Neurocomputing 163 (2015) 137–150

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Semantic query suggestion using Twitter Entities

Ioannis Anagnostopoulos [a,*], Gerasimos Razis [a], Phivos Mylonas [b], Christos-Nikolaos Anagnostopoulos [c]

[a] Computer Science and Biomedical Informatics Department, University of Thessaly, Greece
[b] Department of Informatics, Ionian University, Greece
[c] Cultural Technology and Communication Department, University of the Aegean, Greece

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form
22 December 2014
Accepted 25 December 2014
Available online 21 March 2015

Keywords:
Query suggestion
Microblogging
Viral content
Twitter

ABSTRACT

There are many web information management methods and techniques that help search engines and news services to provide useful suggestions with respect to queries, thus facilitating the users' search. However, the penetration of microblogging services in our daily life demands to also consider social sphere as far as query suggestion is concerned. Towards this direction, we introduce an algorithmic approach capable of creating a dynamic query suggestion set, which consist of the most viral and trendy Twitter Entities (that is hashtags, user mentions and URLs) with respect to a user's query. For evaluation purposes, we firstly compare the results derived from two case studies, against the suggestions of popular services like Google News, Yahoo! News, Bing News, and Reuters. In addition we further evaluate our approach with subjective user ratings against Google Trends service. Finally, we provide comparative results that clearly show that our proposal outperforms other methods and baselines in the respective literature.

1. Introduction – motivation and research contribution

Microblogging – a "light", rather live, version of blogging – is considered to be one of the most recent social raising issues on the Internet, being one of the key concepts that brought Social Web to the broad public. The main characteristic of microblogging is the fact that posts are produced almost in real-time and are strictly limited to a specific and rather small amount of characters, such as short sentences, term concatenation or tinyURLs[1] that point to hyperlinks with web and/or even multimedia content. It comprises many very brief updates that are presented to the microblog's readers in reverse-chronological order. Motivated by its increasing popularity, among many microblogging services we focus herein on the Twitter[2] social network, where microblogs are known as ***tweets***.

In web information retrieval, the effectiveness of search engines strongly depends on whether users can express their information needs through the terms they submit. However, submitting queries is not an easy task. Queries are short, not written in natural language, and – mostly – their terms are ambiguous. Many proposed methods offer meaningful query suggestions, usually by employing knowledge extraction methods from browsing history records or search logs. However, very few consider time as an important parameter related to the actual meaning of a query term. Thus, in this work, we do not tackle query suggestion in the traditional way, but we provide time-aware suggestions according to the most viral terms that appear in Twitter along with the user's query.

In other words, the main contribution of this work is the effective suggestion of microblogging social content (called hereafter as *Twitter Entity*/*Entities* – TE/TEs) that manage to become viral in time, given a user query; the more viral the social content is, the more relevant are the suggestions. Our ultimate goal is to provide users a way to enter any type of query and retrieve accurate, relevant and popular (viral) Twitter Entities suggestions that would semantically "fit" to their information needs. In order to measure virality, we extend the capture–recapture methodology, which is mainly used for estimating population properties (e.g., birth/survival rates) in real-life biological experimentations; in our work, the concept of virality in social content is considered equal to survivability in animal populations under study. The concept of social content on the other hand is directly related to Twitter Entities (hashtags (#), user mentions (@) and URLs) and should not be related with named entities or Wikipedia concepts as considered in

* Corresponding author.
E-mail addresses: janag@dib.uth.gr (I. Anagnostopoulos), razis@dib.uth.gr (G. Razis), fmylonas@ionio.gr (P. Mylonas), canag@ct.aegean.gr (C.-N. Anagnostopoulos).
URL: http://www.anagnostopoulos.name (I. Anagnostopoulos).
[1] http://en.wikipedia.org/wiki/TinyURL.
[2] http://www.twitter.com.

http://dx.doi.org/10.1016/j.neucom.2014.12.090
0925-2312/

most papers in the related literature of text mining and information retrieval[3] (e.g., see the work described in [20]).

It is rather true that research works on microblog posts analysis and extraction of meaningful information from them in a (semi-) automated manner have been considered recently in the literature, yet we have reason to believe these approaches are quite different to ours. As the interested reader will see within next Section 2 of this manuscript, related research on query suggestion [15] is highly related with query expansion [14] query substitution, query recommendation or query refinement tasks. In this work, we deviate from the traditional query suggestion proposals in a sense that users have their queries expanded directly from Twittersphere,[4] and without having their queries or browsing history processed by search engines. In addition, another important difference against related query suggestion techniques, focused on web search and the real-time variant of the problem at hand, is the narrow time frame considered herein in which suggestions have maximal impact. For this work we were extra motivated by the facts that (a) microblogging social content annotation is provided directly in real-time by users worldwide and (b) the more this annotation becomes important or so-called "*viral*", the more semantically related it becomes with a recent trend, top news, thematic categorization, etc.

