Context Comparison of Bursty Events in Web Search and Online Media

Yunliang Jiang
University of Illinois
Urbana, IL, 61801
jiang8@illinois.edu

Cindy Xide Lin
University of Illinois
Urbana, IL, 61801
xidelin2@illinois.edu

Qiaozhu Mei
University of Michigan
Ann Arbor, MI, 48109
qmei@umich.edu

Abstract

In this paper, we conducted a systematic comparative analysis of language in different contexts of bursty topics, including web search, news media, blogging, and social bookmarking. We analyze (1) the content similarity and predictability between contexts, (2) the coverage of search content by each context, and (3) the intrinsic coherence of information in each context. Our experiments show that social bookmarking is a better predictor to the bursty search queries, but news media and social blogging media have a much more compelling coverage. This comparison provides insights on how the search behaviors and social information sharing behaviors of users are correlated to the professional news media in the context of bursty events.

1 Introduction

Search is easy. Every day people are repeating the queries they have used before, trying to access the same web pages. A smart search engine tracks the preference and returns it next time when it sees the same query. When I search for "msr" I always try to access *Microsoft research*; and even if I misspelled it, a smart search engine could suggest a correct query based on my query history, the current session of queries, and/or the queries that other people have been using.

Search is hard. I search for "social computing" because there was such a new program in NSF; but the search engine might have not yet noticed that. People use "msg" to access *monosodium glutamate* in most of the cases, but tonight there is a big game in *Madison square garden*. *H1N1* suddenly became a hot topic, followed by a burst of the rumor that *it was a hoax*, and then the *vaccine*. The information need of users changed dramatically during such a period. When a new event happens, the burst of new contents and new interests make it hard to predict what people would search and to suggest what queries they should use.

Web search is easy when the information need of the users is stable and when we have enough historical clicks. It becomes much more difficult when a new information need knocks the door or when there is a sudden change of the information need. Such a shift of the information need is usually caused by a burst of new events or new interests.

When we are lack of enough historical observations, why don't we seek help from other sources? A bursting event will not only influence what we search, but hopefully also affect *what we read*, *what we write*, and *what we tag*. Indeed, there is already considerable effort in seeking help from these sources, by the integration of news and blogs into search results or the use of social bookmarks to enhance search. These conclusions, however, are mostly drawn in a general context (e.g., with general search queries). To what extent are they useful when dealing with busty events? How is the bursting content in web search, news media, social media, and social bookmarks correlating and different from each other? Prior to the development of desirable applications (e.g. enhancing search results, query suggestion, keyword bidding on advertisement, etc) by integrating the information from all these sources, it is appealing to have an investigation of feasibility.

In this work, we conduct a systematic comparative study of what we search, what we read, what

Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pages 1077–1087,
MIT, Massachusetts, USA, 9-11 October 2010.

we write, and what we tag in the scenarios of *bursty events*. Specifically, we analyze the language used in different *contexts* of bursty events, including two different query log contexts, two news media contexts, two blog contexts, and an additional context of social bookmarks. A variety of experiments have been conducted, including the content similarity and cross-entropy between sources, the coverage of search queries in online media, and an in-depth semantic comparison of sources based on language networks.

In the rest of this paper, a summary of related work is briefly described in Section 2. We then present the experiments setup in Section 3, The results of the experiments is presented in Section 4. Finally, our major findings from the comparative analysis are drawn in Section 5.

2 Related Work

Recently, a rich body of work has focused on how to find the bursting patterns from time-series data using various approaches such as time-graph analysis (Kleinberg, 2003; Kuman et al., 2003), context-based analysis (Gabrilovich et al., 2004), moving-average analysis (Vlachos et al., 2004), and frequency analysis (Gruhl et al., 2005), etc. These methods are all related to the preprocessing step of our analysis: detecting bursty queries from the query log effectively.

The comparison of *two* web sources at a time is widely studied recently. (Sood et al., 2007) discussed how to leverage the relation between social tags and web blogs. (Lloyd et al., 2006; Gamon et al., 2008; Cointet et al., 2008) investigated the relations between news and blogs. Also some work has aimed to utilize one external web source to help web search. For example, (Diaz, 2009) integrated the news results into general search. (Bao et al., 2007; Heymann et al., 2008; Krause et al., 2008; Bischoff et al., 2008) focused on improving search by the social tags. Compared with the above, our comparison analysis tries to explore the interactions among *multiple* web sources including the search logs.

Similar to our work, some recent work (Adar et al., 2007; Sun et al., 2008) has addressed the comparison among multiple web sources. For example, (Adar et al., 2007) did a comprehensive correlation study among queries, blogs, news and TV results. However, different from the *content-free* analysis above, our work compares the sources based on the *content*.

Our work can lead to many useful search applications, such as query suggestion which takes as input a specific query and returns as output one or several suggested queries. The approaches include query term cooccurrence (Jones et al., 2006), query sessions (Radlinski and Joachims, 2005), and click-through (Mei et al., 2008), respectively.

