Web Intelligence and Agent Systems: An International Journal 11 (2013) 107–124
DOI 10.3233/WIA-130265
IOS Press

Investigating query bursts in a web search engine

Ilija Subašić [a,*] and Carlos Castillo [b]

[a] *Department of Computer Science, K.U. Leuven, Celestijenlaan 200a, 3001, Leuven-Heverlee, Belgium*
E-mail: ilija.subasic@cs.kuleuven.be
[b] *Yahoo! Research, Diagonal 177 8th floor, 08018 Barcelona, Catalonia, Spain*
E-mail: chato@acm.org

Abstract. The Internet has become for many the most important medium for staying informed about current news events. Some events cause heightened interest on a topic, which in turn yields a higher frequency of the search queries related to it. These queries are going through a "query burst". In this paper we examine the behavior of search engine users during a query burst, compared to before and after the burst. We are interested in how this behavior changes, and how it affects other stake-holders in web search.

We analyze one year of web-search and news-search logs, looking at query bursts from multiple perspectives. First, we adopt the perspective of search engine users, describing changes in their effort and interest while searching. Second, we adopt the perspective of news providers by comparing web search and news search query bursts. Third, we look at the burst from the perspective of content providers.

We study the conditions under which content providers can "ride" a wave of increased interest on a topic, and obtain a share of the user's increased attention. We do so by identifying the class of queries that can be considered as an opportunity for content providers that are "late-comers" for a query, in the sense of not being among the first to write about its topic. We also present a simple model for predicting the click share content providers could obtain if they decide to provide content about these queries.

Keywords: Query log analysis, burstiness, news analysis

1. Introduction

To stay in touch with current events people use a variety of sources including television, the Internet, radio, newspapers, etc. On a given day a person typically uses more than one source [25]. Among these sources, television continues to be the most important one. However, since 2008, for the general public in the U.S., the Internet has been more important as a source of news than newspapers, and the most important news source among people under the age of 30 [24]; by 2010 the Internet was the source of news for 61% of users [25] with a rising trend. As search engines are one of the primary tools for online news discovery and access, analyzing their query logs can answer many questions about how people inform themselves.

Users express a heightened interest in queries related to current events. This leads to sharp increases in frequencies of these queries in web search query logs. For instance, on October 18, 2008, after being parodied several times in the TV show Saturday Night Live, U.S. politician Sarah Palin appeared in the show and met her impersonator, comedian Tina Fey. On that day, the frequency of the query "`snl sarah palin`" was 22 times larger than two days before the event. This is referred to as a query burst [19].

From an economic perspective, this higher attention on a topic, quantified as query frequency, can be regarded as an increase in the "demand" for an informational good. The "supply" that can satisfy this demand are the documents that are relevant to the query

*Corresponding author.

topic. An increase in the demand generates an increase in the "price" users pay for accessing the information (quantified as the effort they spend searching). This is later matched by an increase in the supply of the informational good, as content providers notice the information need and write about the topic. Following this marketplace metaphor, we can measure the market share of the content providers with the number of clicks their contents receive.

During query bursts we know that demand increases. In this paper we investigate how are other components of this "marketplace" affected by the query bursts, motivated by the following set of questions:

- *goods:* What are the types of bursty information?
- *price:* Does the effort users spend change during the burst?
- *supply:* How do bursts affect the production of documents?
- *market share:* How are clicks distributed over the created documents?

We set these high-level research questions to encompass our motivation in investigating query bursts, and further develop them into a number of more specific research questions. In addition, we investigate the origin of the bursts and their relations to actual news events.

In our research, we first detect query bursts, and then go beyond detection into characterizing their effects on the users of search engines. We also realize that not all query bursts are related to what would be considered a newsworthy event by traditional news outlets. To account for this we compare searches in a news portal with general web searches.

Next we look at query bursts from the users' perspective, with the aim of uncovering how higher interest in a query changes user behavior. We are particularly interested in what happens before and after a query burst. To investigate this, we analyze several metrics that describe the effort and attention of users while searching for bursty queries.

Contributions This study contributes to the understanding of the effect of query bursts on web search results by observing that:

- Query bursts can be grouped in classes having distinctive properties.
- During a query burst, not only query frequency, but per-query user effort is higher according to several metrics. At the same time, clicks on query results tend to be more concentrated at the top documents for each search.
- The same query has a higher burst intensity and shorter duration on a news search log than on a general web search log.
- After a query burst, the distribution of clicks among search results is substantially different from that before the burst.
- Publishing early represents a clear advantage for content providers, and for some queries this advantage is unsurmountable; for other queries, a late-comer indeed has the opportunity of obtaining a non-trivial part of the users' attention.

The analysis of user activity logs is a type of field study, and methodologically there are advantages and disadvantages to this approach. In particular, there are many variables that we can neither observe nor infer accurately. We recognize this limitation, and support our findings through careful comparison of multiple independent metrics.

Roadmap The next section describes previous work on temporal aspects of web usage mining. Section 3 formally defines the concepts we use. Section 4 describes in detail our experimental setting, sampling methods and metrics. Section 5 presents a characterization of query bursts based on search logs analysis. Section 6 models changes in click share before, during, and after the query bursts. Finally, Section 7 presents our conclusions.

2. Previous work

Query-log analysis is a research topic that has received substantial attention in the last few years, with even entire venues devoted to the topic, such as the *Web Search Click Data* and the *Query Log Analysis* workshops. Since the early studies of query logs presented in [17,21,28], the field has branched out into several areas, and our coverage of them in this brief section is by no means complete.

Query categories User behavior while searching for different content categories has been studied using different notions of categories and different methods for assigning queries to categories. The analysis of an hourly time series in [4] and a long-term time series in [3] showed the distinct properties in the frequency profile for queries relevant to different editor-assigned topical categories. Conversely, the authors of [2] study

whether the different frequency profiles of queries can be used to improve query classification. In [10,11] instead of topical categories, authors look for differences between common (high frequency) and rare (low frequency) queries. In this study, we do not categorize general queries but only bursty ones, and our categories are based on multiple factors which are neither topics nor overall frequencies.

Temporal query analysis A related study by Adar et al. [1] compared time series from originating different sources. The study resulted in a description of different classes of temporal correlation and a visual tool for summarizing them. Previously, using correlation between query frequency time series [7] uncovered semantic similarity between time-aligned series. Time-based query similarity discovery using clustering of a bipartite graph of queries and pages is described in [35]. In [31] Sun et al. present a method for uncovering possible causal relationships between queries. In contrast to previous work, our paper focuses on differences in user behavior before and after a certain disruptive event, and compares it to user behavior on randomly-chosen queries and on queries that are stable over time.

Our research over a one year period can be considered long-term with respect to a majority of works on query-log mining. Query logs of this length have been shown useful for learning about changes and trends in user interest [27].

Query bursts Burst analysis includes methods for detecting queries currently in a period of increased user interest. In [32], query bursts are detected as outliers in the query frequency series, specifically as moments at which a query shows 1.5 to 2 standard deviations times higher frequency than its average in previous periods. In [27], increases in normalized query frequency are used to discover query bursts; this is the method we use in this paper and impose other constraints to the detection of query bursts (such as having a single burst during a one year period) increasing precision at the expense of recall.

One of the main applications of query-burst detection has been the detection of real-world events, as in [6,36]. One particularly interesting usage of this data is to epidemiology for instance to track the spread of flu [14]. In recent years, several tools that allow for the tracking and comparison of query frequencies have been developed [13,15,34].

Studying evolution of documents There has been a substantial amount of research on the detection and evolution of term bursts in text corpora. Many of these works are based on [19]. Burstiness has been explored with respect to various domains and phenomena including so-called "buzz" in text and news streams in [12,29,33]. In particular, blogs are analyzed in [20], while the method presented in [22] is applied to both blogs and traditional news outlets.

Some of the results presented in this paper appeared in summary form in [30]. We extend this work in several ways and: (a) present deeper background and motivation for this research, (b) widen the scope and interpretation of the initial results, and (c) introduce a new analysis of differences between bursts in news search and general web search, exploring how users search for bursty information using a specialized news search engine, as opposed to a more general web search engine.

