Master Thesis

Spatial Temporal Analysis of Social Media Data

Supervisors: Dr.-Ing. Christian Murphy
Khatereh Polous

Presented By
Smita Singh

Contents

Introduction

1. Motivation

- To identify what users are talking about in social media.

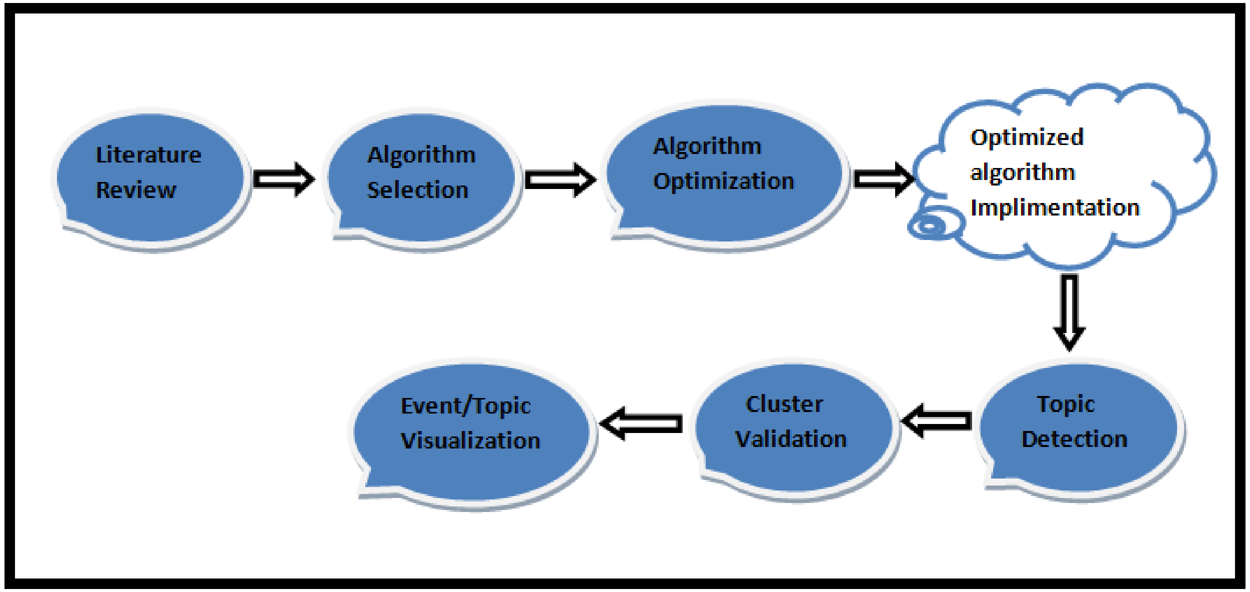
2. Purpose

- To extract local hot topics and thereafter events from the social media data.

3. Research Questions

- Which clustering methods are available?
- Can a suitable algorithm be identified for extracting local hot topics from literature review?
- How to extract local hot topics from spatial temporal data?
- How to validate the clustering result ?
- How to visualize event clusters ?

Workflow of thesis



4. Related work

Birant et al. (2007)

- Detect the cluster in both spatial and non-spatial attributes of dataset.
- Detect noise in varied density.
- Requires three user defined parameter to identify the cluster.

Chen et al. (2009)

- To detect event on Flickr data.
- Wavelet transform approach is used to remove the noise from the data.
- Suitable to detect periodic event.

Kisilevich et al. (2010)

- photo based DBSCAN for event detection through geo-tagged photograph.
- Considered user as density threshold to detect the unique event.
- Requires user defined parameter.