3 Analysis Setup

Tasks of web information retrieval such as web search generally perform very well on frequent and *navigational* queries (Broder, 2002) such like "chicago" or "yahoo movies." A considerable challenge in web search remains in how to handle *informational* queries, especially queries that reflect *new* information need and *suddenly changed* information need of users. Many such scenarios are caused by the emergence of bursty events (e.g., "van gogh" became a hot query in May 2006 since a Van Goghs portrait was sold for 40.3 million in New York during that time). The focus of this paper is to analyze how other online media sources react to those bursty events and how those reactions compare to the reaction in web search. This analysis thus serves as an primitive investigation of the feasibility of leveraging other sources to enhance the search of bursty topics.

Therefore, we focus on the "event-related" topics which present as bursty queries submitted to a search engine. These queries not only reflect the suddenly changed information need of users, but also trigger the correlated reactions in other online sources, such as news media, blog media, social bookmarks, etc. We begin with the extraction of bursty topics from the query log.

3.1 Bursty Topic Extraction

Search engine logs (or query logs) store the history of users' search behaviors, which reflect users' interests and information need. The query log of a commercial search engine consists of a huge amount of search records, each of which typically contains the following information: the query submitted by

a user, the time at which the query was submitted, and/or the URL which the user clicked on after the query was submitted, etc. It is common practice to segment query log into search sessions, each of which represents one user's searching activities in a short period of time.

We explore a sample of the log of the Microsoft Live search engine[1], which contains $14.9M$ search records over 1 month (May 2006).

3.1.1 Find bursty queries from query log

How to extract the queries that represent bursty events? We believe that bursty queries present the pattern that its day-by-day search volume shows a significant spike – that is, the frequency that the user submit this query should suddenly increase at one specific time and drop down after a while. This assumption is consistent with existing work of finding bursty patterns in emails, scientific literature (Kleinberg, 2003), and blogs (Gruhl et al., 2005).

Following (Gruhl et al., 2005), we utilize a simple but effective method to collect bursty topics in the query log data as follows:

• We choose bigrams as the basic presentation of bursty topics since bigrams present the information need of users more clearly and completely than unigrams and also have a larger coverage in the query log comparing to n-grams ($n \geq 3$).

• We only consider the bigram queries which appear more frequently than a threshold s per month. This is reasonable since a bursty event usually causes a large volume of search activities.

• Let $f_{max}(q)$ be the maximum search volume of a query q in one day (i.e., day d). Let $\hat{f}_{-5}(q)$ be the *upper* bound of the daily search volume of q outside a time window of 5 days centered at day d. If $f_{max}(q)$ is "significantly higher" than $\hat{f}_{-5}(q)$ (i.e., $r_m = \hat{f}_{max}(q)/f_{-5}(q) > m$), we consider q as a query with a spike pattern (m is an empirical threshold).

• The ratio above may be vulnerable to the query that has more than one spike. To solve this, we define $\bar{f}_{-5}(q)$ as the *average* of daily search volume of q outside the same time window. This gives us an alternative ratio $r_a = f_{max}(q)/\bar{f}_{-5}(q)$. We further balance these two ratios by ranking the bursty queries using

$$score(q) = \alpha \cdot r_m(q) + (1 - \alpha) \cdot r_a(q) \quad (1)$$

By setting $s = 20$, $m = 2.5$, $\alpha = 0.8$ (based on several tests), we select the top 130 bigram queries which form the pool of bursty topics for our analysis. Table 1 shows some of these topics, covering multiple domains: politics, science, art, sports, entertainment, etc.

ID	Topic	ID	Topic
1	kentucky election	66	orlando hernandez
2	indiana election	75	daniel biechele
8	van goph	81	hurricane forecast
24	north korea	92	93 memorial
34	pacific quake	113	holloway case
52	florida fires	128	stephen colbert
63	hunger strike	130	bear attack

Table 1: Examples of News Topics

3.2 Context extraction from multiple sources

Once we select the pool of bursty topics, we gather the contexts of each topic from multiple sources: query log, news media, blog media, and social bookmarks. We assume that the language in these contexts will reflect the reactions of the bursty events in corresponding online media.

3.2.1 Super query context

The most straightforward context of bursty events in web search is the query string, which directly reflects the users' interests and perspectives in the topic. We therefore define the first type of context of a bursty topic in query log as the set of surrounding terms of that bursty bigram in the (longer) queries. For example, the word *aftermath* in the query "haiti earthquake aftermath" is a term in the context of the bursty topic *haiti earthquake*.

