TWEEPROFILES2: REAL-TIME DETECTION OF SPATIO-TEMPORAL PATTERNS IN TWITTER

João Henrique Alves Pereira

Dissertation accomplished under the orientation *Prof. Carlos Soares* and co-orientation of Tiago Cunha
at *Sapo Labs FEUP*

1. Motivation

Social networks play a big part in contemporary societies. Its influence is felt in several aspects of our lives, ranging from the way we do marketing, for instance, to the way we interact with our loved ones. Twitter is one of the top social networks in existence, both in popularity (worldwide public awareness) and monthly active users (around 250 million [1]). TweeProfiles [2] is a Twitter analysis tool that enables the visualization of the results of a clustering methodology applied to a dataset of tweets. The data is processed over 4 dimensions of the data: spatial, temporal, social and content. It enables analysis giving different weights to each of the dimensions, producing different clustering models with the same dataset. It lacks, however, the ability to produce real-time visualizations of the evolution of the data stream, which with the growing volume of data can rapidly become an issue as it is unable to capture emerging trends in the patterns of the stream.

2. Objectives

The main objective of this project is to develop a tool that allows real-time analysis of Twitter data over 3 dimensions: spatial, temporal and content. In order to achieve that, we will:

- Develop a method that allows the clustering of a stream of data;
- Develop a way to visualize the produced results;
- Test the tool and analyze its results.

3. Project Summary

The work developed in this dissertation can be divided in 3 distinct parts: the distance function development, the clustering algorithm choosing and adaptation, and the visualization tool. In this section wee will explain each of them in more detail.

3.1. Distance function

In order to be able to assess different tweet dimensions, we must use a composed distance function. Consider a tweet to be tuple $TW(lat, lon, hou, wkd, date, \overline{txt})$, where lat and lon are the tweet's geographical coordinates, hou and wkd are the hour and weekday attributes, $date$ is the date of the tweet and $\overline{txt}$ is the vector with the tweet's words. For two tweets TW_i and TW_j, our distance function is defined as:

$$dist(TW_i, TW_j) = \begin{cases} haversineformula(lat, lon) \\ euclideanformula(hou, wkd) \\ cosinesimilarity(\overline{txt}) \end{cases}$$

As we can see from the formula above, our composed function uses the haversine distance [3] for the spatial dimension (with the tweets' coordinates), the euclidean distance [4] for the temporal dimensions (with the tweet' hour and weekday) and the cosine similarity [4] for the content dimension (using the tweets' texts).

3.2. Clustering methodology

As any other data mining process, there are some steps to be taken before applying a data mining algorithm. In this case, before feeding the stream information to our clustering algorithm, we do some data pre-processing tasks. These tasks can be divided in text processing and date processing. The text processing involves tokenization, removing URLs and punctuation, detecting the language of the text, and finally removing the stopwords and stemming (according to the language). The date processing just extracts the hour and weekday attributes from the date.

In order to cluster data streams, we must use stream clustering algorithms. As there was no algorithm that combines stream clustering of numerical, categorical and textual attributes, we had to develop one. We used DenStream [5] as a framework, adding the necessary mechanisms to allow it to perform the task at hand. We called our adapted algorithm HybridDenStream, as it is able to cluster data points with different types of information, such as tweets. Before we could adapt DenStream to cluster these three dimensions, we also had to adapt the underlying micro-cluster structures so they could summarize information of several types. So, we will first specify our hybrid micro-cluster structures and then explain the functioning of HybridDenStream.

The DenStream framework works in two steps: the online, micro-clustering step and the offline, macro-clustering step. The micro-clustering step is meant to summarize the information coming through the stream, creating small clusters (micro-clusters) that are stored and later used in the macro clustering step in order to generate the final clustering results.

In order to process tweets' information, besides altering the DenStream default distance function, we also had to change the micro-clusters' structure, so we

are able to summarize all the information concerning tweets that we need for the clustering process So we created a HybridMicroCluster structure, that is meant to represent a small aglomeration of tweets.

Like the original DenStream's micro-clusters, hybrid micro-clusters also are characterized by a weight, a center and a radius, with the weight and radius calculated in the same way. However, the concept of center in a data point with categorical and textual attributes is quite abstract. So we consider the center of a hybrid micro cluster to contain not only its numerical attributes (in this instance, the sums of the coordinates, weekday and hour values divided by the weight of the cluster) but also a term-frequency vector (containing the words of the clustered tweets paired with their respective relative frequencies).

3.3. Visualization

A simple visualization tool was developed to help analyze and make sense of the obtained results. It allows three types of cluster visualizations: a spatial visualization, a temporal visualization and a content visualization.

The spatial visualization, consists of a world map where the red dots represent micro-clusters; the clusters are represented by a blue circular area, not visible in the picture. The clusters are plotted according to their center's geographical coordinates.

The temporal visualization, consists of a x-y graph where the x axis represents the day of the week (Sunday through Saturday) and the y axis represents the hour. The clusters are represented by bubble and are plotted according to their center's hour and weekday values, which define the clusters center in the graph; the radius of the spheres represents the number of points (tweets) in that cluster.

The content visualization, consists of a word cloud, where the size of each word represents its frequency. This means that the bigger the word is, the more times it appears in the texts of the tweets in the cluster.

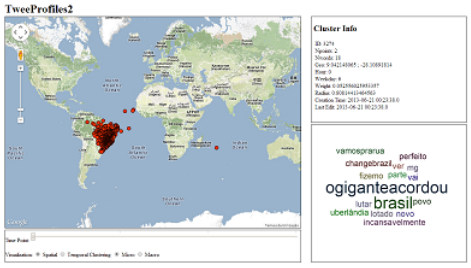

Fig. 1 – Visualization Tool

3.4. Testing

To test our algorithm, we used a dataset with 50,735 geo-located tweets from SocialBus (formerly TwitterEcho [6]), from June to August of 2013. Most of these tweets are related to the social uprising that occured in Brazil in 2013.

The results obtained are very extensive and therefore cannot be summarized in this document. For a complete results analysis we advise to read chapter 4 of the dissertation

4. Conclusions

The objective of this work was to produce a tool that allowed real-time, multi-dimensional clustering of a Twitter stream. In order to achieve that, we developed a data-mining process composed of several phases: Data Extraction, Data Pre-processing, Clustering and Visualization

This project is a natural "evolution" of the TweeProfiles tool, with its focus being on improving it so it could generate clustering results in real-time. Therefore, we can say that the main objective of this project was achieved.

References

[1] Twitter. About twitter - https://about.twitter.com/company, 2014. [Online; accessed 1-July-2014].

[2] Tiago Cunha. TweeProfiles : detection of spatio-temporal patterns on Twitter, 2013.

[3] J.C. Hannyngton. *Haversines Natural and Logarithmic Used in Computing Lunar Distances for the Nautical Almanac*. Great Britain. Nautical Almanac Office. G.E. Eyre and W. Spottiswoode, 1876.

[4] Jiawei Han, Micheline Kamber, and Jian Pei. *Data Mining: Concepts and Techniques, Third Edition*. Morgan Kauffman, 3rd edition, 2011.

[5] Feng Cao, Martin Ester, W Qian, and A Zhou. Density-Based Clustering over an Evolving Data Stream with Noise. *SDM*, pages 328–339, 2006.

[6] Matko Boanjak, Eduardo Oliveira, José Martins, Eduarda Mendes Rodrigues, and Luís Sarmento. TwitterEcho: a distributed focused crawler to support open research with twitter data. In *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 1233, New York, New York, USA, 2012. ACM Press.