Predicting Event-Relatedness of Popular Queries

Seyyedeh Newsha Ghoreishi
School of Computer Engineering
Nanyang Technological University, Singapore
se0001hi@e.ntu.edu.sg

Aixin Sun
School of Computer Engineering
Nanyang Technological University, Singapore
axsun@ntu.edu.sg

ABSTRACT

Many but not all popular queries are related to ongoing or recent events. In this paper, we identify 20 features including both contextual and temporal features from a small set of search results of a query and predict its event-relatedness. Search results from news and blog search engines are evaluated. Our analysis shows that the number of named entities in search results and their appearances in Wikipedia are among the most discriminative features for query event-relatedness prediction. Our study also shows that contextual features are more effective than temporal features. Evaluated with four classifiers (*i.e.,* Support Vector Machine, Naïve Bayes, Multinomial Logistic Regression, and Bayesian Logistic Regression) on two datasets, our experiments show that query event-relatedness can be predicted with high accuracy using the proposed features.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*

Keywords

Query event relatedness, Query classification, Event detection

1. INTRODUCTION

The popularity of various social platforms (*e.g.,* blogs, Twitter, and Facebook) make information sharing a much easier task for common Web users. Examples include sharing a tweet message or posting a short Facebook status update through Web or mobile devices. Many pieces of such information are shared with very limited context, because of the *principle of least effort*. That is, people used to communicate information with the least context, especially in the situation where a short message with free style is allowed [14]. The lack of context often triggers many searches for more relevant information, particularly on some *ongoing or recent events and emerging topics*. On the other hand, searches may be triggered by many other reasons and not all popular queries to news/blog search engines are related to ongoing or recent events. Correct prediction of whether a query is event-related is therefore critical for better understanding and addressing users' search intent.

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.

http://dx.doi.org/10.1145/2505515.2507853.

In this paper, we focus on the prediction of event-relatedness of popular queries submitted to dedicated search engines for time-sensitive document streams (*e.g.,* search engines for news articles and blog posts). It was reported that, when an event happens, many users search for related information which makes the event-related queries popular [10]. On the other hand, many extremely popular queries are not related to any events; examples include queries referring to major websites like Google, MySpace, and YouTube [11]. It is therefore interesting to predict event-relatedness of popular queries so as to guide a search engine to provide relevant search results. This problem, however, is challenging because of at least two reasons. First, given a query, very limited information can be directly derived from the query string and the user who issues this query. A query can be issued to a search engine through multiple interfaces/platforms, without any additional contextual information. Second, the prediction should ideally be in near real-time. That is, the prediction should avoid expensive computations given the volume of the search queries.

In this paper, we adopt a feature engineering approach. More specifically, we study the effectiveness of the features derived from the query string and a small set of initial search results of the query. Two sets of the features are studied, namely, *contextual features* and *temporal features*. The former is based on the assumption that the meaning of a query is defined collectively by its top-ranked search results at the time of the search. Example contextual features are derived from the cohesiveness of the top-ranked results, and the named entities contained in the query string and the search results. Temporal features, on the other hand, quantify the temporal distribution of the search results. In total 20 features are evaluated for popular queries collected from a major blog search engine for 3 months, with search results from blog and news search engines respectively. The features are evaluated using Chi-square for their discriminative power in predicting query event-relatedness. Four classifiers (*i.e.,* Support Vector Machine, Bayesian Logistic Regression, Multinomial Logistic Regression, and Naïve Bayes) are then constructed using these features for query event-relatedness prediction. Our experimental results show that both contextual and temporal features are effective in predicting query event-relatedness and contextual features are more effective than temporal features. More importantly, query event-relatedness can be predicted with high accuracy with F_1 values greater than 0.85.

2. RELATED WORK

In this section, we briefly review the related works in the areas of event detection, query classification, time-sensitive query and recency ranking.

Event Detection Using Queries. Parikh [8] studied the role of query in bursty event detection in e-commerce systems by using

daily query streams. By applying an infinite 2-state automaton, they show that temporal patterns of query streams using frequency of each unique query lead to detection of burst queries which are likely to be event. In [4], events are detected based on the reaction from variety of online medias and Web content such as news media, blogging, and social bookmarking. Extracting bursty topics from user's search behavior is a major issue.