3. Preliminaries and notation

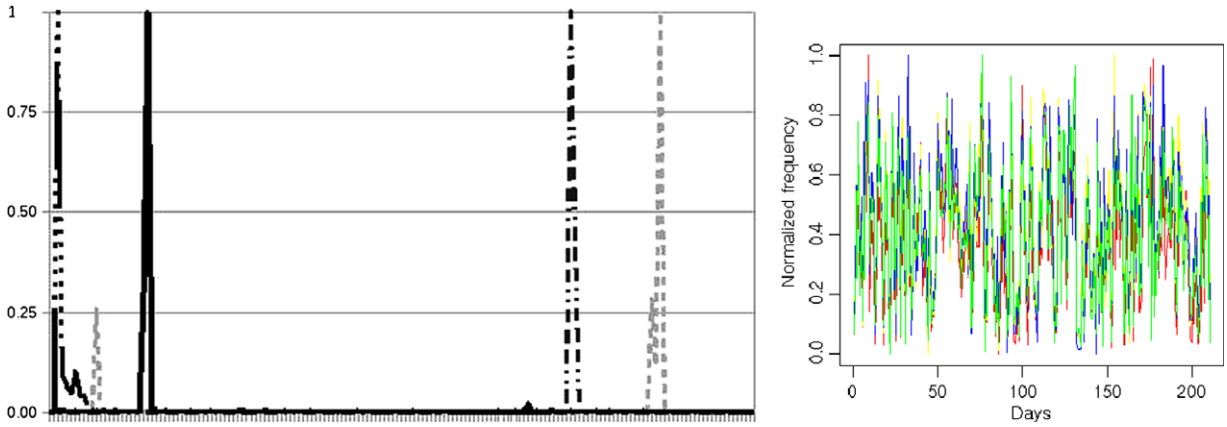
This section introduces some concepts and the notation that is used in the rest of the paper.

3.1. Query bursts

There is no standard or widely-accepted test for query burst detection. This largely depends on the application for which the test is developed. In the case of this paper, we are interested in precisely identifying query bursts. Therefore, we define our burst measure to be precision-oriented, and include the queries which are clear outliers from a stable frequency, possibly at the expense of missing some query bursts that are not so pronounced.

Specifically, we apply a burst measure based on normalized lift in query probability. This measure has been used for discovering bursty queries in query logs [27] as well as bursty keywords in news documents [29]. We impose a large increase in frequency, and the property of having a single distinctive burst during the one-year observation period. In practice and with the parameter setting we use, this turns out to be more restrictive than the test shown in [32]. As a consequence, the query bursts we sample are very clear (some examples are in Fig. 1) and would be detected as bursts by any reasonable test.

Query burstiness Let Q be the set of all queries. Let $T = \{t_0, t_1, \ldots, t_{|T|-1}\}$ be the set of observation periods, in which each period represents an interval of



(a) Bursty queries (b) Stable queries

Fig. 1. Examples of bursty and stable queries time series; x-axis is time in days, y-axis is normalized frequency (thus, the large variation for stable queries).

time. In this study, each $t \in T$ corresponds to one day. Let $f : (Q \times T) \to \mathbb{N}$ be such that $f(q,t)$ is the number of occurrences of query q in the period t.

For each query q and period t we derive a BURST INTENSITY index $b(q,t)$ which tells us how "bursty" this query is in that period, by measuring its relative increase in frequency compared to the past. This is obtained by computing:

$$b(q,t) = \frac{\frac{f(q,t)}{\sum_{q \in Q} f(q,t)}}{\frac{\sum_{u \le t} f(q,u)}{\sum_{q \in Q} \sum_{u \le t} f(q,u)}}. \quad (1)$$

Whenever $b(q,t) \geq \beta \sum_{u \in T} b(q,u)/|T|$, we say that the query q is going through a query burst at time t. If a query has no bursty period, we say that the query is *non-bursty*.

If the query has bursty periods that are not contiguous, we say the query is *bursty during multiple episodes*. If all the periods in which the query is bursty are contiguous, we say that the query is *bursty during a single episode*.

In the following, we refer to a sample of bursty queries during a single episode as the BURSTY queries. We also built a sample of queries having a very small variation of $b(q,t)$ in the observed series. In the following we refer to this sample as the STABLE queries. Figure 1 shows some of the queries from both sampled subsets. The parameters for this specific sample are presented in Section 4.

These samples represent extremes; most of the queries are neither STABLE nor BURSTY, therefore for some experiments we introduce a third sample of RANDOM queries chosen uniformly at random, having at least K appearances during the year.

3.2. Pre-episode, episode, and post-episode

For each query that is *bursty during a single episode*, i.e. in the BURSTY sample, we let $E_q = \{s_q, s_q + 1, s_q + 2, \ldots, s_q + d_q - 1\}$ be the set of consecutive periods in T where the query is undergoing a query burst. We name s_q the *start* of the episode, and d_q the *duration* of the episode. In our experiments we select only queries having a minimum duration $d_q \geq \delta$.

We also obtain time intervals before and after the episode for comparison, and refer to them as *pre-episode* and *post-episode*. These time intervals are obtained in such a way that they (i) are not too close to the episode, and (ii) comprise a number of occurrences of a query that is at most the occurrences of the query in the episode.

Formally, the pre-episode of a query ends at the time period $s_q - d_q$, and starts at a time $\mathrm{pre}(q)$ such that

$$\sum_{\mathrm{pre}(q) \le t \le s_q - d_q} f(q,t) \approx \sum_{t \in E_q} f(q,t) \quad (2)$$

in which the approximation is due to the time granularity of one day, so we approximate $\mathrm{pre}(q)$ to the nearest possible whole day. If there are not enough query occurrences before the episode, we set $\mathrm{pre}(q) = t_0$. We do the same for the post-episode period, starting at $s_q + 2d_q$ and ending at $\mathrm{post}(q)$ so that the total frequency during the post-episode period is at most

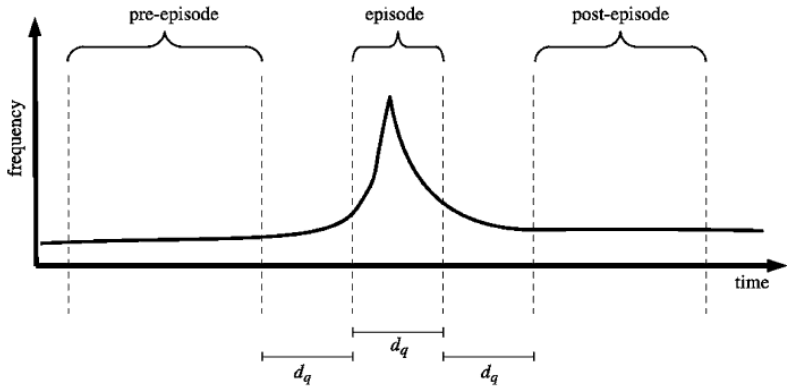


Fig. 2. Depiction of pre-episode, episode, and post-episode.

the total frequency during the episode. If there are not enough query occurrences, we set $\mathsf{post}(q) = t_{|T|-1}$.

Figure 2 depicts graphically the relationship between pre-episode, episode, and post-episode.

3.3. Pseudo-episodes

For some experiments we want to study whether a phenomenon is related to the bursty nature of the query or not. In the case of STABLE and RANDOM queries, we create *pseudo-episodes* that have the same query volume as the episodes of BURSTY queries, but usually have a longer duration.

Specifically, for each of the queries in these samples, we select a starting date uniformly at random (leaving the first three and the last three months out), then pick the volume of queries in the *pseudo-episode* according to the distribution of query volume in the episodes of the BURSTY sample. The pre- and post-episode periods are created in the same manner as for the bursty queries.

We decided to sample based on volume, rather than time. This certainly introduces a time bias on our results, and we can not say how much the different lengths of time periods affect our results. However, for most of our analysis we needed to have approximately the same query volume during, before, and after a burst, making volume-based sampling a reasonable choice. Due to the short length of the bursts, we observed that time-based sampling would have produced samples of largely disproportional query volumes, and for our analysis we regard time-bias as having less effect than volume-bias would have if we employed time-based sampling.

4. Experimental framework

4.1. Dataset and sampling

We processed an in-house query log[1] to obtain one year of web searches originating in the US.

The activity of each user in the query log is first divided into logically-coherent groups of queries, using the method in [5]. In the following, when we refer to *sessions* we always mean groups of related queries, known in the literature as query chains [26] or search missions [18].

From this log we sampled three subsets, the BURSTY queries subset, the STABLE queries subset and the RANDOM query subset.

4.1.1. Selecting bursty queries

Given the huge amount of data to process, we used an iterative process in which we started by sampling uniformly at random sessions that contained at least one of 1,400 "torso" queries (having frequencies that were neither too low nor too high), and continued by rounds – deepening (sampling more sessions) and narrowing (sampling less queries) our sample. The process was completed with a full sample of all the user sessions during 13 months containing 190 queries that are *bursty during a single period*. In our experiments we set $\beta \geq 3.5$, meaning that the $b(q,t)$ index must be 3.5 times larger than the average. We also set $\delta \geq 3$, meaning that the duration of the single episode must be of 3 days or more. Figure 1(a) shows the normalized frequency of a few queries in this sample.