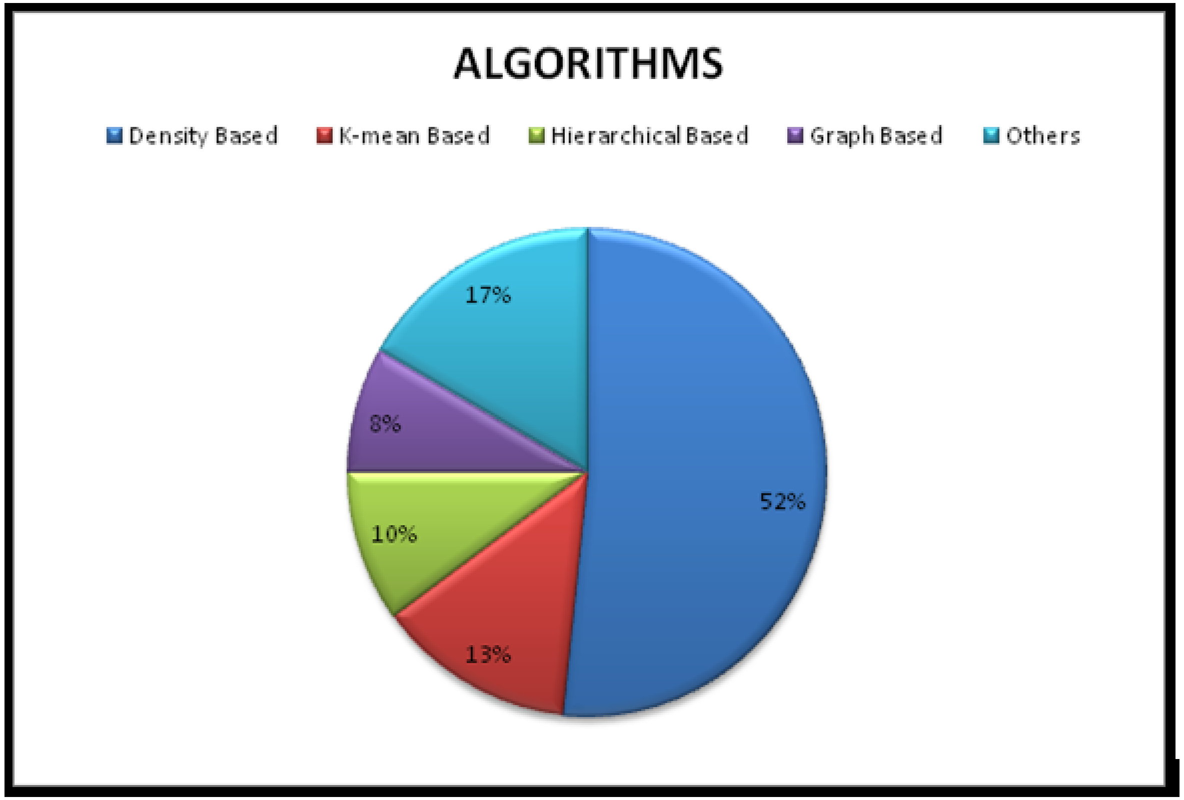


Figure 4.2 Percentage of different clustering algorithms used in reviewed literature

Selected paper "Density-based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Geo referenced Documents'' (DSC) scientific paper published by Keiichi Tamura and Takumi Ichimura in Oct 2013.

Reason to choose the variant of DBSCAN.

- Text based spatial temporal dataset .
- It can handle noisy data
- Can reveal arbitrary shape clusters.
- Prior knowledge of clusters is not needed e.g. K-Means.
- Suitable for different type of data.
- Simplicity.