Most related to this work is the study on query-guided event detection [10]. In their approach, each query is described by a query profile which consists of its top-ranked results from a search engine. Events are detected by grouping similar queries together. The query similarity is based on many aspects including query string similarity, query profile content similarity, and temporal distance between queries. Two features known as *clarity* and *recency* are proposed to predict whether a query is event-related (see Section 3 for more details). However, because of the simple threshold based approach, it is not clear whether these two features are effective in the prediction of query event-relatedness and no classification model has been built on these two features.

Query Classification. Our problem can be considered as a binary classification problem. Different query classification schemes have been investigated in the literature. KDD-CUP 2005 defines the task of classifying queries to 7 *topical categories* and their 67 sub-categories. In many of the proposed solutions, queries are enriched by their search results from search engines [9]. Other than search results, other information sources like query log have been considered in helping topical query classification [1]. In [5], the authors classify queries into three *temporal classes*: atemporal queries, temporally unambiguous queries, and temporally ambiguous queries. A decision tree classifier is utilized to classify the queries into the three categories using content clarity and temporal features of queries. The temporal features of a query are derived from a collection of timestamped documents that contain the query keywords [2]. Another *temporal query classification scheme* introduced in [13] categories queries into three classes: explicit timestamp, implicit timestamp, and no timestamp. These categories are based on existence of any time indicators (*e.g.,* year, month or seasonal terms) in the query.

Time-sensitive Query and Recency Ranking. Several studies are related to time-sensitive queries and Web page ranking. Temporal and contextual features are extracted to predict how likely a query is a *recurrent event query* (REQ) [13] or *year qualified query* (YQQ) [7, 12]. Results from the most recent years or most influential years may be ranked higher than others. Freshness and topical relevancy are two major factors considered in recency sensitive queries [3]. Twitter data is used to promote the ranking of fresh URLs related to ongoing or recent events.

In summary, the problem of query event-relatedness prediction has not been well studied in literature. Nevertheless, features used in related studies could be useful in predicting query event-relatedness.

3. FEATURES FOR PREDICTION

Let $\mathcal{S}$ be a document stream (*e.g.,* news articles or blog posts) indexed by a dedicated search engine in which a document $d \in \mathcal{S}$ is associated with its timestamp $d.t$. Given a query q with query keywords $q.w$ submitted at time $q.t$, the search engine returns a set of top-ranked documents $D_q \subset \mathcal{S}$ as the search result. The problem of *query event-relatedness prediction* is to predict whether query q is related to an *ongoing or recent event*.

The set of results D_q is also known as the *query profile* of query q as in [10]. In this paper, the top-ranked 50 results are considered, *i.e.,* $|D_q| = 50$. Besides publication time, each document $d \in D_q$ is expected to have a title and a snippet. Because of the time-sensitivity in news/blog search, submitting the same query keyword to the same search engine at different time points very likely leads to different search results, or different query profiles.

In the following, we describe the possible contextual and temporal features derived from the query string $q.w$ and query profile D_q for query event-relatedness prediction.

3.1 Contextual Features

We derive three sets of contextual features. The first set measures the topic specificity and cohesiveness of the search results; the second set is derived based on the named entities in the query string and query profile; and the third set is based on word matching with predefined recurrent event-related seed words.

Topical Specificity and Cohesiveness. If a query is event-related, then many news articles (or blog posts) would provide detailed description or comments about the event, highlighting the persons or organizations involved in the event. That is, the documents in D_q are expected to be topically specific as well as topically cohesive. We use 3 features to quantify the topic specificity and cohesiveness.

Topic specificity is quantified using query profile clarity [10]. It is the Kullback-Leibler (KL) divergence between the word distribution estimated from the query profile and the word distribution of the document stream, shown in Equation 1.

$$Clarity(D_q) = \sum_{w \in D_q} P(w|D_q) \log_2 \frac{P(w|D_q)}{P(w|\mathcal{S})} \quad (1)$$

Following [10], $P(w|D_q)$ is estimated by the relative document frequency of word w in query profile D_q; similarly for $P(w|\mathcal{S})$.

