

# Temporal Analysis of Web Search Query-Click Data

Sitaram Asur  
Ohio State University  
asur@cse.ohio-state.edu

Gregory Buehrer  
Microsoft Corporation  
buehrer@microsoft.com

## ABSTRACT

Understanding the intent of a query is paramount to building an effective web search engine. This is challenging, due to the sparsity in the number of words provided by the user (typically less than three), and also because the nature of queries evolves over time. In particular, distinguishing requests specifically for recently published content such as blog post requests or news queries is a very difficult task. This is due primarily to the transient nature of the topics for these queries. To improve query classification, we propose to understand and extract specific temporal signatures for different types of queries. In this regard, we construct a time series of query-clicks as a sequence of bipartite graphs, and develop a click rank score measure to capture the difference in terms of rank entropy across the time series. We demonstrate the benefits of offline temporal analysis in studying and isolating different types of queries such as navigational, adult and news queries.

## 1. INTRODUCTION

Conservative estimates place the size of the web in terms of number of links to be 300 billion links<sup>1</sup>. More aggressive estimates suggest a number in the hundreds of billions. The goal of a search engine is to match user requests with a small relevant subset of these pages. Search engines service millions of users and hundreds of millions of queries daily and this number is constantly growing, and thus the challenge of engineering an effective engine is particularly daunting. One source of information that can aid in this engineering is query-click data.

Search engine query-click data is constructed from sets of daily interactions between users and the search engine. It represents the pages clicked for particular queries by users and can provide large scale statistics of search engine performance in terms of answering queries. Also, such data can provide information useful for gaining insight into the tem-

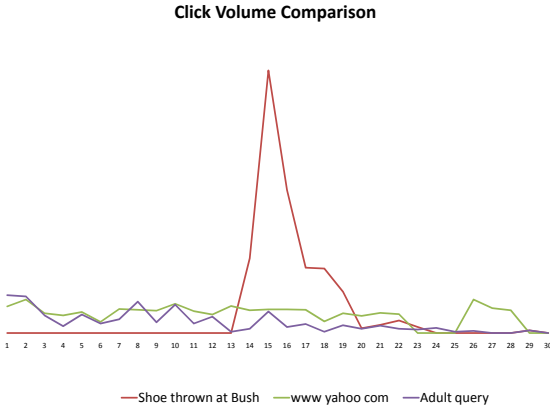
<sup>1</sup><http://www.pandia.com/sew/383-web-size.html>

poral dynamics of users and query patterns. The task of analyzing the evolution of query patterns over time is important from the perspective of developing efficient search services because it can significantly affect the optimal choice of documents to display to the user. For example, if a query is known to be a news query, the most recent documents on the topic are likely to be preferred. However, if the query is a reference query, the publication date of the document is less likely to be a dominant factor in its underlying relevance.

It has been shown in previous work that web queries can be categorized into taxonomies [1, 3, 5, 6]. However, these approaches have been based almost entirely on utilizing semantic and topical information. It has been discussed that the relative small sizes of query words makes this task considerably difficult. Our hypothesis is that the temporal behavior of these queries, both on a per user basis, as well as in aggregate, can provide interesting clues to filter different categories of queries. Furthermore, these temporal query signatures can be useful in designing features for augmenting classification models for web queries.

To illustrate the benefits of temporal analysis, we consider three example queries; 1) a news query {shoe thrown at bush}, 2) a navigational query [5] {www yahoo com}, and 3) an adult query. The relative daily click volumes for these queries for the month of December is shown in Figure 1. In terms of total click volume for the month, the query {www yahoo com} has the most, followed by the news query and then the adult query. When we consider the trends for these queries, we can observe an obvious difference between the news query and the other two, in that it is represented as a brief spike of high mass. Most news stories can be considered to occur in brief busy chunks over time, restricted to a few time intervals once they break, as opposed to navigational and adult queries, which may be more frequent and are not characterized by spikes of very high magnitude over short durations. Our goal in this paper is to analyze the temporal patterns for these different classes of queries and attempt to isolate their behavior.

Discovering patterns for particular types of queries has tremendous applicability and importance in web search. Satisfying news queries is challenging for at least three reasons. First, the pages matching the query are typically recently published, and so crawling to discover the page in sufficient time is difficult. Second, ranking the discovered page is difficult. This is because the page has little to no links associated



**Figure 1: Temporal trends for 3 different queries - navigational, news and adult.**

with it (due to its recency), which lowers its HITS [9] and PageRank [4] scores, and lessens the signal of the associated anchor text. Third, the query must be correctly categorized as a news query, which is difficult because it is a function of current events. In this article, we only address the third challenge.

Accordingly, we examine large amounts of query-click data in a sequence of temporal snapshots using two scales of granularity - hours and days. Our primary focus is on identifying temporal signatures that can enable us to extract particular types of queries with high precision. We develop a rank difference score that can lead to a temporal stability measure which affords good separability for different types of queries. In particular, from our analysis of the temporal dynamics of queries, we discover that the patterns of news queries are different from other queries. They are far more irregular and occur in chunks over time, which can be captured by considering the data temporally. Our preliminary experiments in using temporal features for classification indicate that they can be employed to augment traditional classification features to good effect. Currently, we do this in an offline manner, and are investigating online techniques.

In short, the contributions of this paper include

- Temporal analysis of a large time-series of query-click data using two scales of granularity - hours and days.
- Development of a Rank Difference Score and a temporal Stability measure to compute trends in query-clicks over time.
- Study of temporal signatures for 3 different types of queries - Navigational, Adult and News queries.
- Preliminary classification results demonstrating the benefits of temporal analysis and temporal features for query classification.

