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

L. Iliadis, H. Papadopoulos, and C. Jayne (Eds.): EANN 2013, Part II, CCIS 384, pp. 20–29, 2013.

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 inter-est 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 suit-able microblog posts corpus. Even in our own recent previous work

[13], 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 with the original query and thus expanded the query with the new hashtags. Finally, another broad related research category is the one formed by the observation that microblog posts are created during an actual event and contain comments or in-formation directly related to it, thus leading to event detection research efforts [14] based on posts and/or hashtags.

3 Proposed Query Expansion Mechanism

The microblogging service we use is Twitter and the Query Expansion (QE) mechanism is based on concatenated terms, known as hashtags (prefixed with "#"). Hashtags are actually a way Twitter users can semantically annotate the tweeted content, while there are no complicated syntactic rules, so Twitter users can annotate the information according to their will. This freedom of expression provide the best way to create a vast pool of crowd-sourced meta-data, leading to trends that best describe a social aware issue.

Prior to describing our QE method, we need to briefly introduce the context of capture-recapture experiments used in wildlife biological studies [15]. In these experiments, animals, birds, fishes or insects (subjects of investigation) are captured, marked and then released. If a marked individual is captured on a subsequent trapping occasion, then it is mentioned as “recaptured”. Based on the number of marked individuals that are recaptured, and by using statistical models, we can estimate the total population size, as well as the birth, death and survival rate of each species under study. The sampling process is divided into k primary sampling periods, each of them consisting of l secondary sampling periods. Among primary sampling periods we assume that in the population we can have births, deaths and/or migration incidents. This population is called “Open”. On the other hand, among secondary sampling periods the population is assumed closed, meaning that there are not gains of losses in the population [15]. During a secondary sampling period a set of different species is randomly selected, marked and keeping in parallel a history record of them, and then released back to nature. After a specific time interval, the second secondary sampling periods occurs and so forth, until the end of the last l secondary sampling period. Secondary periods are near and very short in time, while trapping occasions are considered instantaneous in order to assume that the populations under study are closed, meaning that no losses or gains occur during these time intervals. However, longer time intervals between primary sampling periods are desirable so evolution events can occur (e.g. survival, movement, and growth), as defined in the basic structure of Pollock’s robust design model [16], which extends the Jolly-Seber model [17].

In wild-life experiments this model is applied to open populations, in which death, birth, and migration incidents possibly occur in the populations under study. In our case, birth means the appearance of a new Twitter hashtag, while death and migration incidents corresponds to evolution of hashtags. Moreover, a basic evolution metric we

want to employ in our methodology is the survival rate of the subject under investigation (the expanded term in our paradigm). This actually highlights how durable in time hashtags are correlated to each other in the created real-time graph.