[1] http://search.yahoo.com/

4.1.2. Selecting stable and random queries

For the STABLE set, we set the maximum standard deviation of b to 0.5 during the entire year, obtaining a set of 768 stable queries candidates, and subsampled 200 queries from this set using the empirical frequency distribution from our BURSTY sample. Figure 1(b) shows the normalized frequency of queries in this sample.

To select the RANDOM queries we first binned the bursty queries based on their frequency during the episode. Then from each bin we randomly selected queries having a one year frequency at most three times larger to ensure that pseudo-episodes have complete pre-episodes and post-episodes periods. Using this process we created a sample of 340 queries.

4.2. Metrics

To characterize the queries we chose to use a broad set of different metrics that covers different aspects of the search. The first three groups are computed for each particular period (pre-episode, episode, and post-episode), while the last group is computed for the entire time series.

- *Activity/effort metrics* capture in general how much effort users invest in locating information.
- *Attention metrics* show the concentration of user clicks.
- *Comparative metrics* compare the behavior of users between two periods.
- *Global metrics* include general properties of the query being analyzed.

4.2.1. Activity/effort metrics

The first group of metrics captures the users' effort in finding the information they sought. Most of these metrics are session-level, in which a session is a set of related queries obtained using the method in [5].

For a given query q, these metrics include:

- SESSION DURATION: average duration in seconds of sessions containing q, this is the time from the first query in the session to the last query (or click on a search result).
- DWELL-TIME: average time in seconds from an occurrence of q to the next query done by the user, limited to 30 minutes.
- QUERIES/SESS.: average number of queries in sessions containing q.
- CLICKS/SESS.: average number of clicks on search results in sessions containing q.
- EVENTS/SESS.: average number of events per session, including queries, clicks on search results, and clicks on the pagination links "previous-page/next-page".
- CLICKS/QUERY: number of clicks on search results, on average, after a query q and before the next query in each the session.
- NON-CLICKS %: fraction of issued queries that are not followed by a click on a search result (either because none of the results was relevant, or because the user found the information directly in the document snippets shown in the search results).
- ASSISTANCE %: fraction of query reformulations that were the result of a search suggestion. Most search engines display for some queries a few suggested queries, usually with a label such as "also try" or "related searches". This variable measures how often, when doing a reformulation, users click on one of these suggestions instead of typing a new query themselves.
- USERS/QUERY: number of distinct users issuing q, divided by number of occurrences of q. A small number indicates that a small group of users is repeatedly issuing the same query. A large number indicates that the query is of interest to a larger audience.

4.2.2. Attention metrics

The second group corresponds to a variety of metrics describing how concentrated or dispersed users clicks are on the search results. For a particular period (episode or pre/post-episode) and a specific query, we sort the URLs clicked for that query during the period in decreasing order according to the observed click probability. In the following, the "top URL(s)" for a period are the most clicked search results. This usually, but not always, matches the ordering in which URLs are shown to users, because of positional bias [9]. These metrics include:

- DISTINCT URLS: number of distinct search results clicked.
- TOP-1 SHARE: fraction of clicks on the search result with the highest number of clicks for a query. For example, if a query q appears ten times in the query log, and the highest clicked-on returned page has six clicks for query q, then the TOP-1 SHARE is 60%.
- URLS 90%: minimum number of search results required to cover 90% of users' clicks.

– RANK-CLICK DROP: steepness of rank-click frequency curve, measured by the exponent resulting of fitting a power-law to the curve of click probability.
– CLICK ENTROPY: entropy of the distribution of clicks on search results, as used in [23], for every query q and a set of clicked results U. This is defined as:

$$H(q) = \sum_{url \in U_q} p(url|q) \times \log p(url|q). \quad (3)$$

The first three attention metrics are straightforward and calculated directly from the query log, while the last two are slightly more complex and encompass the full click share distribution. The motivation for using RANK-CLICK DROP as a measure of attention is in the long-tailed nature of clicked distribution. If we fit a power law function of a form $y = ax^{-\alpha} +$ to the click distribution, the value of the (positive) parameter α suggests the steepness of a power law curve. The steeper a curve is the head of the distribution has more clicks, and therefore we can say that users attention is focused on a section of the results. Similarly to this, CLICK ENTROPY tells us how much information bits of a query a URL "carries". It has previously been used for measuring how difficult it is to satisfy the information need behind a query [23]. A higher CLICK ENTROPY indicates more disperse clicking (users click on more different documents) suggesting a more complex search, since users need to read more documents in order to satisfy their information need. The converse is also assumed: lower entropy indicates that users click on a smaller subset of the search results, suggesting that their information need is somehow easier to satisfy.

4.2.3. Comparative metrics

The third group of metrics compares different periods of time (e.g.: pre-episode and post-episode), focusing on the changes in their click probability distributions. The goal of these metrics is to discover the impact query burst have on the share of users' attention received by different search results.

– CLICK DIVERGENCE: KL-divergence[2] of click distributions. For a query q, a set of URLs U, and two periods, t_1, t_2, the KL-divergence is defined as:

$$D_q\ t_1||t_2 = \sum_{url \in U} P(url|q, t_1) \times \log \frac{P(url|q, t_1)}{P(url|q, t_2)}. \quad (4)$$

– TOP-1 CHANGE: difference in the probability of the URL with the highest click share in the first period with respect to the second period.
– TOP-N OVERLAP: overlap of URLs sorted by click share, at position $n = 1$ and $n = 5$, between the two periods.

We also considered variations in the activity/effort and attention metrics, e.g.: differences in DISTINCT URLS.

4.2.4. Global metrics

The fourth group of metrics considers the entire time-series:

– PEAK BUILD-UP RATIO: for a URL u, this is the difference between the episode peak, and the first date in which u is seen. This is normalized using the difference between the episode peak and the start of the dataset.
For instance, a value of 1 indicates the URL has existed since the beginning of the observation period, and a 0 indicates it was created the day of the peak of the query burst. Other cases are simply linearly interpolated, as described in Section 6.2.
– BURST INTENSITY: the b index described in Section 3.1.

5. Characterizing query bursts

The broad variety of topics that are covered by bursty queries (as can be seen in the Appendix A), suggest that the nature of the underlying events which caused the bursts, and the way they develop, are also different. We wish to discover the different patterns of query bursts based on user search behavior during these bursts. Apart from topical categories of queries, we would expect differences between query bursts related to new entities, e.g.: a criminal case involving a previously not-well-known person; and query bursts related to existing entities, e.g.: a new movie by a known director. We would also expect differences between query bursts occurring periodically, e.g. every year, and query bursts occurring for the first time.

Our main goal is a descriptive analysis of bursty queries, with the goal of discovering features that point to different classes of bursty queries. Therefore, the

[2] Kullback-Leibler divergence.

first application of the metrics described in Section 4.2 is to the characterization of different types of query bursts. Since there is no ground truth for this type of classification, we choose to discover different types of bursts using an unsupervised approach. For this we apply k-means clustering algorithm using all extracted metrics as the input features.

We experimented varying the number of clusters from two to 30 and found no clear evidence of an inherent number of clusters in the data (e.g.: looking at the sum of distance square from clusters centroids, there is no steep drop when increasing the number of clusters).

We use a partition into three clusters because it uncovers clusters with distinct features and an easy-to-grasp interpretation, and because it is also useful in practice for the predictive task of Section 6.3. A high-level depiction of the clusters and the relative influence of the features to each cluster is shown in Fig. 3. The distribution of queries over three clusters was: 76 in cluster A, 66 in cluster B, and 48 in cluster C. The list of queries on each cluster is included in Appendix A.

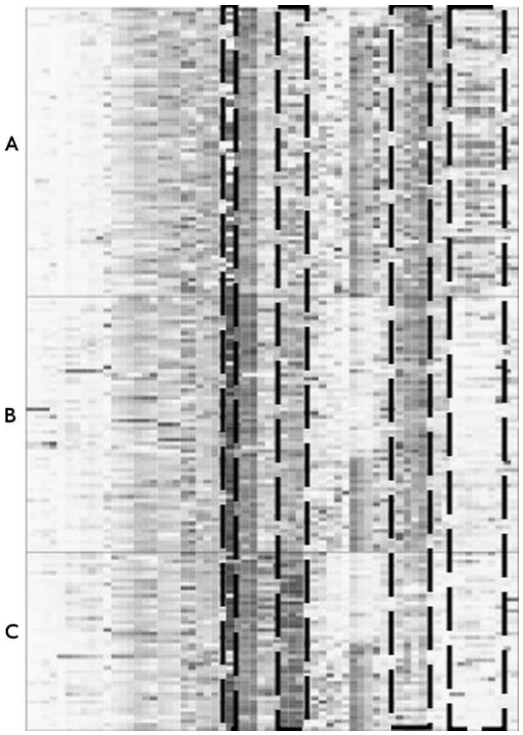


Fig. 3. Depiction of the relative influence of features in the obtained clusters. Each row represents a bursty query (rows are sorted by similarity), and each column a feature. The most important features are marked by the rectangles in the following order (from the left): PEAK BUILD-UP RATIO, TOP-1 SHARE (for 3 periods), CLICK ENTROPY and RANK-CLICK DROP (for 3 periods each), Top-5 and all CLICK DIVERGENCE (all comparisons).

