Visual Analytics: Online clustering on disaster events using Twitter

David Mauricio Guzman Delgado. 1909580
Hamza Bouzidi. 1909207

June 7, 2021

Abstract

This paper proposes an approach to cluster social media posts. It aims at taking full advantage of this recent source of newsworthy information and at facilitating the work of users who need to monitor public events in real-time. The present work is an attempt to approach different situations of awareness which involves the perception, comprehension and projection of different events happening in live time by means of clustering and visualization.

1 Introduction and motivation

Social media and news reporting platforms represent a valuable source of information to monitor events occurring around the world in real-time. Because of the size and speed at which this information is generated and the fact that this information is often produced in a free-text form, it can be an overwhelming challenge to fully benefit from such data sources.

Our focus is on posts from the popular microblogging service Twitter, since it's easy to use but extremely powerful for spreading information in real-time fashion. This becomes quiet relevant during and after emergency events since now we have better telecommunications infrastructure allowing the connectivity almost anywhere.

It is believed that Twitter is a powerful tool for sharing data and information during disasters. Usually during the disasters, the volume of tweets increases in the affected areas, and these tweets contain on-the-ground information, expression of fear and relevant information for describing the situation. Confirming and analysing tweets is quiet important for understanding the emergency situations, we can draw valuable conclusions about real life disasters, just by some amount of data.

The task of understanding and analyzing the tweets manually in a real-time

fashion can be draining task that should be automatized. That's why is so important the extraction of information from social media in order to have a timely and accurate response.

We approach the problem of our work by using techniques of clustering, and then we provide some different ways of visualizations from the informations gotten from the techniques. Specifically, in this work we faced this problem by providing an online-algorithm helping gathering information from Twitter to improve the emergency management. The term online meaning that at every time-step we receive some new samples that either form a new sequence or are a continuation of some previously observed sequence, in our case the sequence is a set of incidents or disasters, and then exploring the impact of those incidents and monitoring the evolution of those.

2 Dataset

In order to proceed with our work, we used the ***crisisNLP*** dataset found in [1] which contains tweets related to several disasters as follow :

- Earthquake
- Typhoon
- Volcano
- Floods

and many others, we combined all the dataset in one so that it would help us for checking how our proposed online-algorithm behaves. This dataset will perform the task of data capture and will simulate the capture of information from Twitter which involves burst-detection and classification.

Moreover, the chosen dataset is also helpful to make our proposal work with very well-known techniques in clustering such as k-mean, which is an offline algorithm, we use K-means because of its simplicity and efficiency on the reduced space, compared to the original space. This comparison was not made in the article in which this work is based on.

3 Online clustering for discovery of information

As said before, in order to discover and inspect information from Twitter an online algorithm is proposed here. The online clustering algorithm works as follow : is a mapping

$$f(S(t), k) \rightarrow C(t) = \{C1(t)...Ck(t)\}$$

that for each t $\in$ N gives a partitioning of the index-set $1..N(t)$ into k disjoint subsets $Ci(t), i = 1..k$.

For this, the algorithm should be able to handle high volume of data and not requiring beforehand information about the number of clusters (like in the case of K-means for offline clustering).

Briefly, we implemented the online algorithm by vectorizing first the tweets using the tf-idf approach. After getting this representation of the tweets, we proceeded to the clustering task for which we used the idea an incremental approach.

The incremental approach takes first coming tweet from a stream and uses it to build a cluster. Then, for each incoming tweet it uses a similarity measure with any existing cluster. If the similarity is greater than a threshold, the tweet is assigned to that cluster and the centroid is recomputed with the tweet members of that cluster. Otherwise, a new cluster is created using the incoming tweet. The similarity measures considered here were Jaccard and cosine.

For evaluating our approach, we took the silhoutte score for the online algorithm and compare it with the offline algorithm k-means. The results are the next:

	Online-algorithm (Jaccard)	Online-algorithm (Cosine)	K-means
Silhoutte	0.64	0.59	0.75

According to our results, the online-algorithm based on Jaccard similarity is the best approach and very similar to k-means which also goes along with the results gotten in [2].

4 Visualizations

The information gotten with the online-algorithm can be useful only if we can show it in a proper manner by means of a meaningful visualization, for this purpose we will use tools like concept map, bar chart, and a 2D scatter plot, each one of them has different visualization characteristics. Our goal with the proposed visualizations is not only to show relevant informations from the tweets but also to compare the proposed-online algorithm with the offline one in order to check the difference between them and check how well these two algorithms group the information.

