Visual Analytics: Online clustering on disaster events using Twitter

David Mauricio Guzman Delgado. 1909580
Hamza Bouzidi. 1909207

May 4, 2021

Abstract

Nowadays, social media is the biggest source of information among the world enabling current and rapid communication. Twitter is one of these social media tools used for the broadcast of short textual messages containing (tweets) containing information that usually contain valuable information for the detection of meaningful events such as disasters. 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

Twitter has become one of the most popular social media tools 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.

There's a lot of information supporting that Twitter usually is the preferred tool for sharing data and information during disasters. Usually during the disasters, the volume of tweets increases in the affected areas. Usually the tweets involved in the affected areas contain on-the-ground information, expression of fear and relevant information for describing the situation. Confirming and inspecting the tweets is quiet important for the understanding the emergency situations, specially in a real-time fashion.

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.

This work approach this problem by using techniques of clustering and also

providing helpful visualizations of the information gotten from these techniques. Specifically, in this work we faced this problem by providing an online-algorithm helping gathering information from Tweeter to improve the emergency management. The term online refers to a real-time technique for dealing with high-volume text streams which assist identifying early indicators of incidents, exploring the impact of those incidents and monitoring the evolution of those.

2 Material

In order to assist this work, we used a dataset found in [1] which contains different tweets related to disasters and that will 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 Tweeter which involves burst-detection and classification.

Moreover, the chosen dataset is also helpful to compare our proposal with very well-known techniques in clustering such as k-mean, which is an offline algorithm, in order to get a better understanding on the behavior of the proposed algorithm. This comparison was not made in the article in which this work is base 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. This algorithm is online-incremental which means that it automatically groups similar tweets from a stream into topic clusters. The aim of this is that each cluster groups event-specific topic.

For this, the algorithm should be scalable to handle high volume of data and not requiring a priori-information of the number of clusters (offline clustering, eg. k-means).

Briefly, the online algorithm was implemented by means of vectorized representation of the tweets using tf-idf approach. After getting these 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 meaningful visualization. Our goal with the proposed visualizations is not only to show relevant information from the tweets but also compare the propose-online algorithm with and offline algorithm in order to check the difference between them and check how well this two algorithm 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 Tweeter. Since we are working with the online-algorithm 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:

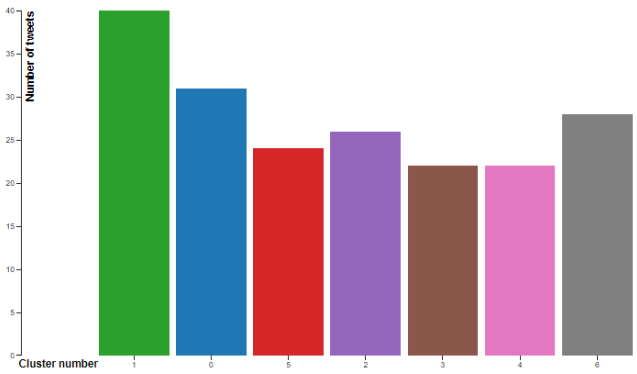

Figure 1: Bar chart showing accumulation of tweets on clusters found

We also use a distinct color for 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:

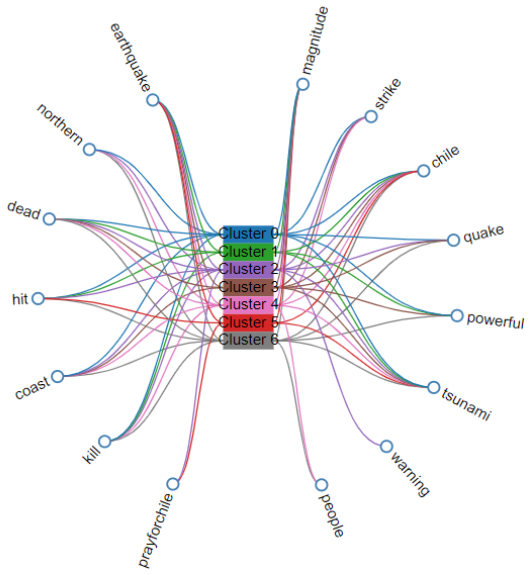

Figure 2: Concept map

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:

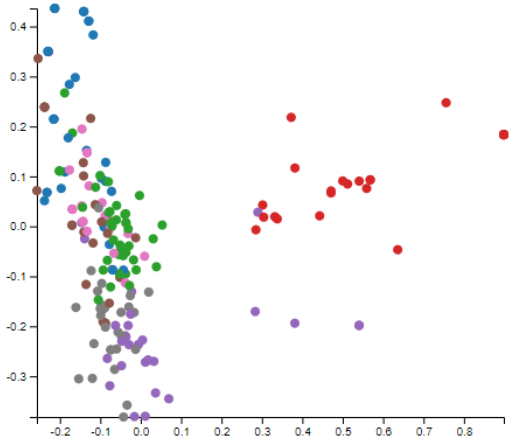

Figure 3: PCA projection on 2D space

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.

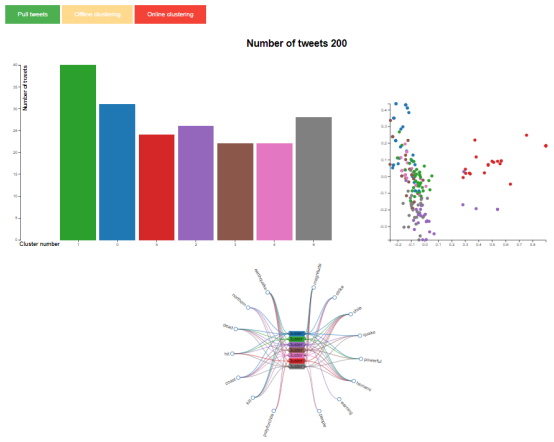

Figure 4: Total interface

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. In our opinion, this visualization 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.