Exploiting Social #-Tagging Behavior in Twitter for Information Filtering and Recommendation

[TREC 2011 – Microblog Track]

Ernesto Diaz-Aviles
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
diaz@L3S.de

Patrick Siehndel
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
siehndel@L3S.de

Kaweh Djafari Naini
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
naini@L3S.de

ABSTRACT

We present a ranking approach for Twitter documents that exploits social hashtagging behavior. We first map topics of user interest, represented by keywords, to a set of twitter hashtags that we use as query terms to retrieve twitter documents (*tweets*) based on tf-idf scores, with the additional restrictions that the documents retrieved should occur before the query timestamp. We show that this simple method performs significantly better than a disjunctive baseline based on the topic description.

Keywords

microblog, ranking, recommender systems, social tagging, web 2.0

1. INTRODUCTION

The Social Web is successfully established and poised for continued growth. Social networking sites such as Twitter (twitter.com), a microblogging service, have experienced an explosion in global Internet traffic over the past years. Twitter itself is considered as one of the most-visited sites worldwide [3]. It is estimated to have over 200 million users, generating more than 100 million short messages (*tweets*) every day, handling over 800,000 search queries daily [8, 6]. This vast amount of information, exchanged in real-time, brings critical challenges in applying traditional Information Retrieval (IR) or Collaborative Filtering (CF) techniques to social media streams.

Despite the amount of research attracted by Twitter in the last years, search and online ranking on Twitter have not yet been addressed extensively. The first TREC2011 Microblog Track aims to fill this gap.

TREC2011 Microblog track addresses a realtime search task where a user wishes to see the most recent but relevant information at a specific time. User's information need is represented by 50 topics, each of the topics is represented by a set of keywords. The system should answer a query by providing a list of relevant tweets ordered from newest to oldest, starting from the time the query was issued.

A particular characteristic of Twitter messages is the use of hashtags. A hashtag is the specific name for a tag in Twitter. Hashtags derive their name from the fact that they are preceded by the symbol '#', also known as a hash mark, e.g., #TREC2011. Tagging has proved to be an intuitive and flexible Web 2.0 mechanism to facilitate search [2], navigation (e.g., tag clouds) [1] and to improve the performance of recommendation systems [7, 4], but tagging practices in Twitter are different from those in other Web 2.0 systems such as Flickr[1] or Delicious[2]. Tagging in Twitter is more about filtering and directing content so that it appears in certain streams, the tag itself is not just metadata but integral part of the message, and can either serve as a label in the traditional sense of a tag, or it can serve as a prompt for user comment [5]. Can hashtags help us to search and retrieve relevant tweets?

The main contribution of this paper, is an approach that exploits the social hashtagging behavior in Twitter to rank tweets for a topic of interest. The performance achieved makes it specially attractive for information and collaborative filtering tasks, where a personalized lists of items (e.g., tweets) needs to be computed based on the user-item interactions in the system.

We argue that our simple approach captures part of Twitter's social dynamics and should be used to compare the performance of more complex systems.

2. OUR APPROACH

Dataset Collection and Preprocessing

The data used for the TREC2011 microblogging track is the *Tweets2011* corpus[3]. The corpus is comprised of 2 weeks of tweets (from January 24th until February 8th, 2011, both dates inclusive) and it is considered to be a representative sample of the *twittersphere*. The tweets comprising the corpus have to be downloaded directly from Twitter using a tool provided by the track organizers. This implies that even though the lists of tweets provided by the track are the same to all participants, the snapshot of the data, i.e., the

Text REtrieval Conference (TREC) 2011. Microblog Track.
NIST in Gaithersburg, Md. USA.
.

