Understanding Temporal Query Dynamics

Anagha Kulkarni
Carnegie Mellon University
Pittsburgh, PA, USA
anaghak@cs.cmu.edu

Jaime Teevan, Krysta M. Svore, Susan T. Dumais
Microsoft Research
Redmond, WA, USA
{teevan, ksvore, sdumais}@microsoft.com

ABSTRACT

Web search is strongly influenced by time. The queries people issue change over time, with some queries occasionally spiking in popularity (e.g., *earthquake*) and others remaining relatively constant (e.g., *youtube*). Likewise, the documents indexed by a search engine change, with some documents always being about a particular query (e.g., the Wikipedia page on earthquakes is about the query *earthquake*) and others being about the query only at a particular point in time (e.g., the New York Times is only about earthquakes following a major seismic activity). The relationship between documents and queries can also change as people's intent changes (e.g., people sought different content for the query *earthquake* before the Haitian earthquake than they did after). In this paper, we explore how queries, their associated documents, and the query intent change over the course of 10 weeks by analyzing query log data, a daily Web crawl, and periodic human relevance judgments. We identify several interesting features by which changes to query popularity can be classified, and show that presence of these features, when accompanied by changes in result content, can be a good indicator of change in query intent.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – Query formulation, Search process.

General Terms: Human Factors, Measurement.

Keywords: Query dynamics.

1. INTRODUCTION

There are temporal aspects to Web search queries that search engines must account for in order to provide the most relevant results to their users. As an example, in the middle of March 2010, the query *march madness* suddenly became very popular, occurring thousands of times when one month before it occurred infrequently. The rise in popularity was a result of the popular annual college basketball championship in the United States.

In addition to changes in query frequency, there were other changes associated with the query *march madness* during the championship period. For one, the National Collegiate Athletic Association (NCAA) homepage (http://ncaa.com) became very relevant to the query. The page provides comprehensive coverage of US college sports, but does not typically focus on basketball – except in March during March Madness. Other results were more relevant to the query *march madness* during March because they provided dynamic content. For example, the CBS Sports college basketball page (http://www.cbssports.com/collegebasketball), which provides real time game information, became relevant to people seeking to learn the score of a game in progress. In contrast, relatively static pages, like the Wikipedia page about March Madness, became less relevant during this period of high interest. Such pages are useful for learning about March Madness in general, but not for actively monitoring the event, and thus are better suited to satisfy the need of searchers when the query is not spiking. The changes in which pages were relevant to the query *march madness* during the month of March reflects the fact that people's query intent was also changing.

Understanding changes to query popularity, Web content, query intent, and relevance is fundamental to understanding the search experience. By looking at a broad picture of all of these factors over time, search engines can know more about what people are looking for than they can with just a static snapshot. Analysis of change can be useful in identifying the most relevant results, as how a webpage's content changes over time provides insight into what the page is fundamentally about and how temporally sensitive it is, and this can be appropriately matched to the user's intent. Change can also help search engines decide when it is appropriate to interject news into the result page, how strongly to incorporate older behavioral data compared to more recent behavioral data, and whether content change should be incorporated into search result representations (e.g., snippets) so as to enable users to make the best possible relevance decisions.

To support the development of temporally-aware approaches to search, in this paper we paint a picture of the interaction between temporal changes in query popularity, document content, and query intent. We use *observables* such as changes in query popularity attributes (e.g., does the query spike or trend in a particular direction?) and changes in associated document content (e.g., does the frequency of query terms in the document increase?) to model changes in the *latent* entity – user's query intent. We use human relevance judgments and real-world Web search behavior gathered over a period of time as proxies for query intent.

After a discussion of relevant research studying query and Web content dynamics, we present the approach we took to understand the relationship between the two, especially as it relates to changes in query intent. Our analysis focuses on 100 temporally interesting queries, and we describe the rich, multifaceted temporal information we collected for each query. We then discuss what we learned from analyzing these queries, including:

- The identification of several interesting features by which changes in query popularity can be classified, including spikiness, spike shape, periodicity, and overall trend.
- The finding that the presence of several of the popularity features, when associated with changes in result content, can be a good indicator of change in query intent.

WSDM'11, February 9–12, 2011, Hong Kong, China.

- The finding that an upward trend in query popularity is generally accompanied by an informational intent for most queries, but certain types of queries (aperiodic and spiking) are navigational for the duration of the rise.

We discuss the implications of these findings on the design of temporally-aware search systems and conclude.

2. RELATED WORK

We begin by discussing three lines of work that are relevant to the general topic of the temporal dynamics of search and relevance: 1) query dynamics, 2) Web content dynamics, and 3) how these dynamics can be used to improve Web search.

Several groups have examined *query dynamics*, or how query volume changes over time. For example, some researchers have investigated queries in aggregate over time to understand changes in popularity [27] and the uniqueness of topics at different times of day [6]. Vlachos et al. [26] were among the first to examine and model periodicities and bursts in Web queries using methods from Fourier analysis. They also developed a method to discover important periods and to identify query bursts. Jones and Diaz [19] identified three general types of temporal query profiles: atemporal (which have no periodicities), temporally unambiguous (which contain a single spike), and temporally ambiguous (which contain more than one spike). They further showed that query profiles were related to search performance, with atemporal queries having lower average precision. Chien and Immorlica [8] used the temporal patterns of queries to identify similar queries. Others examined the relationship between query dynamics and events. Ginsberg et al. [18] showed how query dynamics reflect real-world events such as H1N1 influenza outbreak, and Adar et al. [3] identified when changes in query frequency lead or lag behind mentions in both traditional media and blogs.

Other researchers have described characteristics of *Web content dynamics* and user interaction with dynamic content. Fetterly et al. [17] and Ntoulas et al. [23] conducted large-scale analyses of the extent to which Web content changes over time, and examined the implications for Web crawling policy and more generally for search engines. Cho et al. [9] investigated how changes in link patterns could be used to identify high quality pages. Adar et al. [1, 2] looked in more detail at the change of DOM elements and individual terms over time (specifically, the longevity of terms within documents and the general collection). In addition, their Web crawl focused on pages that had different user re-visitation patterns, and they found that the popularity of a page was positively correlated with the frequency of change but not the amount of change. Kleinberg [20, 21] developed general techniques for summarizing the temporal dynamics of textual content and for identifying bursts of terms within content. From a Web search perspective, a particularly interesting kind of evolving content has to do with breaking news events. Diaz [11] developed algorithms for identifying queries that are related to breaking news and for then blending relevant news results into core search results.