The paper is organized as follows. In the next section, we review the related work. In Section 3, we describe the proper-

ties of the click-through data that we consider. In Section 4, we describe our temporal analysis on the click-through graph in detail, followed by experimental results. We conclude in Section 5 and detail future directions from this study.

## 2. RELATED WORK

Recently there has been some interest in temporal analysis of query logs and click-through data. The focus of the work has been mainly in two directions - identifying semantically similar queries, and capturing burst trends and periodicity in query time-series. Vlachos and others [14] have described methods for discovering semantically meaningful queries from query log data. They build a timeseries for the occurrences of query words in a day. They use Fourier coefficients to compress the data and a measure of temporal correlation using Euclidean distance to identify similar queries. They index the extracted features and discuss methods for burst detection and discovering important periods in the timeseries. Beitzel *et al* [2] have analyzed hourly query log data for a week to identify changes in topical popularity and uniqueness of queries. Their main focus is on illustrating trends in terms of popularity and traffic patterns for queries belonging to different topics. Their main conclusions are that query volume is highest and query sets most stable during peak hours in the day. Chien and Immorlica [8] have devised a temporal correlation measure for identifying semantically similar queries. Their measure is based on the correlation coefficient of the frequency functions of the queries. They have presented an efficient implementation to identify top-correlated queries with space and time savings.

Zhao and others [16] discuss a scheme for using temporal characteristics of click-through data to identify semantically similar queries. They use a marginalized kernel technique to construct their similarity model. In other work, Zhao *et al* [17] propose an approach to detect events from web query logs. They represent the query-url pairs as a vector graph and cluster its dual to obtain semantically relevant clusters of query-url pairs. They develop a two-phase graph cut algorithm using semantic and temporal information for clustering. Li *et al* [10] show how to infer class memberships for unlabeled queries on the click graph, using proximity measures. Their method regularizes the learning by content-based classification to improve the efficiency. Chen and others [7] propose a method to identify real-world events by performing subspace analysis on click-through data. They extract subspaces using a Generalized Principal Component Analysis algorithm and perform non-parametric clustering to isolate interesting events.

There has been considerable work in query categorization [1, 5]. Broder [5] introduced a taxonomy for web searches classifying queries as informational, navigational or transactional. He also presented an analysis and survey on Altavista users. Beitzel and others [1] classified navigational queries from a web search log using edited titles in web directories to automatically evaluate navigational web search. In later work, Broder and others [6] attempt to classify rare queries using search engine results. Dou Shen and others [13] discuss using an intermediate taxonomy to build a bridging classifier that can categorize queries in an online mode. The main benefit is that it reduces training costs, since the classifier does not need to be retrained for new category structures.

### 3. THE QUERY-CLICK GRAPH

The Query Click-through data consists of sets of queries entered by users in the web search engine and the corresponding clicks on the pages returned. The information that is available includes the pages clicked on for a query  $P$ , the number of clicks on each of the pages  $C$  and the time during which the click occurred. This can be modeled as a bipartite graph  $G = (Q, P, E)$  with weighted edges between queries and the corresponding web pages clicked.

The Query Click graph evolves over time, with new queries generated and existing queries repeated. The pages clicked for a particular query changes as new pages are created and the interests of users change. In order to analyze the evolution of this graph, we need a *temporal interval*  $T$  for partitioning. The evolving bipartite graph can then be represented as a series of bipartite graph snapshots, one for each chosen time interval, with weighted edges between queries and the corresponding web pages. The weights on the edges represent the number of clicks on that url for each query in the interval being considered. The sequence of bipartite graphs is represented as  $G = \{G_1, G_2, \dots, G_n\}$ , where each snapshot graph is given as  $G_i = (Q_i, P_i, E_i)$ . Here  $Q_i$  represents the queries,  $P_i$  representing the pages and the edges between them is given as  $E_i = \{(q, p, c) : |q \in Q_i, p \in P_i, c > 0\}$ . Each edge  $(q, p, c)$  is weighted by the number of clicks  $c$  on page  $p$  for query  $q$ .

The data that we used consisted of hundreds of millions of queries drawn from the December 2008 logs of a major commercial search engine. We considered two temporal intervals - one hour and one day. For the hourly time scale, we extracted data for 30 hours from 12:00 AM Dec 21st. For the daily data, we considered data from Dec 1st to Dec 30th. We pre-processed the data to remove all non-english queries. Also, for each time scale, we considered only queries that had at least 3 clicked pages, in that time interval. The query distributions for the two time scales are shown in Figure 3(a) and (b). We can see that for the hourly data, the peak hours are from 8AM - 8PM, where the number of queries is more than 4% the total number of queries <sup>2</sup>.

For the daily data, we see that there are more queries in the first half of the month than the second. <sup>3</sup> In our analysis, we use only the temporal information and statistics and do not incorporate semantic features, although that is a part of our proposed future work.

### 4. TEMPORAL ANALYSIS

In this section, we describe our temporal analysis framework. We are motivated to discover how changes occur across time with respect to queries and the click rate.

As shown in Algorithm 1, we consider each set of successive timestamps and analyze the changes across the corresponding bipartite graphs. For this, we consider queries that are common across timestamps and compare their urls and clicks. An interesting observation from the data was that, although the queries common across two timestamps (referred

<sup>2</sup>These are queries that have at least 3 pages that are clicked.

<sup>3</sup>We attribute this to the fact that most people are on vacation in the second half of December.

