Introducing Social Distance to ST-DBSCAN

Olga Yanenko
University of Bamberg
Chair of Computing in the Cultural Sciences
An der Weberei 5
96047 Bamberg, Germany
olga.yanenko@uni-bamberg.de

Abstract

ST-DBSCAN is a widely used algorithm for event detection in social media. It computes clusters of point data based on the spatial and temporal proximity of the points. However, the social distance between the observers provides additional information, especially when validating the time, location and duration of an event. Observations of an event are more reliable if they are provided by different observers rather than originating from the same source. This paper presents STS-DBSCAN, a variant of the ST-DBSCAN algorithm, that takes into account the social distance between the data creators by introducing a social distance threshold. An evaluation of the proposed clustering method is performed on a dataset with photos about five events from the Flickr platform. The results show an accuracy improvement of the predicted temporal and spatial extents for the five events compared to the original algorithm.

Keywords: VGI, event detection, spatio-temporal proximity, social distance, spatio-temporal clustering, ST-DBSCAN

1 Introduction

Spatio-temporal crowdsourced data also known as Volunteered Geographic Information (VGI, Goodchild 2007) has become a valuable source for automated event detection research by providing large and highly available datasets. Nevertheless, data produced by non-professionals comes with new challenges regarding the data quality (Goodchild 2007, Kisilevich et al. 2010, Schlieder & Yanenko 2010). In classical journalism the data quality issue is addressed by adhering to the *principle of independent confirmation*: information has to be verified by a number of independent sources before being published (Schlieder & Yanenko 2010). The goal of this work is to develop automated approaches that enhance the accuracy of spatio-temporal clustering results by implementing this kind of confirmation principle into the clustering methods.

The main contributions of this work are (1) a characterization of common tagging errors on the Flickr platform, (2) the definition of STS-DBSCAN, a version of the ST-DBSCAN algorithm originally introduced by Birant and Kut (2007) that takes the social distance between the data creators into account, (3) the evaluation of the event detection results produced by STS-DBSCAN compared to ST-DBSCAN and a simple method based on the yearly mean and one standard deviation.
The remainder of the paper is organized as follows: Section 2 presents related work in the field of event detection in social media. Section 3 provides an overview of common data quality issues of crowdsourced data. The STS-DBSCAN algorithm is described in section 4. Section 5 describes the experimental setup for the evaluation of event detection as well as a performance comparison between ST-DBSCAN and STS-DBSCAN. Section 6 concludes the paper and gives an overview of future work.

2 Related Work

Spatio-temporal DBSCAN (ST-DBSCAN, Birant & Kut 2007) is a widely used algorithm for event detection in social media. It extends the density-based clustering algorithm DBSCAN proposed by Ester et al. (1996) by taking into account the temporal proximity between two points besides the spatial proximity. Its performance was demonstrated in different social media event detection scenarios, such as rainfall (Feng & Sester 2017) or crime detection (Huang et al. 2018) in Twitter[1] data.

However, most of the event detection research that relies on spatio-temporal clustering with ST-DBSCAN concentrates on preprocessing methods that improve the data basis before the clustering in order to produce better results. But there is also a considerable amount of ST-DBSCAN variants, that address different shortcomings of the original algorithm (Tork 2012, Arcaini et al. 2016). P-DBSCAN (Kisilevich et al. 2010) which

[1] http://www.twitter.com

stands for Photo-DBSCAN is probably the most similar algorithm to the STS-DBSCAN method presented in this work. P-DBSCAN was developed for finding interesting places by analyzing the photo activity in a specific area. The authors extended the original DBSCAN algorithm which only operates on spatial data without taking the temporal component into account with an ownership function that requires that a photo has at least one neighbor that was created by a different owner, otherwise it is considered as noise. Although the social distance used in this work is similar to the ownership function of P-DBSCAN, a more general definition of social distance seems necessary that accounts for arbitrary social distance measures.