Difference between DSC algorithm and (ϵ, k, t)-DBSCAN algorithm

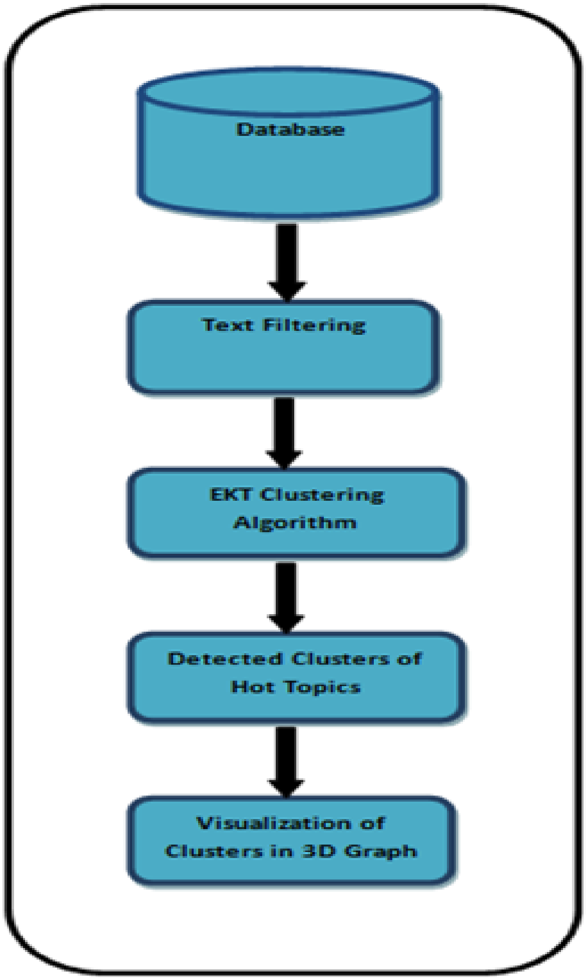
DSC Algorithm

- Extracts spatial and temporal clusters.
- Parameters: radius(ϵ) , timestamp (t) and threshold value (min doc)

(ϵ, k, t)-DBSCAN

- Extracts semantically similar, spatial and temporal clusters.
- Semantically similarity - cosine similarity k between 2 documents.
- MinDoc is enhanced to $\text{MinDoc}_{\text{DifferentUsers}}$

Workflow of (є, k, t)-DBSCAN



Experiment

Input Parameter	Various Values Considered					
Radius (km)	0.7	1	2	3		
Cosine Similarity	0.55	0.60	0.65	0.70	0.75	0.80
Time (Hours)	24 (converted in to seconds)	48(converted in to seconds)	72 (converted in to seconds)			
Minimum Number of documents of different users	6	7	10			

- 216 combinations of above parameter were verified through cross validation in weka tool.

- Finally chosen input parameter for (ϵ,k,t)-DBSCAN

  radius (ϵ) = 2 km

  Inter arrival time (t) = 48 hrs (172800 in seconds),

  cosine similarity (k) = 0.70 and

  minimum number of documents of different users = 10

Hardware used in this setup was as below:

Processor :Intel Core i7

RAM: :4GB and

Operating system :Linux Fedora 20.

Language :Python 2.7.8

Weka result on different parameters

T1 (One day time arrival = 86400 seconds)					Classified		Average Weight		
S.No.	Radius	Cosine Similarity	Minimun User	InterarrivalTime	Correctly	Incorrectly	Precision	Recall	F-Meaure
1	0.7	0.6	10	One day	78.38	21.62	0.81	0.82	0.8
2	2	0.6	7	One day	77.45	22.55	0.8	0.76	0.76
3	0.7	0.55	6	One day	77.3	22.7	0.82	0.77	0.74
4	0.7	0.6	7	One day	76.77	23.23	0.78	0.77	0.76
5	0.7	0.8	10	One day	76.48	23.52	0.77	0.77	0.77

T2 (Two days time arrival = 172800 seconds)					Classified		Average Weight		
S.No.	Radius	Cosine Similarity	Minimun User	InterarrivalTime	Correctly	Incorrectly	Precision	Recall	F-Meaure
1	2	0.7	10	2days	82.83	17.17	0.87	0.83	0.81
2	3	0.6	7	2days	80.93	19.07	0.79	0.81	0.78
3	0.7	0.7	7	2days	80.87	19.13	0.79	0.8	0.79
4	3	0.55	10	2days	80.63	19.38	0.8	0.81	0.79
5	2	0.55	10	2days	80.59	19.41	0.82	0.81	0.79

