Classification in Twitter via Compressive Sensing
Dimitrios Milioris

To cite this version:
Dimitrios Milioris. Classification in Twitter via Compressive Sensing. IEEE International Conference on Computer Communications (INFOCOM), Apr 2015, Hong Kong, Hong Kong SAR China. hal-01138337

HAL Id: hal-01138337
https://hal-polytechnique.archives-ouvertes.fr/hal-01138337
Submitted on 27 May 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Abstract—In this paper we introduce a novel low dimensional method to perform topic detection and classification in Twitter. The proposed method first employs Joint Complexity to perform topic detection. Then, based on the nature of the data, we apply the theory of Compressive Sensing to perform topic classification by recovering an indicator vector, while reducing significantly the amount of information from tweets. In this paper we exploit datasets in various languages collected by using the Twitter streaming API, and achieve increased classification accuracy when comparing to state-of-the-art methods based on bag-of-words, along with several reconstruction techniques.

I. INTRODUCTION

During the last decade, the information within social networks has increased dramatically. The ability to study the interaction and communication between users in these networks can provide real time valuable prediction of the evolution of the information.

In this paper, first we use the theory of Joint Complexity (JC) to perform topic detection. The evaluation of the proposed method is based on the detection of real world topics like the categories of a mainstream news portal. We use large datasets, which are tweets from politics, economics, sport, technology and lifestyle. We decompose the tweets in linear time into a memory efficient structure called Suffix Tree and by overlapping two trees, in linear or sublinear average time, we obtain the JC defined as the cardinality of factors (subsequences) that are common in both trees. Then we classify new tweets into these categories with the power of Compressive Sensing (CS) by taking advantage of the nature of the data. The method is simple, context-free, with no grammar and no language assumptions, and does not use semantics. Therefore, there is no need of any specific dictionary or stemming process which will explode the complexity.

The paper is organised as follows: Section II introduces the JC method. Section III describes the application for classification via CS, while Section IV evaluates the performance with real data obtained from Twitter.

II. JOINT COMPLEXITY

In a prior work, we extend JC estimate to Markov sources of any order on a finite alphabet. Markov models are more realistic and have a better approximation for text generation than memoryless sources [1], [2]. We derived a second order asymptotics for JC of the following form

$$\gamma \approx \frac{n^\kappa}{\sqrt{\alpha \log n + \beta}},$$

for $\kappa < 1$, $\gamma > 0$, $\alpha > 0$, $\beta > 0$, with $n$ being the length of the sequence. This estimate has a faster convergence, and is preferred for texts of order $n \approx 10^2$; Therefore JC is an efficient method to capture the similarity degree of short texts, e.g. tweets ($< 140$ characters).

The topic detection method based on JC proceeds in two steps. First, we construct the training databases (DBs) by using Twitter’s streaming API (using the basic json format) while filtering for specific keywords. Using these requests we build $C$ classes on different topics. Assume that each class contains $N$ tweets (e.g. $C = 5$, i.e. Classes: politics, economics, sports, technology, lifestyle of $N = 12,000$ tweets). We populate each class by allocating a number of keywords. Assume that we have a dataset of $S$ timeslots with $s = 1, \ldots, S$, and each timeslot is a 15 minutes request in Twitter API. For every tweet $x_s$, where $i = 1, \ldots, N$, with $N$ being the total number of tweets, in the $s$-th timeslot, i.e. $x_s^i$, we build a Suffix Tree.

Then we compute the JC metric, $JC(x_s^i, x_s^j)$ of the tweet $x_s^i$ with every other tweet $x_s^j$ of the $s$-th timeslot, where $j = 1, \ldots, N$, and $j \neq i$ (by convention we choose $JC(x_s^i, x_s^i) = 0$). For the $S$ timeslots we store the JC scores in the matrices $S_1, S_2, \ldots, S_S$ of $N \times N$ dimensions. We then assign a new tweet to the class that maximises the JC metric in that class. In order to limit the size of each reference class we delete the oldest tweets or the least significant ones (e.g. the ones which obtained the lowest JC score). This ensures the low cost and efficiency of the method.

III. COMPRESSIVE SENSING

Let us first describe the main theoretical concepts of CS [3] as applied in the context of classification. Let $x \in \mathbb{R}^N$ denote the signal of interest. Such signal can be represented as a linear combination of a set of basis $\{\psi_i\}_{i=1}^N$. By constructing a $N \times N$ basis matrix $\Psi = [\psi_1, \psi_2, \ldots, \psi_N]$, the signal $x$ can be expressed as $x = \sum_{i=1}^N s_i \psi_i = \Psi s$. In fact the signal is represented as $x = \Psi s + \theta$, with $\theta \in \mathbb{R}^N$ being the noise, where $\mathbb{E}(\theta) = 0$ and $\text{var}(\theta) = O(\|\Psi s\|)$.

The efficiency of a CS method for signal approximation or reconstruction depends highly on the sparsity structure of the signal in a suitable transform domain associated with an appropriate sparsifying basis $\Psi$. The measurement model in the original space-domain is expressed as $g = \Phi x$, where $g \in \mathbb{R}^M$ is the measurement vector and $\Phi \in \mathbb{R}^{M \times N}$ denotes the measurement matrix. The measurement model has the following equivalent transform-domain representation

$$g = \Phi \Psi s + \Phi \theta.$$

In fact when the length of the sequence $n \to \infty$ and $N \to \infty$, $\mathbb{E}(\Psi s) = O(nN)$, with $\text{var}(\theta) = O(nN)$,
F−Score 0
0.2
0.4
0.6
0.8
1
Percentage of Total Number of Measurements

Fig. 1. Classification accuracy of various methods, measured by F-Score as a function of the percentage of measurements (%) by using several reconstruction techniques.

Fig. 2. Classification accuracy of the JCurl+CS method, measured by F-Score as a function of the percentage of measurements (%) by using several reconstruction techniques.

references


For the implementation of methods 1)-5) the MATLAB codes can be found in: http://parselab.stanford.edu/, http://www.acm.caltech.edu/l1magic, http://people.ee.duke.edu/~icarin/BCS.html