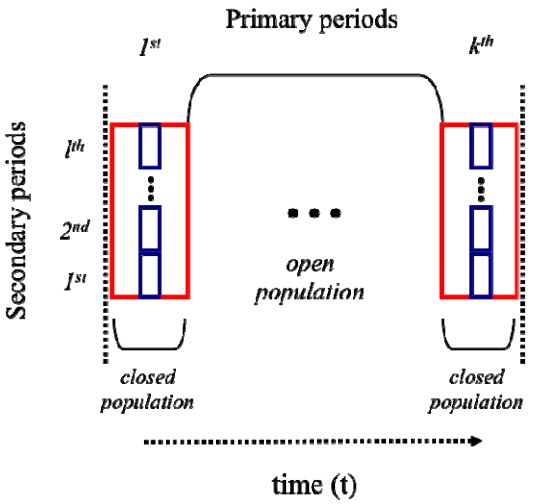

Fig. 1. The modified structure of the sampling scheme used in this work

However, since it is impossible to assume that due time the population of Twitter hashtags is closed, we modified the capture-recapture experimentations, in order to simultaneously conduct the secondary periods. This does not violate the assumptions of the Pollock's model, since in the real-life case, biologists set traps to different locations in the same space universe (e.g. a lake, a national park), while in the Twittersphere the space universe cannot be separated. Thus, Figure 1 depicts the modified structure of the sampling scheme we employ for the web paradigm. In our case, the individuals under study are Twitter hashtags. The trapping occasions are the keywords/seeds to be extended in our query. Each trapping occasion occurs in the primary sampling period, which we split in l secondary and simultaneously sampling occasions. In each of these l sampling instances we capture (and then mark) some hashtags with the same probability value p. By this we ensure that all secondary samplings are made in a "close" pool of instances under the basic principle of the Pollock's model. Then by investigating the recaptured instances in subsequent primary occasions, we calculate the survival probability of the examined hashtag according to Equation 1, where $(M_i - m_i)$ defines the marked hashtags not captured during the i^{th} sampling period, while R_i are the hashtags captured at the i^{th} period, marked, and then released for possible recapture in future samplings.

$$\tilde{\varphi}_i^e = \frac{\tilde{M}_{i+1}^e}{\tilde{M}_i^e - m_i^e + R_i^e} \tag{1}$$

4 Real-Life Experimentation: The Boston Marathon Bombing Case

As case study we use the “Boston Marathon Bombing” case, where during the Boston Marathon on April 15, 2013, two pressure cooker bombs exploded, killing 3 people and injuring 264 individuals[1]. This shocking event was among the breaking news globally for several weeks, since it was emotionally touched millions of people worldwide. Apart from mainstream media (e.g. TV/Radio), social media platforms covered all aspects of the incident disseminating an enormous amount of information, which was created from millions of users. Especially in Twitter, information contained not only shared information, but also personal opinions/thoughts, and constantly new links related to the crime, directly related to user-generated hashtags as semantic annotations. Having the experience for a previous work of ours [13], we selected the words "Boston" and “Marathon” as the basic terms someone enters in a search engine in order to find relevant information. It is worth noticing that even after one week after the event, most search engines in their main front-end environment did not suggest relevant terms after these terms. Our main contribution here is to provide query expansion to the user’s submitted term(s) under the knowledge disseminated in Twitter, without having any other access or use of search engines' query logs.

In Figure 2, we can see the filtered graph that presents the entities with the 10% higher values in respect to eigenvector centrality (EiVC) values measured. This graph corresponds to a captured instance in one of the secondary sampling occasions conducted during our experiments. In this figure edges are classified as:

- Colour: black, name: *searchQuery*-from_user, explanation: This property is applied to edges between a Twitter user that used a queried term and the queried term,
- Colour: green, name: from_user-to_user, explanation: This property is applied in order to create an edge between a Twitter user that replied using a tweet to (an)other Twitter user(s),
- Colour: red, name: from_user-mentioned_user, explanation: This property is applied in order to create an edge between a Twitter user who mentioned at a tweet (an)other Twitter user(s) and the mentioned user(s),
- Colour: blue, name: from_user-tweeted_hashtag, explanation: This property is applied in order to create an edge between a Twitter user that included a hashtag in a tweet and the included hashtag,
- Colour: Yellow, name: from_user-tweeted_URL, explanation: This property is applied in order to create an edge between a Twitter user that included a Url in a tweet and the included Url,
- Colour: Purple, name: hashtag-URL, explanation: This property is applied in order to create an edge between hashtag and a Url in case that both of these hashtags are included in a tweet.

[1] Boston Marathon bombings, `http://en.wikipedia.org/wiki/Boston_Marathon_bombings`, last accessed in May 13, 2013.

