University of Lugano at TREC 2011 Microblog Track

Giacomo Inches
University of Lugano
Faculty of Informatics
Lugano, Switzerland
giacomo.inches@usi.ch

ABSTRACT

In this document we present the participation of University of Lugano in the Microblog track of TREC 2011. We describe our approach based on a time-based filtering algorithm of retrieved documents. We highlight the results and the possible improvement of the described technique.

1. INTRODUCTION AND RELATED WORK

User-Generated Content recently gained lot of attention in the Information Retrieval research community [5, 8, 7]. In the latest years and edition of TREC the focus of the investigations was primary Blogs, while in this edition the attention has been moved to the Twitter microblog platform.

Twitter microblog messages represent a shift in the conventional Information Retrieval documents: their length and language style is completely different, even from other User-Generated documents like Blog and discussion Forums. Blog and discussion forums, in fact, resemble much more traditional documents (newspaper articles like the Wall Street Journal) than Twitter messages, that are closer to the length and style of the online conversation (like the chat rooms or the wall-to-wall communication) [3, 4].

Different techniques have been already applied to some Twitter collection in order to extract the topic of the messages [11], to be later used for different purposes, like Social TV [1] or sentiment analysis toward a public debate [9]. Moreover, some work has already been done in reconstructing and understanding the conversations within the Twitter messages [12].

In the following sections we are presenting the work done for the TREC 2011 Microblog track, mentioning the dataset used (Section 2), our approach (Section 3) and concluding with ideas on how to improve our results (Section 4).

2. DATASET

We downloaded the Twitter Microblog dataset prepared by the organizer of the TREC 2011 Microblog Task with the provided tool[1]. The twitter-corpus-tools software allowed us to crawl twitter.com to screen-scrape and later reconstruct the Twitter messages. This was chosen because we could not rely on the Twitter API Rate Limit to download the corpus on time. Nevertheless, the downloading process was quite long (first downloading the messages ids, then crawling the messages, then repairing, then checking, ...) and took most of our time and resources. From an informative point of view, the messages crawled with the `html` option were missing all the metadata information (user profile, informations on retweets, ...) which are generally available using the Twitter API and the relative returned json code.

We were able to download about 15 millions Twitter messages with about 12 million valid (200 status) messages.

3. APPROACH

We planned different experiments on the collection but due to the limited time and resources we only managed to run a basic retrieval algorithm. We intended this first experiment as a baseline for the future ones and we will identify in the next section possible improvements.

The task of this year Microblog Track was a Realtime Adhoc one[2], where the participants had to retrieve from the given collection the most recent but relevant documents for a particular query (representing the user need).

Since time was an important factor (we had to rank the Twitter messages from the newest relevant to the oldest) we decided to index the messages each day (24th of January 2011 to 8th of February 2011) independently from the others, for a total of 16 indexes. We used the Terrier IR platform [10] with its experimental Twitter plugin[3] to index the messages. In doing this we had to convert our collection from the `html` format into a `json` format.

We then ran the 50 given queries on each index per each day and obtained 16 different ranks per each query. The ranking was computed using the standard `BM25` ranking function with $b = 0.75$. In the next step we read the ranking list from the day the query was issued. We ignored all the documents retrieved before the query time (to exclude the documents in the future) and compared pairwise the scores of the remaining documents against the scores of the documents retrieved the day before. If the score for the document retrieved the day before was higher than any in the query

[1] `https://github.com/lintool/twitter-corpus-tools`

[2] `https://sites.google.com/site/microblogtrack/2011-guidelines`

[3] `http://ir.dcs.gla.ac.uk/wiki/Terrier/Tweets11`

.

day, we discarded all the documents with lower score in the query time list and appended instead all the documents with higher score from the ranking list of previous day. This was done to give more importance to the documents with higher score, that we suppose to be more relevant, but in a time ordered way, to respect the task goals.