T3 (Three days time arrival = 259200 seconds)					Classified		Average Weight		
S.No.	Radius	Cosine Similarity	Minimun User	InterarrivalTime	Correctly	Incorrectly	Precision	Recall	F-Meaure
1	2	0.55	10	3days	80.96	19.04	0.82	0.83	0.81
2	2	0.55	7	3days	80.87	19.13	0.8	0.81	0.77
3	3	0.55	10	3days	80.57	19.43	0.77	0.81	0.77
4	1	0.65	6	3days	79.76	20.24	0.84	0.8	0.78
5	3	0.6	10	3days	79.56	20.44	0.79	0.8	0.75

Comparison of (є, k, t)-DBSCAN with DBSCAN

The parameter of (є,k,t)-DBSCAN algorithm

radius (є = 2 km),

t = 48 hrs (172800 in seconds),

Cosine similarity (k) =.70 and

minimum number of different users = 10

The parameter of DBSCAN algorithm

radius (є) = 2 km,

minimum doc = 10

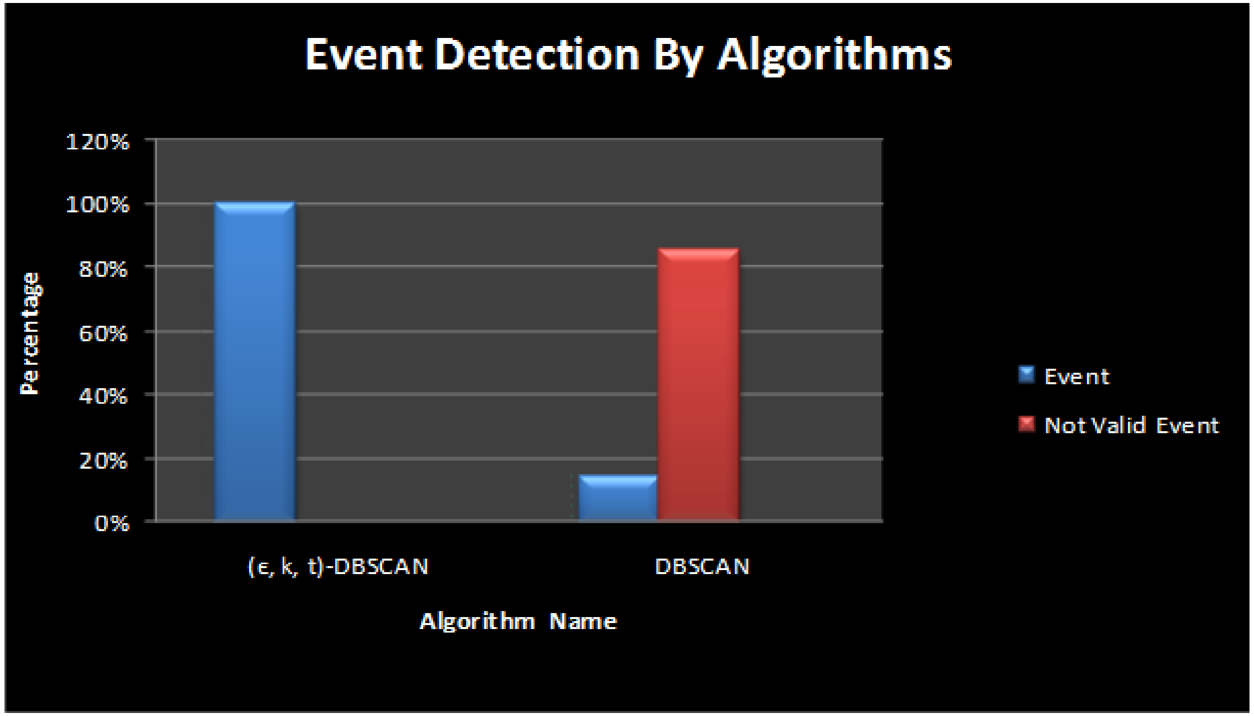
Algorithm Name	Correctly Classified Instances (%)	Incorrectly Classified Instances (%)	Precision (0-1)	Recall (0-1)	F-Measure (0-1)
(є, k, t)-DBSCAN	82.83	17.17	0.87	0.83	0.81
DBSCAN	41.7	58.3	0.57	0.42	0.47

