results in a future work.

11. CONCLUSION

Long-term query logs are an invaluable source of information about the real world. The vast quantity of observational data contained in the logs opens up new opportunities in many fields that are not possible with short-term logs. The potential for benefit extends from the Web to a wide array of disciplines such as sociology, medicine, and economics. Most exciting of all are the new opportunities the logs provide for experimentation and understanding that have previously been impossible. By better understanding people, the way they behave, and the way their interests evolve, we can provide benefit to them, and to the society at large. Though the results presented in this article are preliminary, we believe that they demonstrate significant promise for further research in this area, paving the way for many advances in using query logs for the benefit of users and society.

ACKNOWLEDGMENTS

Many thanks to Mikhail Bilenko, Eric Brill, Silviu Cucerzan, Kristin Miller, and Ryen White for fruitful discussions.

REFERENCES

ADAR, E., WELD, D., BERSHAD, B., AND Gribble, S. 2007. Why we search: Visualizing and predicting user behavior. In *Proceedings of the 16th International World Wide Web Conference.* Banff, Alberta, Canada.

ADAR, E. 2007. User 4XXXXX9: anonymizing query logs. In *Proceedings of the Workshop on Query Log Analysis.* Banff, Alberta, Canada.

AGRAWAL, R. AND SRIKANT, R. 2000. Privacy-preserving data mining. In *Proceedings of the 2000 SIGMOD International Conference on Management of Data*. Dallas, TX.

BAEZA-YATES, R. AND TIBERI, A. 2007. Extracting semantic relations from query logs. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* San Jose, CA.

BEEFERMAN, D. AND BERGER, A. 2000. Agglomerative clustering of a search engine query log. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* Boston, MA.

BEITZEL, S., JENSEN, E. C., CHOWDHURY, A., GROSSMAN, D., AND FRIEDER, O. 2004. Hourly analysis of a very large topically categorized web query log. In *Proceedings of the 27th Annual International ACM SIGIR Conference*. Sheffield, South Yorkshire.

BLUM, A., DWORK, C., MCSHERRY, F., AND NISSIM, K. 2005. Practical privacy: the SuLQ framework. In *Proceedings of the 24th ACM SIGMOD International Conference on Principles of Database Systems*. Baltimore, MD.

BOLLEGALA, D., MATSUO, Y., AND ISHIZUKA, M. 2007. Measuring semantic similarity between words using Web search engines. In *Proceedings of the 16th International World Wide Web Conference.* Banff, Alberta, Canada.

CHIEN, S. AND IMMORLICA, N. 2005. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of the 14th International World Wide Web Conference.* Chiba, Japan.

CHURCH, K., HANKS, P., HINDLE, D., AND GALE, W. 1991. Using statistics in lexical analysis. In *Lexical Acquisition: Exploiting Online Resources to Build a Lexicon*, U. ZERNICK, Ed. Lawrence Erlbaum, Hillsdale, NJ, 115–164.

CUCERZAN, S. AND BRILL, E. 2005. Extracting semantically related queries by exploiting user session information. Tech rep. Microsoft Research.

DOU, Z., SONG R., AND WEN, J. 2007. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of the 16th International World Wide Web Conference*. Banff, Alberta, Canada.

FELLBAUM, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

GRIMES, C., TANG, D., AND RUSSELL, D. 2007. Query logs alone are not enough. In *Proceedings of the Workshop on Query Log Analysis.* Banff, Alberta, Canada.

HIRSCHMAN, L., PARK, J., TSUJII, J., WONG, L., AND WU, C. 2002. Accomplishments and challenges in literature data mining for biology. *Bioinformatics 18*, 12, 1553–1561.

JONES, R., REY, B., MADANI, O., AND GREINER, W. 2006. Generating query substitutions. In *Proceedings of the 15th International World Wide Web Conference.* Edinburgh, Scotland.

KULLBACK, S., AND LEIBLER, R. A. 1951. On information and sufficiency. *Ann. Math. Stat. 22*, 1, 79–86.

REY, B. AND JHALA, P. 2006. Mining associations from web query logs. In *Proceedings of the Web Mining Workshop*, Berlin, Germany.

SHEN, X., TAN, B., AND ZHAI, C. 2005. Context-sensitive information retrieval using implicit feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference.* Salvador, Brazil.

SWANSON, D. R. 1986. Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspect. Biol. Med. 30*, 7–18.

SWANSON, D. R. 1988. Migraine and magnesium: eleven neglected connections. *Perspect. in Biol. Med. 31*, 526–57.

TAN, B., SHEN, X., AND ZHAI, C. 2006. Mining long-term search history to improve search accuracy. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA.

TEEVAN, J., ADAR, E., JONES, R., AND POTTS, M. 2007. Information re-retrieval: repeat queries in Yahoo's logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference*. Amsterdam, The Netherlands.

VLACHOS, M., MEEK, C., VAGENA, Z., AND GUNOPULOS, D. 2004. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the 23th ACM SIGMOD International Conference on Management of Data*. Paris, France.

WEDIG, S. AND MADANI, O. 2006. A large-scale analysis of query logs for assessing personalization opportunities. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Philadelphia, PA.

WEN, J., NIE, J., AND ZHANG, H. 2001. Clustering user queries of a search engine. In *Proceedings of the 10th International World Wide Web Conference*. Hong Kong, China.

Received December 2007; revised July 2008; accepted August 2008