Density based Clustering for Geo-referenced documents – A Survey

Nilang R. Pandey[1], Mr Narendra Limbad[2]
[1]*Computer Engineering Department, L.J.I.E.T, Ahmedabad, India*
[2]*Computer Engineering Department (P.G. Department), L.J.I.E.T, Ahmedabad, India*

***Abstract*— Nowadays, an enormous amount of geo-referenced documents are posted on social media which have their location and posting time information. The objective of this survey paper is to discuss various works that have been carried out on geo-referenced documents. Geo-referenced documents have spatial and location information tagged with them from which we can extract required information. All the works discussed here are based on DBSCAN which finds arbitrary shapes clusters and is appropriate for geo-spatial clusters.**

***Index Terms*— Density based clustering, geo-referenced documents, spatio-temporal clusters, spatial clusters, topic retrieval, location detection.**

I. Introduction

Nowadays, we are witnessing the emergence of new types of spatio-temporal data such as GPS, mobile networks, sensors, geo-tagged images. Location-acquisition technology is becoming very popular among ordinary people, allowing them to record their positions, while the Internet allows them to share their data with others [1]. Geo-referenced documents are typically a type of spatio-temporal data, from which we can extract the information regarding the local events and topics posted in them via Internet.

Extracting local area topics and their location from geo-referenced documents help in various geo-location domain applications like marketing, tourism informatics and local topic retrieval. The goal of this survey paper is to discuss various works that have been carried out on geo-referenced documents for local topic extraction. All these various methods are based on the most basic density based clustering algorithm that is DBSCAN. The most significant impact on many studies is DBSCAN, a density-based clustering method for geo-spatial data. [7.] The shapes of clusters that formed by the geo-spatial data are arbitrary shaped and hence DBSCAN is the most preferred algorithm. Also we do not need to have prior information about the number of clusters. So in these cases the density based clustering is considered as the most appropriate.

A. Density based clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density based clustering algorithm. The algorithm grows regions with sufficiently high density into clusters and discovers clusters of arbitrary shape in spatial databases with noise. [8.] The clusters with high density points are made into a clusters and the lower dense areas are considered as noise. This algorithm finds the clusters of arbitrary shape.

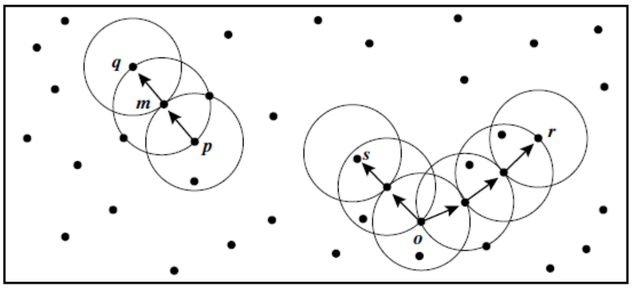

Figure 1: Density reachability and density connectivity[8.]

- The neighborhood within radius ε of a given object is called ε -neighborhood of the object.
- If the ε -neighborhood of an object contains at least a minimum number of points *MinPts*, then object is called a core object.
- Given a set of objects D, and consider an object *p* is directly density reachable from object *q* if *p* is within ε -neighborhood of *q*, and *q* is a core object.
- An object *p* is density connected to *q* with respect to ε and *MinPts* in a set of objects D, if there is a chain of objects p_1,…., p_n, where p_1 = *p* and p_n = *q* such that p_{i+1} is directly density reachable from p_i with respect to ε and *MinPts.*
- An object p is density-connected to object q with respect to ε and MinPts in a set of objects D, if there is an object *o* such that both *p* and *q* are density reachable from *o* with respect to ε and *MinPts*.

The steps involved in this algorithm are as follows.

- Select a random point *p*.

- Extract all density-reachable points from *p* by given parameters *Eps* and *MinPts*.
- If *p* is a core point, then a cluster is formed.
- If *p* is a border point, no points are density reachable from *p* and DBSCAN visits the next point of the database.
- Continue the process until all the points have been processed.

II. LITERATURE SURVEY

1. P-DBSCAN: A density based clustering algorithm for analysis of attractive areas using collection of geo-tagged photos. [1.]

This paper presents a new clustering algorithm based on DBSCAN for analysis of places and events using a collection of geo-tagged photos. Here the authors have introduced two new concepts of density threshold and adaptive density. Density threshold is known as the number of people in the neighborhood, while adaptive density is used for fast convergence towards the high density regions.[1.]

Some of the terminologies used in this paper are.