Query clarity has been used to predict query difficulty and a large KL divergence indicates a clear and unambiguous query [2]. Stated in Sun [10], high clarity score can be achieved if (i) there are a large number of words with high document frequencies in query profile, and (ii) the probability distribution of these words is distinguishable from that of the whole document collection.

Topic cohesiveness is evaluated using two features: *centroid-based cohesion* and *pairwise similarity* of documents in query profile D_q. The former is computed by the averaged cosine similarity between a document $d \in D_q$ and the centroid of D_q. The word feature vectors are weighted by TF-IDF scheme. The latter is the average Jaccard coefficient of the set of words contained in each pair of documents. Note that, as we only consider the title and snippet of a document as its content, all documents in the query profile have similar number of words.

In total, we have 3 features for topical specificity/cohesiveness.

Named Entities and Newsworthiness. An event is something that happens in a certain place at a certain time. Documents returned for an event-related query are highly probable to have some keywords related to location, organization, or person. We therefore consider the number of Named Entities (NE) as a feature for query event-relatedness prediction. More specifically, we use the Stanford NLP package to detect named entities from the query string $q.w$ and the documents in the query profile D_q. Considering the three categories of NEs (*i.e.,* person, location, and organization) we have 8 features: number of NEs under each category and all three categories (4 features), and number of distinct NEs (based on string matching) under each category and all three categories (4 features).

In [6], the authors demonstrated the effectiveness of using string's *newsworthiness* measure to distinguish realistic events from casual conversations. The newsworthiness of a string s is the probability of s appears as anchor text in Wikipedia articles that contain the string s. We compute the newsworthiness of the NEs in the three

Table 1: Recurrent event seed words

American	attack	awards	bonus	collapse
concert	congress	death	earthquake	elections
ending	euro	awards	execution	express
fire	flight	Ford	Friday	gale
gangster	gravel	hanging	highrise	hurricane
idol	invite	liberty	marathon	memorial
minister	Oscars	polls	primary	prince
report	republican	secondary	selamat	wildfire

categories (*i.e.,* person, location, and organization) respectively as three features. The sum of the newsworthiness of all NEs extracted from a query profile is also computed as a feature.

In short, we have 8 features for the number of NEs, and 4 features for the newsworthiness of the NEs.

Recurrent Event Seed Words. Recurrent event queries are periodic and/or repetitive queries which share some common words [13]. These words, combined with some other words, often lead to a recurrent event query. For example, "Oscars 2008" and "hurricane katrina" are event-related queries by two seed words "oscars" and "hurricane". The former is a seed word for an expected time-sensitive recurrent event and the latter is a seed word for an unexpected event.

A small number of seed words are given in [13] as examples. In our study, by manual evaluation of frequent tokens in the popular query strings in our data collection (see Section 4.1), we labeled 73 seed words. Each seed word is a token that is generalized and highly meaningful to be recurrent over time, not particularly for specific places or groups of people, location and nation. In some cases, there are some tokens which belong to a specific place but their impact is worldwide and leading to worldwide searches. One example is seed word "euro" for many annual events such as conferences, sport games, contests, and even dramatic change of euro exchange rate. Table 1 lists 40 of the 73 seed words used in our experiments. These seed words cover major awards ceremony like Oscars, politic events such as election, and unexpected natural events like hurricane and wildfire.

Two features are derived based on the seed words, namely *query string recurrency* and *query profile recurrency*. The former is a binary feature indicating whether any token in a query string $q.w$ matches a seed word. The latter is the total number of appearances of seed words in a query profile D_q.

Query String Frequency. In addition, we evaluated the frequency of a query string in the past 3 days because a major event often last a few days.

3.2 Temporal Features

We consider two features to describe the temporal characteristics of a query profile, namely, *recency* and *temporal clarity*.

Recency. Recency reflects the freshness of the search results in query profile D_q with respect to the query time $q.t$. In [10], recency measure is the average time difference between the query time and the timestamp of the documents in a query profile. Considering that the average value may be adversely affected by outliers, we use median for recency measure.