Recently, several groups have started examining how the temporal dynamics of queries, documents and user interaction can be used to *improve Web search*. Alonso et al. [5] present a method for clustering and exploring search results based on temporal expressions within the text. Alfonseca et al. [4] showed how query periodicities could be used to improve query suggestions, although they seem to have more limited utility for general topical categorization. Dakka et al. [10] developed a question answering system that could generate different answers at different points of time, if appropriate. Zhang et al. [28] explored identifying implicit temporal queries (specifically those that implicitly refer to the current year, e.g., World Cup results meaning World Cup 2010 results) and re-ranking search results. They found that favoring recent documents improved Web retrieval performance, for this subset of queries. Dong et al. [13] explored the use of recency-based features more generally in Web ranking. Li and Croft [22] classified queries based on the temporal distribution of documents and used this classification to establish different document priors within a language modeling framework. However, they did not address the challenge of automatically classifying queries into classes with different priors. Elsas and Dumais [16] also used a language modeling approach to incorporate temporal features into an improved ranking algorithm for navigational queries. They used information about how often and by how much document content changes over time to establish different document priors, and further used differential term weights depending on the longevity of a term in a document to develop a ranking method that was sensitive to temporal dynamics of documents. Efron [15] developed linear time series models based on term occurrences to improve term-weighting estimates for retrieval.

The work presented in this paper differs from previous work in several ways. We extend the three types of temporal query profiles identified by Jones and Diaz [19] to build a rich picture of changes to query popularity. We then consider how changes to query popularity relate to changes in a consistent set of webpages, whereas previous research has explored at the relationship of spikes in queries to changes in news or blog content. We further examine the extent to which changes in user intent (as evidenced in explicit judgments and user click patterns) are related to the dynamics of page content.

3. METHODOLOGY

To understand how time impacts what queries are issued, which documents are relevant, and how the content of those documents change, we gathered a series of rich data for 100 queries and the associated URLs over a period of 10 weeks, from March 25, 2010 to May 28, 2010. We used large-scale query log analysis to track how frequently each query was issued during the study period, and which results were clicked. We also gathered weekly human relevance judgments for the selected query and URL pairs. These data were complemented with a daily Web crawl of the result content. After describing how the queries and associated URLs were selected for study, we describe in greater detail the three sources of data we gathered for the queries and URLs (human relevance judgments, query logs, and Web crawl content).

3.1 Query and Result Selection

Prior to the study period, we selected 100 queries and 2000 documents (with 20 documents associated with each query) to track. Although it is easy with the benefit of hindsight to identify temporally interesting queries, we needed to select queries and results in advance because some of the data we were interested in gathering would be impossible to collect retrospectively. For example, a human judge cannot easily determine whether a result would have been relevant to a query a week ago because it is hard to remember or imagine last week's information context. Thus we undertook the challenging task of predicting which queries would undergo interesting changes in advance of our study.

We took several approaches to selecting queries that were likely to be interesting, summarized in Table 1. Our goal was to identify queries that would be likely to change in frequency and queries where what was relevant was likely to change (represented by the columns in Table 1). We looked for such queries using historical

Table 1. Query selection criteria.

		Type of change targeted	
		Δ query popularity	**Δ relevant results**
Year	**2009**	Queries that spiked during time period last year *april fools*, *tax extension*	Queries with high click entropy during time period last year *miss usa*, *easter baskets*
	2010	Queries related to scheduled upcoming events *ipad*, *crystal bowersox*	Queries guessed to be likely to have breaking news *earthquake*, *lady gaga*

data from the previous year and using predictive data from the current year (represented by the rows).

Using historical data gathered from the same dates last year as the study time period (namely March through May 2009), we identified queries that had previously spiked in popularity and seemed likely to spike again this year. Examples of queries selected in this manner include *april fools* and *tax extensions.* We also looked at queries for which the clicked results changed significantly last year, as measured by click entropy. Click entropy measures the variation in what people click on following the query. Although high click entropy can be the result of many different factors (including how much the results presented for the query change and the average number of clicks following the query) [25], we manually identified queries with high click entropy that seemed likely to have had temporally interesting click patterns, such as *miss usa* and *easter basket*.

Using data gathered from 2010, we also tried to anticipate new events that were likely to happen. We identified queries likely to become more popular by identifying upcoming events, such as product releases, movie premieres ad television episodes. For example, several queries included American Idol contestants (*crystal bowersox, adam lambert*), episode titles from the popular television series 'Lost' (*ab aeterno*), and expected product releases (*ipad*). In addition, we selected queries about people or events likely to be in the news (*lady gaga, eclipse*).

For each query, we then selected 20 results to track during the study period. For the queries that were issued in 2009 (97 of the 100 queries), we included the three most clicked URLs from the period of March through May 2009. These were supplemented with the top ten results as displayed in early March 2010 from two popular Web search engines, Google and Bing, with duplicates removed. We then manually selected several additional pages for each query, with a focus on pages that seemed likely to contain interesting content related to the query.

In our study design, we chose not to focus on understanding when information resources become newly available. For example, prior to the American Idol season 9 there was no Wikipedia page about Crystal Bowersox, but as soon as the page was created it became highly relevant to queries related to her. While quickly identifying relevant new content is an important problem, we chose to focus on the dynamics of existing content.

3.2 Data sources

Over the study period, we collected data for our 100 queries from three different sources: query and click data, human relevance judgments, and a daily crawl of pages we monitored.

3.2.1 Query and Click Data

To gather information about query popularity and clicked results during our 10 week study period, we analyzed the query logs from the Bing search engine. From the logs, we sampled information related to our 100 queries gathered from 3.6 million users. The sample included our queries, the top ten returned results with display position, clicked results, and time stamp information. The sample was filtered to remove bots and spam, and processed so that pagination and back button clicks were treated as the same query. To remove variability caused by geographic and linguistic variation in search behavior, we only included log entries generated in the English speaking United States English locale. The resulting data consisted of 11.7 million query log records.

3.2.2 Human Relevance Judgments

We also obtained weekly human relevance judgments during the study period for each query and URL pair. URLs were judged for relevance to the query on a five-point scale, ranging from "perfect" to "bad". Raters were paid and received training prior to rating. Training focused on the relevance of the URL to the query, and did not explicitly focus on temporality, although they were asked to consider whether content was out of date.

While a number of different judges rated the queries, a single judge provided relevance judgments for all URLs for a given query at a given time. The same judge did not judge the same query every week. Thus a single query was judged by several people over the 10 week period. To measure the consistency across judges, a number of URLs were judged by multiple judges at each time period. According to Cohen's Kappa, there was a fair inter-rater reliability of 0.38. We have attempted to normalize out judge-related bias as best as possible in our analysis by looking at relative judgments as opposed to absolute judgments.