3 Data Quality of Crowdsourced Data

Data produced by volunteers is known to be biased and thus poses new challenges to the research community (Kisilevich et al. 2010). This also applies to the dataset consisting of 10.043 images from the Flickr[2] platform created for this work. The dataset contains images that are tagged with one of five yearly German events chosen for the event detection evaluation, namely *Sandkerwa* (s), *Berlinale* (b), *Oktoberfest* (o), *Fusion Festival* (f) and *Köln Karneval* (k) including the image owner and tags provided by the same. Flickr was preferred over Twitter since it's simpler to verify if a picture actually shows a certain event in contrast to a short text referring to an event. Four of the events were chosen because they are very popular and therefore have a sufficient number of images on the Flickr platform. *Sandkerwa* was chosen despite the fact that there might be not enough photos for an accurate event detection but it can give more insights about tagging errors since it is well known by the author. Another reason for choosing well-known events is that there is a ground truth about the time and location of the events.

In order to identify common tagging errors in the dataset 2158 images were manually reviewed. A tagging error is defined as an image that is tagged with an event but does not actually show this event. Note, that the tag is not necessarily incorrect but since it is not useful for detecting the time and location of the event, it is considered as incorrect for this work.

663 tagging errors were identified, that means that ~ 44% of the images with an event tag did not show the event itself. Table 1 summarizes common tagging errors that were found in the dataset. The errors are combined to error categories which are grouped by two parameters that refer to the actual time and location of an event. For example, images that are taken while travelling to an event are often tagged with this event although they are outside its spatial and temporal extent. One user in the dataset was travelling several weeks through Europe, his journey ended with a short visit at the *Oktoberfest* in Munich, but all pictures of the trip were tagged with this event.

Table 1: Common tagging errors in Flickr

		image location: event	image location: ¬event
image timestamp	¬event	event location (113) bulk tagging (46) analog/historical photos (29) bulk tagging travel (20) similar event (15) promotion (16)	similar event (137) bulk tagging travel (44) promotion (2) other (8)
	event	bulk tagging travel (83) promotion (3)	similar event (87) bulk tagging travel (60)

Most of the tagging errors are part of four big categories:

1. **Bulk tagging**: Bulk tagging in Flickr is maybe the most frequent reason for tagging errors. This includes bulk tagging of whole journeys as mentioned above. But travel and tourism is not the only reason for bulk tagging errors. For example, one user tagged all his pictures of the event *Sandkerwa* with fireworks (german: Feuerwerk) although only a small part of the photos is actually showing the fireworks.
2. **Similar event**: This includes some small event clones as well as unknowingly wrong event designations. For example, some tourists seem to call every German beer fest *Oktoberfest*.
3. **Event location**: Pictures of locations where an event is usually held. For example, photos of the Sony Center in Berlin which is the main location for *Berlinale* are often tagged with the event although they are only showing the event location. The same applies for Theresienwiese as the location for the *Oktoberfest*.
4. **Promotion**: The category promotion includes besides pictures of event posters also photos of trophies, costumes, stamps and other objects related to the event.

4 STS_DBSCAN

ST-DBSCAN is a widely used algorithm when it comes to event detection in social media data. It clusters point data based on the spatio-temporal density of the provided points. Three parameters have to be provided, $Eps1$ (the spatial threshold), $Eps2$ (the temporal threshold) and $minPts$ (the minimum number of neighbors for a point to be considered a member of a cluster). The original work of Birant and Kut (2007) uses a method called $Retrieve\_Neighbors(obj, Eps1, Eps2)$ for finding the neighbors of obj. As pointed out in Schlieder and Yanenko (2010) besides the spatio-temporal proximity, the social distance is also an important factor for data validation. Since most of the tagging errors in Flickr can be seen as subjective errors produced by one user, it is important to have the provided data confirmed by other users than the image owner himself. If the social distance of the users is being ignored, unknowingly wrong tagged pictures of one user are compared to other (probably also wrong tagged) pictures of the same user what leads to less reliable results. Thus, STS-DBSCAN extends the ST-DBSCAN algorithm with a third threshold parameter $Eps3$

[2] http://www.flickr.com

– a social distance threshold. Figure 1 shows the modified version of the *Retrieve_Neighbors* method, the modifications are highlighted in red.