---

#### Algorithm 1 Query Analysis( $G, T$ )

---

**Input:** Query click graph  $G = (Q, P, E)$  and  $T$ , the number of intervals  
Convert graph  $G = (V, E)$  into  $T$  temporal snapshots  $G = \{G_1, G_2, \dots, G_T\}$ .  
**for**  $i = 1$  to  $T - 1$  **do**  
    Identify queries in common (Match Queries)  
    Compute Rank Difference Score (RDS)  
**end for**  
Compute stability for queries using RDS

---

to as *Match* queries) form a small percentage of the total queries for the two timestamps, they represent a large percentage of the click volume across those timestamps. This result is shown in Figure 3(a) and (b) for the hourly and daily data respectively. The Query Match rate refers to the proportion of the queries that are common between the two timestamps considered.

$$Match\ Rate(i) = \frac{|Q_i \cap Q_{i+1}|}{|Q_i|} \quad (1)$$

The Click Volume Rate refers to the percentage of the total click volume for those two timestamps that is contained by these matching queries. In the case of the daily data, around 20% of the queries are common and they contribute 70% of the click volume, which is close to the Pareto 80-20 ratio [12].

We are interested in observing changes in the page set for a query over time. The rank of web-pages displayed for a query is important to evaluate the performance of a search engine. It has been shown that most users examine only the top ranked pages for a query. Hence, there is great emphasis on the top pages displayed being extremely relevant to a given query. In our case, we only have information regarding the clicks on a given query, and not the order of display. Since the rank of pages is important, we choose to measure and quantify changes to queries over time, based on changes in the order of clicked pages over time. We also note that, a small change in the displayed urls is possible, due to slight variability of the index beds.

We illustrate with an example in Table 1, showing 3 queries and their clicked pages. The time periods considered were Dec 22nd and 23rd. The first query {caylee anthony latest news} is a news query and the pages clicked are from different news websites. It can be seen that the pages clicked vary drastically over the two time periods. This is due to the fact that news articles are dynamic and updated based on the real-world events that are occurring. Once a story breaks, users track the progress of the story by accessing recently updated articles. The second query is a general informational query {cheap flights}. It can be seen that the order of the pages for this query remains the same over the two time periods. They point to popular travel websites. The third query {walmart} is a navigational query. And in both timestamps, the homepage for walmart is clicked on the most, followed by the wikipedia article for it. It is apparent from this example that changes in click behavior is a function of the *type of query* that is being serviced. Also, the difference in the top few results by click can be used to quantify the extent of the change across time, and possibly identify the type of query. For this purpose, we devise a measure based on the changes in the top-ranked pages to capture the difference in click behavior across time.

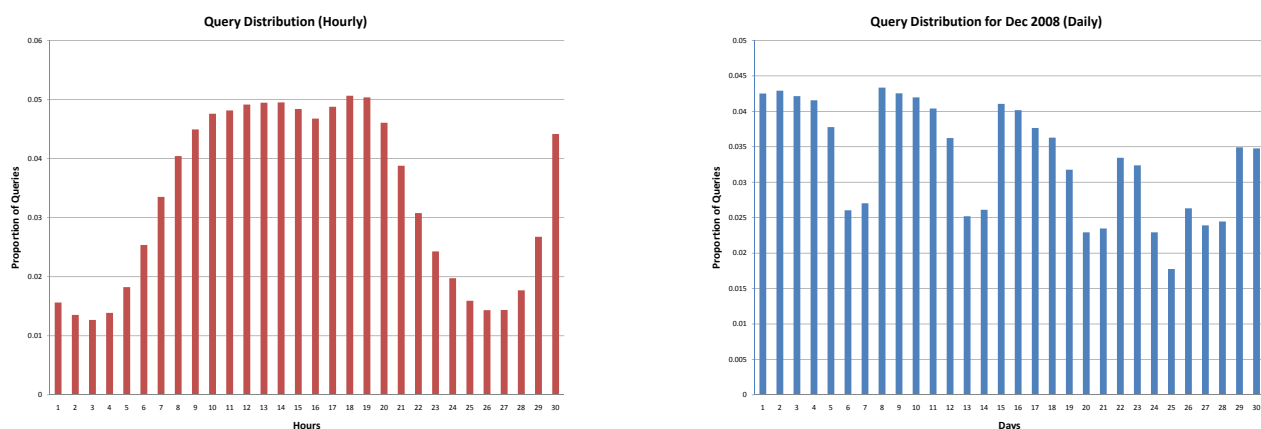


Figure 2: Query Distribution graph showing the proportion of queries for (a) hourly data (b) daily data.



Figure 3: Comparison of Query Match Rate and Click Volume Rate for (a) hourly data (b) daily data.