Comparison of (є, k, t)-DBSCAN and DBSCAN results

Cluster result discussion

Rank	Cluster Number	No of Tweets	Range of latitude	Range of longititude	Time	Top 5 Frequent words	Hot topic description (real world event)
1	Cluster 3	29	48.13035 - 48.15	11.54916667 - 11.5833	2014-09-28 17:30:41 - 2014-10-05 10:04:21	oktoberfest, germany, münchen. fidefesta, control	Oktoberfest
2	Cluster 1	23	48.21878263 - 48.21895031	11.62456288 - 11.62466168	2014-09-17 17:32:55 - 2014-09-17 21:25:41	arena, allianz, manchester, bayern, münchen	Bayern München Vs Manchester match at allianz arena
3	Cluster 5	18	48.21878263 - 48.21878263	11.62466168 - 11.62466168	2014-11-05 16:24:34 - 2014-11-05 21:54:11	bayern, münchen, roma, arena, allianz,	Bayern München Vs Roma match at allianz area
4	Cluster 4	15	48.21878263 - 48.21878263	11.62466168 - 11.62466168	2014-10-04 13:37:25 - 2014-10-04 15:30:54	bayern, münchen, arena, allianz, hannover	Bayern München Vs Hannover match at allianz area
5	Cluster 2	14	48.21878263 - 48.21878263	11.62466168 - 11.62466168	2014-09-23 18:39:50 - 2014-09-23 21:18:20	bayern, münchen, arena, allianz, Paderborn,	Bayern München Vs Paderborn match at allianz area