5.1. Types of bursty queries

Next, we inspected the queries in each cluster, and their feature values, to try to understand which were their key characteristics. Our interpretation of the clusters is the following:

Type A: bursts that fade out completely afterwards These queries are not very frequent during the pre-episode, and fade away quickly in the post-episode. They have a high divergence (high CLICK DIVERGENCE) between the pre- and post-episode, meaning that the episode completely changes the search results for the query. There is also no strong authoritative URL (low TOP-1 SHARE, high CLICK ENTROPY), which partially explains why click share is so strongly affected by the episode.

This cluster contains many queries related to entertainment, some examples are: *katt wiliams, super bowl 2009 commercials, snl sarah palin, jett travolta, air car, kawasaki disease*. Typical behavior of this type can be represented by the query *snl sarah palin*. The mentioned TV show caused a huge increase of the query's frequency, and created a previously non-existing topic without an authoritative source. These are "buzz" topics that after an initial hype quickly lose the interest of the users.

Type B: bursts that create new topics These queries are also not very frequent during the pre-episode, but contrary to Type A, they maintain some presence in the post-episode. They have a less dominant top URL (medium TOP-1 SHARE) and less click concentration (medium CLICK ENTROPY).

This cluster contains many queries related to new scientific/technical developments and events that have long-term effects, for instance: *2008 olympics, joe biden, obama mccain polls*. For example, the information on *2008 olympics* is present long before the games commence, but it is the start of the games that triggers the increased user interest in the topic, and changes the click distribution to, in this case, sporting events result pages.

Type C: bursts on existing topics These queries appear both in the pre-episode and in the post-episode with non-negligible frequency. They have an authoritative top result with a high click share (high TOP-1 SHARE) and a low CLICK ENTROPY, so the users' attention is concentrated. For these queries, the episode

does not change the distribution of clicks, reflected by the fact that the CLICK DIVERGENCE is low.

This cluster contains many queries related to topics that are searched during the entire year, but for which a real-world event triggers heightened user interest. Examples: *teen choice awards, national hurricane center, saturday night live*. For example, the burst of *saturday night live* is caused by the same previously discussed TV appearance of U.S. politician Sarah Palin, but the query itself is present before that particular episode of the show, and its burst does not have long-lasting effects on the search results for the query.

Remark. This classification of query bursts matches the classes of bursts predicted by the model of Crane and Sornette [8] using completely different methods. Type A corresponds to *exogenous sub-critical*, expected in cases of external events that do not propagate well virally. Type B corresponds to *exogenous critical*, expected in cases of external events that are highly viral. Type C corresponds to *endogenous critical*, expected in cases of internally-motivated messages that are highly viral.

5.2. Characteristics of query bursts

Next, we look at specific sets of metrics, studying them during the *pre-episode*, *episode*, and *post-episode* periods defined as in Section 3.2. With respect to query bursts, our main findings can be summarized as follows:

1. Per-user effort/activity is higher during query bursts.
2. Users' clicks are more concentrated during query bursts.

These findings are supported by the changes in multiple query attributes during the query burst, as detailed in the rest of this section.

5.2.1. User effort/activity is higher during query bursts

Table 1 shows an increase in several metrics of activity/effort for bursty queries during the *episode* compared to pre-episode and post-episode. During the *episode*, sessions are not significantly longer in duration, but contain more queries, more clicks, and more events in general; also more individual sessions have clicks.

Bursts of query activity are driven mostly by an increase in the number of users issuing the query, given that the ratio USERS/QUERY does not change significantly. The fact that on average users click on search assistance more often during the episode, may indicate less familiarity with the topic being queried; the comparison with the stable queries also points in that direction.

Table 1
Averages of activity/effort metrics from Section 5.2.1. Statistically significant differences with episode: $p < .01$ (***), $p < .05$ (**), $p < 0.10$ (*)

Metric	Pre-	Episode	Post-	Stable
SESSION DURATION	1768.6	1886.00	1624.10	2238.1**
DWELL-TIME	175.13	178.00	157.80	216.7*
EVENTS/SESS.	5.06***	7.64	4.57***	4.69***
QUERIES/SESS.	2.67***	3.19	2.28***	2.14***
CLICKS/SESS.	2.29***	3.73	1.96***	1.87***
CLICKS/QUERY	0.79	1.81	1.39	0.86***
ASSISTANCE %	11.90***	13.18	12.29***	4.69**
NON-CLICKS %	35.97***	28.22	41.84***	22.25***
USERS/QUERY	1.47*	1.65	1.47	2.87**

Query-sessions during the *episode* are in general more "intense" than regular search sessions. This increase may be due to a number of causes, including increased interest and increased difficulty in locating information. Given that most episodes tend to be short (Table 3), the effect of the episode in effort and activity could be attributed more to increased user interest.

We find that feature ASSISTANCE % which measures the fraction of query reformulations that are the result of clicking on a search suggestion, exhibits an interesting behavior from the point of view of query bursts. Figure 4 shows distributions of ASSISTANCE % for burst episode, pre-episode, post-episode, and stable queries. Higher values for the burst episodes suggest that users click on the search suggestions more during the burst. On the other hand, for the STABLE queries users do not do this as frequently.

One possible hypothesis, for which the empirical analysis goes beyond the scope of this research, is that users who participate in a query burst become "activated" after the signals they receive go beyond an activation threshold (see e.g. [16]). In other words, users who query about a topic for the first time, must be sufficiently interested in the topic to query about it.

Comparison with stable queries Stable queries are part of longer sessions with fewer events, hence with longer dwell times. Stable queries also have much less use of search assistance.

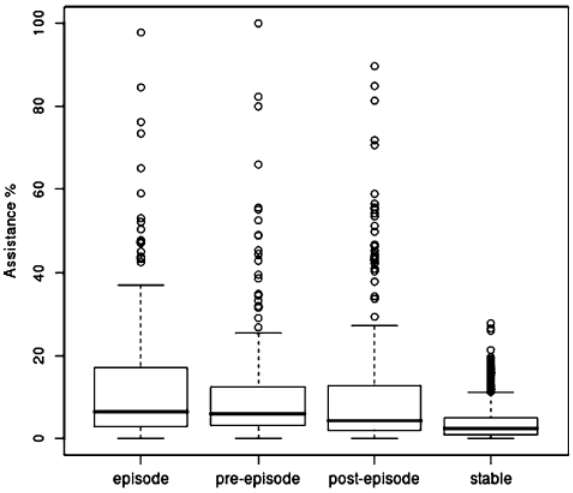


Fig. 4. Distribution of the fraction of query reformulations that are the result of clicking on a search suggestion (feature ASSISTANCE %) for burst episodes, pre-episode, post-episode, and stable queries.

Table 2
Averages of concentration metrics from Section 5.2.2

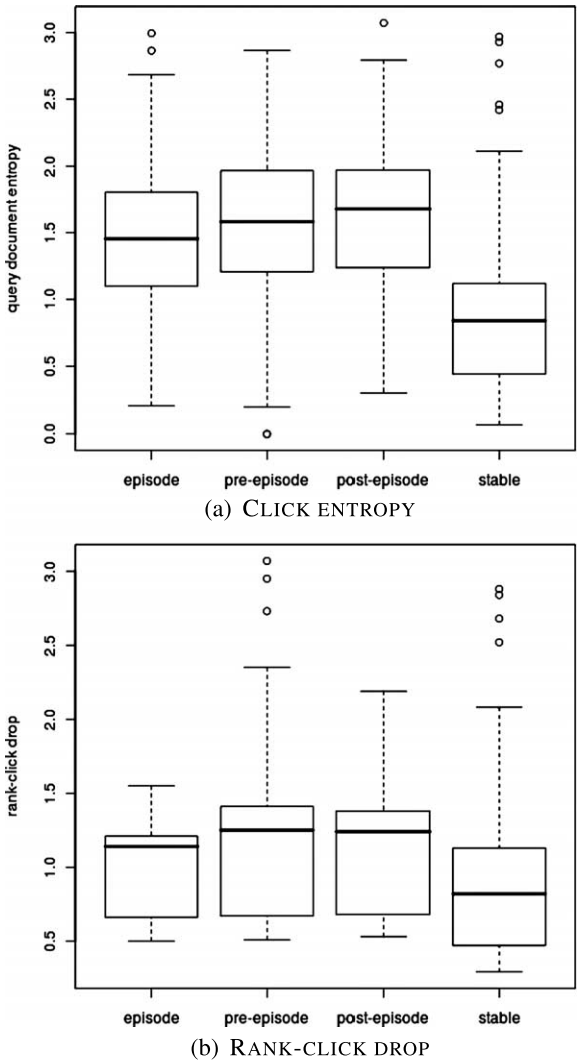
Metric	Pre-	Episode	Post-	Stable
TOP-1 SHARE	0.52	0.56	0.52	0.71***
RANK-CLICK DROP	1.15***	1.01	1.10***	0.55***
CLICK ENTROPY	1.54**	1.44	1.61***	0.93***
URLS 90%	5.12	4.40	5.46**	4.69
DISTINCT URLS	32.95	35.57	41.03*	59.17***

5.2.2. Clicks are more concentrated during episodes

Table 2 shows that clicks tend to be more concentrated during the query burst than in the pre-episode and post-episode periods. The share of clicks of the single top URL does not change significantly, but click probabilities on the top clicked URLs are higher, as evidenced by a steeper rank-click drop and a lower entropy.