| Query         | Dec 22 |   | Dec 23 |   |
|---------------|--------|---|--------|---|
|               | Clicks | Url   | Clicks | Url   |
| caylee        | 17     | http www.msnbc.msn.comid28347850  | 12     | http www.postchronicle.comnewsoriginal article_212195043.shtml  |
| anthony       | 9      | http://www.foxnews.com/story/0,2933,470650,00.html  | 7      | http www.foxnews.comstory 0,2933,471664,00.html   |
| latest news   | 7      | http://www2.tbo.com/content/2008/dec/22/disturbed-evidence-caylee-anthony-site-could-hurt-news-metro/ | 7      | http://www.cfnews13.com/News/Local/2008/12/23/casey_cant_go_to_caylee39s_funeral.html                 |
|               | 7      | http://www.examiner.comSubjectCaylee_Anthony.html   | 5      | http www.abcnews.go.comTheLawstory?id6517089&page1  |
| cheap flights | 6      | http://www.cbsnews.com/stories/2008/12/22/national/main4681516.shtml                                  | 5      | http://www.orlandosentinel.com/news/local/breakingnews/orl-bk-anthony-memorial-122308,0,7555287.story |
|               | 326    | http://www.cheapflights.com/  | 291    | http://www.cheapflights.com/  |
|               | 22     | http://www.bookingbuddy.com/  | 34     | http://www.bookingbuddy.com/  |
|               | 15     | http://farecast.live.com/   | 28     | http://farecast.live.com/   |
| walmart       | 9      | http://www.kayak.com/   | 17     | http://www.kayak.com/   |
|               | 5      | http://www.flightnetwork.com/   | 5      | http://www.flightnetwork.com/   |
|               | 46688  | http://www.walmart.com/   | 41098  | http://www.walmart.com/   |
|               | 17     | http://en.wikipedia.org/wiki/Wal-Mart   | 15     | http://en.wikipedia.org/wiki/Wal-Mart   |
|               | 12     | http://walmartwatch.com/  | 6      | http://walmartwatch.com/  |

Table 1: Example Queries showing corresponding Urls and clicks across two timestamps

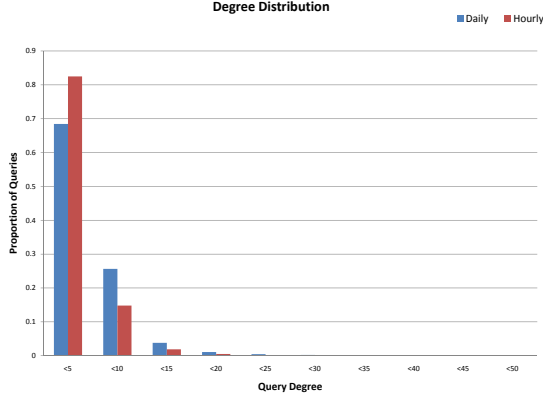


Figure 4: Degree Distribution for daily and hourly data.

#### 4.1 Rank Difference Score

For each pair of timestamps  $i$  and  $i + 1$ , and for a given query, we consider the top  $k$  pages by click rank. Obviously, a difference in the top-ranked page would be more significant than a difference in the lower ranked pages. Hence, we build a measure that provides greater penalty for differences at the top of the rank list. Let the top  $k$  pages at time  $i$  be given as  $P_{top_i}$  with  $R_i$  giving their click rank. For each page  $p$  in  $\{P_{top_i} \cup P_{top_{i+1}}\}$  we compute the Rank Difference Score as :

$$Score(p) = Diff(p) \times (k - \min(R_i(p), R_{i+1}(p))) \quad (2)$$

where

$$Diff(p) = \begin{cases} \frac{|R_i(p) - R_{i+1}(p)|}{2^x \times k} & R_i(p) \leq k \ \& \ R_{i+1}(p) \leq k \\ otherwise & \end{cases}$$

We can see from (1) that if the page  $p$  does not occur in the top  $k$  pages in the other timestamp, it is penalized with a score as :  $(2^x \times k) \times (k - \min(R_i(p), R_{i+1}(p)))$ . Also, queries have different page distributions. Some queries might not contain  $k$  pages at any interval (irrespective of what the value of  $k$  is chosen to be). Hence, we normalize the score by considering the worst score for the given page distribution, which is when all the pages are different across the two timestamps.

Thus the Rank Difference Score for a query  $q$  is given as:

$$RDS(q, i, i+1) = \frac{\sum_{p:|(q,p) \in (E_i \cup E_{i+1})} Score(p)}{2^x k (\sum_{m=1}^{|P_{top_i}|} (k - m) + \sum_{n=1}^{|P_{top_{i+1}}|} (k - n))} \quad (3)$$

To choose an appropriate value of  $k$ , we examined the degree distribution for the queries, shown in Figure 4. The figure shows the distribution of the proportion of node degrees. We can observe that 82.5% of the queries for the Hourly data, and 68% for the Daily data, have degrees less than or equal to 5. Hence, we chose  $k$  to be 5, and considered the top 5 ranked pages for each query for our analysis. We chose the exponent  $x$  to be 1. Higher values represent larger penalties for changes in rank.

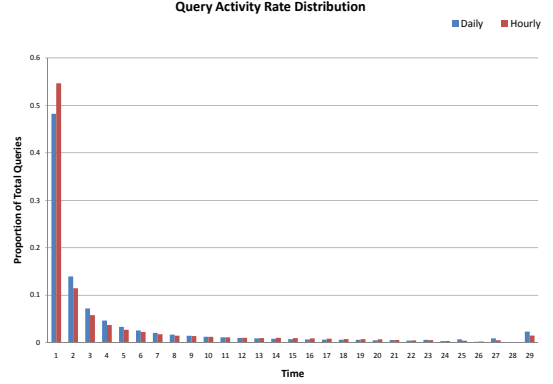


Figure 6: Activity Rate Distribution for Queries

Figure 5(a) and (b) show the histograms of RDS values for queries for the hourly and daily data respectively. We can see that, for most of the queries the score is less than 0.5, which would indicate that there is some similarity in the pages clicked on for these queries.

We can aggregate the RDS values for a query over all timestamps of the graph to measure different types of behavior.

Next, we consider the same 3 queries we illustrated in Section 1 (Figure 1). Now, we examine the relative RDS values for these queries over time, shown in Figure 7. We can observe that when we consider the RDS values, there is a considerable difference between the Adult query and the other two. The Adult query has typically high RDS values indicating high entropy, while the navigational query has low scores. Thus, the RDS score achieves a good separation in this case. We will discuss this more in the rest of this section, where we focus on the behavior of these 3 types of queries.

