Query Suggestion by Concept Instantiation

Jack Wei Sun[1], Franky[1], Kenny Q. Zhu[1], and Haixun Wang[2]

[1] ADAPT-Lab, Shanghai Jiao Tong University
[2] Google Inc.

Abstract. A class of search queries which contain abstract concepts are studied in this paper. These queries cannot be correctly interpreted by traditional keyword-based search engines. This paper presents a simple framework that detects and instantiates the abstract concepts by their concrete entities or meanings to produce alternate queries that yield better search results. [3]

Keywords: Query Suggestion, Concept-based Queries, Search Logs

1 Introduction

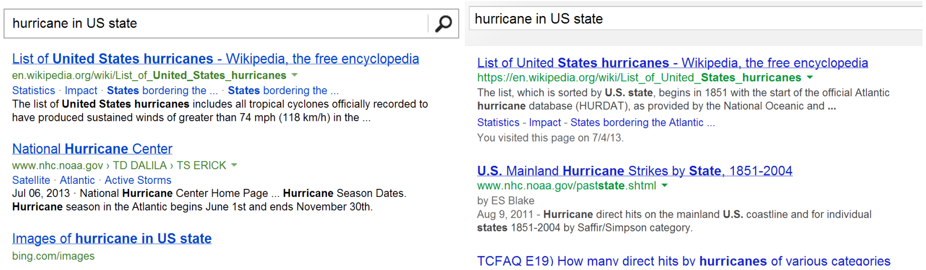
The quality of search results largely depend on the specificity of the keywords they put in search engine, for modern search engine are mostly keyword-based. When user queries do not general good results, search engine may suggest a list of alternate queries for the user to select and re-submit.

Traditionally, the function of query suggestion is implemented by keyword-based methods and partially depends on the information from search log [1, 4, 5, 10], which includes click-through data, session data, etc. With the rapid development of Internet, researchers are increasingly interested in semantic view of the web and do a lot of work on semantic relation extraction from search log [2] and concept search [6, 8]. There are also other attempts on sematic query suggestion without search log [3].

This paper focuses on a class of difficult queries which include abstract concepts, e.g., "*hurricane in US state*". The likely intention of this query is to look for a specific instance of a hurricane that happened to a US state, e.g., Katrina in Louisiana. Here both *hurricane* and *state* are used as abstract concepts which contain many specific entities. The reason for such an abstract query may be because the user has forgotten the name of the hurricane or the state.

Search engines today return pages *about* these abstract concepts, and that usually means a list of huricanes in the US history for the above example (see Figure 1). It takes the user at least a few more clicks and scanning through the pages to find the information needed. Our objectives is to return a list of suggested queries such as: *Katrina in Lousiana*, *Sandy in Connecticut*, *Dolly in Texas.*

[3] Kenny Q. Zhu (corresponding author) is partially supported by NSFC grants 61100050 and 61033002.



hurricane in US state

List of **United States hurricanes** - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/List_of_**United_States_hurricanes**
Statistics · Impact · **States** bordering the ... · **States** bordering the ...
The list of **United States hurricanes** includes all tropical cyclones officially recorded to have produced sustained winds of greater than 74 mph (118 km/h) in the ...

National **Hurricane** Center
www.nhc.noaa.gov › TD DALILA › TS ERICK
Satellite · Atlantic · Active Storms
Jul 06, 2013 · National **Hurricane** Center Home Page ... **Hurricane** Season Dates. **Hurricane** season in the Atlantic begins June 1st and ends November 30th.

Images of **hurricane in US state**
bing.com/images

hurricane in US state

List of United **States hurricanes** - Wikipedia, the free encyclopedia
https://en.wikipedia.org/wiki/List_of_United_**States_hurricanes**
The list, which is sorted by **U.S. state**, begins in 1851 with the start of the official Atlantic **hurricane** database (HURDAT), as provided by the National Oceanic and **...**
Statistics - Impact - States bordering the Atlantic ...
You visited this page on 7/4/13.

U.S. Mainland **Hurricane** Strikes by **State**, 1851-2004
www.nhc.noaa.gov/past**state**.shtml
by ES Blake
Aug 9, 2011 - **Hurricane** direct hits on the mainland **U.S.** coastline and for individual **states** 1851-2004 by Saffir/Simpson category.

TCFAQ E19) How many direct hits by **hurricanes** of various categories

Fig. 1. Search Result of Concept-based Query in Bing and Google

To this end, our main approach utilizes a comprehensive, probabilistic taxonomy and a search log with access statistics (click-through rate, etc.). The probabilistic taxonomy, called Probase[9], provides large number of concepts and their instances in is-a relations (e.g. katrina *isa* hurricane) as well as the *typicality* of an instance belonging to a concept. For example, a robin is a *typical* bird, but an ostrich is not. Given a user query, we use this taxonomy to detect high level concepts in it and translate them into likely instances to produce a new query, and then return the queries from the query log which are closest to the newly transformed query. Next, we present the prototype system in some detail and some preliminary experimental results.

2 The Prototype System

Our system is divided into two parts, an offline part which creates an index on all historical search queries from the search log and an online runtime system which retrieves a number of relevant queries to an input concept-based query and ranks them according to a scoring function. The architecture of the system is shown in Figure 2. Next, we briefly discuss the two key components in the architecture.

