Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 27 (2015) 26 – 29

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns

An Author-Topic based Approach to Cluster Tweets and Mine their Location

Mohamed Morchid[a], Yonathan Portilla[a,b], Didier Josselin[c,a], Richard Dufour[a], Eitan Altman[b,a], Marc El-Beze[a], Jean-Valère Cossu[a], Georges Linarès[a], Alexandre Reiffers-Masson[a,b]

[a]Laboratoire d'Informatique d'Avignon, LIA, 339 chemin des Meinajariès, Agroparc BP 91228, 84911 Avignon cedex 9, France
[b]INRIA, B.P 93, 06902 Sophia Antipollis Cedex, France
[c]UMR ESPACE 7300 ; 74 rue Louis Pasteur, 84029 Avignon Cedex, France

Abstract

Social Networks became a major actor in information propagation. Using the Twitter popular platform, mobile users post or relay messages from different locations. The tweet content, meaning and location show how an event-such as the bursty one "JeSuisCharlie"' happened in France in January 2015 is comprehended in different countries. This research aims at clustering the tweets according to the co-occurrence of their terms, including the country, and forecasting the probable country of a non located tweet, knowing its content. First, we present the process of collecting a large quantity of data from the Twitter website. We finally have a set of 2.189 located tweets about "Charlie", from the 7th to the 14th of January. We describe an original method adapted from the Author-Topic (AT) model based on the Latent Dirichlet Allocation method (LDA). We define a homogeneous space containing both lexical content (words) and spatial information (country). During a training process on a part of the sample, we provide a set of clusters (topics) based on statistical relations between lexical and spatial terms. During a clustering task, we evaluate the method effectiveness on the rest of the sample that reaches up to 95% of good assignment.

Keywords: Author-Topic model, Tweet location

1. Context of the study and state of the art

The exponential growth of available data on the Web enables users to access a large quantity of information. Micro-blogging platforms evolve in the same way, offering an easy way to disseminate ideas, opinions or common facts under the form of short text messages. Depending on the sharing platform used, the size of these messages can

doi:10.1016/j.proenv.2015.07.109

be limited to a maximum number of words or characters. Although Twitter is a recent information-sharing model, it has been widely studied. Many works have focused on various aspects of Twitter, such as social impact [1] event detection [2], user influence [3], sentiment analysis [4], hash-tag analysis [5] or theme classification [6].

The aim of the proposed approach is to locate a given tweet by using the tweet content (a set of words). Nonetheless, the Twitter service does not allow to send messages whose size exceeds 140 characters. This constraint causes the use of a particular vocabulary that is often unusual, noisy, full of new words, including misspelled or even truncated words [7]. Indeed, the goal of these messages is to include a lot of information with a small number of characters. Thus, it may be difficult to understand the meaning of a short text message (STM) with only the tweet content (words). Several approaches have been proposed to represent the tweet content. The classical bag-of-words approach [8] is usually used for text document representation in the context of keyword extraction. This method estimates the Term Frequency-Inverse Document Frequency (TF-IDF) of the document terms. Although this unsupervised approach is effective for a large collection of documents, it seems unusable in the particular case of short messages since most of the words occur only once (*hapax legomena* [9]).

Other approaches propose to consider the document as a mixture of latent topics to work around segments of errors. These methods build a higher-level representation of the document in a topic space. All these methods are commonly used in the Information Retrieval (IR) field. They consider documents as a bag-of-words without taking account of the words order. Nevertheless, they demonstrated their performance on various tasks. Several approaches were proposed such as Probabilistic LSA (PLSA) [10] or Latent Dirichlet Allocation (LDA) [11]. LDA is a generative model of statistics which considers a document, seen as a bag-of-words, as a mixture probability of latent topics. In opposition to a multinomial mixture model, LDA considers that a theme is associated with each occurrence of a word composing the document, rather than associating a topic with the complete document.

Thereby, a document can belong to different topics from a word to another. However, it is noted that the word occurrences are connected by a latent variable which controls the global respect of the topic distribution in the document. These latent topics are characterized by a distribution of word probabilities which are associated with them. During the LDA learning process, distribution of words into each topic is estimated automatically. Nonetheless, the location associated with the tweet is not directly taken into account in the topic model. As a result, such a system considers separately the tweet content (words), to learn a topic model, and the labels (location) to train a classifier. Thus, the relation between the tweet content and its location (country) is crucial to efficiently locate (unknown) new tweets.

In this paper, we propose to build a topic model, called author-topic (AT) [12,13] that takes into consideration all information contained in a tweet: the content itself (words), the label (country) and the relation between the distribution of words into the tweet and the location, considered as a latent relation. From this model, a vector representation in a continuous space is built for each tweet. Then, a supervised classification approach, based on Support Vector Machines (SVM) [14] is applied. For mathematical and methodological details, see [6, 13].

2. Experimental protocol applied on the tweet "Charlie"

We propose to evaluate the approach on a Twitter corpus. This corpus is composed of tweets from 16 countries. A classification approach based on Support Vector Machines (SVM) is performed to find out the most likely country of a given tweet, in two stages: a training and an assignment [14]. The table 1 lists the corpus of tweets. This data set is split in three parts depending of the tweet emission day in January 2015: 887 tweets between the 7th and the 8th, 471 tweets between the 9th and the 10th and 881 tweets between the 11th and the 14th.

We obtain 1.520 tweets for the training phase of the AT models and 669 for the validation (testing) phase which corresponds to a corpus of 2.189 tweets for the whole 16 countries (roughly 137 tweets for each country). The number of topics contained in the AT model strongly influences the quality of this model. Indeed, an AT model with only few topics will be more general than one with a large number of classes (granularity of the model). For a sake comparison, a set of 100 AT models is learnt (between 5 and 105). As the classification of tweets requires a multi-class classifier, the SVM (one-against-one) method is chosen with a linear kernel. This method gives a better accuracy than the (one-against-rest) one [15].