#### 4.2 Query Stability

A query can be said to be stable over time, if its page set does not change much over time. Note that, a low value of RDS for a query over time indicates high temporal stability for that query, since the corresponding webpages clicked do not change significantly over time. Stability can be formulated as:

$$Stability(q) = Act(q) - \sum_{i=1}^{T-1} RDS(q, i, i+1) \quad (4)$$

Here,  $Act(q)$  represents the number of interval-pairs that the query is common to. The Activity rate distribution for the queries is shown in Figure 6. We can observe that around 50% of the queries occur only once.

We computed stability for the queries over the Daily data. We considered only frequent queries ( $Act(q) > 25$ ) and computed the Stability score. Next, we will discuss how the Stability measure can be employed to study and distinguish between different types of queries - Navigational, Adult and News queries. We choose these three classes of queries for

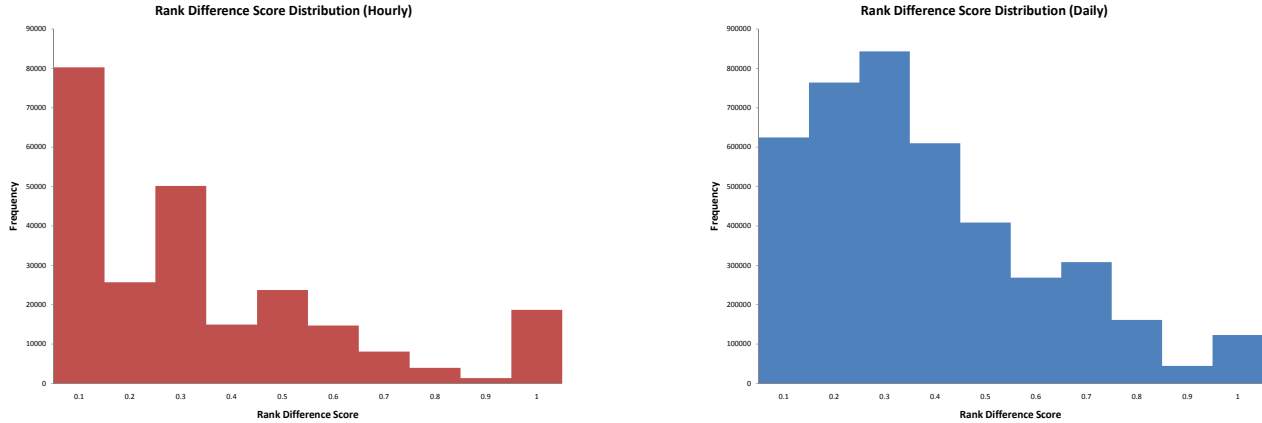


Figure 5: Rank Difference Score distribution for (a) hourly data (b) daily data.

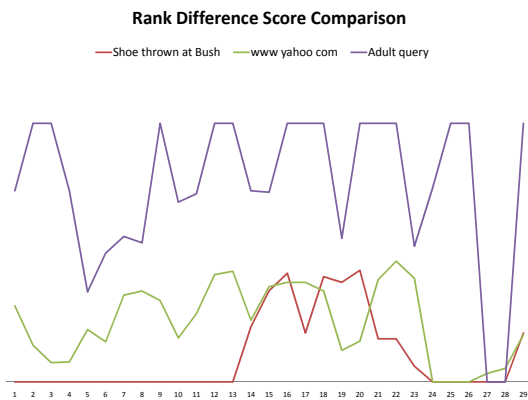


Figure 7: Query Trend Comparison using RDS values.

our analysis since they represent 3 different user-intent/document-entropy combinations. Navigational queries represent low intent, since the user is interested in getting to a particular page. The destination set entropy is low as we have seen from the example shown previously. In the case of adult queries, the intent has typically low variance, while the document result set has high entropy. News queries encompass many different user intents and have destination url-sets with high entropy. We will discuss these three query classes in detail in the rest of this section.

### 4.3 Navigational Queries

Navigational Queries are those where the user is interested in accessing a particular website and types either the name of the website or the description to access it. Examples of navigational queries include {walmart}, {google} and {microsoft.com}. According to research, approximately 18% of queries in Web search are navigational queries [11]. Therefore, correctly identifying navigational queries has a great potential to improve search performance.

From a temporal perspective, navigational queries are endowed with low entropy in the clicked set, which means they have stable temporal signatures. The top pages clicked for these queries are not likely to differ much over time. When we inspected the queries which were frequent and had high Stability values, we found that most of them were navigational queries, which are queries where the user wants to access a particular web-page. Out of the top 100 Stability values, 70 were navigational queries<sup>4</sup>. When we considered the top 200, 136 of them were Navigational queries. The remainder were mainly informational queries, which linked to articles such as Wikipedia. Only 7 out of the top 200 stable queries were Adult queries<sup>5</sup>. This result is unsurprising since adult queries can be expected to be characterized by reasonable entropy of webpages.

When we considered the hourly data, the results were even better. 97 out of the top 100 were navigational queries and 193 out of the top 200, giving a precision of 96.5. There were 76 queries having a stability value of 1. We have shown 25 of them in Table 2.

To evaluate the results, we also computed the entropy for the top 200 High Stability values in terms of their links.

$$Entropy(q) = \sum_{\forall u:|(q,u) \in E} p(q \rightarrow u) \log(p(q \rightarrow u)) \quad (5)$$

The results are shown in Figure 8. We can observe that 193 out of the 200 have entropy below 0.5.