Figure 1: Modification of ST-DBSCAN

```
Retrieve_Neighbors(obj, Eps1, Eps2, Eps3):

Neighbors = []

For i=1 to |D|:

    If dist1(obj, oi) <= Eps1 &
       dist2(obj, oi) <= Eps2 &
       dist3(obj, oi) >= Eps3:

        Neighbors += oi

Return Neighbors
```

Since $minPts$ is used to decide if a point obj shall be considered as part of a cluster or as noise, the modified algorithm ensures that obj has to be confirmed by points created by users that have at least the predefined social distance $Eps3$ to the owner of obj. The social distance function has to be provided based on the underlying data and use case. For example, a graph-based distance measure can be defined as the minimum number of edges between two users of a social network. In this case, $Eps3 = 0$ means that all points in the dataset will be considered for computing the neighbors for an object obj while $Eps3 = 1$ only takes points from different users into account and $Eps3 = 2$ only points that were created by users that are not directly connected to the owner of obj.

5 Experimental Setup

In order to evaluate the results produces by STS-DBSCAN, a simple experiment was conducted with the data collected from Flickr.

5.1 Data

The experimental evaluation was performed on the dataset described in section 3. The amount of images per event and year are summarized in table 2. The number after the slash indicates the number of users that provided the images. It can be clearly seen that the pictures of an event are mostly created by a small number of users.

The red numbers indicate data which was not sufficient to compute all five methods used for evaluation. The methods are described in the following section.

Table 2: Number of images per event tag and year

	s	b	o	f	k
2008	3/1	355/28	2/1	-	-
2009	50/1	695/29	8/2	-	152/1
2010	65/3	293/33	5/4	-	-
2011	-	647/36	3/1	-	1/1
2012	33/3	134/21	-	-	480/19
2013	27/2	372/27	24/3	87/10	880/20
2014	8/2	178/19	47/7	325/10	1295/18
2015	3/3	37/4	355/45	154/17	494/21
2016	-	83/8	1262/91	208/12	68/16
2017	-	43/3	724/71	59/4	384/11

5.2 Method

For the evaluation of the proposed STS-DBSCAN algorithm, the spatial and temporal extent of an event were predicted on a yearly basis by five different methods:

1. **MEAN & SD**: The simplest and most obvious method for detecting outliers is to compute the annual mean and one standard deviation of the spatial and temporal footprint of the images.
2. **ST-DBSCAN**: ST-DBSCAN clusters were computed for the five events. Since it is known that all events in this study are yearly events, all clusters of one year were considered related to this event and grouped together. The following parameters were used: $Eps1 = 1km$; $Eps2 = 1d$. As proposed in the original paper (Birand & Kut 2007) $minPts$ was set to $ln(n)$ where n is the total number of images to process.
3. **ST-DBSCAN-BIGGEST**: Since part of the event data (especially *Oktoberfest* and *Fusion*) contained besides photos of the desired event also photos about smaller events tagged with the same name, another ST-DBSCAN based method was computed. In contrast to method 2 only the biggest cluster is considered as being related to the main event.
4. **STS-DBSCAN**: Same as method 2 but with the modified algorithm described in section 4. The social distance between two images was defined as 0 if the images were taken by the same user, and 1 otherwise. The social threshold $Eps3$ was set to 1.[3]
5. **STS-DBSCAN-BIGGEST**: Same as method 3 but with the modified STS-DBSCAN algorithm and the social threshold from method 4.

Based on these five methods, five different clusters were computed for each event tag and year. The minimum and maximum values of the timestamps and geotags of each cluster were taken to estimate the event's start date, end date and location. In order to compare the results of the three methods the temporal error err_{temp} was computed as the mean of the distances between the minimum date of the cluster $date_{min}$ and the real start date $odate_{start}$ of the according event and respectively the maximum date of the cluster $date_{max}$ and the real end date $odate_{end}$. The ground truth for this evaluation was obtained from event announcements by official resources:

[3] Unfortunately, the dataset was too sparse and not suitable for testing a more complex social distance measure as proposed in section 4.

$$err_{temp} = \frac{|odate_{start} - date_{min}| + |odate_{end} - date_{max}|}{2} \quad (1)$$