We then iteratively repeated this process, using as reference the retrieved documents one day earlier the query date and comparing them to the documents of the day before, until the end of the collection. At the end, we obtained one rank list per query, each of a different size where documents where ordered by date and time and with decreasing scores in each day. This strategy allowed us to obtain a run that was fully automatic, without the use of any future or external evidence.

We drew also another approach that aimed at reformulating and resubmitting each query, based on the Kullback-Leibler (KL) divergence between the Language Model (LM) of the retrieved documents and the LM of all the other previous (to the query) documents. The intuition behind this approach was the enlargement of the query terms to be able to detect more relevant and diverse documents. Unfortunately we were not able to submit the results for this run due to time constraints.

4. RESULTS

The results for our run reflect the nature of the approach: a simple one that was intended as baseline for improvements. In most of the queries we were below the median for all the metrics (P@30, MAP, R-precision) and only for few queries (9, 26, 39, 42), in the case of all relevant topics, and just query 26, in the case of high relevant topics, we were above the median for both MAP and R-precision.

We think that the the critical point of our approach resides in the usage of the BM25 scoring algorithm on the short, unstructured and noisy Twitter messages without any additional preprocessing or post-processing of the ranking list based on additional features (e.g. KL divergence of the different LM).

An interesting extension could be the introduction of smoothing factors [6] or the introduction of additional but external features like the prioritization of the sources (e.g. the users) based on their authoritativeness or the informativeness of their messages [2]. The introduction of smoothing factors could help the retrieval process in enlarging the informative content of the Twitter messages. The prioritization of the sources, instead, could be interesting in giving more relevance to those sources which are verified by Twitter and therefore more reliable than others. In doing this we automatically select the best messages to the user and augment the probability of retrieving relevant messages.

5. ACKNOWLEDGMENTS

We want to thank the TREC organizers for their hard work and for their support in proving the collection and the software tools to crawl and index it.

6. REFERENCES

[1] O. Dan, J. Feng, and B. Davison. Filtering microblogging messages for social tv. In *Proceedings of the 20th international conference companion on World wide web*, WWW '11, pages 197–200, New York, NY, USA, 2011. ACM.

[2] G. Inches, A. Basso, and F. Crestani. On the generation of rich content metadata from social media. In *Proceedings of the 3rd international workshop on Search and mining user-generated contents*, SMUC '11, pages 85–92, New York, NY, USA, 2011. ACM.

[3] G. Inches, M. Carman, and F. Crestani. Investigating the statistical properties of user-generated documents. In H. C. et al., editor, *Flexible Query Answering Systems*, volume LNAI 7022 of *Lecture Notes in Computer Science*, pages 198–209, Ghent, Belgium, 10 2011.

[4] G. Inches, M. J. Carman, and F. Crestani. Statistics of online user-generated short documents. In C. Gurrin, Y. He, G. Kazai, U. Kruschwitz, S. Little, T. Roelleke, S. M. Rüger, and K. van Rijsbergen, editors, *ECIR*, volume 5993 of *Lecture Notes in Computer Science*, pages 649–652. Springer, 2010.

[5] J.Codina, A.Kaltenbrunner, J.Grivolla, R. E.Banchs, and R.Baeza-Yates. Content analysis in web 2.0. In *18th International World Wide Web Conference*, Barcelona, Spain, April 2009.

[6] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: topic tracking in tweet streams. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 422–429, New York, NY, USA, 2011. ACM.

[7] C. Macdonald and I. Ounis. The TREC Blogs06 collection: Creating and analysing a blog test collection. *Department of Computer Science, University of Glasgow Tech Report TR-2006-224*, 2006.

[8] C. Macdonald, I. Ounis, and I. Soboroff. Overview of the trec-2007 blog track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*, 2007.

[9] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International AAAI Conference on Weblogs and Social Media*, 2010.

[10] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*, 2006.

[11] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *ICWSM*, 2010.

[12] A. Ritter, C. Cherry, and B. Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 172–180, Los Angeles, California, June 2010. Association for Computational Linguistics.

.