Temporal Clarity. Inspired by the temporal KL-divergence in [2], we define the temporal clarity of a query profile in Equation 2.

$$TClarity(D_q) = \sum_{min(d.t-q.t)}^{max(d.t-q.t)} P(t|D_q) \log_2 \frac{P(t|D_q)}{P(t|\mathcal{S})} \quad (2)$$

Table 2: Dataset statistics

Number of queries and event-related queries	GN	TR
Number of queries in training set	5251	6417
Number of event-related queries in training set	2268	2667
Number of queries in test set	3106	3560
Number of event-related queries in test set	1101	1150
Total number of queries	8357	9977
Total number of event-related queries	3369	3817

In this equation, $P(t|D_q)$ is the relative document frequency of documents created in time window t, which is set to be a day in our experiments.

Temporal clarity measures the difference between the temporal distribution of documents retrieved for a query and the temporal distribution of documents in the document stream [2]. If a query is event-related, it is expected that its matching documents to be distinguishable temporally from the whole document stream.

4. EXPERIMENTS

4.1 Data Collection

The dataset in our experiments was constructed by collecting the top-15 popular queries published by Technorati (TR for short) for every 3 hours for 3 months from Nov 2007 to Feb 2008. Each collected popular query was then submitted separately to TR and Google News (GN) search engines to get the top-50 search results for query profile construction. A search result in a query profile contains title and snippet.

The queries were manually labeled by two undergraduate students for their event-relatedness, mainly based on the search results with the help of general search engines and Wikipedia. In the labeling process, queries that are not in English and the queries with fewer than 50 search results were ignored. Note that a query can be labeled as event-related based on results from GN, but not necessarily event-related based on results from TR. For this reason, the data collected from TR and GN are considered as two datasets. Recall that the same query string $q.s$ may be searched at different times $q.t$ and get different sets of results. Each search is considered as one instance and the event-relatedness label is assigned to each instance. For simplicity, in our discussions, a query refers to one search instance.

Queries from the first two months are used as training samples and queries from the third month are used as test samples. Table 2 reports the statistics of the data.

4.2 Feature Evaluation

One of the key objectives of this study is to evaluate the effectiveness of the features in predicting query event-relatedness. For this purpose, we applied feature selection techniques including Chi-square, information gain, and gain ratio to evaluate the discriminative power of features described in Section 3.

From the feature evaluation results, it is observed that the three feature selection techniques give very similar discriminative power ranking of the features. We choose to report the ranking by Chi-square only on the two datasets GN and TR respectively, in Table 3. We made the following observations from the result.

First, newsworthiness of all named entities in the query profile is the most effective feature for both GN and TR. To the best of our knowledge, this is the first to use newsworthiness for query event-relatedness prediction. This result suggests that famous people and/or organization that have strong presence in Wikipedia easily attract public attention and are often involved in emerging events.

Table 4: The F_1 values of query event-relatedness predication on two datasets with different sets of features

Dataset	Google News dataset				Technorati dataset			
Feature/Classifier	SVM^{Light}	Bayesian LR	Multinomial LR	NB	SVM^{Light}	Bayesian LR	Multinomial LR	NB
Contextual features	0.745	0.802	0.829	0.792	0.772	0.832	0.797	0.776
Temporal features	0.792	0.792	0.749	0.614	0.671	0.763	0.758	0.685
All features	0.804	0.841	**0.855**	0.821	0.794	**0.870**	0.824	0.808

Table 3: Ranking of the top-10 most discriminative features

GN	TR	Feature
1	1	Total newsworthiness of all NEs in query profile
2	3	Topic cohesiveness (centroid-based cohesion)
3	6	Topic cohesiveness (pairwise similarity)
4	2	Query string frequency in past 3 days
5	7	Query profile recurrency
6	4	Topic specificity (query profile clarity)
7	5	Total number of NEs in query profile
8	9	Query profile recency
9	10	Query string recurrency
10	8	Temporal clarity

Second, query profile topical cohesiveness is effective in event-relatedness prediction. Both the centroid-based measure and pairwise similarity measure give strong discriminative power on both datasets. This is expected because a large number of documents describing an event often emerge within a short period of time, given the easy access of the Internet and information sharing platforms. These documents, however, often share very similar content.

Third, on both datasets, event seed words play a worthy role reflected by the two features for recurrency. Because seed words can be predefined for both expected and recurrent events and unexpected events, search results can be arranged or further processed to best match the information needs for such events.

