# Microblog Search Task at CLEF 2017: Query Generation using IR and LDA Topic Modeling Combination

Malek Hajjem and Chiraz Latiri

LIPAH research Laboratory, Faculty of Sciences of Tunis,
Tunis EL Manar Univeristy,Campus Universitaire Farhat Hached
B.P. n94, 1068 Tunis ,Tunisia

**Abstract.** The microblogs search task at CLEF 2017 consists of developing a system to search the most relevant microblogs for cultural query in a collection about festivals in all languages. Our general approach to get this objective is the following: we propose to generate from the initial tweet queries, provided for the task, extended queries able to get an answer-rich set of microblogs. This is achieved using a thematic representation of tweet query extracted from microblog corpus. We investigate in this paper a novel method to improve topics learned from Twitter content without modifying the basic machinery of LDA. This latter is based on Information Retrieval (IR) process to generate a query-specific set of similar tweets. The result then represent the input of a basic LDA topic modeling process. Finally, the output thematic cluster serves as our source of expansion for the initial queries.

**keywords**: CLEF, Microblogs Search, LDA, Information Retrieval, Aggregation

## 1 Introduction

The microblog search is the second task [1] at CLEF 2017 [8](Conference and Labs of the Evaluation Forum) from Cultural Microblog Contextualization track. This task consists in developing a system to search the most relevant microblogs for cultural query in a collection about festivals in all languages. Topics, announcing some cultural event, were gathered from different sources[2][5, 4]. The goal is to retrieve relevant and diverse tweets related to each event from a dataset of 70 000 000 microblogs. This corpus dates from May to September 2015 and is about the keyword "Festival".
We consider that using topic models such as Latent Dirichlet Allocation (LDA) could be useful in this microblog search task. Taking into account that "topic model is often employed to mine "latent topics" from high dimensionality of terms in text"[2], these topics can be used to describe the content of a collection

[1] https://mc2.talne.eu/spip/Tasks/2-microblog-search/
[2] spanish query: http://www.jornada.unam.mx/2017/05/26/

or a query. In fact, the high probability topics and words within the topics can be viewed as a loose description of a text data.
The contribution of this work is to identifying the "topic" or topics being discussed in a query, and then using this knowledge of topics to include semantically related words. To better match with ambiguous nature of tweet query, topics are extracted from a microblog corpus. A novel method to improve topics learned from Twitter content without modifying the basic machinery of LDA is investigated. Following this introduction, the paper is organized as follows. In Section 2, a state of the art is shown. In Section 3, the methodology used is presented. The Conclusion section wraps up the paper.

## 2 Related works

### 2.1 Topic modeling for short texts

Topic models are used to uncover the latent semantic structure from text corpus. A topic consists of a cluster of words that frequently occur together. Using contextual clues, topic models can connect words with similar meanings and distinguish between uses of words with multiple meanings. Traditional topic models, like LDA, rely on co-occurrence patterns of words in documents to learn latent topics [2]. Due to the messy nature of short texts, applying directly conventional topic models (e.g. LDA and pLSA) on such short texts is not efficient. Indeed, naive topic models implicitly capture the document-level word co-occurrence patterns to reveal topics. Thus, to avoid data sparsity, works such as [7, 12, 9] have applied topic models to tweets based on a pooling strategy. It consists in aggregating similar short texts in one document. In [12], authors proposed tweet pooling startegy which was based on user aggregated messages. Authors in [7] also have experimented several schemes to train a standard topic model and compare their quality and effectiveness. In [1], authors have proposed to gather tweets occurring in a same user-to-user conversation and show that this new technique outperforms other pooling methods in terms of clustering quality and document retrieval. Closer to our work, authors in [9] have proposed a new method of tweet pooling using hashtags where documents with common hashtag were gathered. All these works have proved that by training a topic model on aggregated messages, they obtained a higher quality of learned model.

### 2.2 Topic model for IR: query expansion

There are two obvious approaches to including topic models in IR. In the first, a document is represented by itself and the topics to which it belongs. A second approach is to calculate a query related topic by using topic models and use it for query expansion. In this case, queries are reformulated (i.e. usually expanded) to improve the retrieval effectiveness. Authors in [13] proposed a method to find a good query-related topic based on LDA where experiments confirm that query expansion based on the derived topics achieved statistically significant

improvements. Others, in [11] implement one of most common local approach of query expansion - Pseudo Relevance Feedback (PRF). In this last, top k documents are considered to be relevant and extracts their topic's terms to extend queries.