[1] http://flickr.com/

[2] http://delicious.com/

[3] https://sites.google.com/site/microblogtrack/

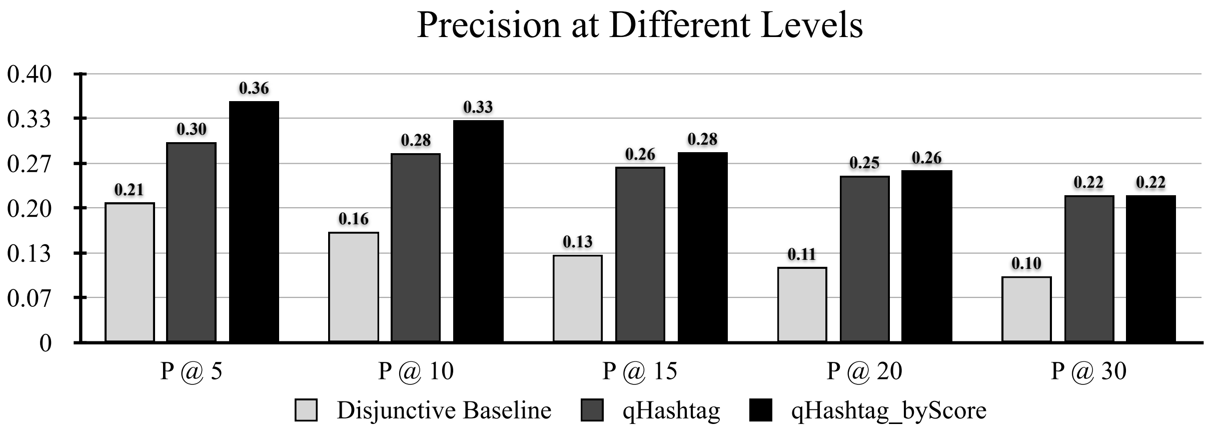

Figure 1: Precision at different levels for our approach using query terms with hashtags (qHashtag) versus the official disjunctive baseline. The figure also presents the results obtained after ensuring that the run is sorted by its score, and not by the id of the document (qHashtag_byScore).

exact number of tweets and their content, is not necessarily the same to everybody given the dynamics of Twitter, where users may change their screen names, cancel their accounts, or delete posted statuses.

We collected the HTML variant of the corpus. The crawling completed on July 8th, 2011. In total we were able to retrieved over 14 million tweets. Table 1 shows the number of tweets collected and the corresponding HTTP status codes.

After completing the crawling, we filtered out tweets that were not in English. We used the language detection tools provided by Cybozu Inc.[4]. We had a final data set of 4,681,523 tweets for our experiments.

Table 1: Number of Tweets collected via HTML screen-scraping and the corresponding HTTP status codes.

HTTP Status Code	No. of Tweets Collected
200	12,409,257
302	986,181
404	629,084
403	219,028
Total	**14,243,550**
Tweets in English	**4,681,523**

Our Strategy

Our strategy can be summarized as follows:

1. We map each topic to a set of representative hashtags. We do this by considering the topic's keywords, or a combination of them, as candidate hashtags that will be used as query terms.

2. We index the corpus. We load the tweets into a MySQL database and then create a full text index on the tweet message text.

3. We use the hashtags as query terms, and for each topic retrieve the top-30 tweets based on tf-idf scores that have a timestamp less than the query time. We finally sort the tweets chronologically. Note that the top-30 list does not include retweets, as they were assumed to be non-relevant according the guidelines of the competition.

Table 2 shows for each topic the associated query terms with the corresponding hashtags used in the track.

Results

The results are presented in Figure 1. Topic 50 did not have any relevant tweets in the dataset, and so was dropped from the evaluation. The evaluation considers all relevant and highly relevant tweets as relevant and is over 49 topics.

We can see that our approach, that relies on the social behavior captured by hashtags, largely outperforms the official baseline.

The complete evaluation files of our run are available at `http://www.L3S.de/~diaz/trec2011/` .

3. CONCLUSION

We contributed a simple method that makes use of hashtags as part of query terms to retrieve relevant tweets. The results obtained significantly improve over the official disjunctive baseline. Given the improvements in precision of short lists, e.g., precision at 5 or 10, we plan to extend this method to a recommender system and personalized setting.

Acknowledgments

This work was funded, in part, by the European Commission Seventh Framework Programme (FP7/2007-2013) under grant agreements No.247829, No.231854 and No.231126.

[4] `http://code.google.com/p/language-detection/`

Table 2: Topics and associated query terms with hashtags.