Overall, among all features evaluated, contextual features are more effective than temporal features. More importantly, we argue that analysis of the content of the search results is important, particularly on named entities (*e.g.,* number of named entities, and their newsworthiness values). However, we note that categorizing the named entities into person, organization, and location, does not benefit query event-relatedness prediction.

4.3 Prediction Accuracy

Next, using all the 20 features, we evaluate the prediction accuracy of query event-relatedness on two datasets. We report the classification accuracy of four classifiers on two datasets using F_1 measure. More specifically, we used the SVM^{light} implementation[1] of Support Vector Machines, and the Weka[2] implementations of Bayesian Logistic Regression, Multinomial Logistic Regression, and Naïve Bayes classifiers. For SVM, polynomial and radial basis kernels were used for TR and GN datasets respectively for their better accuracy over other kernels. Default settings were adopted for other parameters.

Reported in Table 4, contextual features achieve better prediction accuracy than temporal features in general with the only exception of using SVM^{light} classifier on GN dataset. This result is consistent with our observations made in Section 4.2. Combining both contextual and temporal features leads to the best prediction accuracy in F_1 on both datasets. Specifically, F_1 value of 0.855 is achieved on GN with multinomial logistic regression and 0.870 is achieved on TR with Bayesian logistic regression. The relatively high F_1 values on both datasets indicate the effectiveness of the proposed features and suggest that query event-relatedness can be predicted with high accuracy. This paves way of processing event-related queries for better search experiences on time-sensitive document streams.

[1] `svmlight.joachims.org/`

[2] `www.cs.waikato.ac.nz/ml/weka/`

5. CONCLUSION

With the popularity of various social media platforms, users receive the most timely information, but with limited context. Searching time-sensitive document streams like news and blogs becomes an effective way of getting more detailed information. To be able to better process event-related queries, we evaluate 20 features derived from the query string and the top-ranked search results of popular queries for the task of query event-relatedness prediction. Our results show that contextual features are more effective than temporal features, particularly features related to named entities. Our experiment results also show that query event-relatedness can be predicted with high accuracy. We believe our study will benefit search engines in handling event-related queries in both result ranking and in providing more comprehensive information about events, so as to improve user search experiences.

6. REFERENCES

[1] R. Campos, A. M. Jorge, and G. Dias. Using web snippets and query-logs to measure implicit temporal intents in queries. In *SIGIR 2011 Workshop on Query Representation and Understanding*, 2011.

[2] F. Diaz and R. Jones. Using temporal profiles of queries for precision prediction. In *SIGIR*, pages 18–24, 2004.

[3] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: Improving recency ranking using twitter data. In *WWW*, pages 331–340, 2010.

[4] Y. Jiang, C. X. Lin, and Q. Mei. Context comparison of busrty events in web search and online media. In *EMNLP*, pages 1077–1087, 2010.

[5] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3), 2007.

[6] C. Li, A. Sun, and A. Datta. Twevent: Segment-based event detection from tweets. In *CIKM*, pages 155–164, 2012.

[7] D. Metzler, R. Jones, F. Peng, and R. Zhang. Improving search relevance for implicitly temporal queries. In *SIGIR*, pages 700–701, 2009.

[8] N. Parikh and N. Sundaresan. Scalable and near real-time burst detection from ecommerce queries. In *KDD*, pages 972–980, 2008.

[9] D. Shen, R. Pan, J.-T. Sun, J. J. Pan, K. Wu, J. Yin, and Q. Yang. Qc@ust: Our winning solution to query classification in kddcup 2005. *SIGKDD Explorations*, 7(2):100–110, 2005.

[10] A. Sun and M. Hu. Query-guided event detection from news and blog streams. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, 41(5):834–839, 2011.

[11] A. Sun, M. Hu, and E.-P. Lim. Searching blogs and news: a study on popular queries. In *SIGIR*, pages 729–730, 2008.

[12] R. Zhang, Y. Chang, Z. Zheng, D. Metzler, and J. yun Nie. Search engine adaptation by feedback control adjustment for time-sensitive query. In *HLT-NAACL*, pages 165–168, 2009.

[13] R. Zhang, Y. Konda, A. Dong, P. Kolari, Y. Chang, and Z. Zheng. Learning recurrent event queries for web search. In *EMNLP*, pages 1129–1139, 2010.

[14] G. K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Hafner Pub. Co, 1949.