### 4.4 Adult Queries

Adult queries are very common in web search. The expected behavior for these queries includes high volatility in the clicked pages. Among the queries inspected above in our analysis of high-stability values, we found that a small percentage were adult navigational queries. Next, we examined the queries with the lowest stability values to evaluate

<sup>4</sup>We evaluated the queries manually.

<sup>5</sup>Interestingly, 4 out of the 7 were navigational adult queries.

| Query                   | Stability |
|-------------------------|-----------|
| bestbuy.com             | 1         |
| aeropostale             | 1         |
| craigslist boston       | 1         |
| philly.com              | 1         |
| careerbuilder           | 1         |
| intel                   | 1         |
| remax                   | 1         |
| cingular                | 1         |
| www.wachovia.com        | 1         |
| komo 4 news             | 1         |
| weather channel.com     | 1         |
| xbox.com                | 1         |
| weather bug             | 1         |
| best buy                | 1         |
| career builder          | 1         |
| us magazine             | 1         |
| googlemaps              | 1         |
| www.nfl.com             | 1         |
| wunderground            | 1         |
| earthlink webmail       | 1         |
| usairways               | 1         |
| fandango                | 1         |
| whole foods             | 1         |
| www.walmartbenefits.com | 1         |
| watchtvsitcoms.com      | 1         |

Table 2: Top 25 Stable queries for Hourly data.

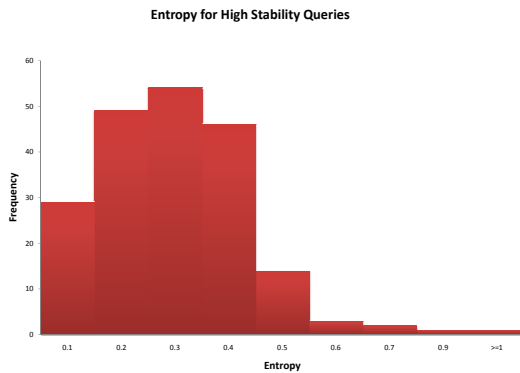


Figure 8: Distribution of entropy for top 200 stable queries.

the percentage of Adult queries. For the hourly data, we observed that around 60% of the low-stability values were Adult queries. For the daily data, when we used the same high frequency filter as for the high-stability we found 55% of the queries were Adult. The remainder were combinations of celebrity gossip and a few informational queries. This is consistent with our understanding of the temporal behavior of Adult queries, which involves consistently high entropy values.

## 4.5 News Queries

Next, we focus on identifying news queries. These include queries on celebrities in the news and real-world events such as political and sporting events. Such queries are hard to classify for most commercial search engines, due to the fact that they do not follow any obvious semantic structure. Temporally however, they can be associated with a particular signature. The intuitive expectation would be for them to occur over a short span of time, such as a week. And during that time, they would have high activity. The entropy would be higher than for navigational queries but not as high as Adult queries. Such queries would be more easily discernible using the daily data, as there is no particular hour in which they are more likely to occur. In the daily data space however, we can expect them to occur over chunks spread over a few days. Accordingly, the desiderata we considered for the extraction of news queries with high precision were :

- Frequency : As mentioned above, such queries occur infrequently globally in time but locally with high frequency. We chose queries that occur only in short periods of 3 to 7 successive days.
- Volume : We expect these queries to have high volume of clicks in the days that they occur, since news stories generate considerable interest among users. Infrequent queries with low click volume are likely to be random queries and not very useful. Accordingly, we used a click volume threshold and considered only queries that generated more than 200 clicks.
- Stability : We expect some degree of volatility in pages clicked for news queries but maybe not as much as Adult queries. To prune out the Adult queries, we considered queries having Average Stability (over the days active)  $< 0.75$ .

We used the above metrics to filter the queries and extracted the top 200 queries for evaluation. We discovered that a large percentage of them were news-related queries. Out of the top 100, 82 were directly related to news articles on those days. The rest comprised of holiday shopping deals from stores. We have shown the top 30 news queries in Table 3. In the table, we have also shown the range of the news story in December and the real-world event that it described.

All the queries that correspond to a particular news item can be extracted by performing a single step walk of the query click graph for the timestamps concerned. We begin with the extracted query and visit the urls that were clicked for it. From each of those urls we find all queries that pointed

