Query Expansion with a Little Help from Twitter

Ioannis Anagnostopoulos\(^1\), Gerasimos Razis\(^1\), Phivos Mylonas\(^2\), and Christos-Nikolaos Anagnostopoulos\(^3\)

\(^1\) Computer Science and Biomedical Informatics Dpt., University of Thessaly
\(^2\) Electrical & Computer Engineering School, National Technical University of Athens
\(^3\) Cultural Technology and Communication Dpt., University of the Aegean
{janag,grazis}@ucg.gr, fmylonas@image.ntua.gr, canag@ct.aegean.gr

Abstract. With the advent and rapid spread of microblogging services, web information management finds a new research topic. Although classical information retrieval methods and techniques help search engines and services to present an adequate precision in lower recall levels (top-k results), the constantly evolving information needs of microblogging users demand a different approach, which has to be adapted to the dynamic nature of On-line Social Networks (OSNs). In this work, we use Twitter as microblogging service, aiming to investigate the query expansion provision that can be extracted from large graphs, and compare it against classical query expansion methods that require mainly prior knowledge, such as browsing history records or access and management of search logs. We provide a direct comparison with mainstream media services, such as Google, Yahoo!, Bing, NBC and Reuters, while we also evaluate our approach by subjective comparisons in respect to the Google Hot Searches service.

Keywords: Query expansion, Microblogging services, Twitter, Social Data Mining.

1 Introduction

Microblogging is considered to be one of the most recent social raising issues of Web 2.0, being one of the key concepts that brought Social Web to the broad public. In other words, microblogging could be considered as a "light" version of blogging, where messages are restricted to less than a small amount of characters. Regarding their actual message content, this may be either textual data (e.g. short sentences), or even multimedia content (e.g. photos or hyperlinks to video sources). Yet, its simplicity and ubiquitous usage possibilities have made it one of the new standards in social communication; i.e., there is already a large number of social networks and sites that appear to have incorporated few or more microblogging functionalities; Twitter and Facebook being the most famous. The task of analyzing microblog posts and extract meaningful information from them in a (semi-)automated manner has been considered recently by some works in the literature, yet we believe their approaches are quite different to the one presented herein. Being part of a vast amount of information
disseminated on the Web, it is very crucial for users to find relevant information in blogs or having recommendation in respect to their queries. Thus, modern information services provide a lot of mechanisms for suggestions in respect to users’ information needs expressed by mostly syntactic queries. Research on query suggestion is highly related with query expansion [1], query substitution [2], query recommendation [3] or query refinement [4]. All are considered as similar procedures, aiming to adjust an initial user query into a revised one, which then returns more accurate results. In this work, we deviate from the traditional query suggestion proposal in a sense that users have their queries expanded directly from Twitter, and without having their queries or browsing history processed by search engines.

The remainder of this paper is organized as follows. In the next section we provide an overview over the related work on query analysis and expansion issues that need addressing in microblogging services. Section 3 provides the methodology we use, as well as the basic steps of our proposed algorithm. In Section 4, we describe a real case study in order to clearly show how our query expansion mechanism works. Finally, in Section 5, we evaluate our results against Google, Yahoo!, Bing, NBC and Reuters, while we also evaluate our approach by subjective comparisons in respect to the Google Hot Searches service. Section 6 concludes our work by summarizing the derived outcomes, providing in parallel some of our future directions.

2 Related Work

In general, microblogging posts [5] form a special category of user-generated data containing two major characteristics, that seriously affect linguistic analysis techniques [6], namely: a) they contain strong vernacular (acronyms, spelling changes, etc.) and, b) in principle they do not include any memorable repetition of words. Motivated by the observation that a microblog user retrieves information through queries formulation in order to acquire meaningful information, researchers focus on each post's characteristic features [7], whose quantitative evaluation could potentially affect the way in which the relevance between the user query and its returned results may be calculated. A first step towards this direction is discussed in [8], where authors identify two feature categories, i.e., features related to the user query and thus calculated as soon as the latter is formed and features that are not related to the specific query, but are inherent posts and thus calculated when the latter are modified, updated or added. In the context of social networking, query expansion techniques are of great interest using either previously constructed language models [9] or by taking into account personal user preferences, such as those resulting from user microblog posts and hashtags analysis [10]. The fact that microblog posts contain hashtags is also exploited in the literature towards query expansion methodologies 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" [11]. In [12], given a query, authors attempt to identify a number of hashtags relevant to the given query, that may be used to expand it and lead to better results; the proposed method is based solely on statistical techniques by building probabilistic language models for each available hashtag and by using a suitable microblog posts corpus. Even in our own recent previous work