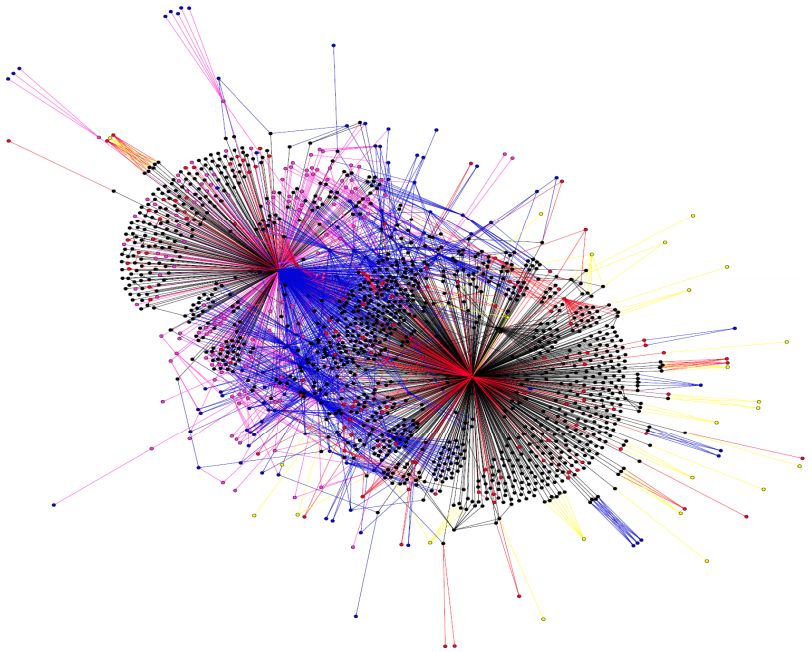

Fig. 2. A random selected secondary sampling occasion (April 18, 2013)

Table 1. Query expansion and suggested terms / case study evaluation

Query terms {Boston, Marathon}	**Metrics**		**Also suggested by (in the same time period)**			
Extended Term	**Entity type**	**Score**	**Google**	**Ya-hoo!**	**Bing / NBC**	**Reu-ters**
bostonstrong	#	0.2082	✗	✗	✗	✗
bostonbombing	#	0.1571	✔	✔	✔	✔
prayforboston	#	0.08359	✗	✗	✗	✗
news	#	0.04921	✔	✗	✗	✗
manhunt	#	0.04425	✗	✗	✗	✗
syria	#	0.04232	✗	✗	✗	✗

Similarly, nodes are classified as:

- Colour: Black, name: Trend/searchQuery, explanation: The queried term(s)
- from_user, colour: Black, explanation: A Twitter user that used the queried term(s)
- Colour: Green, name: to_userexplanation: A Twitter user that received one or more tweets as replies to a created tweet

- Colour: Red, name: mentioned_user, explanation: A Twitter user who is mentioned at other users' tweets
- Colour: Blue, name: tweeted_hashtag, explanation: A hashtag that is included in a tweet
- Colour: Yellow, name: tweeted_URL, explanation: A Url that is included in a tweet

5 Evaluation

In order to evaluate our QE mechanism we firstly compare the results derived from our case study in respect to the query suggestions of well-known search engines, like Google, Yahoo!, Bing, as well as mainstream Web media services, such as Reuter News and NBC. The second part of our evaluation, describes a generic evaluation, which involves subjective user ratings for results obtained from our QE mechanisms and Google.

5.1 Case Study Evaluation

Evaluation results for the "Boston Marathon" case are provided in Table I. Initial query terms ("Boston", "Marathon") are on the left side of the table, and then follow the expanded term(s). All suggested extra terms are according to a normalized survival probability of hashtag e during primary period i ($\tilde{\varphi}_i^e$), as defined in Equation 1. Finally, the last four columns of Table I indicate whether the specific expanded query has been suggested (even in different order of terms with respect to the seed term) by Google, Yahoo!, Bing, NBC and Reuters[2]. The date we performed this evaluation was April 18, 2013, just three days after the bombing incident. The trend analysis and the expanded terms in respect to the initial provided, performed between subsequent primary periods, where each of them consisted of two simultaneously secondary periods with capture probability equal to 0.3. We notice that apart a very good suggestion on related web sources for direct information (e.g. BBC World, 7news, etc.), our mechanism proposes two other really noteworthy acquisition terms like "manhunt" and "Syria", which are not suggested by the other four search services. It is worth noticing that during the next day (April 19, 2013) where one of the suspects was arrested, our algorithm "captured" two hashtags related with the suspect name (#dzhokhartsarnaev, #tsarnaev) having a quite high normalized scoring value (0.079 and 0.1302 respectively).