Formally, we define a *Super Query* of a bursty topic t, $sq(t)$, as the query which contains the bigram query t lexically as a substring. For each bursty topic t, we scan the whole query log Q and retrieve all the super queries of t to form the context which is represented by $SQ(t)$.

$$SQ(t) = \{q | q \in Q \quad and \quad q = sq(t)\}$$

[1]Now known as Bing: www.bing.com

$SQ(t)$ is defined as the super query context of t. For example, the super query context of "kentucky election" contains terms such as "2006," "results," "christian county," etc. These terms indicate what aspects the users are most interested in Kentucky Election during May 2006.

The super query context is widely explored by search engines to provide query expansion and query completion (Jones et al., 2006).

3.2.2 Query session context

Another interesting context of a bursty topic in query log is the sequence of queries that a user searches after he submitted the bursty query q. This context usually reflects how a user reformulates the representation of his information need and implicitly clarifies his interests in the topic.

We define a *Query Session* containing a bursty topic t, $qs(t)$, as the queries which are issued by the same user after he issued t, within 30 minutes. For each bursty topic t, we collect all the $qs(t)$ to form the query session context of t, $QS(t)$:

$$QS(t) = \{q | q \in Q \quad and \quad q \in qs(t)\}$$

In web search, the query session context is usually utilized to provide query suggestion and query reformulation (Radlinski and Joachims, 2005).

3.2.3 News contexts

News articles written by critics and journalists reflect the reactions and perspectives of such professional group of people to a bursty event. We collect news articles about these 130 bursty topics from Google News[2], by finding the most relevant news articles which (1) match the bursty topic t, (2) were published in May, 2006, and (3) were published by any of the five major news medias: CNN, NBC, ABC, New York Times and Washington Post.

We then retrieve the *title* and *body* of each news article. This provides us two contexts of each bursty topic t: the set of relevant news titles, $NT(t)$, and the set of relevant news bodies, $NB(t)$.

3.2.4 Blog contexts

Compared with news articles, blog articles are written by common users in the online communities, which are supposed to reflect the reactions and opinions of the public to the bursty events. We collect blog articles about these 130 topics from Google Blog[3], by finding the most relevant blog articles which (1) match the bursty topic t, (2) were published in May, 2006 (3) were published in the most popular blog community, Blogspot[4]. We then retrieve the *title* and *body* of each relevant blog post respectively. This provides another two contexts: the set of relevant blog titles, $BT(t)$, and the set of relevant blog bodies, $BB(t)$.

3.2.5 Social bookmarking context

Social bookmarks form a new source of social media that allows the users to tag the webpages they are interested in and share their tags with others. The tags are supposed to reflect how the users describe the content of the pages and their perspectives of the content in a concise way.

We use a sample of Delicious[5] bookmarks in May, 2006, which contains around $1.37M$ unique URLs. We observe that the bursty bigram queries are also frequently used as tags in Delicious. We thus construct another context of bursty events by collecting all the tags that are used to tag the same URLs as the bursty topic.

Formally, we define $DT(t)$ as the context of social tags of a topic t,

$$DT(t) = \{tag | \exists url, \quad s.t. \quad tag, t \in B(url)\},$$

where url is a URL and $B(url)$ stands for the set of all bookmarks of url.

3.3 Context Statistics

Now we have constructed the set of 130 bursty topics and 7 corresponding contexts from various sources. We believe that these contexts well represent the various types of online media and sources.

For each context, we then clean the data by removing stopwords and the bursty topic keywords themselves. We then represent it as either the set of unigrams or bigrams from this context. Table 2 shows the basic statistics of each context:

From Table 2 we observe the following facts:

• The query session context covers more terms (both unigrams and bigrams) than the super query

[2]http://news.google.com/

[3]http://blogsearch.google.com/

[4]http://www.blogspot.com/

[5]http://delicious.com/

	N	T_S	M_S	A_U	M_U	A_B	M_B
SQ	130	76k	5.3k	32.7	390	24.3	235
QS	126	108k	5.8k	224	1.5k	150	1062
NT	118	4.7k	411	105	627	102	722
NB	118	4.7k	411	4.7k	22k	22k	257k
BT	128	5.8k	99	184	459	169	451
BB	128	5.8k	99	4.1k	15k	12k	69k
DT	71	2.3k	475	137	2.0k	N/A	N/A

N: The number of topics covered
T_S: The total number of records/documents
M_S: The max number of records/documents per topic
A_U: The avg number of unique unigrams per topic
M_U: The max number of unique unigrams
A_B: The avg number of unique bigrams
M_B: The max number of unique bigrams

Table 2: Basic statistics of collections

context. In both contexts, the average number of unique bigrams is smaller than unigrams. This is because queries in search are usually very short. After removing stopwords and topic keywords, quite a few queries have no bigram in these contexts.

• News articles and blog articles cover most of the bursty topics and contain a rich set of unigrams and bigrams in the corresponding contexts.

• The Delicious context only covers less than 60% of bursty topics. We couldn't extract bigrams from bookmarks since delicious provides a "bag-of-tags" interface.