3.2.3 Results Content Crawl

To understand how the content of the documents associated with the queries changed over time, we crawled each URL daily during our 10 week study period. The full HTML text of the page was collected and stored for each retrieved version.

4. TYPES OF CHANGE

In this section we characterize the change in query popularity, document content, and query intent to our 100 queries, and provide detailed analyses of the relationship between the three different measures of change. Analysis is performed at the granularity level of a day.

4.1 Changes to Query Popularity

We begin by looking at changes to how often a query was issued to the search engine on a given day (i.e., the query's frequency). This value can change over time. We normalize the raw query frequency with the overall query volume of the day in order to isolate the trends in query frequency from those caused merely by a change in overall query volume. We refer to this normalized query frequency as *query popularity*.

Understanding the patterns in query popularity can help us group queries with similar patterns and enable the search experience to be tailored to different query groups. For example, a search engine might add news content to search queries that are spiking in frequency. Additionally, changes to query popularity over time as they relate to changes to other components of the search experience (such as document content and relevance) can inform us about the searcher's underlying goals. In subsequent sections we explore these relationships. Here we focus on patterns of change to query popularity in isolation.

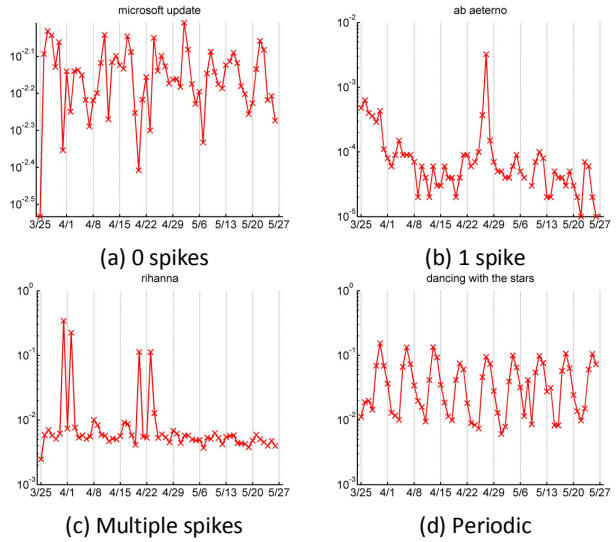

Figure 1. Different queries had different numbers of spikes in query popularity during the study period.

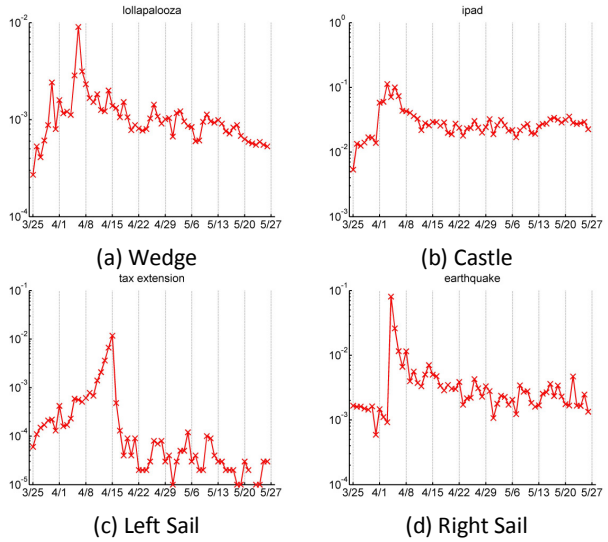

Figure 2. When a query spiked in popularity, the spike could occur in a variety of different shapes.

4.1.1 Measures of Query Popularity Change

We characterize the distribution of query popularity along four dimensions: the number of spikes, the shape of the spikes, the query's periodicity, and the overall trend in popularity. This classification scheme was iteratively developed using an affinity diagramming technique [7]. The 100 queries were categorized in a two-phase process, first analyzing all of the query popularity shapes to develop a categorization scheme, and then re-analyzing all queries to assign them to a category.

Number of spikes (0, 1 or multiple): A spike occurs when there is a sudden increase followed by a corresponding decrease in query popularity. This attribute captures the number of times such a change occurs during the 10 week study period. Figure 1 (a, b, c) gives examples of queries with 0, 1 and multiple spikes. The three temporal query profiles proposed by Jones and Diaz [19] correspond to the above three possible values (0, 1, or M) of this query popularity feature.

Periodicity (yes, no): A consistent repetitive pattern of spikes during the time-frame of the study is considered a periodicity. Figures 1 (d) shows an example of a periodic query. Most other queries shown, Figures 1 (a, b, c), are not periodic.

Shape (wedge, sail, castle): When a query spikes, the spike can have one of the following shapes.

Wedge: The popularity rises over time at the same rate that it later falls off. (See Figure 2 (a).)

Castle: The popularity changes (rises or drops), and stays at the new level for a relatively long period of time (roughly a few weeks or more). (See Figure 2 (b).)

Sail: The query's popularity rises somewhat slowly (roughly over a week) and then dramatically drops off over a short period of time (roughly 1-2 days) or conversely rises sharply but drops off slowly after reaching the peak popularity. (See Figure 2 (c, d).)

Queries that did not spike or that had multiple spikes that exhibited inconsistent patterns in shape were not classified according to spike shape. An example can be seen in Figure 1 (a).

Trend (up, down, flat, up-down): The query popularity can exhibit an overall increase or decrease over the duration of the study. The trend attribute encodes this property of popularity. Examples are given in Figure 3, showing downward, upward, flat and up-down trends.

4.1.2 Understanding Query Popularity Change

We now look more deeply at the groups of queries that emerged from the above classification of popularity, focusing first on the number of spikes and the spike shapes, next on periodicities, and finally on overall trends in popularity.

