Technische Universität München
Faculty of Civil, Geo and Environmental Engineering
Department of Cartography
Prof. Dr.-Ing. Liqiu Meng

Master Thesis

Spatial Temporal Analysis of Social Media Data

Submitted by

Smita Singh

Course of study: Master of Science in Cartography

Date of Submission: 25.02.2015

Supervisors: Dr.-Ing. Christian Murphy
Khatereh Polous

Acknowledgements

I would like to express my gratitude to my Supervisors Dr.-Ing. Christian Murphy and Khatereh polous for their guidance in conducting scientific research and for their time and the patience they showed during our brainstorming sessions.

My gratitude goes to Dr. - Ing. Stefan Peters, for his precious help and support during Master's course, not only in matters of studies but also with administrative issues. My gratitude also goes to Juliane Crone, the coordinator of this Master's program,

A special thank you goes to my family for their constant encouragement and support and without the help of whom I could not have pursued this Master's degree.

Munich, 25 Feb 2015 Smita Singh

Abstract

Social media is a platform where people are sharing their views, ideas, sentiments, and emotions. The popularity of social media has been growing over the past few years. Extracting and interpreting information from user-generated content is a trending topic in the scientific community and the business world. Numerous web applications that deal with processing and the visualization of user-generated content have proved the importance of spatial-temporal data.

In this thesis, the (ϵ, k, t)-density-based spatial temporal clustering algorithm is proposed for extracting local hot topics discussed among the social media users in georeferenced documents. The (ϵ, k, t) neighborhood of geo-referenced documents is defined to extract semantically similar spatial and temporally separated clusters. Real world events are manually verified from the detected local hot topics.

The (ϵ, k, t)-density based spatial temporal clustering algorithm is an extension of the DSC algorithm from Tamura et. al. (2013). From this algorithm, the existing parameters spatial radius 'ϵ', timestamp 't' and minimum number of documents (MinDoc) are extended by a new dimension. The cosine similarity constant 'k' is added as an additional criteria to the algorithm in order to find new clusters. The cosine similarity concept is used to compare the similarity between two text sentences. This new dimension 'k' helps localizing the semantically similar highly discussed local hot topics among different social media users, which are located in a particular small radius of geographical area and time. In addition, the definition of MinDoc is changed to the minimum number of documents of different users ($MinDoc_{DifferentUsers}$), which has a significant impact to get more meaningful cluster results. The input parameters of the proposed algorithm are configurable by the user in order to receive refined clusters of different local hot topics under discussion among social media users. The detected local hot topics are then visualized in 3D-scatter diagram.

In this thesis, an experiment is done on geo-tagged tweets from Twitter from the Munich area recorded during 9 weeks. For the validation of the clustering results, the data mining tool WEKA is used. For benchmarking of the proposed algorithm, the clustering result is compared with the base Density-based spatial clustering of applications with noise (DBSCAN) algorithm's cluster result. The comparison shows that the proposed (ϵ, k, t)-density-based spatial temporal clustering algorithm produces very promising results in comparison to DBSCAN. The three promising results are: (1) It is able to reveal all the events from the datasets on the bases of user defined algorithm input parameters. The input parameters have a decisive impact on the cluster result. (2**)** It can extract spatial, temporal and semantically separated clusters. (3) It is suitable for any text based social

media dataset to reveal the local hot topics and further revealing the events. Certain extra preprocessing might be required for some input datasets other than Twitter and Instagram in order to remove the noise.

The detected local hot topics discussed among the social media users are visualized using a 3D-scatter diagram, text visualization, Google Maps and the CartoDB online tool.

Contents

List of Figures

List of Tables

List of Photos and logos

Social media photo https://media.licdn.com/mpr/mpr/p/7/005/078/3cd/1e8a48b.jpg

Twitter image https://about.twitter.com/sites/all/themes/gazebo/img/ios_homescreen_icon.png

Flickr image http://www.vallistic.gr/media/images/transparent-flickr-logo-icon.png

Instagram image http://scottkleinberg.com/wp-content/uploads/2014/03/instagram-logo-kgo.png

Facebook image http://www.stepaheadinc.com/wp-content/uploads/2011/02/Facebook.jpg

LinkedIn http://upload.wikimedia.org/wikipedia/commons/c/ca/LinkedIn_logo_initials.png

Four square https://playfoursquare.s3.amazonaws.com/press/2014/foursquare-logomark.png

World map http://isghd.com/post/blank-political-world-map-background-1-hd-wallpaper.html

TUM logo http://www.lrr.in.tum.de/~grafs/tum-lbl.gif

Civil engineering Department logo http://www.bgu.tum.de/fileadmin/w00blj/www/_migrated_pics/ou_02.jpg

Cartography Department logo http://www.lfk.bgu.tum.de/fileadmin/w00bti/layout/LFK_Logo.jpg