In this sense we consider that showing the number of tweets is quiet relevant in the online-clustering since it is a clear manner to reveal tendencies on Twitter. Since we are working with the online-algorithm it is quiet important to know which cluster is grouping the majority of the information gotten from the stream, and therefore, containing the most relevant information. Because of

that we designed a bar chart as follows:

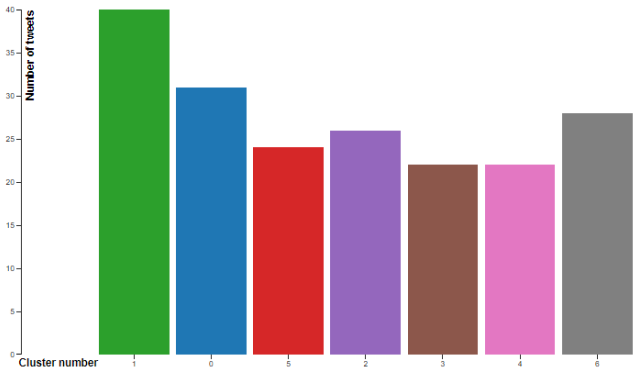

Figure 1: Bar chart showing accumulation of tweets on clusters found

The bar chart represents the multiple clusters, each with a different color, the difference of shapes between the bars reflect clearly the difference of gathered information from each cluster, so we can easily relate the information revealed in other visualizations.

Now that we have the number of tweets for each cluster, we get an idea of which are the dominant clusters found so far. However, each tweet is a short text (approximately 140 characters) that usually contains many words. Some of these words can not be meaningful or relevant for on-the-ground information behind a tweet. That's why we also proposed a conceptual map which allows to easily get the most important words for each cluster. The concept map is as follows:

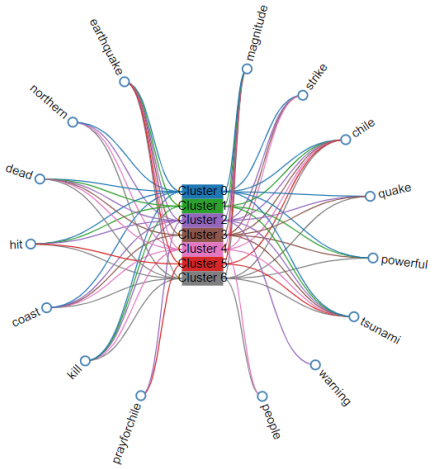

Figure 2: Concept map

Our idea is based on finding the most relevant words at each cluster found so far. Then, we proposed a visualization that relate this words or concepts with the clusters in bidirectional way. It means, that when you mouse over the cluster, you can find the most relevant concepts of it. But also, when you mouse over a concept, you can find the clusters associated with that concept.

Finally, we would like to mention also that with purposes of comparison with an offline algorithm (e.g. k-means), we proposed another visualization to show how the tweets are distributed in a 2D scatter plot. For that, we used PCA over the tf-idf representation and extract from it the first 2 components that maximize the variation in a 2D projection. The visualization is the next one:

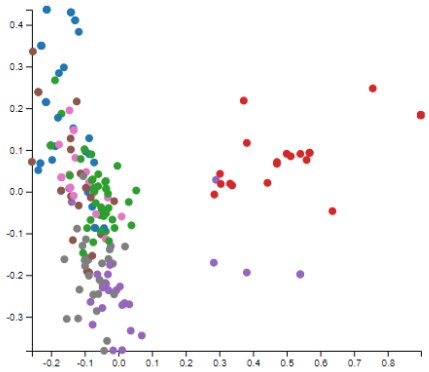

Figure 3: PCA projection on 2D space

As you can see, we play here with the colors so we can distinguish the clusters at which each tweet belongs to in the 2D scatter plot. As we said before, we tried to make a comparison between the online and offline algorithms for which we developed the next interface containing all the visualization showed before.

Figure 4: Total interface

We add three buttons. Pull tweet (green button) will request another 200 tweets from the stream. The other two buttons (yellow and red) will allow us to switch between the visualizations for offline and online algorithm, so we can compare both algorithms in terms of the visualization that we proposed.

5 Comparison with other works

In terms of visualization, the paper on which this work is based propose a visualization like word clouds for each cluster, The word cloud representation is not so efficient in terms of visualization, in the sense that it doesn't give a good overview over the information gathered from the data apart from showing the most important words, by visualizing data using tools like concept map and the bar chart, we can clearly compare clusters with each other in a simple and colorful, we can also compare the two algorithms we used. In our opinion, the visualization way of the paper is not so helpful since a word can be related to more than one cluster and this visualization doesn't help to analyze this kind of situations. We consider important this visualization since an emergency event can be related at the same time to more than one area affected and it helps to identify other clusters that can have meaningful information for one single disaster.

6 Conclusions

We proposed an algorithm for clustering on real-time and also a group of visualization that under our consideration help to analyze in a easy and quick manner the meaningful information gotten through two distinct algorithm for clustering. Moreover, we enabled an interface that allow to compare this two algorithms in term of the information gotten from the proposed visualizations.

References

[1] Twitter datasets from Crisis
`https://crisisnlp.qcri.org/lrec2016/lrec2016.html`

[2] Jie Yin, Andrew Lampert, Mark Cameron, Bella Robinson and Robert Power. *Using Social Media to Enhace Emergency Situation Awareness.* CSIRO ICT Centre.