### 2.3 Topic model for microblog search

With the rapid development of microblogs, microblog search has become one of the most trendy research areas in recent years. In contrast to traditional text retrieval, microblog search significantly differs. In fact, microblogs users often issue short queries to find relevant information. Moreover, the restriction of messages length lead to a problem in discriminating terms within a given item. To improve retrieval effectiveness in microblogs researches tend to use query expansion techniques. The goal is reducing the usual query/document mismatch. In this area, using topic model like, LDA, could be useful. However, unlike regular text, Topic modeling is not good at processing short text. Rare are works which try to enhance microblog search using topic modeling such as LDA. We cite [10], where authors present a method of contextualization of short messages using a thematic representation extracted from Wikipedia. This representation allows to extend the vocabulary of short messages by a set of thematically related words. The results show the contribution of this method to a better understanding of short messages. Other works like [3] propose a novel approach to locating the microblogging experts on a given query. First they define the experts by social influence and content relevance. For the social influence, they present a global-ranking algorithm as GUserRank and a topic-ranking algorithm as TUserRank after applying the LDA topic model.

## 3 Methodology

### 3.1 Ressources and Data pre-proceeding

To build a robust LDA model, a large amount of data is needed. From this perspective, two tweet corpora are used in different runs. We note that we have choose to use explicitly microblog corpus to respect the noisy nature of tweet query. The firt corpus is a comparable tweet corpus about Arab spring collected through Twitter's API[3] in Arabic and French languages. Basically, accessing Twitter data is done by collecting tweets that contain specific keywords. More information about this corpus could be found here [6]. The second tweet corpus is provided to the participants in the evaluation campaign. It is composed of 70 000 000 microblogs. This corpus dates from May to September 2015 and is about the keyword "Festival". Notice that before we applied LDA, redundancy was eliminated by deleting retweets. A language detection was also performed using a java library[4] to remove foreign language tweets.

[3] Otterapi a real time search engine that indexes the most influential tweets search api https://dev.twitter.com/rest/public/search

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

### 3.2 Information retrieval based approach for tweet pooling

In this work, an unsupervised topic model based on aggregating tweets that are thematically closed is presented. The goal is to adapt the LDA basic process to short tweet text. This will lead to improve the quality of topics latent discovered. To perform tweet pooling, we propose to use information retrieval strategy and hierarchical classification in order to avoid data sparsity in short texts as illustrated in Figure 1. Our approach represents an alternative of state of the art methods based on tweet pooling via meta data (hashtags, user information, etc). Indeed, such methods are highly dependent on the meta data content of the tweet corpus. Our approach relies on three main steps, namely:

- **Step 1: Preliminary set generation:** For each $tweet_i$, we propose to retrieve a set $\gamma_i$ of matched tweets out of a large tweet collection $C$ of $n$ tweets, using an information retrieval system. Thus, for a tweet $tweet_i$, performed as a query, several tweets in $C$ may match it with different similarity degrees.
- **Step 2: Pooled set construction:** The idea is to aggregate a set of tweets in $\Gamma$ partitions by gathering preliminary sets, resulting from the IR process, according to a combination criterion. If same search results are assigned to different tweets then the tweets are considered thematically close.
- **Step 3: Topic extraction:** It consists on learning latent topics from aggregating tweets via LDA.

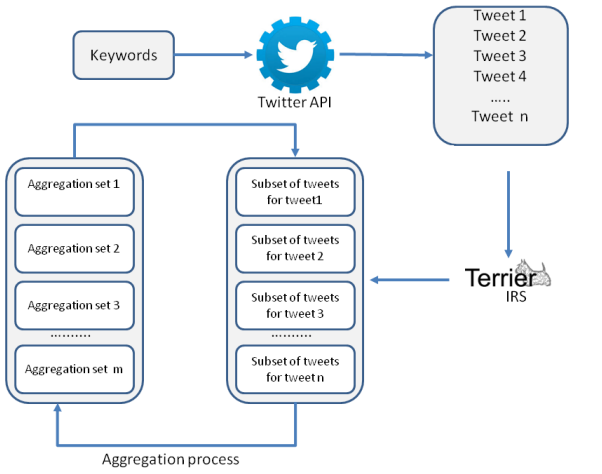


**Fig. 1:** IR for tweet aggregation

### 3.3 Description of Runs submitted

The submitted runs follow three steps (see Figure 2):

– Extract Topics using the combined method of IR and aggregation strategy as described above section 3.2
– Project the resulted Topic on the tweet text to detect the subset of thematic relevant terms
– Reformulate the initial query using the subset of thematic terms as enriched features in form of indri query

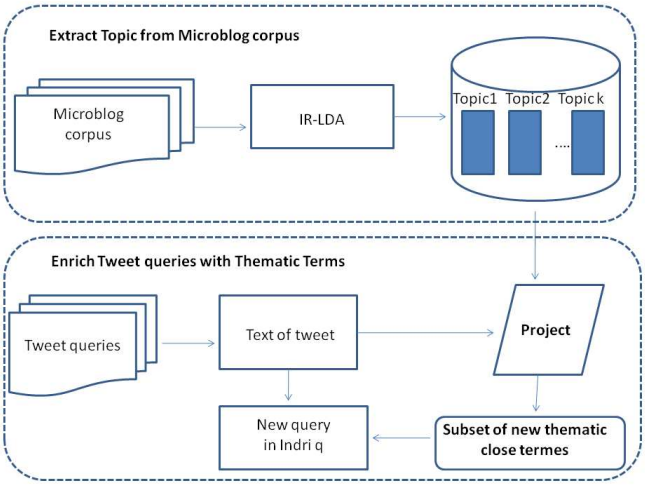


**Fig. 2:** Query expansion using latent topic inferred through IR-LDA method

– RUN 1: Extend the initial query using the latent topic extracted from comparable tweet corpus[5] through IR-LDA
– RUN 2: Extend the initial query using latent topic extracted from the Festival tweet corpus through IR-LDA