During the time-frame of this study, 10% of the queries never spiked, 47% spiked once, and 43% spiked multiple times. A majority of the queries that did not spike could be characterized as broad, general queries, such as *church*, *congratulations* and *wedding*. For these queries, in the absence of an emergent focused intent, the query popularity stayed relatively flat. The queries that spiked once were more focused or clearly defined. Many of these queries referred to an event that repeats every year such as *earth day*, *march madness* and *april fools day*. Others

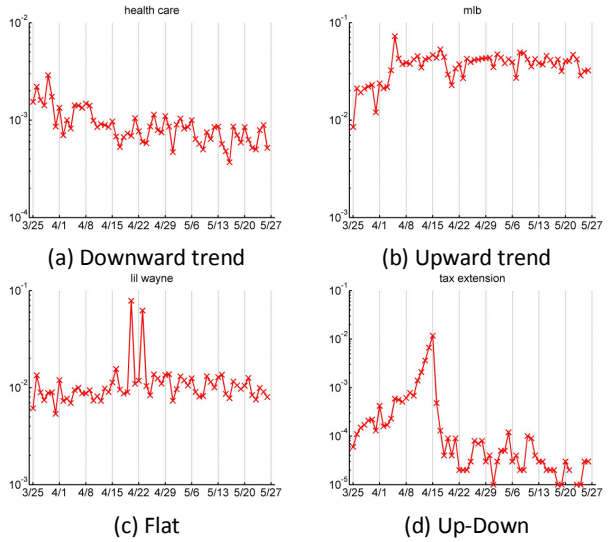

Figure 3. Different trends in query popularity. (y-axis in log scale)

referred to an event that occurred during our study time frame but does not necessarily recur, such as *ipad, health care, kate middleton.* The queries that spiked multiple times were related to television series (e.g., *american idol*, *crystal bowersox*), celebrities (e.g., *beyonce, lady gaga)*, or news (e.g., *iran, japan*). These queries with multiple spikes appear especially important since they account for majority of the query volume. Although they account for only 43% of the unique queries we studied, they account for 75% of the total query volume in our sample.

More than half (54%) of the queries that spiked exhibited wedge-shaped spike(s), while 15% were castle shaped, 11% were sail shaped, and 20% could not be associated with any particular shape. Queries that referred to events or concepts that spanned a relatively longer period of time during the study, such as *march madness, health care,* and *final four*, typically had castle-shaped spikes. Queries based on information needs centered around a particular date (e.g., *tax questions* or *april fools day pranks*) rose gradually as the date approached and then fell sharply, resulting in a left-sail shape. In contrast, queries expressing newly and unexpectedly emerging information needs (e.g., *earthquake*) spiked quickly and then fell off in popularity more slowly in a right-sail spike. However, for most spiking queries, the query popularity rose and fell symmetrically around events of interest (e.g., celebrities: *lady gaga*, television personalities: *crystal bowersox*, television shows: *american idol*, and short-span events: *marathon, lollapalooza*), resulting in a wedge-shaped spike.

While 88% of the queries we studied were not periodic, the remaining 12% were. Examples of aperiodic queries include *carrie underwood* and *justin bieber*, and examples of periodic queries include *american idol* and *pgatour*. Most periodic queries were associated with regularly occurring events such as television serials and sports events. About 79% of the query volume in our query log data comes from aperiodic queries and the remaining 21% comes from periodic queries. Note that celebrity queries like *beyonce* and *jessica simpson* tended to generate sporadic, spiky search traffic, while television shows like *dancing with the stars* triggered surges in search traffic at predictable intervals

Many queries exhibited general temporal trends over the study period; 36% of the queries exhibited an overall upwards trend in query popularity over the duration of the study, 9% showed a downward trend, 42% remained level, and 13% underwent a up-down trend. Queries that referred to new or popular events such as the release of the ipad (*ipad)*, the start of the major league baseball season (*mlb)*, the Miss USA pagent (*miss usa)* exhibited increasing query popularity over time until the day of the event. On the other hand, a query about an American Idol contestant who got eliminated became less popular over time after the elimination (e.g., *andrew garcia*). Many celebrity queries generated relatively flat traffic, excluding sporadic spikes. Other queries, such as *easter ideas* and *final four*, exhibited a downwards trend after the event date, while queries such as *tax extension* and *cma awards* demonstrated an up-down trend where the interest in the query slowly rose until the event date and then fell off after it.

While classifying attributes of query popularity tells us valuable information about changes in user interest, query popularity alone does not fully indicate the intent of the query. For some queries, a spike may occur because people are suddenly interested in finding older information about the query. For example, people who started following American Idol in 2010 might have become interested in old pages about Adam Lambert (a contestant from last year) once he appeared as a mentor this year (query: *adam lambert*). Alternatively, a query may spike because there is a new interpretation for the query and thus new information and content will be available on the Web. For example, the query *ab aeterno* (a term meaning "from the most remote antiquity") acquired a new and a more popular meaning when an episode of the television series 'Lost' was so named. In response to this event, many new pages were created and several existing pages underwent significant changes. A query might also spike at different time points in response to very different events. For example, the query *opening day* spikes during early April due to increased interest in the baseball season, whereas the same query spikes during early May in response to increased interest in the boating season. For such temporally ambiguous queries analyzing the temporal changes in results content can assist in inferring the query intent more accurately. To better understand these differences, we begin by looking at how changes to results content relate to changes in query popularity.

4.2 Changes to Results Content

Although the majority of the content on the Web does not change, the high quality pages, especially those that cater to time-sensitive events and are revisited often, undergo changes frequently [1, 17]. As the content on the Web pertaining to a particular information need evolves, new relevant pages are created, old pages die, and others are updated. For example, after the release of the iPad (query: *ipad*) in April, PC Magazine ran an article about it, effectively creating a new high-quality webpage that did not exist previously. On the other hand, a webpage hosting a forum about tax filing related questions during the tax filing season was discontinued after April 20th. And many existing pages neither come nor go, but rather are updated in response to some event. For example, several pages covering the American Idol *crystal bowersox* underwent substantial change after she was eliminated.

4.2.1 Measures of Result Content Change

Previous work in the area of content change has focused on measuring textual differences between subsequent versions of Web documents by, for example, calculating the differences between blocks of text [17] or word frequencies [1, 23]. We employ two measures of page content change to understand the degree of result change for each query over time:

- **A query *dependent* measure**, based on the term frequency (TF) of the query on the page over the study period, as a measure of how query relevant the document is over time.
- **A query *independent* measure**, based on the Dice coefficient, as a measure of how much the overall page content changes over time.

Like Adar et al. [1], we use the Dice coefficient to represent the amount of textual change. The Dice coefficient, which measures the overlap in text between various document versions, allows us to develop a high-level model of page change over time:

$$Dice(W_i, W_j) = \frac{2|W_i \cap W_j|}{|W_i| + |W_j|},$$

where W_i and W_j are sets of words for the document at time i and j respectively. A high Dice coefficient (i.e., 1) reflects high similarity and vice-versa.

We use the daily crawl data described in Section 3.2.3 to compute the above measures for each query. To aggregate daily document-level term frequency change to the query-level, we follow a two-step process. In the first step, we create an aggregate measure of how much the query TF values change in a document. We do this by computing the average and standard deviation of the daily

Table 2. Correlation between two measures of page content change, one based on page content (*Change in Dice*) and the other on query term occurrence (*Change in TF*). Significant differences ($p < .05$) are shaded.