Topic	Title	Query Using Hashtags
1	BBC World Service staff cuts	#bbc staff cut
2	2022 FIFA soccer	#fifa 2022
3	Haiti Aristide return	#aristide
4	Mexico drug war	#mexico drug war
5	NIST computer security	#nist
6	NSA	#nsa
7	Pakistan diplomat arrest murder	#pakistan diplomat arrest murder
8	Phone hacking British politicians	#uk phone hacking
9	Toyota Recall	#toyota Recall
10	Egyptian protesters attack museum	#Egypt protesters attack museum
11	Kubica crash	#kubica
12	Assange Nobel peace nomination	#assange nobel
13	Oprah Winfrey half-sister	#oprah half-sister
14	Release of “The Rite”	#theRite
15	Thorpe return in 2012 Olympics	#olympics thorpe 2012
16	Release of “Known and Unknown”	#rumsfeld #knownAndUnknown
17	White Stripes breakup	#whiteStripes
18	William and Kate fax save-the-date	#william #kate #williamAndKate
19	Cuomo budget cuts	#cuomo budget
20	Taco Bell filling lawsuit	#tacoBell lawsuit
21	Emanuel residency court rulings	#emanuel residency court rulings
22	healthcare law unconstitutional	#healthcare law unconstitutional
23	Amtrak train service	#amtrak train service
24	Super Bowl, seats	#superBowl seats
25	TSA airport screening	#tsa airport screening
26	US unemployment	#us unemployment
27	Reduce energy consumption	#energy consumption
28	Detroit Auto Show	#detroit Auto Show
29	Global warming and weather	#globalwarming weather
30	Keith Olbermann new job	#olbermann
31	Special Olympics athletes	#specialOlympics
32	State of the Union and jobs	#stateOfTheUnion job
33	Dog Whisperer Cesar Millan’s techniques	#dogWhisperer #cesarMillan
34	MSNBC Rachel Maddow	#msnbc #rachelMaddow #rachel #maddow
35	Sargent Shriver tributes	#shriver
36	Moscow airport bombing	#moscow airport bombing
37	Giffords’ recovery	#Giffords recovery
38	Protests in Jordan	#jordan protest
39	Egyptian curfew	#Egypt curfew
40	Beck attacks Piven	#Beck #Piven
41	Obama birth certificate	#Obama birth certificate
42	Holland Iran envoy recall	#holland #iran envoy recall
43	Kucinich olive pit lawsuit	#kucinich olive pit lawsuit
44	White House spokesman replaced	#whiteHouse spokesman replaced
45	Political campaigns and social media	#politics campaign social media
46	Bottega Veneta	#bottega #veneta #bottegaVeneta
47	Organic farming requirements	#organic #farming requirements
48	Egyptian evacuation	#egypt evacuation
49	Carbon monoxide law	#carbon #monoxide law
50	War prisoners, Hatch Act	#Hatch Act #war prisoners

4. REFERENCES

[1] M. Alrifai and D. Skoutas. Tag clouds revisited. In *Proceeding of the 20th ACM conference on Information and knowledge management*, CIKM '11, New York, NY, USA, 2011. ACM.

[2] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu. Can all tags be used for search? In *Proceeding of the 17th ACM conference on Information and knowledge management*, CIKM '08, pages 193–202, New York, NY, USA, 2008. ACM.

[3] comScore. Indonesia, brazil and venezuela lead global surge in twitter usage. `http://www.comscore.com/`, 2010.

[4] E. Diaz-Aviles, M. Georgescu, A. Stewart, and W. Nejdl. Lda for on-the-fly auto tagging. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, pages 309–312, New York, NY, USA, 2010. ACM.

[5] J. Huang, K. M. Thornton, and E. N. Efthimiadis. Conversational tagging in twitter. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, HT '10, pages 173–178, New York, NY, USA, 2010. ACM.

[6] M. Shiels. Twitter celebrates its fifth birthday. `http://www.bbc.co.uk/news/technology-12805216`. *BBC News*, 2011.

[7] A. Stewart, E. Diaz-Aviles, W. Nejdl, L. B. Marinho, A. Nanopoulos, and S.-T. Lars. Cross-tagging for personalized open social networking. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, HT '09, pages 271–278, New York, NY, USA, 2009. ACM.

[8] C. Twitter. San Francisco. #numbers. `http://blog.twitter.com/2011/03/numbers.html`. 2011.