In Section 4, we present a comprehensive analysis of these different contexts of bursty topics, with three different types of comparison.

4 Experiment

In this section, we present a comprehensive comparative analysis of the different contexts, which represent the reactions to the bursty topics in corresponding sources.

4.1 Similarity & Predictability analysis

Our first task is to compare the content similarity of these sources. This will help us to understand how well the language usage in one context can be leveraged to predict the language usage in another context. This is especially useful to predict the content in web search. By representing each context of a bursty topic as a vector space model of unigrams/bigrams, we first compute and compare the average *cosine similarity* between contexts. We only include contexts with more than 5 unigram/bigrams into this comparison. The results are shown in Table A and Table B, respectively. Each table is followed by a heat map to visualize the pattern.

To investigate how well one source can predict the content of another, we also represent each context of a bursty topic as a unigram/bigram language model and compute the *Cross Entropy* (Kullback and Leibler, 1951) between every pairs of contexts. Cross Entropy measures how certain one probability distribution predicts another. We calculate such measure based on the following definition:

$$H_{CE}(m||n) = H(m) + D_{KL}(m||n)$$

We smooth the unigram language models using Laplace smoothing (Field, 1988) and the bigram language models using Katz back-off model (Katz, 1987).

The results are shown in Table C and Table D, followed by the corresponding heat maps. For each value $H_{CE}(m||n)$ in the table cell, m stands for the context in the row and n stands for the context in the column. Please note that in Figure 3, 4, a larger H_{CE} value corresponds to a lighter cell.

4.1.1 Results

From the results shown in Table A-D, or in Figure 1- 4 more visually, some interesting phenomena can be observed:

• Compared with other contexts, query session is much more similar to the super query. This makes sense because many super queries would be included in the query session.

• Compared with news and blog, the delicious context is closer to the query log context. In fact, delicious is reasonably close to all the other contexts. This means social tags could be an effective source to enhance bursty topics in web search in terms of query suggestion. However, as Table 2 shows, only less than 60% of topics can be covered by delicious tag. We have to explore other sources to make a comprehensive prediction.

• In the news and blog contexts, the title contexts are more similar to the query contexts than the body contexts. This may be because titles usually concisely describe the topic while bodies contain much more details and irrelevant contents.

Context	SQ	QS	NT	NB	BT	BB	DT
SQ	1.0	0.405	0.122	0.072	0.119	0.061	0.188
QS	0.405	1.0	0.049	0.062	0.066	0.054	0.112
NT	0.122	0.049	1.0	0.257	0.186	0.152	0.120
NB	0.072	0.062	0.257	1.0	0.191	0.362	0.114
BT	0.119	0.066	0.186	0.191	1.0	0.242	0.141
BB	0.061	0.054	0.152	0.362	0.242	1.0	0.107
DT	0.188	0.112	0.120	0.114	0.141	0.107	1.0

Table A: Cosine similarity for unigram vectors

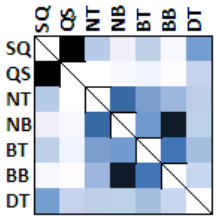


Figure 1: Heat map of table A

Source	SQ	QS	NT	NB	BT	BB
SQ	1.0	0.290	0.028	0.024	0.047	0.027
QS	0.290	1.0	0.004	0.010	0.011	0.009
NT	0.028	0.004	1.0	0.041	0.026	0.011
NB	0.024	0.010	0.041	1.0	0.023	0.040
BT	0.047	0.011	0.026	0.023	1.0	0.044
BB	0.027	0.009	0.011	0.040	0.044	1.0

We do not build bigram vector for DT

Table B: Cosine similarity for bigram vectors

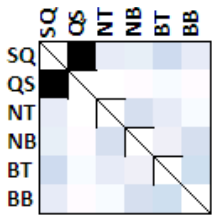


Figure 2: Heat map of table B

Source	SQ	QS	NT	NB	BT	BB	DT
SQ	1.698	4.911	7.538	8.901	7.948	9.050	7.498
QS	7.569	3.842	9.487	11.130	9.997	11.546	8.972
NT	8.957	10.868	3.718	7.946	9.006	9.605	8.825
NB	11.217	12.897	11.317	7.241	12.282	11.582	11.739
BT	9.277	11.084	9.085	10.295	4.637	9.365	9.180
BB	11.053	12.842	11.593	11.742	12.001	7.232	11.525
DT	8.457	9.794	8.521	9.511	8.831	9.473	2.990

Table C: Cross entropy for unigram distribution

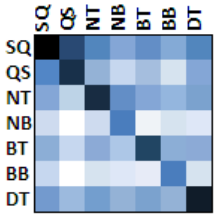


Figure 3: Heat map of table C

Source	SQ	QS	NT	NB	BT	BB
SQ	1.891	2.685	4.290	4.540	4.319	4.607
QS	6.800	3.430	8.144	9.049	8.528	9.304
NT	5.444	5.499	3.652	4.733	5.106	5.218
NB	11.572	11.797	11.254	8.731	11.544	11.073
BT	5.664	5.674	5.503	5.495	4.597	5.301
BB	10.745	10.796	10.517	10.455	10.526	8.518