- Neighborhood of photo point – For a photo point p which belongs to owner o_i, we find at least one point *q* whose owner is not o_i in a neighborhood of radius ϵ. [1.]
- Core photo – This is a photo point where at least a minimum number of owners *MinOwners* not including the owner photo *p* took photos.[1.]
- Directly ownership reachable – A photo point *q* is directly ownership reachable from a point *p* when photo point *q* belongs to neighborhood of *p*. [1.]
- Ownership reachable – A photo point is ownership reachable if there is a chain of photo points $p_1, p_2, ..., p_n = p$ such that p_{i+1} is directly ownership reachable from p_i. [1.]
- Adaptive density threshold – The ratio of current density of a photo point *p* and the previous density. The neighbors of the photo are assigned to the current cluster until density ratio is greater than density threshold. [1.]

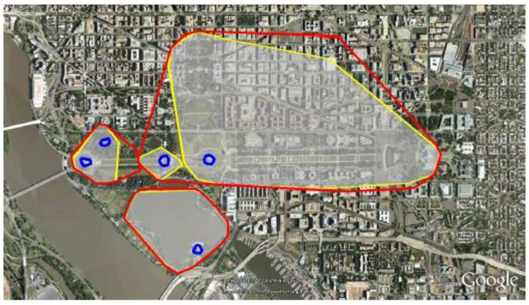

Figure 2 : Experimental results for P-DBSCAN [1.]

The inputs to the algorithm are dataset of points with coordinate and ownership attributes, ϵ - neighborhood radius, adaptive density flag, and adaptive density drop threshold. The output will be the set of clusters.

The authors have shown three types of clusters. First of all, the output that is set of clusters by generic DBSCAN algorithm. After then they have considered their P-DBSCAN algorithm and accordingly shown its output. Another consideration is of P-DBSCAN with adaptive density where more packed clusters are made hence showing highly dense clusters with high photo activity.

The density based clustering based on ownership tends to create small clusters, while adaptive density leads to creation of small 'packed' clusters with high density. The clusters given as output have places where these photos were taken hence analyzing the attractive areas using the geo-tagged photos.

2. Density based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Georeferenced Documents [2.]

This clustering algorithm is used in extracting bursty areas associated with local topics and events from georeferenced documents. The authors have proposed a new clustering algorithm that will make clusters that are not only spatially separated but also temporally separated. To evaluate the proposed clustering algorithm the authors have used geo-tagged tweets posted on Twitter site. Hence this clustering algorithm will extract bursty areas from geo-tagged tweets that also takes into account during the time they were posted on Twitter. The proposed (ϵ, τ) density based spatiotemporal clustering algorithm is extension of DBSCAN algorithm.

Some of the terminologies used in this paper are.

- (ϵ,τ)-neighborhood *N*(*dp*) – This is point that is in given range from *dp* and also within given inter arrival time.
- *MinDoc* – This is the minimum number of documents around a particular document *dp*.
- (ϵ,τ)- density based directly reachable – If a document *dq* is in the neighborhood of *dp* and $N(dp) > MinDoc$ then *dq* is said to directly density reachable from *dp*.

The inputs to the algorithm are datasets with coordinate values that is tweets with latitude and longitude value, ϵ - neighborhood radius, τ–inter arrival time, *MinDoc* is minimum number of document within the specified radius. Here the authors have collected tweets from Twitter API as their geo-referenced documents.

Steps of algorithm:

1. *IsClustered* – First of all the algorithm checks that whether the selected document is already in the cluster or not.

2. *GetNeighborhood* – This gets the documents in the neighborhood of the considered document *dp*. The documents obtained by this method are placed in the queue *Q* for further processing. The assignment and processing is done until the queue gets empty.
3. *EnNniqueQueue* – This function is used when the dequeued document is a core document and its neighborhood documents are queued to queue *Q*.

The first type of cluster is of DBSCAN algorithm cluster and the other type of cluster is of the proposed (ϵ-τ) density based spatiotemporal clustering algorithm. The clusters are ranked according to the number of tweets in each cluster.

Hence the results show for a particular keyword, clusters given by the (ϵ-τ) density based spatiotemporal clustering algorithm are both spatially and temporally separated.

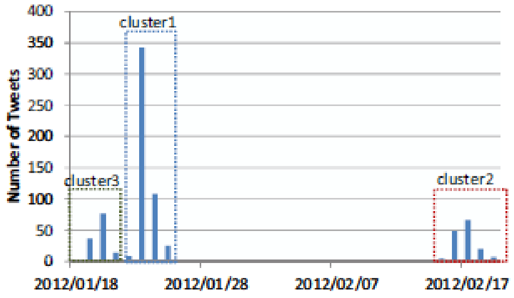

Figure 3 : Cluster with different durations [2.]

3. Extracting Attractive Local-Area Topics in Geo-referenced Documents using a new Density–based spatial clustering algorithm. [3.]

