Mining Unusual Search Behavior Related to Physical Events

Abhay Prakash, Parag Agrawal, Kedhar Nath Narahari, Puneet Agrawal
Microsoft, India
{abprak,paragag,kedharn,punagr}@microsoft.com

ABSTRACT

Search logs of commercial search engines like Bing or Google reflect the topics of interest in humans at any given point of time. While many of these topics are predictable information needs, some are unusual and point to a curious human mind. In this paper, we propose a novel solution for mining unusual search behaviors triggered by some recent physical events using Bing search logs. Our algorithm successfully identified unusual search behaviors related to physical events during 2016. We define unusual-relatedness and use crowdsource judgment to evaluate quality of our results along with a qualitative analysis. Our results show a 9.5 times improvement over baseline.

1. INTRODUCTION

With the widespread growth of internet enabled devices, search engines have become part of our day-to-day life. We heavily rely on search engines for answers to our queries. Consequently, search logs have emerged as a rich source of information depicting human behavior. Sometimes a physical event causes an unusually related search behavior e.g. as per Bing logs, *when Donald Trump won the U.S. presidential elections*, users in the U.S. searched for queries related to *Canada migration* 711% more than average of the last week. We believe interpreting search behavior in context of physical world will lead to deeper insights into human society and help improve products, services and even policies around us. Search engines could also present these unusual search behaviors to their users as trivia or as related facts to another user search query.

In this paper, we introduce the problem of mining unusual search behaviors, caused by some physical event. While several physical events and search behaviors co-exist, the perception of *unusual* search behavior is subjective and one can not establish causation definitively. We however believe that based on the context of ongoing sentiments in society, most humans aware of an event, can perceive the causation. For instance, in aforementioned example, a majority will agree that search for *Canada migration* intent is caused by the physical event of Trump's presidential win. However, another intent *what is meme?* was also popular during the same time duration, but most humans will agree that it is unrelated to Trump. We restrict our definition of *unusual-relatedness* to such majority based agreement of causation.

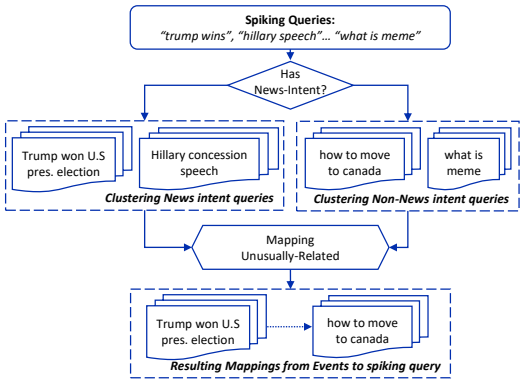

Figure 1: Our algorithm maps physical events and spiking queries while pruning out unrelated queries.

2. RELATED WORK

Search trends and their application to various fields have been widely studied. [1] studied how user queries can help predict economic activity in areas of retail, automobile and travel. In field of epidemics, [2] showed that search trends can successfully predict outbreak of influenza. In a recent work, [5] presents the fashion trends that can be mined from search logs. Our work targets to mine search queries which are unusually related to a recent physical event.

3. OUR APPROACH

A popular physical event must trigger a spike in both directly related, as well as unusually-related search queries. Hence for a given day, we obtain all queries which spiked more than 300% compared to last week's average. These spiking queries are then further used to identify recent events, and to prepare candidate set of unusually-related queries.

We categorize our problem statement into three sub-tasks. First is to identify physical events on a day, second to prepare candidate search queries which could be unusually-related to one of the events, and third one to map candidate queries to events. Figure 1 depicts our algorithm, as described below:

©2017 International World Wide Web Conference Committee (IW3C2), published under Creative Commons CC BY 4.0 License.
WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054232

I. Identifying physical events: Assuming all significant physical events trigger news results on a search engine, we separate those spiking queries which trigger a news result on Bing. These news triggering queries are then clustered based on their semantic similarity using Word2Vec[3]. Each of these clusters now represent a physical event and contain common variations of search queries depicting that event. To increase clustering recall, those spiking queries which did not fire news, but are semantically very similar, are merged with existing clusters. For further discussions, we denote an event cluster by E. To keep only significant events, we discarded those clusters which do not have minimum 5000 cumulative query impressions count[1]. In each cluster, top 5 queries by impression are taken as representative queries for that cluster, denoted as E_Q henceforth.