	Change in Dice	
	Average	Median
Change in TF	**0.38**	**0.79**

values for each {query, URL} pair, and then combining the two values ($100 \times std\,dev/avg$) to obtain a single number representing its TF change. In the second step, we combine the TF change values for each URL for the query into a query-level mean and median, which we call the *change in TF* value for a query. We also apply the above two-step aggregation process to create a measure of daily document-level content change using the Dice similarity score between versions of the document from consecutive days.

In the following section, we describe the amount of observed change for our queries and URLs, in terms of both term frequency and overall content change, and then describe the relationships discovered between query popularity features and content change.

4.2.2 Understanding Result Content Change

Many of the documents we tracked showed little change in their relationship to the query over our 10 week study, as measured by the change in TF; 39% exhibited no query-related change. In some cases this was because the page content did not change, but not all of these pages were entirely static. Only 16% of all the documents showed no overall change as measured by Dice. For the remaining 23% the content changed, but the TF overlap with the query remained the same. The largest overall change in content, as measured by Dice was 30%, however 95% of all the documents underwent less than 15% overall change in content.

Although our query-dependent TF measure and query-independent Dice measure capture related information about content change, the two measures examine somewhat different subsets of page content. Table 2 shows the correlation between the two measures, and we see a strong positive correlation, especially for the median measure. We analyze the relationships between these two measures and the query popularity features in the next section.

4.2.3 Query Popularity and Results Content

For the analyses in this section, we divide our query set using the query popularity features discussed in Section 4.1.1. The resulting query groups are then compared in terms of the amount of content change that each undergo. The results can be found in Table 3. We perform statistical analyses using a two-tailed *t*-test at the 95% confidence level. Significant values are shaded with the significant pairing listed in brackets.

One trend that emerges is that the more a query spiked, the more likely its content was to change, in terms of both the occurrence of query terms and overall content change. Table 4 shows examples of queries that follow this trend as well as exceptions. For example, the query *church* neither spiked nor had a lot of change to associated pages, while the query *mlb* displayed more than twice as much change in both the measures, presumably because the baseball season began during our study period. However, there were some notable exceptions to this trend. Some queries exhibited relatively low content change despite spiking significantly in popularity. These queries differed from the typical spiking queries that exhibited high content change in that they related to annual events like Easter (*hard boiled eggs*) or the

Table 3. Relationships between query popularity features and measures of result content change. Significant differences ($p < .05$) are shaded.

	Changes in TF		Changes in Dice	
# Spikes	Average	Median	Average	Median
0 (10%)	5.26	**[M] 1.70**	**[M] 2.04**	**[M] 1.15**
1 (47%)	7.52	2.95	**[M] 3.01**	1.80
M (43%)	8.00	**[0] 3.47**	**[0,1] 4.12**	**[0] 2.54**
Shape				
Castle (15%)	7.90	2.67	3.12	2.06
Sail (11%)	9.88	**[W] 1.07**	**[W] 2.24**	1.06
Wedge (54%)	7.58	**[S] 3.90**	**[S] 3.94**	2.45
Periodicity				
No (88%)	7.23	2.88	**[Y] 3.19**	**[Y] 1.83**
Yes (12%)	9.72	4.44	**[N] 4.98**	**[N] 3.83**
Trend				
Down (9%)	11.31	3.08	3.88	2.04
Flat (42%)	7.59	3.78	**[UD] 3.83**	**[UD] 2.60**
Up (36%)	7.53	**[UD] 3.58**	**[UD] 3.67**	2.28
Up-Down (13%)	7.52	**[U] 1.43**	**[U,F] 2.43**	**[F] 1.04**

April 15 US tax day (*taxes online*) (see Table 4) (many sail-shaped queries). In these cases, interest in the topic was not new – it was yearly recurring, and old content remained relevant. The other type of exception were the queries that did not spike during the period of the study but were associated with pages that underwent significant changes (*john mayer, apple*). One possible explanation for this trend is that when analyzing the query log data we counted only exact matches to the query, we did not fold-in the occurrences of *synonymous* queries. For example, we did not count *john mayer musician* or *john mayer singer* and *apple computers* when we analyzed the query logs. This trend could also indicate a 'zoom-in' behavior of the query intent (Section 5) where a new event related to the original query (release of a new album by john mayer or launch of a new product by apple) might have triggered the changes in the document content but also could have changed the original query to include more specific terms (*apple ipad*).

In terms of the shape of the spike, queries that spiked symmetrically in a wedge shape (*acm*, *brad paisley, ellen*) exhibited more change in document content than queries that spiked asymmetrically in a sail shape (*earth day, tax extension, earthquake*). This may arise from the fact that change is correlated with spikes (previous paragraph), and with wedge-shaped spikes, content changes both before and after the event of interest, while with sail-shaped spikes, it only happens on one edge. It may also relate to the fact that many low-content change spiking queries tended, as described earlier, to be sail-shaped queries.

With respect to a query's periodicity, we saw a strong significant trend that periodic queries change more in overall content than aperiodic queries. Periodic queries were often related to popular television shows or sports events and were associated with highly dynamic content (Table 5). Although the same general trend holds, the difference in the query term occurrences on the page was not significantly affected by periodicity. This may be because results for periodic queries remained on the query topic and only provided updates. Aperiodic queries that underwent substantial content change tended to be celebrity queries which

Table 4. Examples of queries that did/did not spike and the corresponding amount of content change.

		Content Change	
		Low	**High**
Spikes	**No**	*leaf* *church* *microsoft update*	*john mayer* *apple*
	Yes	*hard boiled eggs* *easter basket* *taxes online* *easter bunny*	*lost* *robert pattinson* *pgatour* *mlb*

typically trigger content change following aperiodic events that the celebrity is involved in.

The queries for which the overall popularity was an up-down trend (*april fool*, *augusta*, *cma awards*) were associated with low content change while the queries for which the overall popularity either rose (*glee*, *japan*, *justin bieber*) or stayed flat (*adam lambert*, *giada de laurentis*) exhibited higher change in content. One p‹ ible explanation for this pattern could be that up and flat queries tend to have many different aspects associated with them while up-down queries have more focused intents. As a result, the volume of content change for up and flat queries is more than that for up-down queries.

4.3 Changes to Query Intent

Given the understanding we have built of patterns of query popularity change and result content change, we turn now to how these two types of change interact with changes in query intent.

4.3.1 Measures of Query Intent Change

We measure query intent in two ways: one based on human-assessed relevance judgments, and the other based on search result click-through behavior.