5.2 Evaluation against Google Hot Searches

In this sub-section we describe an evaluation of our QE mechanism in comparison with Google Hot Searches[3] service. For the purposes of this evaluation 17 individuals

[2] NBC search service is powered by Bing.
[3] `http://www.google.com/trends/hottrends`

were engaged. Their task was to subjectively rate the expanded queries against the respective service from Google. Each individual (postgraduate students from an MBA course class at National Technical University of Athens) was asked to select three different events from Google Hot Searches for a specific testing period (a specific week during March 2013). Each individual had to explicitly rate the suggested entities/hashtags derived by our QE mechanism, against their selected events as appear in Google Hot Searches (in U.S.). The rating performed upon a five-point Likert scale as indicated below:

1. Strongly disagree (totally irrelevant suggestion)
2. Disagree (not so good suggestion)
3. Neither agree nor disagree (nearly same suggestion)
4. Agree (potentially better)
5. Strongly agree (surely better)

After processing the one-week results we ended up with 87 unique related terms (as these were provided by Google Hot Searches) in 31 distinct events (20 out of the 51 events were identical). The average amount of suggested terms per tested event was 2.81, which practically means that nearly 3 terms in average expand the basic term that describe an event. At Figure 3, we can see some points that indicate the average evaluator rating for suggested hashtags, as derived from our QE mechanism in respect to scoring value φ. We noticed that the larger the φ is, the higher the mean subjective rate appears in the five-point Likert scale. More specifically, suggested hashtags that appeared as expanded terms having $\tilde{\varphi}_i^e \geq 0.7$ (even as suggested in concatenated format), were subjectively evaluated as more relevant presenting nearly

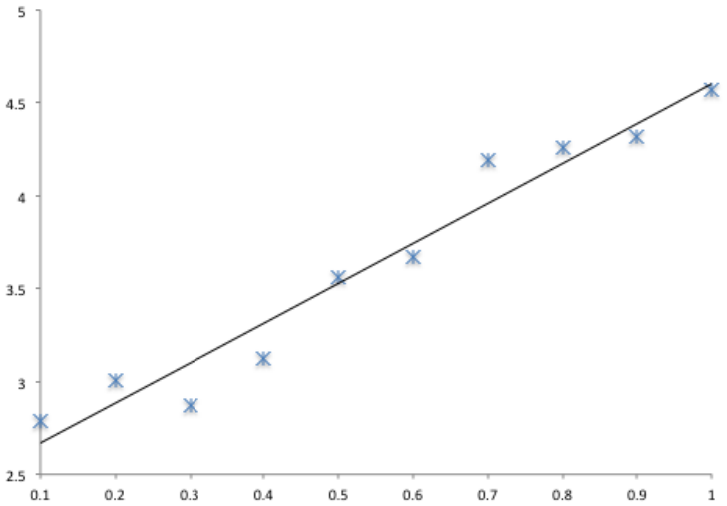

Fig. 3. Mean evaluator ratings vs. proposed hashtags

one-point higher level in the Likert scale. In other words, this means that through subsequent samplings in Twitter, several user-generated terms (hashtags) are more descriptive in comparison to a related query term in Google's log. This was somehow expected, since we performed short-term trend analysis rather than long-term log analysis, yet it is an indicative assumption that our QE method is in the right direction. We can also notice that the majority of subjective rates in average values (more than 60%) were close or slightly higher in the third level (point 3) in the Likert scale, thus indicating a "nearly same suggestion" in comparison to Google search service.