In contrast to the temporal error, the evaluation of the spatial error is less trivial, because there is no accurate ground truth. Thus, an approximation of the correct spatial extent of an event was constructed by reviewing the official announcements for each event and drawing a bounding box represented by four parameters $olat_{min}$, $olat_{max}$, $olon_{min}$ and $olon_{max}$ that fully contains the official event locations. The overall spatial error was defined as the mean of the latitude error err_{lat} and the longitude error err_{lon} computed as follows:

$$err_{geo} = \frac{err_{lat} + err_{lon}}{2} \quad (2)$$

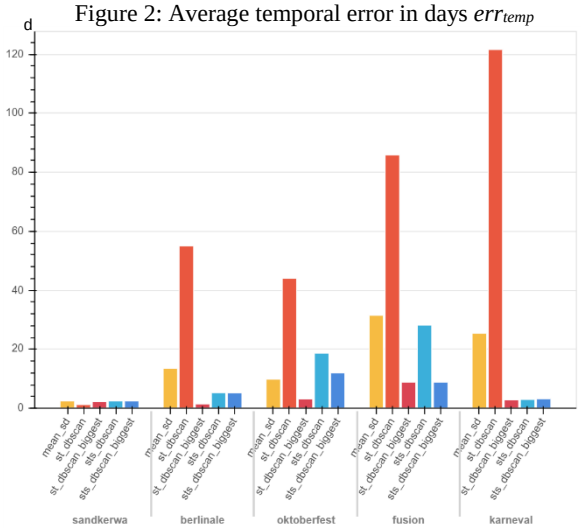
$$err_{lat} = \frac{|olat_{min} - lat_{min}| + |olat_{max} - lat_{max}|}{2} \quad (3)$$

$$err_{lon} = \frac{|olon_{min} - lon_{min}| + |olon_{max} - lon_{max}|}{2} \quad (4)$$

5.3 Results

Figure 2 shows the overall temporal error results by event and clustering method.

Figure 2: Average temporal error in days err_{temp}

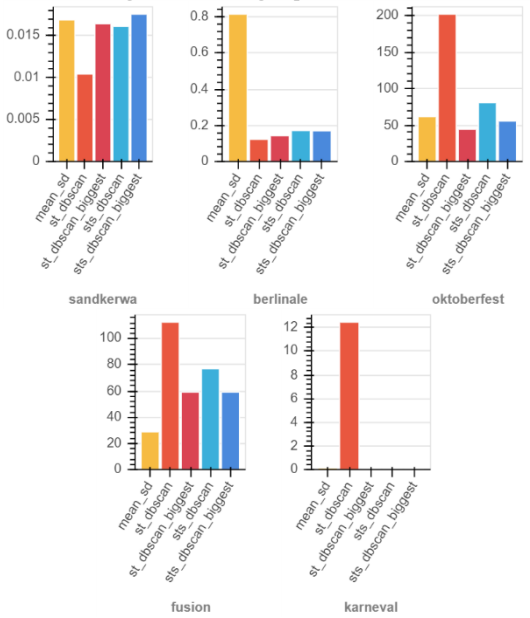


Due to the different characteristics and data amount of the analyzed events, none of the five methods can be regarded as best in predicting the temporal extent of the events. Nevertheless, when comparing methods 2 and 3, except for the *Sandkerwa* event the results computed by the proposed STS-DBSCAN algorithm clearly outperform the results of ST-DBSCAN. The difference between methods 4 and 5 is less distinctive but there is still an improvement when using STS-DBSCAN compared to the original ST-DBSCAN.

The overall results of the spatial error evaluation are presented in figure 4. They were plotted per event since there were considerable differences in the outcome.

Although the application of the proposed ST-DBSCAN algorithm produces smaller errors in the whole, the results of the spatial error evaluation are less conclusive than those of the temporal error evaluation. One of the explanations is probably the fact that the tagging errors have a wider distribution in the spatial dimension than in the temporal. The biggest errors can be observed for *Oktoberfest* and *Fusion* because the data regarding these events had a considerable amount of photos that were referring to smaller events with the same name. For the event *Sandkerwa*, the temporal as well as the spatial error is smallest when applying method 2 based on the original ST-DBSCAN algorithm. That is presumably the result of the sparse data obtained for this small event. In this case, taking the social distance into account leads to very small clusters that do not cover the whole event and therefore have a bigger error than clusters computed by the original ST-DBSCAN algorithm.