In the post-episode, there is an increase in the number of distinct URLs, and the number of search results required to cover 90% of the clicks. This indicates that new relevant search results are present after the query burst.

Table 2 shows that there are no statistically significant differences between TOP-1 SHARE before, during the burst episode, or after it. We investigated the concentration of users on all results. For this we used RANK-CLICK DROP and CLICK ENTROPY measuring concentration of users on a portion of search results. Figure 5 shows in more detail the distribution of the two measures. For both, the results are aligned and show that during the burst episode users



(a) CLICK ENTROPY

(b) RANK-CLICK DROP

Fig. 5. Distribution of concentration measures CLICK ENTROPY (a) and RANK-CLICK DROP (b).

attention is more concentrated than before and after it. This suggests that during the bursts users are interested in some specific information relevant to the query. As expected, for the stable queries users clicks are less dispersed than for the bursty ones. There is a larger number of documents that are clicked (DISTINCT URLS), but the share of clicks most documents receive is small.

Comparison with stable queries Stable queries have clicks that are even more concentrated at the top than in the case of bursty queries, according to all metrics we examined. Information relevant to stable queries changes rarely, and thus the top documents satisfy user information needs by themselves.

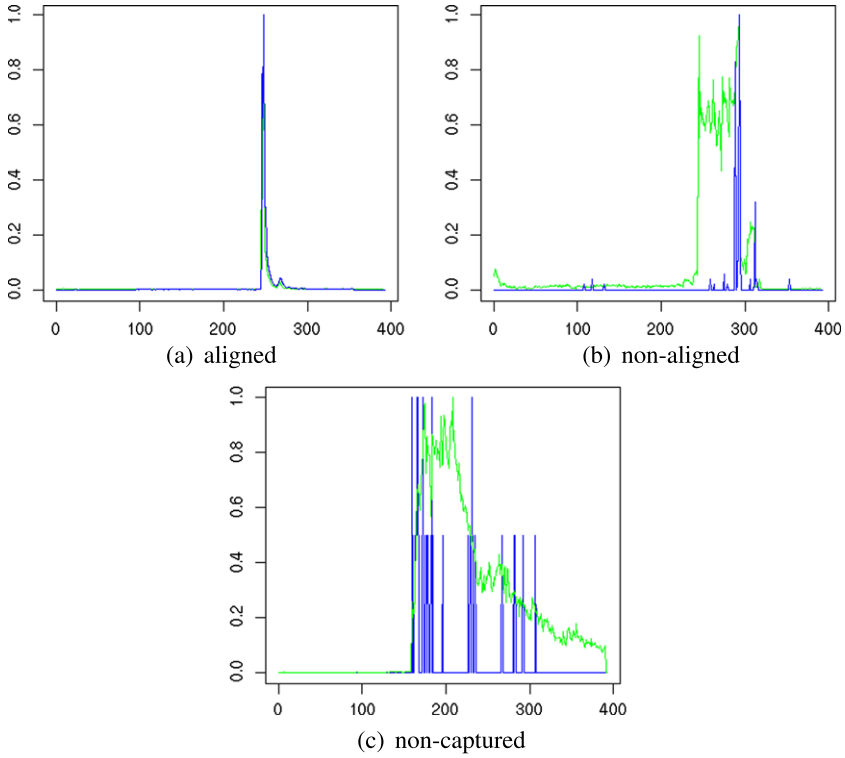


Fig. 6. Normalized frequencies for three queries in web searches (light) and news searches (dark). Three distinct cases are shown: (a) aligned bursts, (b) non-aligned bursts, (c) non-captured burst.

5.3. Relationship with news searches

In this section we introduce data obtained from a news search engine[3]. This search engine searches over an index of documents from an editorially-selected list of thousands of news providers such as CNN, BBC, etc. In the following, we refer to general web search logs as "web searches" and to news search logs as "news searches". We use one year of news searches (from the same year as web searches).

Specifically, we seek to uncover (1) whether there is a correlation of the query frequencies in web search and news search; (2) whether there is a dependency between bursts; and (3) whether there are differences in query burst intensity and duration in the two logs.

Intuitively, in the case of news searches, one may expect that query bursts would tend to appear after an event is reported by traditional media. However, in our BURSTY sample from web search, we observe many queries about subjects that would not be considered as newsworthy by traditional media (e.g.: "`fallout 3 walkthrough`", "`big brother spoilers`", etc.). Hence, we believe that in the case of web searches, query frequencies are often not related to the presence of a topic in news reports.

Looking at their entire one-year time series, we checked whether the frequencies in web searches and new searches are correlated. Measuring the Pearson correlation coefficient between these series for each query, we find values that vary widely from very strong correlation to very weak correlation (median $r = 0.7$).

Burst alignment Alignment between time series of related searches in different systems is not perfect, as observed in [1]. The measure we used for capturing the intensity and the length of a query burst does not guarantee that the captured bursts in the two logs are in the same time period. We identified three possible cases of alignment between the web and the news queries: "non-captured", "aligned", and "non-aligned". Figure 6 shows different cases of burst alignment.

We analyzed the queries from the BURSTY sample and observed their occurrences in the news search log. First we observe if the bursty queries from the web

[3]http://news.search.yahoo.com/

Table 3
Burst intensity and burst duration in Web and News search logs. The intersection marks the restriction of queries in Web search to queries discovered in News search logs

	Frequency		Intensity			Duration (days)		
Cluster	Web	News	Web	∩	News	Web	∩	News
ALL	190	131	4.8	4.9	5.5	7.7	7.9	5.2
A	76	54	5.0	4.9	5.5	7.2	7.4	5.3
B	66	41	4.5	4.6	5.4	7.4	7.7	4.9
C	48	36	5.1	5.3	5.5	7.6	8.8	5.6

searches appear in the news searches at all. To indicate appearance we set a threshold of two occurrences per day during the observed year. All queries whose frequency was bellow this threshold were labeled as non-captured. In total we found 59 (out of 190) non-captured queries in news searches.

For the queries that were present both logs (131 out of 190), we analyzed their burstiness. To discover whether they are bursty, we applied the method from Section 4.1.1. We consider bursts to be aligned when the burst peak in web searches and news searches occur within ten days of each other. Out of 131 queries that appear in both logs, we found 94 to be aligned according to this definition. The rest of the queries were labeled as non-aligned (37 out of 131).

Burst intensity and duration For the bursts that were captured in news searches, we compared the burst intensity and duration in both types of searches. Burst intensities are measured using the peak of the BURST INTENSITY $b(q,t)$ (defined in Section 3.1), and duration is measured in days.

Table 3 compares these indicators, incorporating per-cluster values for the clusters from Section 5.1. We observe that differences in intensity between web searches and news searches are minor, but statistically significant at $p \leq 0.01$; they show that bursts in news searches are slightly more intense. Differences in duration are substantial, and indicate that in news searches the average duration of the burst is shorter by at least 2 days. The news searches peaked 0.78 days (≈18 hours) before web search on average. Users expect to see results about many emerging topics first in traditional news, consistently with findings in [22] showing that traditional news sites mention new "memes" on average 2.5-hours before other sites. A few days after the initial news event, users will stop using the news search engine to get information about the event. Apparently, after this period the query is no longer perceived as "news" by users.

6. Search results and click share

Next we investigate the effect of the query burst on the distribution of clicks on search results, referred in the following as simply the "click distribution". This distribution is a function of both search engine ranking and page quality.