We do not build bigram distribution for DT

Table D: Cross Entropy for bigram distribution

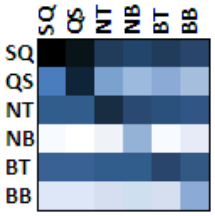


Figure 4: Heat map of table D

$H_{CE}(SQ||n)$

n:	NT	BT	NB	BB
Uni:	**7.538**	7.948	**8.901**	9.050
Bi:	**4.290**	4.319	**4.540**	4.607

$H_{CE}(m||SQ)$

m:	NT	BT	NB	BB
Uni:	**8.927**	9.277	11.217	**11.053**
Bi:	**5.445**	5.664	11.572	**10.745**

Table 3: Cross-entropy among three sources

• News would be a better predictor of the query than blog in general. This is interesting, which indicates that many search activities may be initialized by reading the news.

• News and blogs are much more similar to each other than query logs. We hypothesize that this result reflects the behavior how people write blogs about bursty events – typically they may have read several news articles before writing their own blog. In the blog, they may directly quote or retell a part of the news article and then add their opinion.

• Table 3 reveals the generation relations among three sources: query, news and blog. From the upper table, we can observe that queries are more likely to be generated by news articles, rather than blog articles. From the lower table, we can observe that queries are more likely to generate blog articles(body), rather than news articles(body). This result is quite interesting, which indicates the users' actual behaviors: when a bursty event happens, users would search them from web *after* they read it from some *news articles*. And users would write their own blogs to discuss the event *after* they retrieve and digest information from the web.

• From Table 3 we also find that queries are more likely to generate news *title*, rather than blog *title*. It is natural since blogs are written by kinds of people. The content especially the title part contains more uncertainty.

4.1.2 Case study

We then conduct the analysis to the level of individual topics. Table 4 shows the correlation of each pair of contexts, computed based on the similarity between topics in SQ and corresponding topics in these contexts. We can observe that $News$ and $Blog$ are correlated with each other tightly. If one is a good predictor of bursty queries, the other one also tends to be.

	QS	NT	NB	BT	BB	DT
QS		0.46	0.59	0.58	**0.75**	0.46
NT			**0.73**	**0.79**	0.59	0.61
NB				**0.71**	**0.68**	0.61
BT					**0.78**	0.59
BB						0.48

Table 4: Correlations of the similarity with SQ

For some topics like "stephen colbert," and "three gorges," both $News$ and $Blog$ are quite similar to the queries, which implies some intrinsic properties (*coherence*) of these topics: users would refer to the same content when using the topic terms in different sources.

We also find that a few topics like "hot dogs," "bear attack," for which the similarity of $(SQ, News)$ and $(SQ, Blog)$ are both low. It is probably because these topics are too diverse and carries a lot of *ambiguity*.

Although in most cases they are correlated, sometimes $News$ and $Blog$ show different trends in the similarity to the queries. For example, $News$ is quite similar to the queries on the topics such as "holloway case" and "jazz fest" while $Blog$ is dissimilar. For these *unfamiliar* topics, users possibly search the web "after" they read the news articles and express their diverse opinions in the blog. In contrast, on the topics like "insurance rate" or "consolidation loans," $Blog$ is similar to the queries while $News$ is not. For these *daily-life-related* queries, users would express the similar opinions when they search or write blogs, while news articles typically report such "professional" viewpoints.

4.2 Coverage analysis

Are social bookmarks the best source to predict bursty content in search? It looks so from the similarity comparison, *if they have a good coverage of search contents*. In this experiment, we analyze the coverage of query contexts in other contexts in a systematic way. If the majority of terms in the super query context would be covered by a small proportion of top words from another source, this source has the potential.

4.2.1 Unigram coverage

We first analyze the coverage of unigrams from the super query context in four other contexts: QS, DT, $News$ (the combination of NT and NB) and $Blog$ (the combination of BT and BB) to compare with SQ. For each source, we rank the unigrams by frequency. Figure 5(a) shows the average trend of SQ-unigram coverage in different sources. The x-coordinate refers to the ratio of top unigrams in one source to the number of unigrams in SQ. For example, if SQ contains n unigrams, the ratio 2 stands for the top $2n$ unigrams in the other source. The y-coordinate refers to the coverage rate of SQ. We can observe that:

• Query Session naturally covers most of the super query terms (over 70%).