6 Conclusion – Future Directions

In this paper, we proposed a Query Expansion (QE) mechanism, which employs trend analysis issues from microblogging services (case study in Twitter). In order to efficiently analyze trends in terms of retrieving suggest terms for query expansion and reducing the sampling cost, we used the well-known capture-recapture methodology, which is mostly applied in biology experiments. For evaluating our proposal, we presented a recent real event, while we further evaluated it through subjective rates against Google Hot Search service, having as pool a class of 17 postgraduate students. Ongoing research is performed on how other Twitter entities like Twitter mentions (@s), URIs and other related information (e.g. images, geo-location, replies) can be part of suggested term expansion.

References

1. Xu, J., Croft, W.B.: Query expansion using local and global document analysis. In: 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996), pp. 4–11. ACM, New York (1996)
2. Jones, R., Rey, B., Madani, O., Greiner, W.: Generating query substitutions. In: 15th International Conference on World Wide Web (WWW 2006), pp. 387–396. ACM, New York (2006)
3. Baeza-Yates, R., Hurtado, C.A., Mendoza, M.: Query recommendation using query logs in search engines. In: Lindner, W., Fischer, F., Türker, C., Tzitzikas, Y., Vakali, A.I. (eds.) EDBT 2004. LNCS, vol. 3268, pp. 588–596. Springer, Heidelberg (2004)
4. Kraft, R., Zien, J.: Mining anchor text for query refinement. In: 13th International Conference on World Wide Web (WWW 2004), pp. 666–674. ACM, New York (2004)
5. Efron, M.: Information Search and Retrieval in Microblogs. Journal of the American Society for Information Science and Technology 62(6), 996–1008 (2011)
6. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating Query Expansion and Quality Indicators in Searching Microblog Posts. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 362–367. Springer, Heidelberg (2011)
7. Huberman, B.A., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (January 2009)

8. Tao, K., Abel, F., Hauff, C., Houben, G.-J.: Twinder: A Search Engine for Twitter Streams. In: Brambilla, M., Tokuda, T., Tolksdorf, R. (eds.) ICWE 2012. LNCS, vol. 7387, pp. 153–168. Springer, Heidelberg (2012)
9. Packer, H.S., Samangooei, S., Hare, J.S., Gibbins, N., Lewis, P.H.: Event Detection using Twitter and Structured Semantic Query Expansion. In: Proceedings of the 1st International Workshop on Multimodal Crowd Sensing (CrowdSens 2012), Sheraton, Maui Hawaii, pp. 7–14 (2012)
10. Zhou, D., Lawless, S., Wade, V.: Improving search via personalized query expansion using social media. Information Retrieval 15(3-4), 218–242 (2012)
11. Bouadjenek, M.R., Hacid, H., Bouzeghoub, M., Daigremont, J.: Personalized Social Query Expansion Using Social Bookmarking Systems. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2011), Beijing, China, pp. 1113–1114 (2011)
12. Efron, M.: Hashtag retrieval in a microblogging environment. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, Geneva, pp. 787–788 (2010)
13. Anagnostopoulos, I., Kolias, V., Mylonas, P.: Socio-semantic query expansion using Twitter hashtags. In: Proceedings of the 2012 7th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2012), Luxembourg, pp. 29–34 (2012)
14. Packer, H.S., Samangooei, S., Hare, J.S., Gibbins, N., Lewis, P.H.: Event Detection using Twitter and Structured Semantic Query Expansion. In: Proceedings of the 1st International Workshop on Multimodal Crowd Sensing (CrowdSens 2012), Sheraton, Maui Hawaii, pp. 7–14 (2012)
15. Pollock, K.H., Nichols, J.D., Brownie, C., Hines, J.E.: Statistical inference for capture-recapture experiments. Wildlife Monographs 107 (1990)
16. Schwarz, C., Stobo, W.: Estimating temporary migration using the robust design. Biometrics 53, 178–194 (1997)
17. Jolly, G.: Explicit estimates from capture-recapture data with both death and immigration stochastic model. Biometrika 52, 225–247 (1965)