Basically, we aim to discover if the query burst presents an opportunity for publishing a web page about the topic of the query burst. We expect that documents that exist before the query burst will have the largest share of clicks, but that perhaps new documents can also capture some clicks. Specifically, we investigate the following questions:

1. How much is the click distribution changed by the query burst?
2. Is it necessary to have a page that existed before the burst to have a large share in the click distribution?
3. Is it possible to predict the share of new documents during the burst?

6.1. Changes in click share

We measure the effect the *episode* has on the click distribution using the previously defined CLICK DIVERGENCE measure. We compared the click distributions of pre-episode, episode, and post-episode for the BURSTY sample, and *pseudo-episodes* (as defined in Section 3.3) for the RANDOM and STABLE samples.

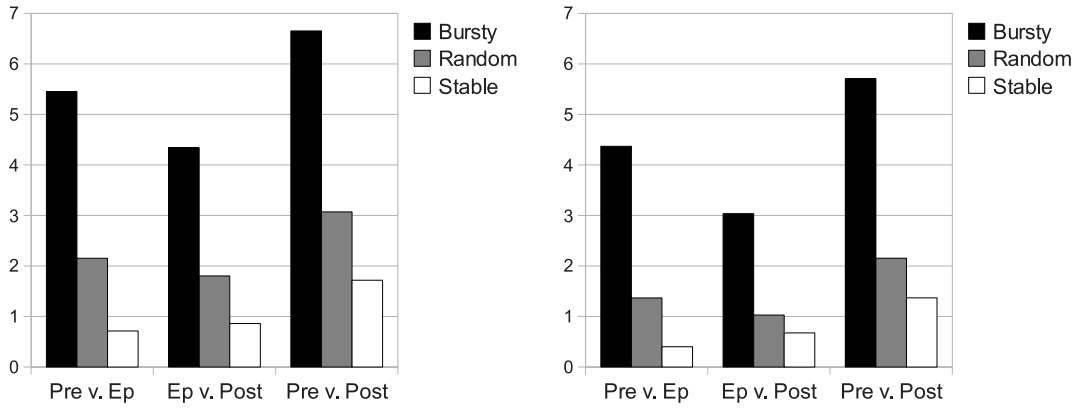
The results shown in Fig. 7(a) confirm the intuition with respect to the effects of query bursts. According to KL-divergence, the click distribution of BURSTY queries changes on average about 3× and 6× more than for RANDOM and STABLE queries respectively.

If we focus on the top-5 results only, as in Fig. 7(b), we see that the changes are smaller but the separation between BURSTY queries and the rest is even larger.

6.2. Click share of late-comers

When the frequency of a query increases, most content providers that already have pages on the topic will receive an increased number of visits and will thus benefit from the heightened user interest. Our observations confirm that publishing early represents an advantage.

To quantify how early a URL is published with respect to a query burst, we use the metric PEAK BUILD-UP RATIO of a URL u in query q. It measures how soon



(a) All results (b) Top-5 results

Fig. 7. Change in click distributions for BURSTY, RANDOM, and STABLE queries, measured using KL-divergence.

the URL appears in the query log compared with the peak of the query burst. Let $t_{u,q}^{first}$ be the first time the URL u is clicked for query q, and let t_q^{peak} be the time of the peak of the query burst of q. Let t_0 be the beginning of the observation period, then this metric is equal to:

$$\max \quad \frac{t_q^{peak} - t_{u,q}^{first}}{t_q^{peak} - t_0}, 0 \quad . \tag{5}$$

A value close to 1 means the URL's first click was close to the beginning of the observation period, while a 0 indicates the URL's first click occurred on the day of the peak. The first click in a specific URL could be observed *after* the episode peak, but this is a rare event and for simplicity of the presentation we truncate those values to zero. In the following, we will refer to documents whose PEAK BUILD-UP RATIO is non-zero as *old pages* (as they existed before the burst) and to documents whose PEAK BUILD-UP RATIO is close to zero as *new pages*.

Figure 8(a) indicates that 61% of the top-URLs have existed since the beginning of the observation period, while only 16% of the top-URLs are *new pages* created on or after the query burst.

When examining the top-5, top-10, and bottom-10 results (Figs 8(b), 8(c), and 8(d)), we see that publishing late, i.e.: having PEAK BUILD-UP RATIO close to zero, means a lower share of clicks during the episode. For instance in the case of top-10 results, on average about 3 results are new pages, while in the bottom-10 results, on average about 5 results are new pages.

Next, we consider the *share* of clicks the new pages will obtain. This information is presented in Table 4 which shows the click share of the *new pages* in the Top-1, Top-5, Top-10, and All. In general, new pages obtain a minority of clicks during the episode (27.5%), and this is distributed among many queries: even the Top-10 most clicked new pages (when considered together) obtain only 8.9% of the clicks.

Our findings from Section 5.1 suggest that the click share of at least the top-URL is different across clusters. Therefore, Table 4 also includes per-cluster results.

The per-cluster analysis shows that there wide variability among the clusters. The best opportunity for publishing new pages are queries of type A (bursts that fade out completely afterwards) for which they obtain 52.1% of the clicks. Next, for queries of type B (bursts that create new topics) the new pages obtain 25.2% of the clicks. Finally, for queries of type C (bursts on existing topics) the new pages obtain only 9.8% of the clicks; in this last cluster, it is in practice hopeless for a publisher that wants to profit from a query burst to publish an article about the topic of the burst.

6.3. Finding opportunities for late-comers

From the content-providers' perspective, the question of finding *which* are the "waves" that should be ridden is the central one. The resources of the content-providers are limited so they can not write a new page for any bursty query related to their expertise, and moreover the time they have to react is very short given that query bursts do not last for long.

Assuming that not all query bursts can be predicted (some can be predicted, e.g. when they are related to newsworthy events that are planned well in advance), a system that were to help content providers in deciding

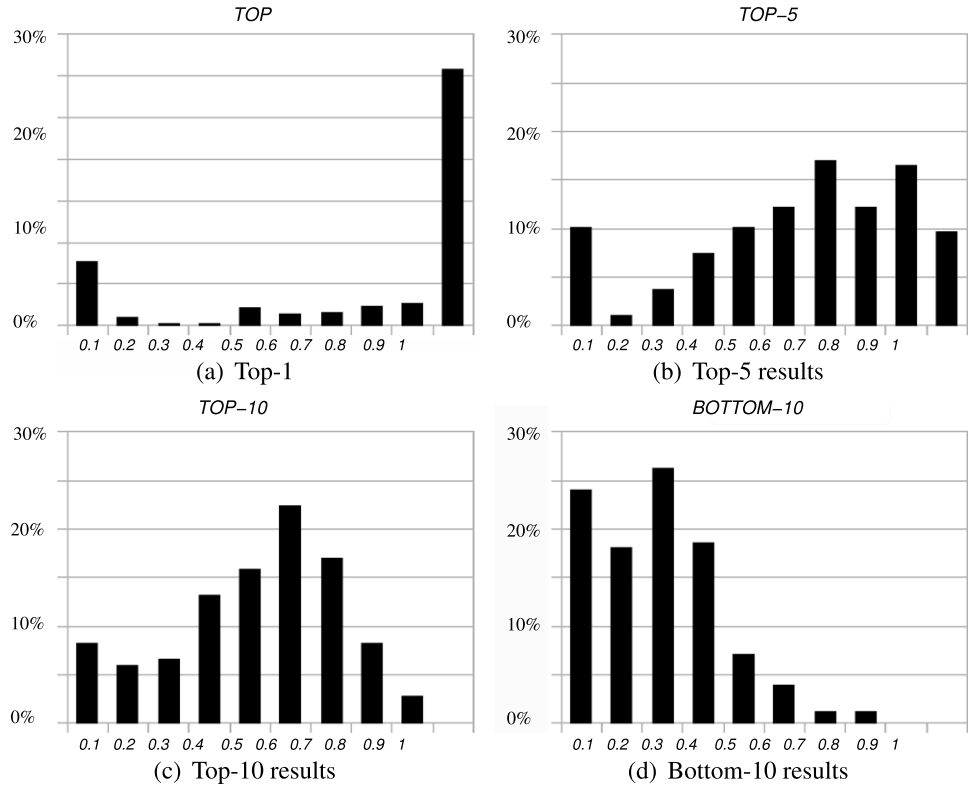


Fig. 8. PEAK BUILD-UP RATIO for the (a) the top result, (b) the top-5 results, (c) the top-10 results, (d) the bottom-10 results.

what to write about, should be capable of (a) identifying query bursts and (b) predicting the expected benefit for content providers. Question (a) was the subject of Section 3.1 while (b) turns out to be more difficult.