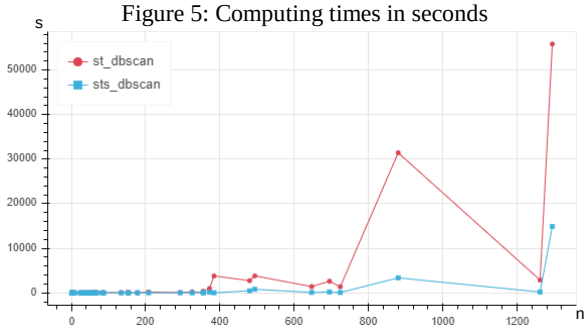
Figure 4: Average spatial error err_{geo}



In some cases, even the simple mean and standard deviation method produced better results than the ST-DBSCAN clustering. Due to its fast computation, this method can be sufficient depending on the use case. For example, if only an estimation of the temporal and spatial extent is needed and performance is more important than accuracy or if the data is to sparse for the computation of spatio-temporal clusters. However, the main limitation of the mean_sd method for event detection is that it only works for periodically repeated events where the period is known.

Another advantage of the STS-DBSCAN algorithm is that it can reduce the computing times for the clusters. Figure 5 shows the computing times in seconds that were recorded within the experiment. However, the main reason for that is the simple social distance model used in this work. Using a more complicated graph-based distance measure to compute the social distance between two data providers will presumably lead to higher computation times.

Figure 5: Computing times in seconds



6 Conclusion and Future Work

This work presented STS-DBSCAN, an extension of the ST-DBSCAN algorithm that takes into account the social distance of the data providers besides the spatial and temporal proximity of the data points. A first experiment demonstrated the application of the proposed method for event detection in social media data. The evaluation shows that the proposed STS-DBSCAN algorithm can produce more accurate results than the original ST-DBSCAN when clustering spatio-temporal social media data, especially from the Flickr platform. However, in order to explore the benefits of a more complex social distance measure, further studies with different and bigger datasets are needed. Besides, the influence of varying thresholds for the spatial, temporal and social distances have to be analyzed. A more detailed research of use cases that can benefit from the social distance extension as well as some experiments with event-related Twitter data will be performed in the near future.

References

Arcaini, P., Bordogna, G., Ienco, D. and Sterlacchini, S. (2016) User-driven geo-temporal density-based exploration of periodic and not periodic events reported in social networks. In: *Information Sciences* 340, pp.122-143.

Birant, D. and Kut, A. (2007) ST-DBSCAN: An algorithm for clustering spatial–temporal data. In: *Data & Knowledge Engineering* 60 (1), pp. 208–221.

Ester, M., Kriegel, H.-P., Sander, J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Data Mining and Knowledge Discovery*, pp. 226-231.

Feng, Y. and Sester, M. (2017) Social media as a rainfall indicator. In: Bregt, A., Sarjakoski, T., Lammeren, R. van, Rip, F. (eds.): *Societal Geo-Innovation: short papers, posters and poster abstracts of the 20th AGILE Conference on Geographic Information Science*.

Goodchild, M. F. (2007) Citizens as sensors: the world of volunteered geography. In: *GeoJournal* 69 (4), pp. 211–221. DOI: 10.1007/s10708-007-9111-y.

Huang, Y., Li, Y. and Shan, J. (2018) Spatial-temporal event detection from geo-tagged tweets. In: *ISPRS International Journal of Geo-Information*, 7 (4), p.150.

Kisilevich, S., Mansmann, F. and Keim, D. (2010) P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos. In: *Proceedings of the 1st international conference and exhibition on computing for geospatial research & application* (p. 38). ACM.

Schlieder, C. and Yanenko, O. (2010) Spatio-temporal proximity and social distance. In: Xiaofang Zhou, Wang-Chien Lee, Wen-Chih Peng und Xing Xie (eds.): *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* - LBSN '10. the 2nd ACM SIGSPATIAL International Workshop. San Jose, California, 11/2/2010 - 11/2/2010. New York, New York, USA: ACM Press, pp. 60-67.

Tork, H.F. (2012) Spatio-temporal clustering methods classification. In: *Doctoral Symposium on Informatics Engineering* 1 (1), pp. 199-209.