Figure : (є, k, t)-DBSCAN cluster results



Event detection By Algorithms

Visualization methods

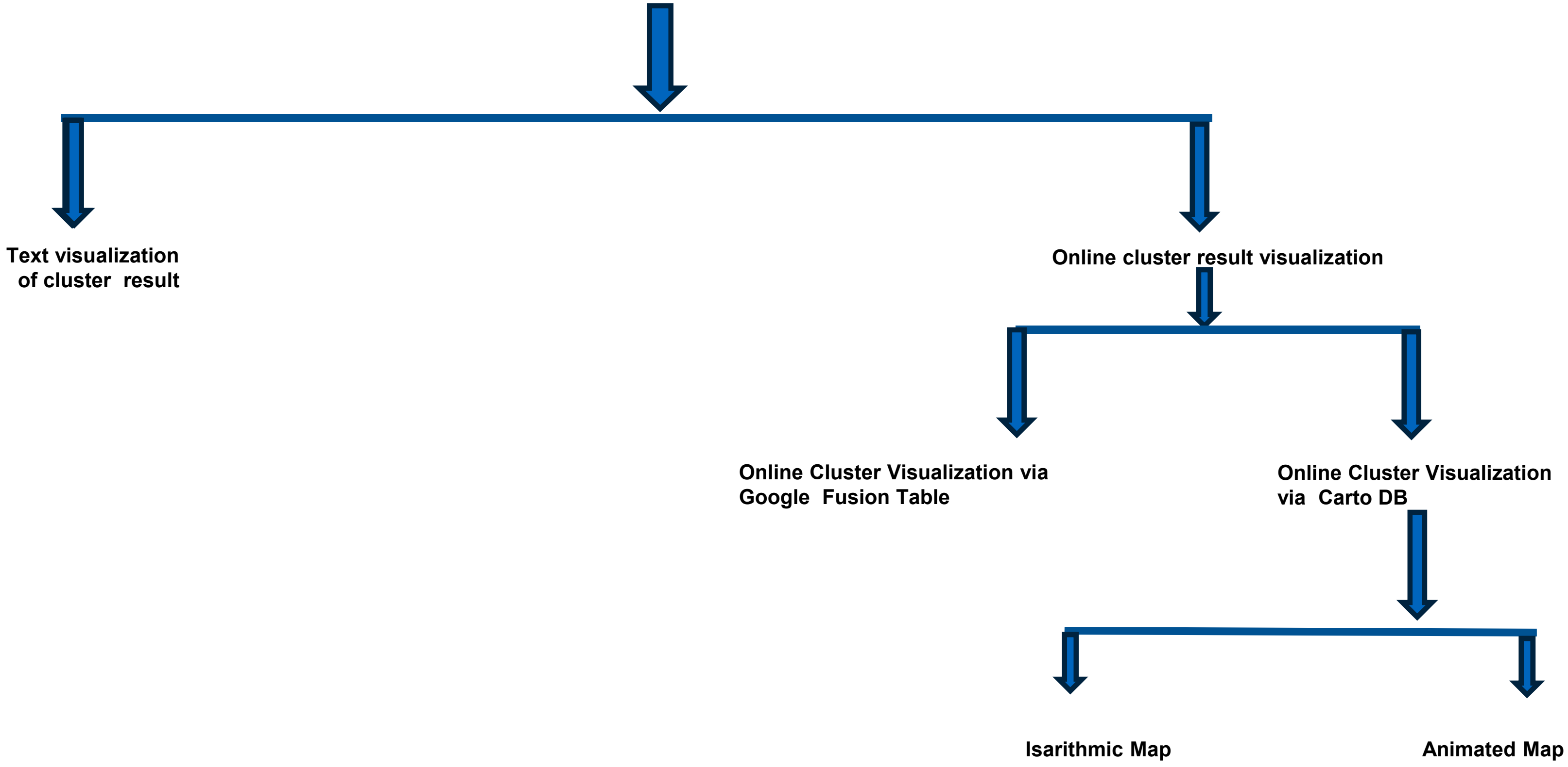


Figure: (Left side) Text visualization of (ϵ, k, t)-DBSCAN with number of counts
(Right side)Text visualization DBSCAN with number of counts