As mentioned in the previous sections, the target of this prediction task is the click share of new pages. We first use a logistic regression model (M_p) with the features from the pre-episode and episode described in Section 4.2. Its performance, measured using the correlation coefficient between the predicted click share and the actual click share for a held-out test set of queries is reported in Table 5.

The insights from Table 4 can be used to improve this prediction, given that the average share of newly published pages depends clearly on the cluster to which the query belongs. Thus, we build a model (M_c) that first computes the probability of a query belonging to each cluster using a Naïve Bayes classifier, and then includes these predictions in the logistic regression model. Table 5 shows the correlation coefficients between the original and predicted values and the improvement that the cluster prediction brings. The results show that it is hard to predict the values for all the pages and for the very first page, while a fair performance can be obtained with Top-5 and Top-10 results.

7. Conclusions

Query bursts are observed in a search engine log whenever there is increased interest in a certain topic. Looking back at our metaphor of a search "marketplace" for information, for the main market components we discovered that:

- Not all queries are equal and there are distinct types of query bursts (*goods*). Our research over one year of query log uncovered different types of query bursts, including (A) bursts that fade out completely afterwards, (B) bursts that create new topics, and (C) bursts on existing topics.
- The analysis of several metrics indicates that during query bursts users invest more effort (pay a higher *price*) in search, and that their clicks are concentrated on a smaller group of search results.

Table 4

Click share of the new URLs as a percentage of total clicks. Top-k indicates the k most clicked new URLs. "All" indicates all the new URLs

Query cluster	New URLs	Click share
All queries		%
	Top	3.1
	Top-5	5.5
	Top-10	8.9
	All	27.5
A: bursts that fade out completely afterwards		%
	Top	37.8
	Top-5	41.1
	Top-10	20.2
	All	52.1
B: bursts that create new topics		%
	Top	5.9
	Top-5	5.6
	Top-10	5.2
	All	25.2
C: bursts on existing topics		%
	Top	2.5
	Top-5	3.5
	Top-10	4.2
	All	9.8

Table 5

Correlation coefficient between predicted and actual click share of new documents

Model	Top	Top-5	Top-10	All
Simple model M_p	0.59	0.71	0.69	0.42
Cluster-based model M_c	0.64	0.77	0.77	0.46

- Publishing documents (*supply*) early, before the burst, is the only way towards obtaining a large proportion of the increased user attention. However, for some queries, content providers that are not among the first to publish can also obtain a non-trivial increase in click share.
- After the query burst, the distribution of clicks (*market share*) in search results for a query is substantially different from that before the query burst.

Based on these findings the main stakeholders in a search market may take different strategies during the query bursts:

- Content providers that intend to capture users' attention on emerging topics should attempt to publish early. If not, they should target query bursts on topics that did not exist before (types A and B). Writing during a query burst about a previously-existing topic is unlikely to yield a substantial share of clicks.
- Search engines should, according to our findings, treat queries undergoing query bursts differently. For instance, search suggestions are much more important for these queries. A search engine may introduce user-interface changes to support the needs of users entering bursty queries.

We consider this work as a part of a broader effort, which is to provide the right signals about users' needs to the authors of Web content. Search engines should help to detect scarcity of information on certain topics so that content providers can supply this information. A system that is capable of telling a content provider e.g. "if you write about environmental issues, you should be writing about solar energy", would be a big step forward for the Web ecosystem.

This involves creating models that also take into account content providers' features such as topic, influence and authority, and that are able to detect users' unsatisfied needs for information in certain areas. A promising approach to this problem would be to perform a topic-sensitive analysis in which queries (and pages) are classified into topical categories, and then studied independently for each topical category.

Acknowledgements

The authors thank Aris Gionis and Adam Rae for their help, and Bettina Berendt, Yoelle Maarek and Ingmar Weber for helpful comments on an earlier version of this manuscript.

Appendix

A. BURSTY QUERIES PER CLUSTER

Cluster A ("bursts that fade out completely afterwards") criselda volks scandal, kawasaki disease, groundhog day, oj simpson, gi joe, jessica simpson weight gain, hgtv dream home, fiesta bowl 2009, groundhog day 2009, saturday night live sarah palin, cyber monday deals, christian bale, super bowl commercials, polling place, gustav, snl sarah palin, hgtv dream home giveaway, jett travolta autism, superbowl commercials, blackberry storm release date, kimbo

slice vs ken shamrock, michael phelps bong, last day to register to vote, jett travolta, ground hog day, kawasaki syndrome, gi joe trailer, cyber monday sales, is katt williams dead, plaxico burress, go daddy commercial, california propositions 2008, hurricane gustav, brooke satchwell, wwe svr 2009, kelly preston, hurricane hanna, neel kashkari, halle berry baby photos, deborah lin, energy saving tips, cyber monday 2008, super bowl 2009 commercials, caylee anthony update, bristol palin, compressed air car, samantha mumba, marykate olsen, superbowl ads, cyber monday, octuplets, misty may, peanut butter recall, michael phelps smoking, fallout 3 walkthrough, anne pressly, successful resume examples, sarah palin vogue magazine, palin, the strangers true story, josiah leming, super bowl ads, latest presidential polls, michael phelps girlfriend, election map, if i were a boy lyrics, niki taylor, free christmas wallpaper, bernie mac illness, montauk monster, katt williams dead, air car, virginia themadsen, soyouthinkyoucandance, volam.com.vn, brangelina twins.

Cluster B ("bursts that create new topics") obama mccain polls, black friday 2008, pineapple express, ducati 1098, register to vote online free, where to vote, morgan freeman, groundhog, register to vote online, scientology, kimbo slice, lita ford, houston weather, cybermonday, tropic thunder, big brother 10 spoilers, sarah palin, where do i vote, taylor swift, turbo tax online, electoral votes, sophie okonedo, madden 09, presidential polls, brett favre, zuleyka rivera, chinese new year 2009, tina fey scar, voting locations, voting, bill ayers, register to vote, breaking dawn, (redacted: adult query), election polls, free turbotax, kimbo, us open tennis, prop 8, burning man 2008, the curious case of benjamin button, mary mccormack, black friday, gina carano drunk, kathy griffin, hotjobs yahoo com, transformer 2, john travolta, labor day, hurricane katrina, poea open jobs in canada, voter registration, marley and me, olympics, bernie mac, the mummy, labor day 2008, irs refund status, john mccain, www.azmoon.com, 2008 olympics, twilight book, sarcoidosis, anthrax, joe biden, michael phelps, cindy mccain.

Cluster C ("bursts on existing topics") elite xc, tampa bay rays, saw 5, puppy bowl, teen choice awards, cell for cash, taxact, turbotax online, fiesta bowl, hurricane center, special k, christian songwriting, lollapalooza, pixie hollow, rasmussen poll, www.mysoju.com, turbotax.com, www.watch-movies.net, bradley effect, turbotax, can i vote, obama stimulus package, gallup poll, mda telethon, khou, the mole, us open, white sox, mccain, snl, shawn johnson, gallup, hurricane tracker, taxact.com, khou.com, kprc, republican national convention, chicago white sox, gi joe movie, fdic, playatmcd.com, taxact online, click2houston, saturday night live, butterfinger, www.pch.com, national hurricane center.

B. STABLE QUERIES

Sample of queries that seldom fluctuate in frequency holland america, national geographic channel, midas, rheumatoid arthritis, dudetube, baby depot, dereon, jimmy johns, essence, ac moore, tribal tattoos, court tv, zoloft friends reunited, viewpoint bank, redtub, boston market, car payment calculator, heidi klum, chicos, af portal, low income apartments, postsecret, philadelphia, mspace, tiger airways, liberty university, ftvgirls, charmeddisney movie club, photography, hydrocodone, mike in brazil, tribune review, yahooligansl, (redacted: adult query), spiegel, netflex, pal, bitcomet, toutube, mr skin, greek mythology, extenze, ebay motors parts, paint colors, stupid videos, english to french translation, yout, vans shoes, (redacted: adult query), pump it up, spa.gov.my, veterans administration, radisson hotel, myspace music, education, candylist, us navy, the gas company, arizona, mcdonald's, nylottery.org, coke rewards, slacker, googlemap, american airline, valley national bank, sports authority store, new jersey lottery, gimp, commerceonline, west elm, university of chicago, mta nyc, knotts berry farm, dragon fable, flicker photo site, alienware, american signature furniture, intervention, akhbar harian metro, city of houston, south bend tribune, sims, pink eye, tabnak, compaq, shyla stylez, cms, faa, suze orman, crigslistlist, malibu strings, asda, long and foster, democrat and chronicle, acs student loan, la fitness locations, basspro, kiss fm, ethan allen, texas child support red, happy birthday, quixtar, hotmai, dailyniner, adolf hitler, hepatitis, baskin robbins wirefly, usps tracking number, simslots, honolulu star bulletin, department of homeland security adobe acrobat reader, pancreatitis, american standard, alloy, at&t universal card, web, red roof inn, jc penney catalog, lexmark drivers, gsc, genealogy, pc world, qoutes, arby's, press democrat, bentley, penndot, kbr, sony digital camera, whole foods market, belize, sheboygan press wynn las vegas, randy blue, inquirer, baby boy names, el salvador, tampa tribune, ohio university myspace', sexyclips, kementerian sumber manusia, kentucky fried chicken, marriott rewards, ace, sugarland, brazil, cold