In this paper, a novel spatial clustering algorithm for extracting "attractive" local-area topics in georeferenced documents, known as the (ϵ,σ)-density-based spatial clustering algorithm, is proposed.[3] This proposed algorithm will generate clusters that are both spatially and semantically separated. Thus the semantically separated clusters are able to extract closely related local area topics with their respective areas.

Here they defined a new similarity measurement for georeferenced documents on social media sites. On such sites, users typically post georeferenced documents comprising short messages including a local topic. Therefore, if georeferenced documents include the same keyword, they can be considered similar to each other.[3.]

Some of the terminologies used in this paper are

- (ϵ,σ)-neighborhood *GN(gdp)* – This is the neighborhood of a geo-referenced document gdp

$$GN_{(\epsilon,\sigma)}(gdp) = \{gdq \; \epsilon \; GDS | dist(gdp, gdq) \leq \varepsilon \; and \; sim(gdp, gdq) \geq \; \sigma\} \qquad (1.)$$

- (ϵ,σ)-density based reachable – Suppose that there is a geo-referenced documents sequence $(gd_1, gd_{2,} gd_3, \ldots, gd_n)$ and the $(i+1)$th geo-referenced document gd_{i+1} is (ϵ,σ)-density based reachable from gd_i.[3.]
- Word based Simpson's coefficient – This has the feature of cosine similarity between sets. The Simpson's coefficient is given as

$$wsim(gd_i, gd_j) = \frac{|dt_i \cap dt_j|}{|\min\ (dt_i, dt_j)|} \qquad (2.)$$

- Keyword based Simpson's coefficient – This will describe the similarity between the documents on the basis of the keyword.

$$ksim(gd_i, gd_j) = \frac{|key_i \cap key_j|}{|\min\ (key_i, key_j)|} \qquad (3.)$$

Here the authors have collected geo-tagged tweets extracted from Twitter API over a certain period of time. There are total three types of clusters shown in the output. They are DBSCAN cluster, word-based cluster and keyword-based cluster. Each cluster is ranked according to the number of tweets in it. The other results in the cluster are the range of the cluster in terms of latitude and longitude and the top 5 frequent words in the cluster.

The resulting cluster will show that clusters shown by DBSCAN will have larger clusters and a variety of topics, hence this is not sufficient to extract local topics. The word based clusters have lesser number of tweets with much specific topics in the clusters. Now in keyword based clusters the number of tweets in the clusters have decreased and they show specific topic in a particular region.

4. Jasmine: A Real time local-event detection system based on geo-location information propagated to microblogs. [4.]

Here the authors propose a system for detecting local events in the real-world using geolocation information from microblog documents. A local event happens when people with a common purpose gather at the same time and place.

To detect such an event, authors identify a group of Twitter documents describing the same theme that were generated within a short time and a small geographic area.[5.] While collection of tweets from the Twitter API many of the tweets were not geo-tagged which had the local topic event but were not useful because of the lack of the geo-location information embedded in it.

The architecture of the Jasmine system is has four parts. They are as follow.

1. Twitter Stream Collector – This module collects Twitter documents through the Twitter Streaming API. [5.]
2. Geotag Allocator – The Place Name Database is created. Place names with certain pattern are extracted. Then, the geographical variance among geotagged documents that

include each place name is calculated. Then the system searches for non-geotagged documents which contains a distinct place name and allocates the geotag to location in real time. [5.]

3. Popular place extractor – This module finds popular places by detecting geotags observed frequently in Twitter documents posted during the user specified period. [5.]
4. Key term Extractor – This modules extracts key terms that describes an event that might be held at each popular place. Key terms that appear three times in the Twitter documents associated with the popular place are extracted. [5.]

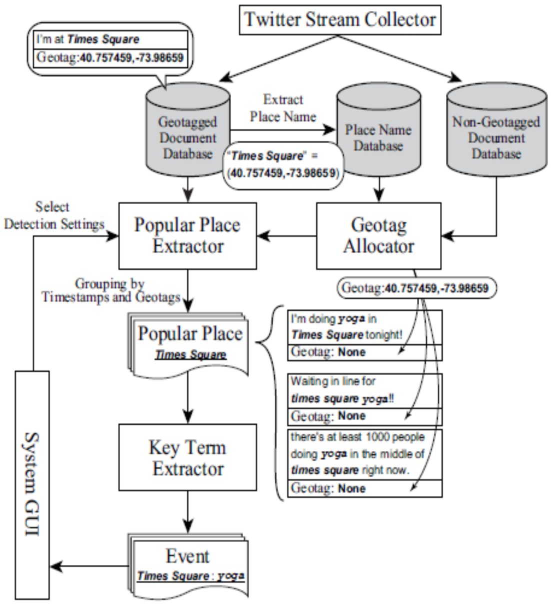
Twitter Stream Collector
I'm at Times Square
Geotag:40.757459,-73.98659
Geotagged Document Database
Extract Place Name
"Times Square" = (40.757459,-73.98659)
Place Name Database
Non-Geotagged Document Database
Select Detection Settings
Popular Place Extractor
Geotag Allocator
Grouping by Timestamps and Geotags
Geotag:40.757459,-73.98659
Popular Place
Times Square
I'm doing yoga in Times Square tonight!
Geotag: None
Waiting in line for times square yoga!!
Geotag: None
there's at least 1000 people doing yoga in the middle of times square right now.
Geotag: None
Key Term Extractor
System GUI
Event
Times Square : yoga

Figure 4 : Architecture and working of Jasmine system [4.]

