Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 761–770

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

# Combining IR and LDA Topic Modeling for Filtering Microblogs

Malek Hajjem[a*], Chiraz Latiri[a]

[a]*LIPAH resaerch Laboratory,Faculty of Sciences of Tunis*
*Tunis EL Manar Univeristy,Campus Universitaire Farhat Hached*
*B.P. n94, 1068 Tunis ,Tunisia*

**Abstract**

Twitter is a networking micro-blogging service where users post millions of short messages every day. Building multilingual corpora from these microblogs contents can be useful to perform several computational tasks such as opinion mining. However, Twitter data gathering involves the problem of irrelevant included data. Recent literary works have proved that topic models such as Latent Dirichlet Allocation (LDA) are not consistent when applied to short texts like tweets. In order to prune the irrelevant tweets, 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 a pooling process which combines Information retrieval (IR) approach and LDA.This is achieved through an aggregation strategy based on IR task to retrieve similar tweets in a same cluster. The result of tweet pooling is then used as an input for a basic LDA to overcome the sparsity problem of Twitter content. Empirical results highlight that tweets aggregation based on IR and LDA leads to an interesting improvement in a variety of measures for topic coherence, in comparison to unmodified LDA baseline and a variety of pooling schemes.

*Keywords:* Microblogs, LDA, Pruning irrelevant tweets, Information Retrieval, Aggregation

## 1. Introduction

The growing number of Twitter users create large amounts of short messages, called *Tweets*. For instance 500 million tweets per day are sent[1]. Tweets are restricted to contain no more than 140 characters of text, including any links. This constraint leads to diverse types of contents and information carried in the tweets. Thus, it is expected that tweets generated from all over the world in different languages will be covering a variety of topics. Matching a tweet to its topic(s) is far from being a trivial task, mostly because of the huge number of potential topics that can be drawn from a tweet corpus as well as the problems of sparsity and drift for effectively filtering in microblogs.

* Malek Hajjem
* Corresponding author
*E-mail address:* malek.hajjem@gmail.com

[1] http://www.internetlivestats.com/twitter-statistics/

10.1016/j.procs.2017.08.166

In this paper, we consider the problem of topic modeling of tweets. This task raises diverse challenges given the short, noisy and ambiguous nature of tweets. It is worth noting that the naive application of topic modeling techniques like LDA to Twitter content produces mostly incoherent topics. It needs to be enriched with complementary techniques such as pooling in order in order to work well with the tweet corpus[1].

In order to enhance tweet topic modeling task and filtering microblogs, we encode in this paper, our intuition into a novel approach which relies on the combination of an Information Retrieval and probabilistic topic model, *i.e.*, Latent Dirichlet Allocation (LDA). Our goal is better understand the thematic diversity available in the collected tweet corpus described in[2]. Thus, we propose to aggregate short texts into homogeneous clusters. The aggregation process is based on IR used to reduce the tweets sparsity issue.

In this work, we conduct an exploratory study on short texts such as Tweets. The goal is to extract from a whole tweet corpus only the most reliable part concerning our topic of interest. Through such pooling, we merge similar tweets together and consider them as single documents used by LDA model. The goal is twofold: (1) extract better topics from twitter using standard LDA and (2) filter out irrelevant tweets from our comparable corpus based on topics learnt via LDA. We examine a variety of topic coherence evaluation metrics, including the ability of the learned LDA topics to reconstruct known clusters and the interpretability of these topics via statistical information measures.

This paper is organized as follows. Section 2 describes the problem statement and our motivations. A brief literature review is given in Section 3. In section 4, our topic modeling medthod based on IR is introduced. The experimental validation is described in section 5. The Conclusion section wraps up the article and outlines future work.

## 2. Problem statement

### 2.1. Impact of keywords used for tweet gathering

Many researches in literature addressed the problem of collecting and analyzing Twitter content in different languages to promote multiple computational tasks such as opinion mining, machine translation of user-generated content and social information retrieval. However texts need to be enough reliable to make correct conclusions. In Twitter, as most of social networks platforms, texts contain specific writing style. Messages such as tweets are short and give rise to a vocabulary mismatch between typically chosen keywords for a tweet collection and words used to describe the topic of interest. Often, selected keywords for crawling process are based on human intuition. To target a specific subject, we crawl tweets that are referring our topic of interest using hashtags. Especially, hashtags are keywords assigned to information that describes a tweet and helps in searching. However, not only relevant tweets includes hashtags. In fact, spammers exploit popular hashtags without considering their semantics to promote their content. For example, #revolution was a trending hashtag concerning our topic of interest. It was used in sentences such as ‘‘IPad 2: 2 times more useless #revolution’’. Manual inspection of such tweet shows that this hashtag does not provide any additional meaning to the tweet.

Table 1. Translated french Tweets containing the keywords ‘‘revolution’’

| Tweet | Content | Relevance |
|---|---|---|
| T1 | The army will not allow ”an against-*revolution*” in Egypt | Relevant |
| T2 | Freebox *revolution* : download is free today | Irrelevant |
| T3 | Peaceful *Revolution* in Iceland #revolution #anticapitalism | Irrelevant |
| T4 | New MacBook Pro ranges & Mabook. It is a *revolution*... or not! | Irrelevant |

Thus, to create a complete dataset, we use multiple keywords that are referring our topic of interest, *i.e.*, Arab spring. This way of gathering data tends to increase the amount of possible relevant content, nevertheless it risks to retrieve noisy tweets. Indeed, a keyword can be used several times in different tweets but that does not mean that it is representative. An illustrative examples of this drawback is depicted in Table1, related to the word **revolution** which is linked to Arab spring protests and it is often used in other contexts. This ambiguity problem is more accentuated due to the access strategy of twitter data. Since tweets are accessed through Twitter’s APIs where search is mainly based on boolean match, no ranking is performed on the retrieval results. Hence, these limitations hamper the thematic quality