Paper Summary

David Miller
CIS 5930: Social Network Mining

February 7, 2018

The advent of geo-tagged data has allowed computer scientists to implement a cariety of learning systems that involve taking a users location into account when learning and predicting. However current learning and predicting models have poor accuracy. The method proposed in this paper, `TRIOVECEVENT`, learns and predicts events with much higher results than those currently employed. At the very core of `TRIOVECEVENT` it mas ll the regions, times, and keywords into a latent space.. This is done via multimodal embedding where spatial, temporal, and textual units are mapped to lower-dimensional spaced with their correlations preserved. Mathematically we can describe the algorithm with the following notation. We let $\mathcal{D} = \{d_1, \ldots, d_n, \ldots\}$ be a continuous stream of tweets. Each tweet d is a tuple (t_d, l_d, x_d) where t_d is its post time, l_d is its location, and x_d is a bag of keywords in the tweet. If we consider a query window $Q = [t_s, t_e]$ where $t_{d_1} \leq t_s < t_e \leq t_{d_n}$ the local event detection aims at extracting all local events that occur during Q and updating the event list online as Q shifts continuously [1]. A novel Bayesian mixture clustering model is used to divide the tweets in the query window Q into a number of geo-topic clusters to generate candidates. Clusters contain tweets that are close in latent space due to there spatial and textual correlation. These candidates are then classified based on six properties: spatial unusualness, temporal unusualness, spatiotemproal unusualness, semantic concentration, spatial and temporal concentration, and burstiness. The results for this method are significantly better than other methods. Figure 1 shows the results of `TRIOVECEVENT` vs two other methods; 'P' is precision and 'R' is pseudo recall. Evidently the `TRIOVECEVENT` outperforms its competitors. In terms of future work I feel like `TRIOVECEVENT` should also be able to cluster events that are not just in one location, such as presidential elections or global movements. The latent space mapping will map these not too close to each other since their geo-location is not close. Another direction that can be taken is using this in disaster prevention technology. If `TRIOVECEVENT` can detect oncoming emergencies based on tweets, this can be incorporated in communities to detect them of oncoming dangers.

Method	LA			NY		
	P	R	F1	P	R	F1
EvenTweet	0.132	0.212	0.163	0.108	0.196	0.139
GeoBurst	0.282	0.451	0.347	0.212	0.384	0.273
GeoBurst+	0.368	0.483	0.418	0.351	0.465	0.401
TrioVecEvent	**0.804**	**0.612**	**0.695**	**0.765**	**0.602**	**0.674**

Figure 1

Three strengths I found with the paper are

1. The noticeable gains in accuracy compared to other methods.
2. The time complexity is linearly dependent on the number of new tweets.
3. `TRIOVECEVENT` lends itself to a myriad of applications.

Three weaknesses I found with the paper are

1. The paper only compared its results to two other methods, leading me to think there are methods that may come close to its accuracy.

2. The paper makes no mention of using other social streams other than Twitter. This can lead to some bias due to Twitter's posting rules and the limit of tweets the API feeds you. Essentially the sample space may not be fully representative.

3. The algorithm relies on social data when in reality there are events that may not be allowed to be broadcast. Examples include countries or regions that impose restrictions on their citizens.

Questions for the reader

1. How dos the model deal with spoofed locations, if at all?

2. How does the model deal with multiple tweets from the same user?

References

[1] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, Jiawei Han *TrioVecEvent: Embedding-Based Online Local Event Detection in Geo-Tagged Tweet Streams*, KDD17, August 13-17, 2017, Halifax, NS, Canada.