| Query                          | Start | End | Clicks | Real-world event  |
|--------------------------------|-------|-----|--------|---|
| melissa britos                 | 4     | 10  | 341    | Alex Rodriguez Cheating on Madonna with Model Melissa Britos?   |
| burger king fragrance          | 17    | 23  | 501    | Burger King Launches Perfume  |
| o. j. simpson                  | 3     | 9   | 227    | O.J. Simpson sentenced to prison in armed robbery case  |
| doug marrone                   | 10    | 16  | 234    | Syracuse hires Doug Marrone as football coach   |
| anna hansen lance armstrong    | 23    | 29  | 288    | Lance Armstrong and Anna Hansen Expecting a Baby  |
| madoff suicide                 | 23    | 29  | 324    | Bernard Madoff found dead Tuesday after committing suicide  |
| anna hansen                    | 23    | 28  | 497    | Lance Armstrong and Anna Hansen Expecting a Baby  |
| shoe thrown at president       | 14    | 18  | 211    | Muntadhar al-Zeidi throws a shoe at President George W. Bush during a news conference Sunday in Baghdad |
| paul weyrich                   | 18    | 22  | 247    | Conservative activist Paul Weyrich, who coined the phrase "moral majority," died                        |
| robert mulligan                | 20    | 25  | 430    | Robert Mulligan died  |
| adolf hitler cake              | 16    | 21  | 322    | Shop refused to write a child's name (Adolf Hitler) on a birthday cake                                  |
| seattle bus                    | 16    | 22  | 328    | Bus crash in seattle  |
| dakota culkin photo            | 11    | 16  | 241    | Dakota Culkin killed after struck by car.   |
| bush shoe incident             | 14    | 18  | 230    | Muntadhar al-Zeidi throws a shoe at President George W. Bush during a news conference Sunday in Baghdad |
| lpga q school                  | 2     | 7   | 394    | LPGA Qualifying School begins   |
| shinseki                       | 6     | 10  | 464    | Barack Obama introduced retired Army Gen. Eric K. Shinseki as Veterans Affairs Secretary-designate.     |
| dakota caulkin                 | 11    | 17  | 886    | Dakota Culkin killed after struck by car.   |
| microsoft patch                | 16    | 21  | 552    | Microsoft to patch security hole in Internet Explorer   |
| shoes thrown at president bush | 14    | 18  | 262    | Muntadhar al-Zeidi throws a shoe at President George W. Bush during a news conference Sunday in Baghdad |
| santa shooting                 | 25    | 29  | 389    | Man in Santa Suit Kills 8, Self on Christmas Eve  |
| andy kennedy                   | 18    | 22  | 365    | Mississippi coach Andy Kennedy arrested for punching cab driver   |
| death map                      | 16    | 20  | 208    | U.S. natural hazard death map is produced   |
| f18 crash san diego            | 8     | 11  | 329    | F18 crashes into San Diego homes  |
| hemlock semiconductor          | 12    | 18  | 230    | Hemlock Semiconductor Corp announced 1 billion expansion  |
| ponzi scheme                   | 12    | 18  | 367    | Veteran Wall Street broker arrested over giant ponzi scheme   |
| adam walsh murder case         | 16    | 22  | 831    | Adam Walsh Murder Case Closed   |
| bruce pardo                    | 25    | 29  | 631    | santa shooter   |
| ms08-078                       | 17    | 23  | 1651   | Microsoft patch   |
| aiko                           | 10    | 16  | 295    | Scientist Builds Female Android Aiko  |
| anna hansen                    | 23    | 29  | 19664  | Lance Armstrong and Anna Hansen Expecting a Baby  |

Table 3: Top 30 news queries for Dec 2008.

| Query                                      |
|--|
| bush ducks shoe                            |
| bush shoe attack                           |
| bush shoe video                            |
| president bush getting shoe thrown at him  |
| president bush getting shoes thrown at him |
| president bush shoe                        |
| president bush shoe attack                 |
| president bush shoe video                  |
| president bush, shoe                       |
| president shoe attack                      |
| shoe thrown at bush video                  |
| shoe thrown at president bush              |

**Table 4: Queries corresponding to the shoe incident having number of clicks greater than 5 for the time period.**

| Classifier | Acc % | F-Measure Nav | F-Measure Adult | AUC   |
|------------|-------|---------------|-----------------|-------|
| J48        | 80.31 | 0.828         | 0.769           | 0.847 |
| Bagging    | 81    | 0.834         | 0.776           | 0.881 |

**Table 5: Navigational vs Adult**

to that page in the time period of the news story and extract all those queries. The queries for the shoe incident having clicks  $> 5$  are shown in Table 4.

## 4.6 Query Classification

We have shown that temporal information can be useful in differentiating queries. We now present some preliminary experiments in using temporal features for query classification. We consider the three types of queries discussed previously and attempt to develop binary classifiers using temporal features. We use the Weka classification package [15] for the experiments performed.

### 4.6.1 Adult vs Navigational

As we have observed previously, adult queries are characterized by high entropy (low stability) across time intervals. On the other hand, navigational queries mostly refer to the same pages, which reduces the entropy of their clicked pages. We consider a feature space of size 10, based on the RDS measure. Each entry corresponds to a range of 0.1 for the RDS measure and represents the frequency of that score range. For instance, the first feature represents the proportion of the time intervals the query is active that it receives an RDS score between 0 and 0.1. For the navigational queries, we would expect the values of the first 5 features to be higher than the latter 5, whereas the reverse is expected for the Adult queries.

Table 5 shows the results for the application of two classifiers - J48 Decision Tree and Bagging using the temporal features and performing 10-fold cross-validation. From the results, we find that the temporal features used can provide reasonable separation and we achieve AUC values of 0.847 and 0.881 for the Decision Tree and Bagging classifiers.

### 4.6.2 News vs Adult

| Classifier | Acc % | F-Measure News | F-Measure Adult | AUC   |
|------------|-------|----------------|-----------------|-------|
| J48        | 84.06 | 0.678          | 0.894           | 0.84  |
| Bagging    | 85.3  | 0.69           | 0.90            | 0.884 |

**Table 6: News vs Adult**

| Classifier | Acc % | F-Measure News | F-Measure Nav | AUC   |
|------------|-------|----------------|---------------|-------|
| J48        | 90.23 | 0.626          | 0.944         | 0.858 |
| Bagging    | 90.53 | 0.64           | 0.946         | 0.882 |

**Table 7: News vs Navigational**

In this experiment we compare news and adult queries. For this purpose, we considered a feature vector of size 4, based on the criteria discussed in the previous subsection. The features we considered were *Activity*, *Number of consecutive sequences*, *Average RDS* and *Average Click Volume*. For news queries, we would expect the activity to be low and the click volume high for the active time intervals. We used the data that we filtered out in the previous subsection along with the queries extracted by the graph walk from these nodes to train the classifiers. The results are shown in Table 6. The two types of classifiers constructed result in AUC values of 0.84 and 0.884. We find that the F-measure is higher for Adult rather than news. This is due to the fact that the recall for the news queries is low. For many news-related queries, the entropy in RDS values is quite high which is hard for the classifier to discriminate from Adult queries, leading them to be mis-classified as adult. This is due to the fact that there is a high degree of entropy in the exact phrases used by people for news. A few queries are commonly used but the rest are infrequently used. It is for this reason news query classification is a difficult task.