This system that automatically geotags the documents successfully assigns location information to non-geotagged documents and increases the number of popular places detected by about 115 times and increasing the accuracy by 25.5%

Here this algorithm does not use any density based clustering method hence the efficiency for showing clusters arbitrary shape is not that good.

5. Finding arbitrary shaped clusters with related extents in space and time [5.]

This paper shows the local event detection using density based clustering with the use of spatio-temporal clusters. This workis based on Tobler's Law which says, events in a small area can be sparse in time and still connected. And on the other hand, events in large areas are likely to be connected if they are close in time. [5.]

Here clustering is done in two steps. In first step by considering the user defined spatial distance threshold the spatial clusters are determined. While in second step, for each spatial cluster DBSCAN is applied again by considering the temporal threshold.

There are several parameters and terminologies used in this paper.

$$\varepsilon_i^t = T(C_i) = \min\left\{1, \frac{N}{n_i}\right\} \cdot \frac{D}{d_i} \cdot \varepsilon^t \tag{4.}$$

$$n_i = \|\{e \in C_{i,j(\tau)}\}\| \tag{5.}$$

$$d_i = \sqrt{(\max\{e.x\} - \min\{e.x\})^2 + (\max\{e.y\} - \min\{e.y\})^2} \tag{6.}$$

$$t_i = \sqrt{(\max\{e.t\} - \min\{e.t\})^2} \tag{7.}$$

$$\rho_{d_i} = {}^{d_i}/_{n_i} \tag{8.}$$

$$\rho_{t_i} = {}^{t_i}/_{n_i} \tag{9.}$$

Where ε_i^t is the temporal cluster threshold, ε^t is the maximum temporal distance between the entries, N is the minimum number of entries in cluster, D is the spatial extent. n_i is the number of entries and d_i is the spatial extent of a particular cluster i.

Here the authors have considered Flickr dataset as their input to the algorithm. First of all the spatial cluster with other temporal clusters within it are considered. The extent of the spatial cluster is large and the time interval between entries should be less. All these temporal clusters describe the same event. Next the spatial extent with less area is considered where the time interval is large. So with the effect of small spatial extent and large time interval the generated temporal clusters describe about the same event.

III. CONCLUSION

This paper surveys on various work carried out on density based clustering for geo-referenced documents. Different works have considered different inputs as their geo-referenced documents and have considered density based clustering as their base for forming clusters. These clusters are helpful in retrieving the local topics and extracting respective locations. Density based clustering method was considered here to make clusters, as rendering clusters geographically asks for a method that could make clusters of arbitrary shape. Some work considered the temporal parameter and made clusters that were both spatially and temporally separated. Semantic parameter was also taken into account while making clusters that were spatially separated. Most of the works considered DBSCAN, and compared their work with DBSCAN. Local topic retrieval shows the prevalence of various topics over a local area which could be helpful in many ways.

REFERENCES

[1] Kisilevich, Slava, Florian Mansmann, and Daniel Keim. "P-DBSCAN: a density based clustering algorithm for exploration and analysis of attractive areas using collections of geo-tagged photos." *Proceedings of the 1st International Conference and Exhibition on Computing for Geospatial Research & Application*. ACM, 2010

[2] Tamura, Keiichi, and Takumi Ichimura. "Density-Based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Georeferenced Documents." *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on*. IEEE, 2013

[3] Sakai, Tatsuhiro, Keiichi Tamura, and Hajime Kitakami. "Extracting Attractive Local-Area Topics in Georeferenced Documents using a New Density-based Spatial Clustering Algorithm." *IAENG International Journal of Computer Science* 41.3 (2014)

[4] Watanabe, Kazufumi, et al. "Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs." *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011

[5] Pölitz, Christian, Gennady Andrienko, and Natalia Andrienko. "Finding arbitrary shaped clusters with related extents in space and time." *EuroVAST 2010: International Symposium on Visual Analytics Science and Technology*. The Eurographics Association, 2010

[6] Han, Jiawei, and Micheline Kamber. *Data Mining, Southeast Asia Edition: Concepts and Techniques*. Morgan kaufmann, 2006.