Google Fusion Table

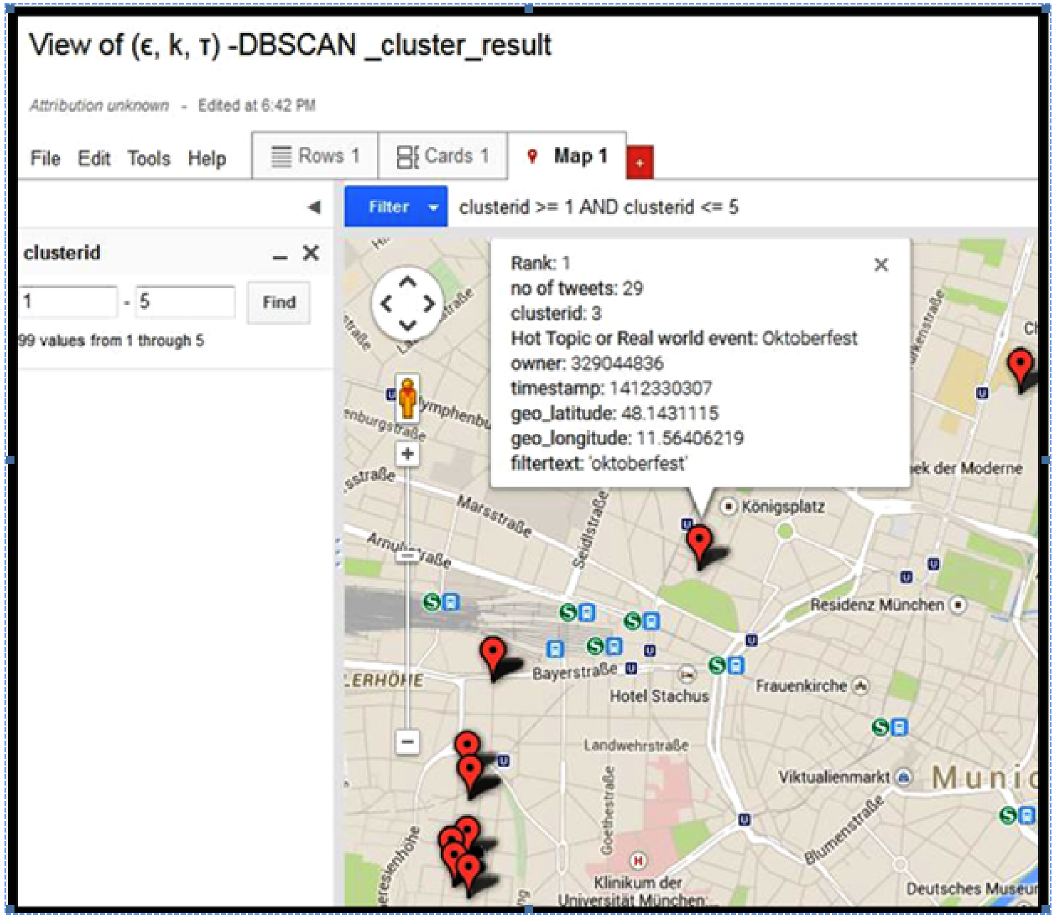


Figure: Screen shot of (ϵ, k, t)-DBSCAN result on Google Map

Isarithmic Map

The maps that represent data sets that have a “continuous distribution and smooth change in value” Kraak et al.(2003).