To understand change to the **human-assessed relevance judgments** for a query, we define the top human-rated count, or *top HR count*. We count the number of results (URLs) with the highest relevance judgment assigned for that query by the human assessor on a given week. Note that this does not mean the results were given the highest rating ("perfect"); if the best judgment given by the assessor for the query was "good", than the number of good results are counted.

The rationale behind this metric is to capture the spread or diversity of the judged query intent. For example, after the 'Lost' episode titled *ab aeterno* was aired the number of intents for query *ab aeterno* increased. Thus we would expect to (and do) see the values for top HR count to go up after the episode. The metric can also capture changes to the overall result quality, since the higher the top rating, the fewer the number of results that are typically given that rating. For example, if an assessor's top rating is "perfect," there are usually very few results labeled "perfect," while if the assessor's top rating is "good," there are more URLs labeled "good."

We also developed a measure that looked at changes in the top human-rated URLs around the time of an event. Specifically we examined the overlap in the top human-rated URLs during and after an event, a measure which we call *Ratio Same URLs*. This measure is possible to compute when an event occurs at one or two intervals during our observation period, which happens for 69% of our queries. Because of this more limited coverage, results for this measure are not available for all comparisons (e.g., no periodic queries have a single event period) and are not statistically significant for the other features, so we focus on the top HR Count measure here.

Table 5. Examples of periodic and aperiodic queries and the corresponding amount of content change.

		Content Change	
		Low	**High**
Periodic	**No**	*easter baskets* *taxes online* *church* *microsoft update* *easter bunny*	*robert pattinson* *miley cyrus* *lil wayne* *justin bieber* *kristen stewart*
	Yes	*new york lottery* *casey james*	*lost* *pgatour* *dancing with the stars*

Changes in judgments over time can also arise from differences in the top HR counts could also be an artifact of differences between the human assessors; these differences are hard to control and isolate. Therefore in addition to this measure we also use a metric based on user-behavior data that has been commonly used in previous work [14].

To capture the variation in **search result click-through behavior**, we use click entropy (CE). The click entropy for a query q measures the variance in the number of results a user clicks. A high $CE(q)$ indicates that many results were clicked for the query while a small $CE(q)$ value indicates fewer clicks and higher user agreement on which pages are relevant. $CE(q)$ is computed as:

$$CE(q) = -\sum_{u \in P_c(q)} p_c(u|q) * \log p_c(u|q),$$

where $P_c(q)$ is the collection of URLs clicked on for query q and $p_c(u|q)$ is the percentage of clicks on URL u among all clicks for query.In addition to the number of results clicked, the click entropy can be influenced by a number of other factors, including how much the results presented for the query change, and the quality of the results [25]. However, because there was relatively little variation among these factors for our queries as compared to typical Web queries, they do not appear to significantly impact click entropy in our case and thus were not accounted for explicitly.

These two measures of query intent (i.e., the top human rated count and click entropy) are negatively correlated (see Table 6). Example queries that have small top HR count but exhibit high click entropy are *tiger woods, ipad*, and *carrie underwood.* Each of these queries has a single or few authoritative websites which have been assigned the highest rating by the human assessors. However, user behavior data shows that each of these queries is *informational*, that is, people explore more than just the authoritative page. The negative correlation between these two metrics underscores the disconnect that is often observed between the human annotator's interpretation of the query intent and the actual query intent of searchers.

The negative correlation between the two measures is also reflected in the relationships seen in the results in Table 7. We see that the two measures do not exhibit significant trends for the same query popularity features; the directions of the trends are in fact opposite. We now look more closely at the relationships described in Table 7 before turning to the relationships between changes in query intent and changes in document content.

Table 6. Correlation between two measures of change in query intent, including click entropy and the number of top rated results. Significant differences ($p < .05$) are shaded.

	Click Entropy	
	Average	**Median**
top HR Count	**-0.28**	**-0.35**

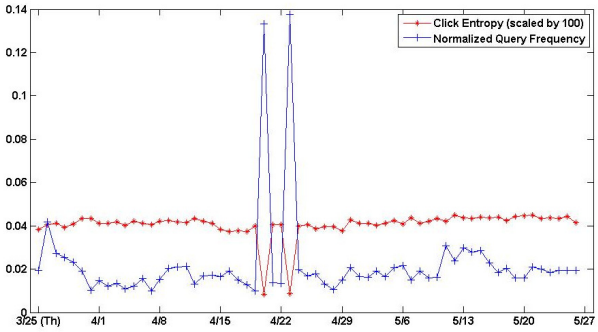

Figure 4. Normalized query frequency and click entropy for the query *lady gaga*.

4.3.2 Query Popularity and Query Intent

Here we look at the relationships between the different features of query popularity and query intent. The analysis procedure is similar to the one described in Section 4.2.3. We measure differences using a two-tailed *t*-test at the 95% significance level.

The queries that spike in popularity once during the time frame of our study experience significantly lower click entropy than those that spiked multiple times. One can think of each popularity spike as an event that generates a new query intent for the same query (though not necessarily unrelated to the previous intent). Naturally, a query that spikes multiple times has a more diverse query interpretation than the one that spikes only once. This diversity is reflected in the user behavior where the diversity of results visited is higher for spiky queries than for flat queries.

Interestingly, celebrity or celebrity-like queries such as *lady gaga, red sox, madonna* exhibit a contrary behavior. When the query popularity spikes (which happens at multiple time points), the click entropy drops significantly (Figure 4). A possible interpretation of this phenomenon could be that these queries, which are in general informational queries (most of these queries have high click entropy (~5) for the majority of the time frame), change to navigational queries during the spike since most users converge to viewing only a few pages that provide information about the breaking news that caused the spike.

The number of results that are assigned top ratings by human-assessors is significantly smaller for the set of queries that exhibit periodic rises and falls in popularity than those that are aperiodic. Many of the periodic queries are about television shows (*american idol, pgatour*) or are related to a periodic event (*crystal bowersox, american idol results today)* and typically have an authoritative webpage which leads to a smaller average top HR count (Table 8).

The click entropy for the set of queries that showed an overall downwards trend in popularity (*andrew garcia*) was significantly lower than the click entropy for the set of queries that exhibit an overall steady trend in popularity (*microsoft update*). Queries that are towards the end of their lifecycle may be less likely to invoke diverse information consumption needs than queries that are still in the prime of their lifecycle.