• Though delicious tags are more similar to queries than news and blog, as well as a relatively higher coverage rate than the other two while size ratio is small, the overall coverage rate is quite low: only 21.28%. Note that this is contradict to existing comparative studies between social bookmarks and search logs (Bischoff et al., 2008). Clearly, when considering bursty queries, the coverage and effectiveness of social bookmarks is much lower than considering all queries. Handling bursty queries is much more difficult; only using social bookmarks to predict queries is not a good choice. Other useful sources should be enrolled.

• As the growth of the size ratio, the coverage rate of news and blogs are both gradually increased. When stable, both of them arrive at a relatively high level (news: 66.36%, blog: 63.80%), which means news and blogs have a higher potential to predict the bursty topics in search. Moreover, in most cases, news is still prior to blog – not only the overall rate, but also the size ratio comparison while the coverage rate reaches 50% (news:$109 <$ blog:183).

4.2.2 Bigram Coverage

Also we analyze the bigram coverage. This time we only have 3 sources (no DT). We rank the bigrams by the pointwise mutual information instead of frequency, since not all the bigrams are "real" collocations. Figure 5(b) shows the results.

Different from the unigram coverage, except that the query session can naturally keep a high coverage rate (66.07%), both news and blog cover poorly. For this issue, we should re-consider the behavior that users search and write articles. News or blog articles consist of completed sentences and paragraphs which would contain plenty of meaningful bigrams. However search queries consist of keywords – relatively discrete and regardless of order. Therefore, except some proper nouns such as person's name, a lot of bigrams in the query log are formed in an ad-hoc way. Since the different expressions of search and writing, detecting unigrams is more informational than bigrams.

4.3 Coherence analysis

The above two experiments discuss the inter-relations among different contexts. In this section we will discuss the inner-relation within each particular context – when it comes to a particular bursty topic, how coherent is the information in each context? Does the discussion keep consistent, or slip into ambiguity?

We represent all the terms forming each context of a bursty topic as a weighted graph: $G = (V, E, W)$, where each $v \in V$ stands for each term, w_v stands for the weight of vertex v in G, and each $e \in E$ stands for the semantic closeness between a pair of terms (u, v) measured by $sim(u, v)$. We define the *density* of such a semantic graph as follows:

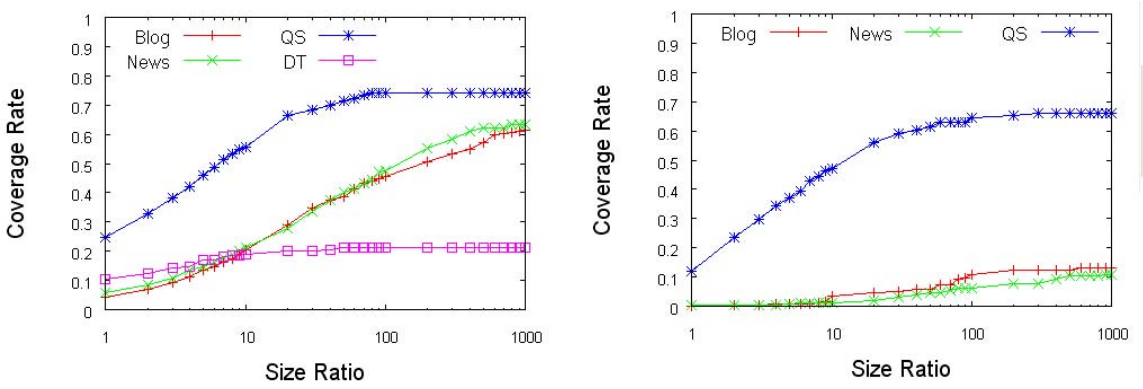
$$Den(G) = \frac{\Sigma_{u,v \in V, u \neq v} sim(u, v) w_u w_v}{\Sigma_{u,v \in V, u \neq v} w_u w_v} \quad (2)$$

If $sim(u, v)$ values the semantic similarity between u and v, a high value of $Den(G)$ implies that the whole context is semantically consistent. Otherwise, it may be diverse or ambiguous.

We build the graph of each context based on WordNet[6]. For a pair of words, WordNet provides a series of measures of the semantic similarity (Pedersen et al., 2004). We use the Path Distance Similarity (*path* for short) and Lin Similarity (*lin* for short) to measure $sim(u, v)$. Both measures range in $[0, 1]$.

For the convenience of computation, we choose the top 1100 unigrams ranked by term frequency in each source (if any) to represent the whole context on one specific topic.

[6] http://wordnet.princeton.edu/



(a) Unigram (b) Bigram

Figure 5: Coverage results

4.3.1 Overall

Table 5 shows the average overall density of each sources over all the topics. From the table we can

Source	*path*	*lin*
SQ	0.098	0.128
QS	0.071	0.082
NT	0.103	0.129
NB	0.109	0.139
BT	0.099	0.109
BB	0.116	0.147
DT	0.102	0.127

Table 5: Overall Density

observe that QS has the lowest density in both of the measures. It is because the queries in one user session can easily shift to other (irrelevant) topics even in a short time.