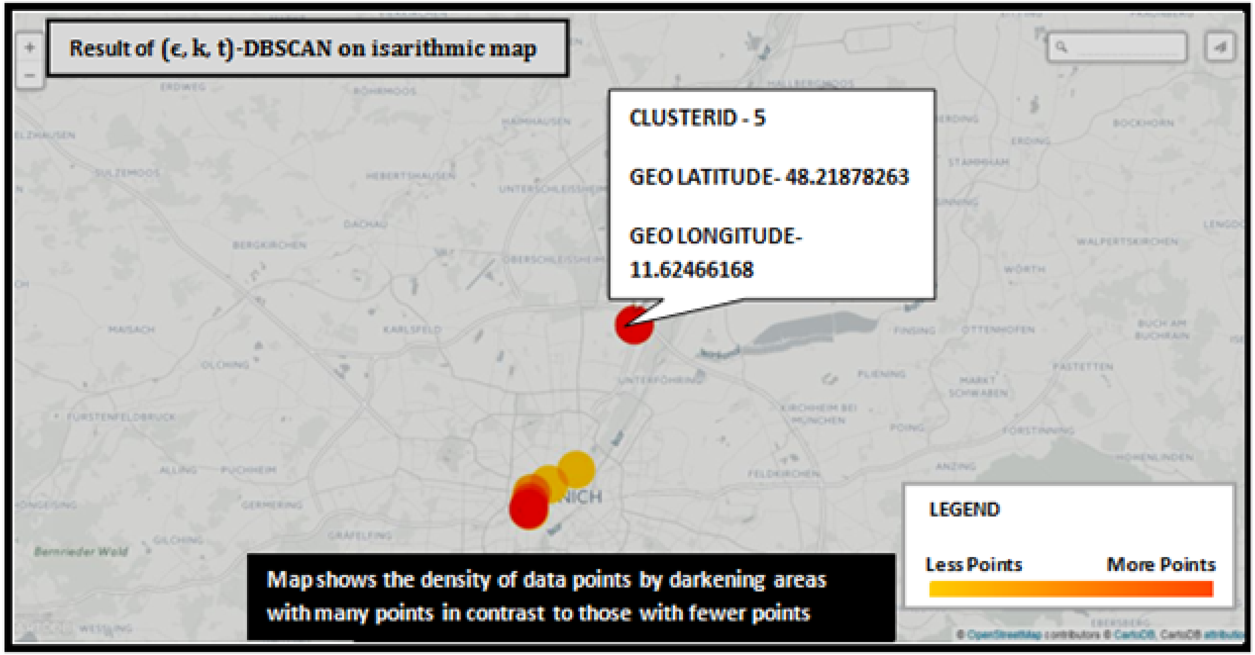


Figure: Screenshot of (ϵ, k, t)-DBSCAN Isarithmic map on CartoDB

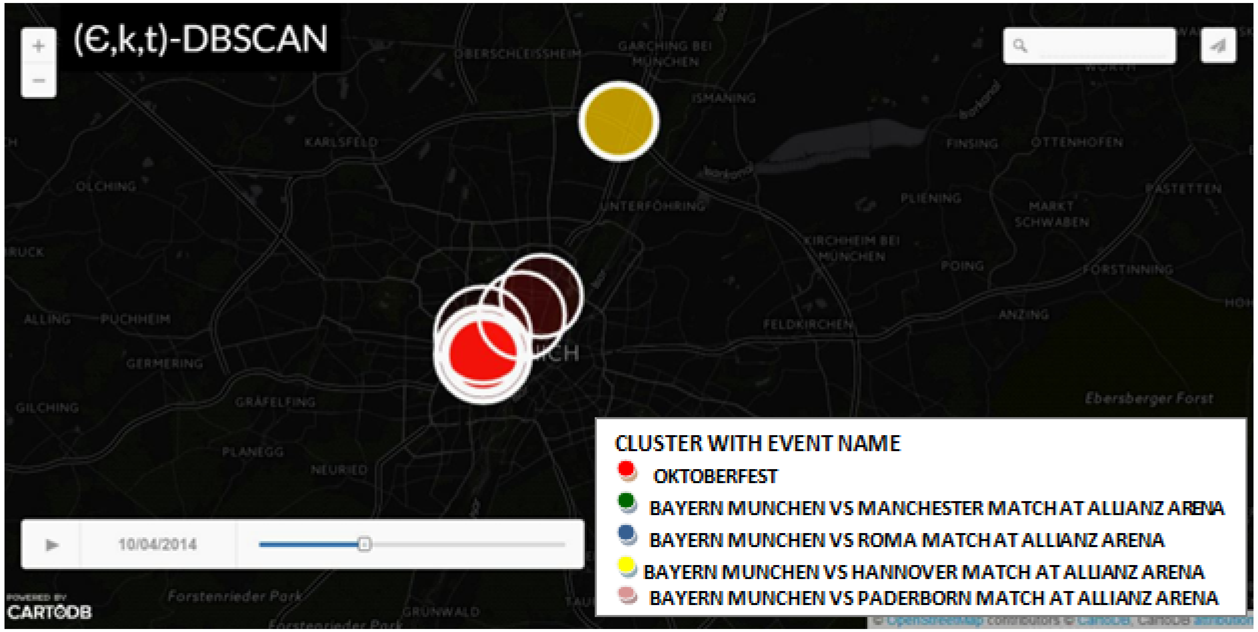


Figure: Screenshot of (ϵ, k, t)-DBSCAN result on Animated map.

Summary

- The proposed algorithm is able to reveal all the events from the dataset .The input parameters have a decisive impact on the cluster result.

- Suitable for any text based social media dataset.

- Real time data downloading capability can be added to the framework as future research.

- Algorithm speed can be enhanced in future research by using Ball Tree or KDTree nearest neighbors learning algorithms instead of brute force algorithm.

References

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

•Chen, L., & Roy, A. (2009). Event detection from flickr data through wavelet-based spatial analysis. In Proceedings of the 18th ACM conference on Information and knowledge management (pp. 523-532). ACM.

• Kraak, m. J. & Ormeling, F. (2003). Cartography: Visualization of Geospatial Data, 2nd ed., Harlow, England: Prentice Hall.

•Kisilevich, S., Mansmann, F., & 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.

•Reuter, T., Cimiano, P., Drumond, L., Buza, K., & Schmidt Thieme, L. (2011). Scalable Event-Based Clustering of Social Media Via Record Linkage Techniques. In ICWSM.

•Tamura, K., & Ichimura, T. (2013). Density-Based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Georeferenced Documents. In Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference on (pp. 2079-2084). IEEE.

Thank you !
Any Question ?