Table 7. Relationships between query popularity features and measures of change in query intent. Significant differences ($p < .05$) are shaded.

	top HR Count		Click Entropy	
# Spikes	**Average**	**Median**	**Average**	**Median**
0 (10%)	4.63	4.00	2.85	2.83
1 (47%)	3.46	3.08	**[M] 2.93**	**[M] 2.83**
M (43%)	2.56	2.15	**[1] 3.50**	**[1] 3.51**
Shape				
Castle (15%)	3.59	3.29	2.89	2.70
Sail (11%)	4.22	3.65	**[W] 2.46**	**[W] 2.30**
Wedge (54%)	2.66	2.21	**[S] 3.48**	**[S] 3.50**
Periodicity				
No (88%)	**[Y] 3.34**	2.91	3.15	3.10
Yes (12%)	**[N] 1.89**	1.59	3.33	3.36
Trend				
Down (9%)	3.33	2.63	**[F] 3.13**	**[F] 3.16**
Flat (42%)	2.74	2.29	**[D] 3.60**	**[D] 3.64**
Up (36%)	2.97	2.56	3.18	**[UD] 3.16**
Up-Down (13%)	3.71	3.21	2.50	**[U] 2.37**

Queries with sail-shaped popularity spikes (*taxes online, pgatour*) lead to significantly lower click entropy than the queries with wedge-shaped popularity spikes (*robert pattinson, kristen stewart*), potentially because the timespan for which the user's interest in the topic is heightened is longer for the wedge than for the sail. It could also be that wedge-shaped queries are more informational than sail-shaped queries.

4.3.3 Query Popularity, Result Content, and Intent

In this section we bring it all together and study the relationships between query popularity, result content and query intent. The analysis between result content and query intent was performed using Pearson's correlation coefficient. Statistical significance was computed at the 95% significance level.

One noticeable relationship is that both, change in query term frequency and change in Dice are significantly negatively correlated with the top HR counts (Table 9). That is the queries for which result content changes are higher are also the queries for which fewer results are assigned top ratings. We saw in the previous section that queries with periodic popularity are associated with lower top HR counts (Table 7). One interpretation

Table 8. Examples of periodic and aperiodic queries and the corresponding number of top rated results.

		top HR Count	
		Low	**High**
Periodic	**No**	*beyonce* *carrie underwood* *cassie* *cma awards* *espn fantasy baseball*	*april fools day pranks* *congratulations* *easter baskets* *time machine*
	Yes	*american idol* *nascar* *pgatour*	*orange* *cassie james*

Table 9. Correlation between measures of change in query intent (*top HR Count*, *Click Entropy*) and change in result content (*Change in TF*, *Change in Dice*). Significant differences ($p < .05$) are shaded.

	Change in TF		Change in Dice	
	Average	**Median**	**Average**	**Median**
top HR Count	-0.16	**-0.41**	**-0.40**	**-0.35**
Click Entropy	0.15	**0.48**	**0.38**	**0.31**

of this trend is that pages that change often are likely to be related to a periodic event such as a weekly television series which have a single or a few authoritative pages that are assigned high relevance score.

$$\text{Periodicity} \propto \frac{1}{\text{top HR Count}} \propto \Delta \text{ Content}$$

Secondly, we know from the analysis in Section 4.2.2 that more spikes in popularity is related to higher change in result content (Table 3). In the previous section, we noted that more spikes is also related to higher click entropy (Table 7). Thus we speculate that changes in result content would positively correlate with click entropy. The statistical analysis provides empirical evidence in support of this hypothesis – the percentage change in Dice is positively correlated with click entropy (Table 9).

$$\#\text{Spikes} \propto \text{Click Entropy} \propto \Delta \text{ Content}$$

We see a similar relationship for queries with flat and up-down trends. Queries that exhibit flat overall trend in query popularity are associated with large change in content and also with high click entropy, whereas queries that trend up and down correlate with low content change and low click entropy.

$$\text{Trend} \propto \text{Click Entropy} \propto \Delta \text{ Content}$$

Wedge- and sail-shaped queries exhibit a similar relationship; wedge-shaped queries correlate with high content change and high click entropy, while sail-shaped queries are associated with pages that undergo little change and low click entropy.

$$\text{Shape} \propto \text{Click Entropy} \propto \Delta \text{ Content}$$

Recall that our TF-based measure of content change only counts exact occurrence matches of the query. As a result, the TF counts can be relatively small, for example, *ncaa.com,* which is one of the authoritative websites for the query *march madness,* especially during the tournament dates, exhibits a maximum term frequency of 8. The two-step aggregation method used to compute the query-level change in content (described in Section 4.2) can wash-out temporal content change trends if only a few URLs for the query exhibit them. This could be one of the reasons why our query-dependent measure does not show as much correlation with the other metrics as the query-independent measure.

5. DISCUSSION

We have explored the relationship between three measures of query dynamics: query popularity changes, result content changes, and query intent changes. We have identified features by which changes in query popularity can be classified, including spikiness, spike shape, periodicity, and overall trend. The presence of some of these attributes (for example, wedge shaped spikes) when accompanied by changes in content can be a good indicator of change in query intent. In addition, for most queries, a rise in query popularity is accompanied by the query becoming or remaining *informational.* We observe, based on our findings, three general types of query intent dynamics:

- **Zoom:** Query intent zooms in on to the current event around the time frame of the event and then zooms out post-event. Figure 5 shows an example in which the intent behind the query *final four* becomes focused on different intents at different points in time. The tickets (click counts for http://four-tickets.com/ increase just before the tournaments), videos of live matches and scores (http://mmod.ncaa.com/ becomes popular destination during the tournaments) are more important than general information and historic data about *final four* at the start of April every year, whereas historic data is more relevant other times in the year.
- **Shift:** Query intent can undergo a shift. For example, the query *opening day* indicates an information need about the baseball opening season when it is issued at the start of April, however, by the start of May the query intent shifts to the opening season of boating events.
- **Static:** Query intent can remain relatively static, for example for queries such as *easter ideas* and *tax questions*.

Our findings suggest that these three different query intent types should be addressed by a search engine in different ways. For zoom queries that represent an increase in focus, people are likely interested in breaking news on the query topic. In these cases, news stories [11] can be interjected and results can be biased toward newly available pages or new content on existing pages [13]. Search engines may also actively crawl [24] and look for Web content related to zoom queries. How the content has changed, can also be useful. Elsas and Dumais [16] found stable content is valuable for navigational queries. Our findings suggest dynamic content may be useful for spiking queries.

The manner in which people zoom into a query may suggest the best way to handle the change in focus, as we saw the shape of the spike impacted intent and changes in result content. Left-sail queries and wedge queries are ones with rising user interest. The search system could use this information to proactively improve the search experience for such queries as the query is gaining popularity. In contrast, right-sail queries are losing popularity and search engines can use this knowledge to reduce the crawl frequency for pages related to these queries, in particular because sail queries had less change in intent and less content change than wedge queries.