Another interesting phenomenon comes out that for either news or blog, the *body* is denser than the *title*, even if the body context contains much more terms. It can be explained by the roles of the title and the body in one article: the title contains a series of words which briefly summarize a topic while the body part would describe and discuss the title in details. When it maps to the semantic word network, the title tends to contain the vertices scattered in the graph, while the context of the body part would add more semantically related vertices around the original vertices to strength the relations. Thus, the body part has a higher density than the title part.

4.3.2 The trend analysis

Figure 6 shows the tendency of the density in each source. The x-coordinate refers to the TopN unigrams ranked by the term frequency in each source. From Figure 6 we can find that in most cases, the average density will gradually decrease while less important terms are added, which implies that the most important terms are denser, and other terms would disperse the topic.

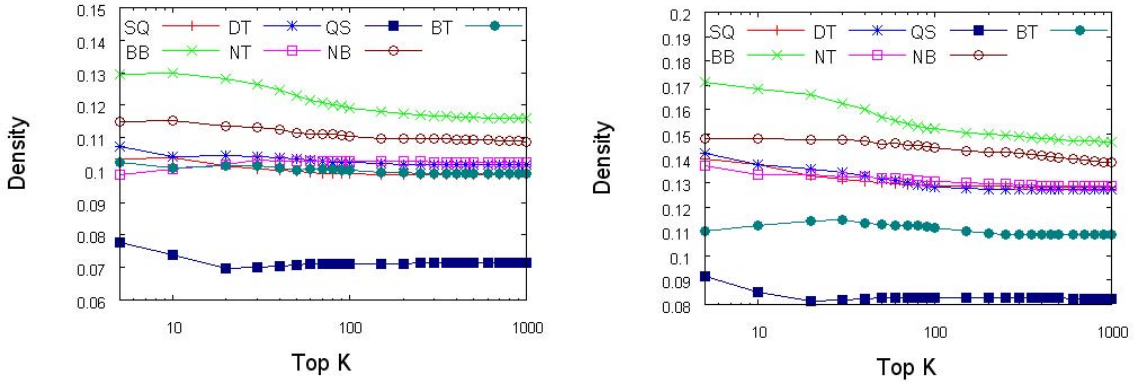
To better evaluate this tendency of each source, Table 6 shows the change rate of the highest density to the overall density measured by *lin*. We can easily find the following facts:

• The highest density is achieved when a small proportion of top terms are counted (6 sources for Top5 and one for Top20), which also supports our hypothesis: the more important, the more coherent.

• BB's density drops the fastest of all (15.1%), following by DT(10.6%). It may be because both blog and delicious tag are generated by many users. And the diversity of the users leads to different prospectives, which dilutes the context significantly.

• Both NT and NB drop quite slowly (5.8%, 6.8%), which means the professional journalists would have the relatively similar prospectives on the same topic. Thus the topic does not disperse too much. BT also keeps a high stability.

• Compared with news, blog is easier to disperse, which can be reflected by the density comparison between NT and BT. Although the density of BB is still higher than NT, we should notice that these two sources are not completely covered – about $3/4$ unigrams in these two contexts are not included in



(a) *path* (b) *lin*

Figure 6: Trend of Density

the semantic networks. The curves clearly shows BB dropped faster than NB. One can expect that NB becomes denser than BB if all the unigrams in both sources are included in the network.

Source	Highest Den.	Overall Den.	Change
SQ	0.140(Top5)	0.128	-8.6%
QS	0.091(Top5)	0.082	-9.9%
NT	0.137(Top5)	0.129	-5.8%
NB	0.148(Top5)	0.139	-6.8%
BT	0.114(Top20)	0.109	-4.4%
BB	0.172(Top5)	0.147	-15.1%
DT	0.142(Top5)	0.127	-10.6%

Table 6: Tendency analysis of Density (Lin)

4.3.3 Case Study

From these 130 news topics, some of them shows a special tendency of coherence. For example, when more words are included, the density of the topic"three gorge" drops rapidly in most of the sources. The topic "florida fires" has the same trend. These topics are typically "*focus*" topics, which means users clearly pursue the unique event while they use these terms. Thus, the density in top unigrams is very high. It drops rapidly since users' personal interests and opinions toward to this event will be enrolled gradually.

In contrast, some topics like "heather mills,", "insurance rate" express differently: their densities gradually increase with the growth of the terms. By observing these topics we find they are usually diverse topics (e.g: famous person name or entity name), which may lead to diverse search intentions of users. So the density of top unigrams is low and gradually increased since one main aspect is probably strengthened.

5 Conclusion and Future work

In this paper, we have studied and compared how the web content reacts to bursty events in multiple contexts of web search and online media. After a series of comprehensive experiments including content similarity and predictability, the coverage of search content, and semantic diversity, we found that social bookmarks are not enough to predict the queries because of a low coverage. Other sources like news and blogs need to be added. Furthermore, news can be seen as a consistent source which would not only trigger the discussion of bursty events in blogs but also in search queries.