II. Preparing Candidate Set of Unusually-related queries: The queries which are not mapped to any event after the above task, are taken as potential search queries which could be unusually-related to some event(E). We represent the queries in this set as Q.

III. Mapping candidate queries to events: The mapping essentially represents the unusual-relatedness of query to the event, which has been calculated using following process. First, we evaluate correlation using:

$$Score_{correlation}(Q, E) = \frac{\#Times(Q, E)}{\#Times(E)}$$

where #Times(Q,E) is number of times the Q and E_Q have spiked together, and #Times(E) is number of times E_Q have spiked in the year. The $\text{score}_{\text{correlation}}$ signifies the temporal correlation of the query and the event. We keep only those mappings which get $\text{score}_{\text{correlation}}$ more than 0.6.

However, temporal correlation is necessary but not a sufficient condition to establish causation or unusual-relatedness. Hence, we further refine the mappings using following score to estimate the relation of query to event.

$$Score_{unusually-related}(Q, E) = \frac{Similarity(Q_T, E_T)}{Similarity(Q, E_Q)}$$

where Similarity(Q_T , E_T) is the semantic similarity[2] of document titles obtained by searching E_Q and Q on Bing. Similarity(Q , E_Q) is the semantic similarity[2] of E_Q and Q. Note that score is directly proportional to similarity of titles, because it shows that the query could be related to the event. Moreover, score is inversely proportional to similarity of E_Q and Q, which captures the unusualness of relationship. We empirically chose a threshold of 0.8 to get final mappings.

4. EXPERIMENTS AND RESULTS

To measure unusually-relatedness defined in Section 1, we showed *event–query pairing* to 5 judges, and asked them to label it as one of – *'Unusually-Related', 'Directly Related'* or *'Not Related at all'*. We decided the label, which 3 or more judges agreed upon. In case of non-majority, we took the label as *'Can not decide'*.

For evaluation of our algorithm, we generated such unusually related pairs for four significant events of 2016, which are also stated in column 1 of Table 1. We took 100 sampled result pairs, and got the labels as described above. Result shows that 38% times pairs are 'Unusually-Related', 52% are 'Not Related at all', 4% are 'Directly Related' and 6% are 'Can not decide'. Table 1 also present examples of good results from our algorithm, along with an interpretation for query being unusually related to physical event.

Event	Example Queries unusually related to the Event	Interpretation of result
Trump won U.S. presidential elections	living in canada; canada immigration office; canada jobs	Peaked interest in Canada as potential place to live due to Trump's stand on various public policies.
Hillary collapsed due to Pneumonia	symptoms of pneumonia in adults; is pneumonia contagious	Hillary collapsed due to pneumonia which developed curiosity about the disease.
Britain exit from EU	culture in switzerland; cost of living switzerland	Switzerland has special relationship with EU and speculation were that Britain might follow the same model.
Terrorist attack in Brussels	belgium people and culture; belgium gun laws	Attack in Brussels developed curiosity about culture and gun laws.

Table 1: Unusually-Related search queries for four significant events in 2016.

For comparison, we also got judgment for 100 sampled pairs obtained by temporal correlation only i.e. mapping Es and Qs (on same day) without using our mapping algorithm. Results show that only 4% of them were 'Unusually-Related', 10% were 'Directly Related' and 70% were 'Not Related at all'. This indicates that on a given day multiple events and queries trend, which may not relate to each other. Thus, 9.5 times increase (4% to 38%) in accuracy of our algorithm is significant.

5. CONCLUSION AND FUTURE WORK

Our work presents a novel method to discover unusual search behavior related to physical events. While we realize the challenges which [4] associated with search logs, we show through various examples that deeper insights can be drawn from search logs, and hope to attract attention of research community on the same.

In future, we will be studying how re-occurrence of same events results in different behavior in search logs e.g. we will study how search trends have changed over time during release of multiple iPhone versions.

6. REFERENCES

[1] H. Choi and H. Varian. Predicting the present with google trends. *Economic Record*, 88(s1):2–9, 2012.

[2] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014, 2009.

[3] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[4] G. C. Murray and J. Teevan. Query log analysis: social and technological challenges. In *ACM SIGIR Forum*, volume 41, pages 112–120. ACM, 2007.

[5] Torrence Boone. Fashion Trends 2016: Google Data Shows What Shoppers Want. *http://bit.ly/2bFy5Vo*, 2016.

[1] An impression count is total number of times a query is issued on search engine.

[2] Using Word2Vec[3].