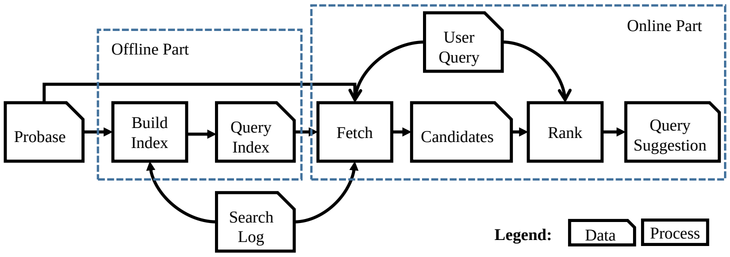


Fig. 2. Architecture of System

2.1 Build Index

First, we parse each historical query in the search log and identify all noun-phrase terms which exist as entities in Probase. Entities are those terms which appear as the instance in at least one is-a pair in the taxonomy, e.g., *katrina*, *louisina*, etc. Then, we conceptualize these instances into their most likely concepts by considering the neighboring instances in the same query, using the technique by Song et al.[7, 8]. For example, query "*katrina* victims from *texas*" maybe conceptualized into "[hurricane] victims from [state]". Finally, we build an index of the search log using the concepts as keys.

2.2 Rank Candidates

Given a query, our system first parses the query and recognizes the concepts in it and then fetch a number of historical queries which are indexed by these concepts as candidate suggestions. Ranking the candidate is the main challenge in this work. We apply a hybrid method to calculate the ranking score. This score takes three factors into consideration, i.e. semantic similarity, context similarity and quality of suggestion query.

The context similarity, $CScore$, is measured by the *edit distance* between the input query and the candidate in terms of the number of insertion, deletion or replacement of words. Matched instances in the candidate and the matched concepts in the input queries which represent semantic information are removed before calculating the edit distance.

The semantic similarity between a candidate query and an input query is the combined distance between the instances in the candidate and the corresponding concept in the input. The distance between an instance and a concept is measured by the *typicality* score between these two terms in Probase. Because *typicality* score is a value between 0 and 1 and can be dense in some range (around 0.01 in practice), we take logistic function on the typical value to expand the range into something more distinguishable. We also use a scalar to make it linear comparable with $CScore$ according to statistics. That is,

$$SScore(t) = \beta \times (-\frac{1}{1 + e^{-\alpha \times t}} + 1)$$

where t is the *typicality* value and α is the factor to locate the dense range and β is a scalar factor. Both factors are learned from some training examples.

We also utilize the click-through rate from the search log, in order to measure the quality or effectiveness of candidate suggestion. In this paper, we use a simple measure defined as

$$QScore(q) = \frac{\text{Num\_clicks(q)}}{\text{Frequency(q)}}$$

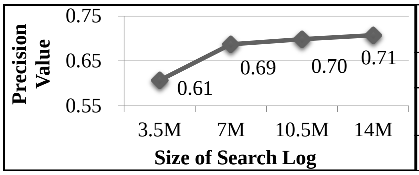
where q is a candidate suggestion query. Besides the click-through rate, the quality score can be extended to include other information from search log.

Currently, the overall score (the lower, the better) is defined as:

$$OverallScore = [CScore + SScore] + e^{-QScore}$$

3 Preliminary Result

We take 1/10 of a 6-month worth of Bing search log as our experimental data source. To evaluate the system, we arbitrarily select 10 concept-based queries related to the events happening during the 6-month time span. All suggested queries from the system are given to 3 human judges who would grade the suggestion on the scale of 1-5, 1 being the least helpful and 5 being the most helpful. The averaged normalized precision score of of these results with different size of search log and the suggestions of two example queries are shown in Figure 3. The results show that having more search log is helpful but the effect saturates at some point. Complete data set as well as evaluation results can be found at `http://adapt.seiee.sjtu.edu.cn/~jack/query/`.



extreme sports in american cities	**diseases found in domesticated animals**
skiing in chicago	tetanus in dog
snowboarding in washington	cirrhosis in horses
sky diving in seattle	herpes in cattle

Fig. 3. Average Precision and Examples of the System

References

1. Baeza-Yates, R., Hurtado, C., Mendoza, M.: Query recommendation using query logs in search engines. In: EDBT 2004 Workshops. pp. 588–596. Springer (2005)
2. Baeza-Yates, R., Tiberi, A.: Extracting semantic relations from query logs. In: Proceedings of ACM SIGKDD. pp. 76–85. ACM (2007)
3. Bhatia, S., Majumdar, D., Mitra, P.: Query suggestions in the absence of query logs. In: SIGIR. pp. 795–804 (2011)
4. Cao, H., Jiang, D., Pei, J., He, Q., Liao, Z., Chen, E., Li, H.: Context-aware query suggestion by mining click-through and session data. In: Proceedings of the 14th ACM SIGKDD. pp. 875–883. ACM (2008)
5. Dupret, G., Mendoza, M.: Recommending better queries from click-through data. In: String Processing and Information Retrieval. pp. 41–44. Springer (2005)
6. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept search: Semantics enabled syntactic search. Semantic Search p. 109 (2008)
7. Song, Y., Wang, H., Wang, Z., Li, H., Chen, W.: Short text conceptualization using a probabilistic knowledgebase. In: IJCAI (2011)
8. Wang, Y., Li, H., Wang, H., Zhu, K.Q.: Concept-based web search. In: Conceptual Modeling, pp. 449–462. Springer (2012)
9. Wu, W., Li, H., Wang, H., Zhu, K.Q.: Probase: A probabilistic taxonomy for text understanding. In: Proceedings of ACM SIGMOD. pp. 481–492. ACM (2012)
10. Zhang, Z., Nasraoui, O.: Mining search engine query logs for query recommendation. In: Proceedings of the 15th WWW. pp. 1039–1040. ACM (2006)