## 4 Conclusion

A large multilingual collection of posts have become publicly available due to the phenomenal growth of using social networks and microblogs all over the world. This makes from Microblogs valuable information sources. However little is known about how search socially-generated content in effective way. In this paper, we present a method to expand short messages using a thematic representation. A novel method for aggregating tweets in order to improve the quality of LDA-based topic modeling for short text is used to improve the quality of latent topics.

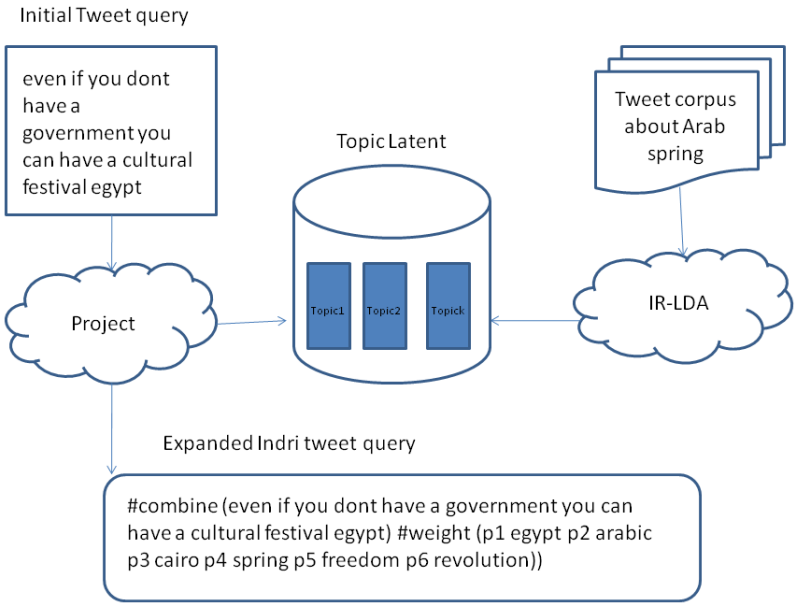


**Fig. 3:** Exemple of query expansion using latent topic inferred through IR-LDA method from Arab spring corpus

## References


1. D. Alvarez-Melis and M. Saveski. Topic modeling in twitter: Aggregating tweets by conversations. In *Tenth International AAAI Conference on Web and Social Media*, 2016.
2. D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.
3. Q. Chen, Y. Yang, Q. Hu, and L. He. Locating query-oriented experts in microblog search. In *Proceedings of Workshop on Semantic Matching in Information Retrieval co-located with the 37th international ACM SIGIR conference on research and development in information retrieval, SMIR@SIGIR 2014, Queensland, Australia, July 11, 2014.*, pages 16–23, 2014.
4. J.-V. Cossu, J. Gaillard, T.-M. Juan-Manuel, and M. El Bèze. Contextualisation de messages courts :l'importance des métadonnées. In *EGC'2013 13e Conférence Francophone sur l'Extraction et la Gestion des connaissances*, Toulouse, France, Jan. 2013.
5. M. Hajjem and C. Latiri. Features extraction to improve comparable tweet corpora building. In *JADT Acte*, nice, France, Juin 2016.
6. M. Hajjem, C. C. Latiri, and Y. Slimani. Twitter as a multilingual source of comparable corpora. In *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia, Kaohsiung, Taiwan, December 8-10, 2014*, pages 342–345, 2014.

7. L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, SOMA '10, pages 80–88, New York, NY, USA, 2010. ACM.
8. J. M. P. M. J.-Y. N. Liana Ermakova, Lorraine Goeuriot and E. SanJuan. Clef 2017 microblog cultural contextualization lab overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages Proceedings,LNCS volume, Springer, CLEF 2017, Dublin.*, 2017.
9. R. Mehrotra, S. Sanner, W. Buntine, and L. Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 889–892, New York, NY, USA, 2013. ACM.
10. M. Morchid, R. Dufour, and G. Linarès. Combinaison de thèmes latents pour la contextualisation de Tweets. In *EGC'2013 13e Conférence Francophone sur l'Extraction et la Gestion des connaissances*, Toulouse, France, Jan. 2013.
11. L. Ren. Implement topic relevance model for query expansion.
12. J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: Finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, WSDM '10, pages 261–270, New York, NY, USA, 2010. ACM.
13. X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*, ECIR '09, pages 29–41, Berlin, Heidelberg, 2009. Springer-Verlag.