### 4.6.3 News vs Navigational

We use the same features discussed above to classify news against navigational. The results are shown in Table 7. We can observe that the partitioning is better than in the experiment above with better AUC values of 0.858 and 0.882 for J48 and Bagging respectively. This is due to the fact that adult queries have higher variance in their temporal behavior than navigational. Note that, in practice, navigational queries are easy to classify due to their low entropy in selected documents. Here, we demonstrate the ease of classifying them using temporal stability.

### 4.6.4 News vs Non-news

Next, we conduct an experiment evaluating the performance of our temporal features in classifying news queries given the entire query distribution. We chose several random samples from the rest of the queries to account for the non-news examples. We used the same two classifiers as in the previ-

| Classifier | Acc % | F-Measure Non-news | F-Measure News | AUC  |
|------------|-------|--------------------|----------------|------|
| J48        | 77.4  | 0.85               | 0.544          | 0.72 |
| Bagging    | 77.65 | 0.852              | 0.543          | 0.75 |

**Table 8: News vs Non-news**

ous experiment, Decision Trees and Bagging from the Weka toolkit. The results are summarized in Table 8. The task of classifying news from non-news is obviously harder and the results show lower F-Measure scores for the news class. Once again, the recall for the news queries are low contributing to the low F-Measure value. Note that, in these experiments, we are using only temporal statistics for classification. The low recall indicates that although these temporal patterns can help extract queries of a particular type, these queries do not necessarily correspond to the entire set of queries of that type. Hence, we believe that temporal patterns can be useful to augment models built with other forms of meta-data. And we expect this to be particularly beneficial for classifying news queries, which are very hard to classify, as we have mentioned before. We intend to perform this evaluation as part of our future work.

## 5. CONCLUSIONS AND FUTURE WORK

Previous research on web query categorization has focused purely on using semantic information, which is low for web queries due to their typically short lengths. To overcome this, they have considered using taxonomies and meta-data to build models for classification. The temporal behavior of web queries has been virtually untapped for this problem. In this paper, we have examined the benefits of temporal analysis in unearthing interesting characteristics for different types of web queries. In particular, we have focused on extracting temporal signatures for three types of queries - navigational, adult and news. We have developed a Rank Difference Score measure which can be used to characterize changes in query behavior over time. From our preliminary analysis on using temporal features for classification, we have observed good precision in extracting navigational, adult and news queries. This indicates that temporal features can indeed be beneficial for query classification. In the future, we would like to extend our work by using temporal features to augment existing classification models for automatic categorization of web queries. The query examples that we have extracted can be useful as training data for such classifiers. Finally, our analysis has been offline thus far, but we are investigating approaches to move to an online mode.

## 6. REFERENCES

- [1] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, and David Grossman. Using titles and category names from editor-driven taxonomies for automatic evaluation. *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 17–23, 2003.
- [2] Steven M. Beitzel, Eric C. Jensen, Abdur Chowdhury, David Grossman, and Ophir Frieder. Hourly analysis of a very large topically categorized web query log. *SIGIR '04: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 321–328, 2004.
- [3] Steven M. Beitzel, Eric C. Jensen, Ophir Frieder, David Grossman, David D. Lewis, Abdur Chowdhury, and Aleksandr Kolcz. Automatic web query classification using labeled and unlabeled training data. pages 581–582, New York, NY, USA, 2005. ACM.
- [4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, pages 107–117, 1998.
- [5] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.
- [6] Andrei Z. Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. Robust classification of rare queries using web knowledge. *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 231–238, 2007.
- [7] Ling Chen, Yiqun Hu, and Wolfgang Nejdl. Using subspace analysis for event detection from web click-through data. *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1067–1068, 2008.
- [8] Steve Chien and Nicole Immorlica. Semantic similarity between search engine queries using temporal correlation. In *Proc. of the 14th Int'l World Wide Web Conference*, pages 2–11, 2005.
- [9] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.
- [10] Xiao Li, Ye-Yi Wang, and Alex Acero. Learning query intent from regularized click graphs. *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 339–346, 2008.
- [11] Yumao Lu, Fuchun Peng, Xin Li, and Nawaaz Ahmed. Coupling feature selection and machine learning methods for navigational query identification. *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 682–689, 2006.
- [12] William J. Reed. The pareto, zipf and other power laws. *Economics Letters*, 74(1):15–19, December 2001.
- [13] Dou Shen, Jian-Tao Sun, Qiang Yang, and Zheng Chen. Building bridges for web query classification. *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 131–138, 2006.
- [14] Michail Vlachos. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 131–142, 2004.
- [15] I. H. Witten and E. Frank. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, 31(1):76–77, 2002.
- [16] Qiankun Zhao, Steven C. H. Hoi, Tie-Yan Liu, Sourav S. Bhowmick, Michael R. Lyu, and Wei-Ying Ma. Time-dependent semantic similarity measure of queries using historical click-through data. *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 543–552, 2006.
- [17] Qiankun Zhao, Tie-Yan Liu, Sourav S. Bhowmick, and Wei-Ying Ma. Event detection from evolution of click-through data. *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 484–493, 2006.