The rest of this manuscript is organized as follows. In the next section, we provide an overview over the literature within the query suggestion field, emphasizing on related works within the social sphere. Section 3 provides an overview of the methodology we use, as well as the basic steps of our proposed query suggestion method. In Section 4 – and in order to clearly show how our query suggestion expansion mechanism works – we describe the results of two real-life scenarios (case studies) that lasted more than a week within January 2014. In addition, we evaluate our results against four famous Internet commercial news services (Google News, Yahoo! News, Bing News, and Reuters). In Section 5 we further evaluate our approach by subjective comparisons with respect to Google Hot Trends service, as well as against a cluster labeling and a microblog retrieval task for a time span of one month (May 9, 2014–June 9, 2014), providing comparative results. Finally, in Section 6 we conclude this work by highlighting its main outcomes, and describing in parallel our future directions based on the experiences we faced.

2. Works on related information search and retrieval tasks

Given the fact that microblogging is increasingly popular, several research methods for organizing and providing access to microblog data have been emerged on this topic since the last few years. In this section, we provide an overview over microblog-related information retrieval research works focusing on the field of query suggestion.

2.1. Information search and retrieval in microblogs

In general, the social sphere consists of the so-called tweets or microblogging posts [5], where the large amount of real-time tweets per day is highly attractive for information retrieval research. Within the *social sphere* context query suggestions must be in real-time, i.e., results need to be temporally relevant and timely [15]. Now, microblogs form a rather special category of user-generated data: they typically contain two major characteristics that seriously affect linguistic analysis techniques, namely: (a) they contain strong vernacular (acronyms, spelling changes, etc.) and (b) they do not include any memorable repetition of words. More specifically, Massoudi et al. in [14] studied a Twitter-based retrieval model by considering the model with textual quality and Twitter specific quality indicators. Naveed et al. [16] combined document length normalization in a retrieval model to resolve short texts sparsity in the case of tweets. Motivated by the observation that in a typical microblog user tends to retrieve meaningful information through queries formulation, researchers focus on each post's characteristic features [8], whose quantitative evaluation could potentially affect the way in which the relevance between the user query and its returned results may be calculated. Even TREC 2011 introduced the Microblog Track which addressed one single pilot task, entitled "real-time search task", where the user wished to see the most recent but relevant information to the query (e.g., [17]). A first step towards this direction is discussed in [22], where Tao et al. identify two feature categories, i.e., features related or not to the user query.

The fact that microblog posts contain *hashtags* is also exploited in the literature in the direction of acquiring information that the user "is not aware of" and formulate queries that the user "does not know how to express" (e.g., [4]). In a representative approach [5] and given a query, Efron attempts to statistically identify a number of hashtags relevant to the given query, that may be used to expand it and lead to better results. Even in our own previous work [2], we proposed the utilization of hashtags as the main source of information acquisition, by searching the specific query terms within microblog posts under the condition that the former need to appear as hashtags; then, we calculated the most common hashtags that co-occur together with the original query, and, thus, expanded the query with the new hashtags. Last but not least, the observation that microblog posts are created during an actual event and contain comments and/or information directly related to it leads to event detection research efforts [18] based on posts and/or hashtags.

2.2. Query manipulation works

Typical microblog query manipulation research problems include both *query analysis and expansion* and *query suggestion* approaches. Still, there are also some distinctive differences between the two. A *query expansion* task is typically used transparently to the end-user and internally within a search engine mechanism, whereas a query suggestion is exposed to its end-users and therefore can use additional explicit information to its aid. In this manner, Bandyopadhyay et al. [3] attempt to improve weak ad-hoc queries through a process they call "web assistance", by exploring standard query expansion approaches and utilizing external corpora as a source for the query expansion terms, namely pages derived from the Web and their titles. Efron [5] showed that for a Twitter microblog collection, hashtags may be predicted using query expansion techniques; he proposed restricting the added query terms to those candidates that are hashtags, stripping candidates of their leading "#". In another more recent approach, Kumar and Carterette [10] take into account the fact that most existing models for Information Retrieval do not take the very important time aspect into account and focus on Twitter search models; they utilize time-based feedback and a simple query expansion by using highly frequent terms in top tweets as their expanded terms. In another detailed approach by Massoudi et al. [14], authors propose an efficient dynamic query expansion model for microblog post-retrieval, utilizing a language modeling approach to search microblog posts by incorporating query expansion and certain "quality indicators" during matching. The latter is very interesting since several typical microblog characteristics may be exploited as quality indicators, such as temporal [12] or topological ones.

In the case of actual *query suggestion* tasks though, the problem at hand becomes slightly different and its complexity increases as all current major web-search engines and most proposed methods that suggest queries rely solely on search engine query logs to determine

[3] Thus, it should be clear that whenever we mention the term "entity" in the manuscript, we refer to Twitter Entity/Entities (TE/TEs), unless otherwise explicitly stated.

[4] http://www.oxforddictionaries.com/definition/english/Twittersphere.