EEST: Entity-driven Exploratory Search for Twitter

Chao Lv, Runwei Qiang, Lili Yao and Jianwu Yang*

Institute of Computer Science and Technology
Peking University, Beijing 100871, China
{lvchao,qiangrw,yaolili,yangjw}@pku.edu.cn

Abstract. Social media has become a comprehensive platform for users to obtain information. When searching over the social media, users' search intents are usually related to one or more entities. Entity, which usually conveys rich information for modeling relevance, is a common choice for query expansion. Previous works usually focus on entities from single source, which are not adequate to cover users' various search intents. Thus, we propose EEST, a novel multi-source entity-driven exploratory search engine to help users quickly target their real information need. EEST extracts related entities and corresponding relationship information from multi-source (i.e., Google, Twitter and Freebase) in the first phase. These entities are able to help users better understand hot aspects of the given query. Expanded queries will be generated automatically while users choose one entity for further exploration. In the second phase, related users and representative tweets are offered to users directly for quickly browsing. A demo of EEST is available at http://demo.webkdd.org.

Keywords: Entity Driven, Twitter Search, Real-Time Exploratory Search

1 Introduction

Social media such as Twitter has become a comprehensive platform for users to obtain information. When searching over the social media, users' initial interest is usually vague. However, their search intents are usually linked to an entity. As related entities can reflect different aspects of a topic, users often choose them to expand their queries. Previous studies mainly focus on how to use related entities as a feedback process. However, they simply adopt entities from single source such as DBpedia, which is not adequate to cover various search intents from our point of view.

Hence, we introduce EEST, a multi-source entity-driven exploratory search engine. For a certain query posted from users, related entities are offered in the first phase, to help users better understand hot aspects of the initial topic. According to the nature of selected source, two kinds of related entities, i.e., real-time entities and historical entities, are adopted. Real-time entities extracted

*Corresponding author.

from real-time source provide latest aspects while historical entities extracted from historical source offer global aspects of the query. Those related entities as well as their relationship will be presented in a graph view, and users can choose a certain entity for further exploration. For each selected entity, EEST will retrieve related users and tweets from Twitter. Moreover, considering the redundancy problem of retrieved tweets, we also apply TTG [1] (tweet timeline generation) to generate a summary that captures relevant information.

2 System Architecture

EEST can be divided into two phases, as illustrated in Figure 1.

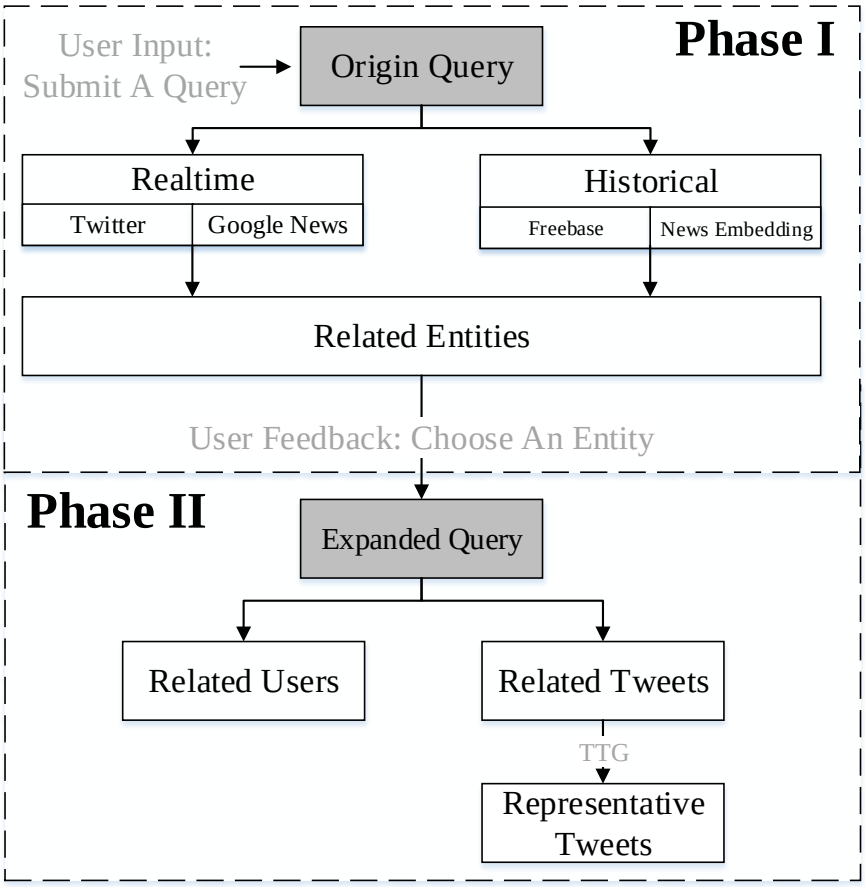

Fig. 1. System Framework

[1] https://github.com/lintool/twitter-tools/wiki/TREC-2014-Track-Guidelines

2.1 Phase I: Entity Extraction

In the first phase, EEST accepts a certain query submitted from a user, and we denote it as *OriginQuery*. To help users better understand hot aspects about the given query, related entities are extracted from multi-source. Extracted entities could be divided into categories of real-time entities and historical entities. Google News and Twitter are chosen as real-time entities source while Freebase and News Embedding are selected as historical entities source.