stone creamery, celebrity hairstyles, coast to coast am, starbucks locations, bargain news, yahoo malaysia, general electric, collections etc, terra, proactiv, cheap ticket, crohn's disease, spanx, entergy, wthr, bipolar disorder, currency calculator, tillys, 1800contacts, galottery, odd news, virginia, albert einstein, (redacted: adult query), (redacted: adult query), trilulilu, adobe photoshop, spybot search and destroy, sean cody, cover letter, hartford courant, citicard, goodyear tires, advanced auto parts, metric conversion mary kay, kaiser permanente california, hotmail email, rapidshare, baby names meaning, sherwin williams wescom credit union, cialis, cathay pacific, livejournal, subaru, net flix.

References

[1] E. Adar, D.S. Weld, B.N. Bershad, and S.S. Gribble, Why we search: Visualizing and predicting user behavior, in: *Proc. of the 16th International Conference on World Wide Web*, WWW '07, ACM, New York, NY, USA, 2007, pp. 161–170.

[2] S. Asur and G. Buehrer, Temporal analysis of web search query-click data, in: *Proc. of WebKDD/SNAKDD 2009: Web Mining and Social Network Analysis Workshop*, Paris, France, ACM Press, 2009.

[3] S.M. Beitzel, E.C. Jensen, A. Chowdhury, O. Frieder, and D. Grossman, Temporal analysis of a very large topically categorized web query log, *Journal of American Society for Information Science* 58(2) (2007), 166–178.

[4] S.M. Beitzel, E.C. Jensen, A. Chowdhury, D. Grossman, and O. Frieder, Hourly analysis of a very large topically categorized web query log, in: *Proc. of the 27th Conference on Research and Development in Information Retrieval*, SIGIR '04, ACM, New York, NY, USA, 2004, pp. 321–328.

[5] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, The query-flow graph: Model and applications, in: *Proc. of the 17th Conference on Information and Knowledge Management*, CIKM '08, ACM, New York, NY, USA, 2008, pp. 609–618.

[6] L. Chen, Y. Hu, and W. Nejdl, Using subspace analysis for event detection from web click-through data, in: *Proc. of the 17th International Conference on World Wide Web*, WWW '08, ACM, New York, NY, USA, 2008, pp. 1067–1068.

[7] S. Chien and N. Immorlica, Semantic similarity between search engine queries using temporal correlation, in: *Proc. of the 14th International Conference on World Wide Web*, WWW '05, ACM, New York, NY, USA, 2005, pp. 2–11.

[8] R. Crane and D. Sornette, Robust dynamic classes revealed by measuring the response function of a social system, *Proceedings of the National Academy of Sciences* 105(41) (2008), 15649–15653.

[9] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, An experimental comparison of click position-bias models, in: *Proc. of the 1st International Conference on Web Search and Web Data Mining*, WSDM '08, ACM, New York, NY, USA, 2008, pp. 87–94.

[10] D. Downey, S. Dumais, and E. Horvitz, Heads and tails: Studies of web search with common and rare queries, in: *Proc. of the 30th International Conference on Research and Development in Information Retrieval*, SIGIR '07, ACM, New York, NY, USA, 2007, pp. 847–848.

[11] D. Downey, S. Dumais, D. Liebling, and E. Horvitz, Understanding the relationship between searchers' queries and information goals, in: *Proc. of the 17th Conference on Information and Knowledge Management*, CIKM '08, ACM, New York, NY, USA, 2008, pp. 449–458.

[12] G.P.C. Fung, J.X. Yu, P.S. Yu, and H. Lu, Parameter free bursty events detection in text streams, in: *Proc. of the 31st International Conference on Very Large Databases*, VLDB '05, VLDB Endowment, 2005, pp. 181–192.

[13] Google Inc. Google Corelate. http://correlate.googlelabs.com, 2009.

[14] Google Inc. Google Flu Trends. http://www.google.org/flutrends/, 2009.

[15] Google Inc. Google Trends. http://www.google.com/trends/, 2009.

[16] M. Granovetter, Threshold models of collective behavior, *The American Journal of Sociology* 83(6) (1978), 1420–1443.

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

[18] R. Jones and K.L. Klinkner, Beyond the session timeout: Automatic hierarchical segmentation of search topics in query logs, in: *Proc. of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, ACM, New York, NY, USA, 2008, pp. 699–708.

[19] J. Kleinberg, Bursty and hierarchical structure in streams, *Data Mining and Knowledge Discovery* 7 (October 2003), 373–397.

[20] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins, On the bursty evolution of blogspace, *World Wide Web* 8(2), Kluwer Academic Publishers, Hingham, MA, USA, June 2005, pp. 159–178.

[21] T. Lau and E. Horvitz, Patterns of search: Analyzing and modeling web query refinement, in: *Proc. of the 7th International Conference on User Modeling*, Springer-Verlag, New York, Inc., Secaucus, NJ, USA, 1999, pp. 119–128.

[22] J. Leskovec, L. Backstrom, and J. Kleinberg, Meme-tracking and the dynamics of the news cycle, in: *Proc. of the 15th International Conference on Knowledge Discovery and Data Mining*, KDD '09, ACM, New York, NY, USA, 2009, pp. 497–506.

[23] Q. Mei and K. Church, Entropy of search logs: How hard is search? With personalization? With backoff? in: *Proc. of the 1st International Conference on Web Search and Web Data Mining*, WSDM '08, ACM, New York, NY, USA, 2008, pp. 45–54.

[24] Pew Research Center, Internet overtakes newspapers as news outlet. http://pewresearch.org/pubs/1066/internet-overtakes-newspapers-as-news-source, 2008.

[25] Pew Research Center, The new news landscape: Rise of the Internet. http://pewresearch.org/pubs/1508/internet-cell-phone-users-news-social-experience, 2010.

[26] F. Radlinski and T. Joachims, Query chains: Learning to rank from implicit feedback, in: *Proc. of the 11th International Conference on Knowledge Discovery and Data Mining*, KDD '05, ACM, New York, NY, USA, 2005, pp. 239–248.

[27] M. Richardson, Learning about the world through long-term query logs, *ACM Transaction on the Web* 2(4) (2008), 1–27.

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

[29] I. Subasic and B. Berendt, Discovery of interactive graphs for understanding and searching time-indexed corpora, *Knowledge and Information Systems* 23(3) (2010), 293–319.

[30] I. Subasic and C. Castillo, The effects of query bursts on web search, in: *2010 IEEE/ACM International Conference on Web Intelligence–Intelligent Agent Technology*, WI-IAT '10, IEEE, Aug. 2010, pp. 374–381.

[31] Y. Sun, K. Xie, N. Liu, S. Yan, B. Zhang, and Z. Chen, Causal relation of queries from temporal logs, in: *Proc. of the 16th International Conference on World Wide Web*, WWW '07, ACM, New York, NY, USA, 2007, pp. 1141–1142.

[32] M. Vlachos, C. Meek, Z. Vagena, and D. Gunopulos, Identifying similarities, periodicities and bursts for online search queries, in: *Proc. of the 2004 International Conference on Management of Data*, SIGMOD '04, ACM, New York, NY, USA, 2004, pp. 131–142.

[33] X. Wang and A. McCallum, Topics over time: A non-markov continuous-time model of topical trends, in: *Proc. of the 12th International Conference on Knowledge Discovery and Data Mining*, KDD '06, ACM, New York, NY, USA, 2006, pp. 424–433.

[34] Yahoo! Inc. Yahoo! Clues. http://clues.yahoo.com, 2011.

[35] Q. Zhao, S.C.H. Hoi, T.-Y. Liu, S.S. Bhowmick, M.R. Lyu, and W.-Y. Ma, Time-dependent semantic similarity measure of queries using historical click-through data, in: *Proc. of the 15th International Conference on World Wide Web*, WWW '06, ACM, New York, NY, USA, 2006, pp. 543–552.

[36] Q. Zhao, T.-Y. Liu, S.S. Bhowmick, and W.-Y. Ma, Event detection from evolution of click-through data, in: *Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, ACM, New York, NY, USA, 2006, pp. 484–493.