When the target is to diversify the search results and query suggestions, blogs and social bookmarks are potentially useful accessory sources because of the high diversity of content.

Our work serves as a feasibility investigation of query suggestion for bursty events. Future work would address on how to systematically predict and recommend the bursty queries using online media, as well as a reasonable evaluation metrics upon it.

Acknowledgments

We thank Prof. Kevin Chang for his support in data and useful discussion. We thank the three anonymous reviewers for their useful comments. This work is in part supported by the National Science Foundation under award number IIS-0968489.

References

Jon Kleinberg 2003. Bursty and Hierarchical Structure in Streams *Data Mining and Knowledge Discovery*, Vol 7(4):373-397

Daniel Gruhl, R. Guha, Ravi Kumar, Jasmine Novak and Andrew Tomkins 2005. The Predictive Power of Online Chatter *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, 78-87.

Ravi Kumar, Jasmine Novak, Prabhakar Raghavan and Andrew Tomkins 2003. On the Bursty Evolution of Blogspace *WWW '03: Proc. of the 12th International World Wide Web Conference*, 568-576.

Michail Vlachos and Christopher Meek and Zografoula Vagena and Dimitrios Gunopulos 2004. Identifying similarities, periodicities and bursts for online search queries *SIGMOD '04: Proceedings of the 2004 ACM SIGMOD international conference on Management of data*, 131-142.

Evgeniy Gabrilovich, Susan Dumais and Eric Horvitz 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty *WWW '04: Proceedings of the 13th international conference on World Wide Web*, 482-490.

Eytan Adar, Daniel S. Weld, and Brian N. Bershad and Steven S. Gribble 2007. Why we search: visualizing and predicting user behavior *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 161-170.

Aixin Sun, Meishan Hu and Ee-Peng Lim 2008. Searching blogs and news: a study on popular queries *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, 729-730.

JeanPhilippe Cointet, Emmanuel Faure and Camille Roth 2008. Intertemporal topic correlations in online media : A Comparative Study on Weblogs and News Websites *ICWSM '08: International Coference on Weblogs and Social Media*

Levon Lloyd, Prachi Kaulgud and Steven Skiena 2006. Newspapers vs. Blogs: Who Gets the Scoop? *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*

Michael Gamon, Sumit Basu, Dmitriy Belenko, Danyel Fisher, Matthew Hurst, and Arnd Christian Konig 2008. BLEWS: Using Blogs to Provide Context for News Articles *ICWSM '08: International Coference on Weblogs and Social Media*

Sanjay Sood, Sara Owsley, Kristian Hammond and Larry Birnbaum 2007. TagAssist: Automatic Tag Suggestion for Blog Posts *ICWSM '07: International Coference on Weblogs and Social Media*

Fernando Diaz 2009. Integration of news content into web results *WSDM '09: Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 182-191.

Beate Krause, Andreas Hotho and Gerd Stumme 2008. A Comparison of Social Bookmarking with Traditional Search *Advances in Information Retrieval*, Vol 4956/2008:101-113.

Paul Heymann, Georgia Koutrika and Hector Garcia-Molina 2008. Can social bookmarking improve web search? *WSDM '08: Proceedings of the international conference on Web search and web data mining*, 195-206.

Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei and Zhong Su 2007. Optimizing web search using social annotations *WWW '07: Proceedings of the 16th international conference on World Wide Web*, 501-510.

Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl and Raluca Paiu 2008. Can all tags be used for search? *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, 193-202.

Rosie Jones, Benjamin Rey and Omid Madani 2006. Generating query substitutions *Proceedings of the 15th international conference on World Wide Web,*, 387-396.

Filip Radlinski and Thorsten Joachims 2005. Query chains: learning to rank from implicit feedback *Proceedings of the 11th ACM SIGKDD international conference on Knowledge discovery in data mining*, 239-248.

Qiaozhu Mei, Dengyong Zhou and Kenneth Church 2008. Query suggestion using hitting time *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, 469-478

Andrei Broder 2002. A Taxonomy of Web Search *SIGIR Forum*, Vol 36(2):3-10.

Solomon Kullback and Richard Leibler 1951. On Information and Sufficience *Annals of Mathematical Statistics*, Vol 22(1):79-86.

David A. Field 1988. Laplacian Smoothing and Delaunay Triangulations *Communications in Applied Numerical Methods*, Vol 4:709-712.

Stephen M. Katz 1987 Estimation of probabilities from sparse data for the language model component of a speech recogniser *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3), 400-401.

Ted Pedersen, Siddharth Patwardhan and Jason Michelizzi 2004. WordNet::Similarity: measuring the relatedness of concepts *PHLT-NAACL '04: Demonstration Papers at HLT-NAACL 2004 on XX*, 38-41.