- **Google News** is our first choice as news is more formal and brief compared to normal text. *OriginQuery* is retrieved in Google News, and the latest related news is returned. We extract related entities from these news text by using named entity recognition (NER) [3]. In particular, three kinds of entities are involved, i.e., Person, Location and Organization.
- **Twitter** is a popular application for users to share and discuss information. Just like Google News, related tweets are retrieved from real-time Twitter steam, then related entities are separated out.
- **Freebase** is a practical, scalable tuple database used to organize general human knowledge [1]. We take advantage of the summary description information in Freebase to get a descriptive text about *OriginQuery*. Similarly, we extract entities from the description content via NER.
- **News Embedding** Currently, the distributed word representations (i.e. word embedding) have attracted more attention in text understanding. The word embedding allows to explicitly encode various semantic relationships as well as linguistic regularities and patterns into the new embedding space [2]. For this purpose, we downloaded pre-trained vectors trained on part of Google News dataset [2]. Then we can compute the cosine similarity distance of *OriginQuery* and other terms in the vocabulary. The top k scored terms are regarded as related entities to the given query.

A new *ExpandedQuery* is generated by combining *OriginQuery* with the chosen entity. This *ExpandedQuery* is going to be transmitted to second phase as input.

2.2 Phase II: Result Presentation

ExpandedQuery is submitted to the Twitter Search [3] in the second phase, and related users and related tweets are obtained. For related users, profile images and lots of statistical information are provided, including followers number, following number, tweets number, etc. For related tweets, TTG is conducted on them with the goal of noise elimination and representative tweets selection. We adopt a star clustering algorithm proposed in [4] as our TTG core algorithm. After that, a clear and representative tweets list will be unfolded in front of users.

[2] https://code.google.com/p/word2vec/
[3] https://dev.twitter.com/rest/public/search

3 Demonstration

We take a query "Obama" as an example to discuss the main modules of EEST briefly, indicated with circled letters in Figure 2.

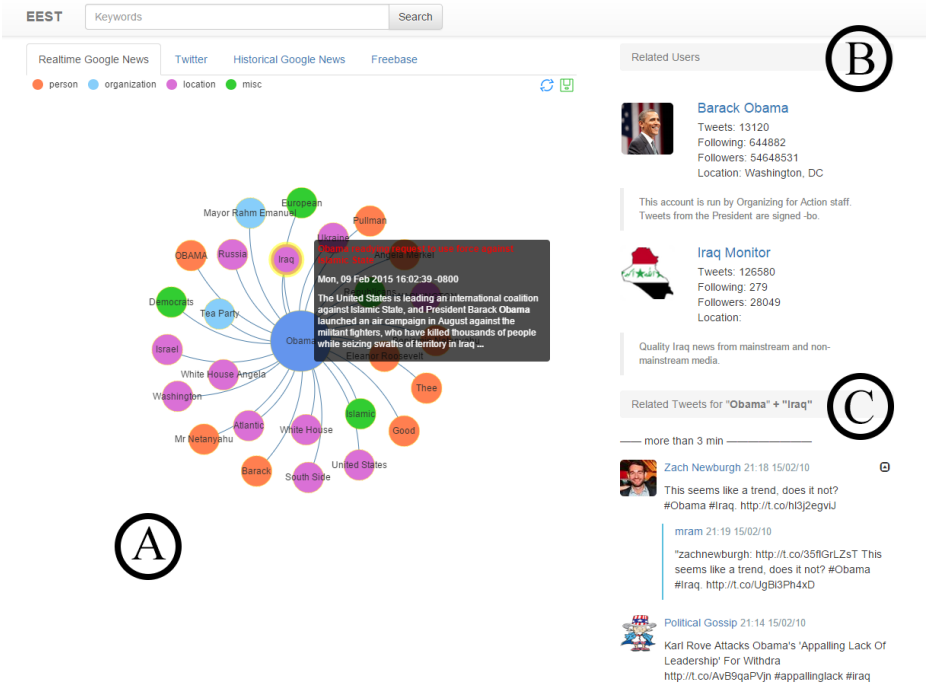

Fig. 2. An example for query "Obama" in EEST.

3.1 A - Related Entities Module

Related entities extracted from multi-source are expressed as a graph in this module. As we can see, "Iraq", "White House" and "European" are highly related to "Obama" in Google News recently. Related news will appear above corresponding entities for users to see what's happening between them. Let us assume that user is interested in entity "Iraq" and choose it for further exploration.

3.2 B - Related Users Module

After the user choose entity "Iraq" for further exploration, two Twitter account, "Barack Obama" and "Iraq Monitor", are extracted and displayed in this module. Statistical information and Twitter account links are offered for users to

easily navigate to their homepages, including followers number, following number, tweets number, etc.

3.3 C - Related Tweets Module

Related tweets talking about "Obama" and "Iraq" are displayed in this module. Redundant tweets are clustered to their representative tweets, which makes the related tweets list readable and clear.

4 Conclusions

In this paper, we have described a demonstration of a multi-source entity-driven search engine for Twitter, called EEST. We presented our initial motivation and the proposed methods, as well as the main functionality of the system. With the help of related entities, users are able to better understand their information need. At the same time, summarized information can save users' time for browsing the retrieval results.

Acknowledgments. The work reported in this paper was supported by the National Natural Science Foundation of China Grant 61370116.

References

1. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: a collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD international conference on Management of data. pp. 1247–1250. ACM (2008)
2. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems. pp. 3111–3119 (2013)
3. Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1), 3–26 (2007)
4. Zeng, H.J., He, Q.C., Chen, Z., Ma, W.Y., Ma, J.: Learning to cluster web search results. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 210–217. ACM (2004)