For static queries, where both the document content and the query intent stay constant over time, we can use significant amounts of long-term information to provide the best search results. Previous page content can be used to understand what the page is about [1, 16]. Past behavioral information remains relevant for static queries, and can be used to better determine the most relevant pages. In some cases, it is worth looking for long-term periodicities (e.g., annual periodicities) and taking advantage of behavior from the appropriate corresponding time even when that time is in the distant past.

6. CONCLUSION

In this paper, we have seen that Web search is strongly influenced by time. The queries people issue change over time, with queries having periodicities, trends, and spikes (of different shapes) in popularity. The content of documents also change with some documents always being relevant to a particular query and others being relevant to it at a particular point in time. The relationship between documents and queries can also change as people's intent changes. We have explored how queries, their associated documents, and query intents change over the course of 10 weeks by analyzing large scale query log data, a daily Web crawl, and

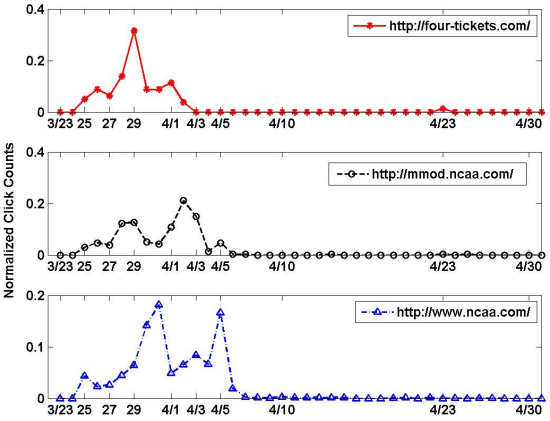

Figure 5. Normalized click count for 3 URLs associated with query *final four*

periodic human relevance judgments. We identified several interesting features by which changes to query popularity can be classified, and showed that the presence of several of these features, when accompanied by changes in result content, could be a good indicator of change in query intent.

In future work, we plan to exploit the relationships we have observed to develop a search algorithm that uses the term history in a document to identify the most relevant documents. We also plan to use temporal patterns to help people make better relevance decisions by creating temporally aware search result page representations (such as snippets) for queries where such representations would be appropriate.

7. REFERENCES

[1] Adar, E., Teevan, J., Dumais, S. and Elsas, J. The Web changes everything: Understanding the dynamics of Web content. In *Proceedings of WSDM 2009*, 282-291.

[2] Adar, E., Teevan, J. and Dumais, S. T. (2009). Resonance on the web: Web dynamics and revisitation patterns. In *Proceedings of CHI 2009*, 1381-1390.

[3] Adar, E., Weld, D., Bershad, B., and Gribble, S. Why we search: Visualizing and predicting user behavior. In *Proceedings of WWW 2007*, 161-170.

[4] Alfonseca, E., Ciaramita, M. and Hall, K. Gazpacho and summer rash: Lexical relationships from temporal patterns of Web search queries. In *Proceedings of EMNLP 2009*, 1046-1055.

[5] Alonso, O. and Gertz, M. Clustering of search results using temporal attributes. In *Proceedings of SIGIR 2006*, 597-598.

[6] Beitzel, S. M., Jensen, E. C., Chowdhury, A., Grossman, D. and Frieder. Hourly analysis of a very large topically categorized Web query log. In *Proceedings of SIGIR 2004*, 321-328.

[7] Beyer, H. and Holtzblatt, K. *Contextual Design: Defining Customer-Centered Systems*. Academic Press, 1998.

[8] Chien, S. and Immorlica, N. Semantic similarity between search engine queries using temporal correlation. In *Proceedings of WWW 2005*, 2-11.

[9] Cho, J., Roy, S. and Adams, R. E. Page quality: In search of an unbiased Web ranking. *SIGMOD 2005*, 551-562.

[10] Dakka, W., Gravano, L. and Ipeirotis, P. G. Answering general time sensitive queries. In *Proceedings of CIKM 2008*, 1437-1438.

[11] Diaz, F. Integration of news content into Web results. In *Proceedings of WSDM 2009*, 182-191.

[12] Diaz, F. and Jones, R. Using temporal profiles of queries for precision prediction. In *Proceedings of SIGIR 2004*, 18-24.

[13] Dong, A., Chang, Y., Zheng, Z., Mishne, G., Bai, J., Buchner, K., Zhang, R., Liao, C. and Diaz, F. Towards recency ranking in Web search. In *Proceedings of WSDM 2010*, 11-20.

[14] Dou, Z., Song, R. and Wen J. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of WWW 2007*, 581-590.

[15] Efron, M. Linear time series models for term weighting in information retrieval. *JASIST, 61(7):*1299-1312, 2010.

[16] Elsas, J. and Dumais, S. T. Leveraging temporal dynamics of document content in relevance ranking. In *Proceedings of WSDM 2010*, 1-10.

[17] Fetterly, D., Manasse, M., Najork, M. and Wiener, J. A large-scale study of the evolution of Web pages. In *Proceedings of WWW 2003*, 669-678.

[18] Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*, 1012-1014, February, 2009.

[19] Jones, R. and Diaz, F. Temporal profiles of queries. *TOIS 25(3)*, 2007.

[20] Kleinberg, J.. Bursty and hierarchical structure in streams. In *Proceedings of KDD 2002*, 91-101.

[21] Kleinberg, J. Temporal dynamics of on-line information systems. In *Data Stream Management: Processing High-Speed Data*. Springer, 2006.

[22] Li, X. and Croft, W. B. Time-based language models. In *Proceedings of CIKM 2003*, 469-476.

[23] Ntoulas, A., Cho, J. and Olston, C. What's new on the Web? The evolution of the Web from a search engine perspective. In *Proceedings of WWW 2004*, 1-12.

[24] Olston, C. and Pandey, S. Recrawl scheduling based on information longevity. In *Proceedings of WWW 2008*, 437-446.

[25] Teevan, J., Dumais, S.T. and Liebling, D.J. To personalize or not to personalize: Modeling queries with variation in user intent. In *Proceedings of SIGIR 2008*, 163-170.

[26] Vlachos, M., Meek, C., Vagena, Z. and Gunopulos, D. Identifying similarities, periodicities and bursts for online search queries. In *Proceedings of SIGMOD 2004*, 131-142.

[27] Wang, S., Berry, M. W. and Yang, Y. Mining longitudinal Web queries: Trends and patterns. *JASIST, 54(8): 743-758*, 2003.

[28] Zhang, R., Chang, Y., Zheng, Z., Metzler, D. and Nie, J.-Y. Search result re-ranking by feedback control adjustment for time-sensitive query. In *Proceedings of NAACL 2009*, 165-168