List of Abbreviations

API	Application Programming Interface
BIC	Bayesian Information Criterion
BIRCH	Balance Iterative Reducing and Clustering using Hierarchies
CB-SMoT	Clustering Based Stops and Moves of Trajectories
CLARA	Clustering Large Applications
CLARANS	Clustering Large Applications based on Randomized Search
CLIQUE	Clustering in Quest
CURE	Cluster Using Representatives
DBCLASD	Distribution Based Clustering Algorithm for Mining in Large Spatial Databases
DBLP	Data Base Systems and Logic Programming
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DCPGS	Density-based Clustering Places in Geo-Social Networks
DEN stream	Density-Based Clustering over an Evolving Data Stream with Noise
DENCLUE	Density-Based Clustering
DSC	Density-Based Spatial Temporal Clustering
DVDBSCAN	Density Variation Based Spatial Clustering of Applications with Noise
(ϵ, k, t)-DBSCAN	Density based Spatial Temporal Clustering Algorithm (where ϵ = Distance, k=Cosine similarity rate constant and t=Inter arrival time)
EM	Expectation Maximization
ET	Events from Tweets
F-DBSCAN	Fast DBSCAN
GAC	Group Average Clustering

G-DBSCAN	Generalized Density based algorithm
HTML	Hyper Text Markup Language
HTTP	Hypertext Transfer Protocol
KML	Keyhole Markup Language
L-DBSCAN	Local-Density Based Spatial Clustering Algorithm with Noise
LTT	Location Time constrained Topic
LOF	Local Outlier Factor
ME-DBSCAN	Memory Effect -DBSCAN
NMI	Normalized Mutual Information
ODBSCAN	An Optimized Density-Based Clustering Algorithm
OPTICS	Ordering Points to Identify the Clustering Structure
PAM	Partitioning Around Mediods
SCAN	Structure Clustering Algorithm for Networks
SED-RHOCC	Social Event Detection with Robust High-Order Co-Clustering
SHC	Similarity Histogram-based Clustering Method
SMM	Social Media Monitoring
SMoT	Stops and Moves of Trajectories
SOM	Self Organizing Map
ST GRID	Spatial Temporal Grid
ST_DBSCAN	Spatial–Temporal Density-Based Spatial Clustering Algorithm with Noise
STING	Statistical Information Grid
TF-IDF	Term Frequency–Inverse Document Frequency

TF-OPTICS	Time Focused version of OPTICS
T-OPTICS	Trajectory-OPTICS
t-SNE	t-distributed Stochastic Neighbor Embedding
VDBSCAN	Varied Density-Based Spatial Clustering of Application with Noise
XML	Extensible Markup Language

1 Introduction

1.1 Introduction and Background

Extracting and interpreting information from user generated content is a trending topic in the scientific community and the business world. Among user generated information, spatial-temporal data have a greater value. This is proved by the numerous web applications that deal with processing and visualization of user-generated content.

The rapid development of social networks has enticed much attention all over the world. This paradigm has attracted the attention of researchers that wish to study the corresponding social and technological problems. Social media is a platform where people are sharing views and ideas, sentiments and emotions. Facebook, Twitter, YouTube, Instagram, Foursquare, LinkedIn, etc. are popular examples of social media.

Social media data are readily available through application programming interfaces (API), which motivates researchers to explore data streams that help to look inside the trend of data. The use of a reasonable clustering algorithm to find events is challenging due to the complexity of clustering algorithms that require a broad knowledge of data mining and data analysis. This master thesis focuses on the spatial and temporal analysis of social media data for event detection to visualize them for further exploration.

1.2 Purpose and Motivation

The purpose of this thesis was to get an insight into social media data to detect any type of significant changes named as “Event or topic under discussion among users” in the data set. According to (Polous et al. 2013) an event may be defined as any anomalous user activity, which happened at a time or within a particular period at a particular location. Local hot topics can be perceived as a superset, which consists of topics under discussion among social media users which are classified as events and others which are not classified as events. For example topics under discussion at a real world event, e.g. job fair, football match, music concert, festival are classified as events, while topics like weather discussions are not likely to be classified as events. To achieve the thesis´s objective a literature review related to the topic is done and the DSC clustering algorithm (Tamura et al. 2013) is selected for further modification and implementation. The modified (ϵ, k, t)-DBSCAN clustering algorithm (section 6.4) is used to detect the local hot topics under discussion among users on the real world social media data set.

Besides the research of finding an optimal clustering algorithm to achieve the objective, a personal motivation was also to discover what people are talking about. Such information might be useful for many applications for example marketing, advertisement, news etc.

1.3 Workflow of the Thesis

Following are the workflow steps to achieve the thesis objective of detecting the local hot topics under discussion among social media users from a real world dataset.

Figure 1 shows the workflow of this research work. It is divided into 6 levels from work point of view:

- First step was to identify the existing clustering algorithms for spatial temporal datasets that can detect the local hot topics under discussion among social media users from the real world dataset. To achieve the above described objective, the theoretical literature review was done from various scientific research papers of similar domain (clustering algorithm). Subsequently, the best suitable research papers were selected. The selection of the algorithm was based on the advantages and disadvantages of different investigated algorithms.

- In the second step, the chosen algorithm was optimized to bring it in line with the thesis's objective for receiving a better clustering result in order to detect local hot topics under discussion among users on the chosen social media platform. The optimization was primarily done in the algorithm's functionality and its input parameters. (More details are available in Chapter 6 for this topic.).

- In the third step, the optimized algorithm is implemented in Python language. The proposed algorithm is an extension of the Tamura et al (2013) base algorithm, where another new dimension is introduced. The cosine similarity constant 'k' is added as an input to the algorithm along with the change of the parameter "MinDoc" to get the desired result.

- In fourth step, the "local hot topics under discussion among social media users" are extracted automatically by the implemented framework from of the dataset as a clustering algorithm output. After that, the events are detected from the extracted local hot topics.

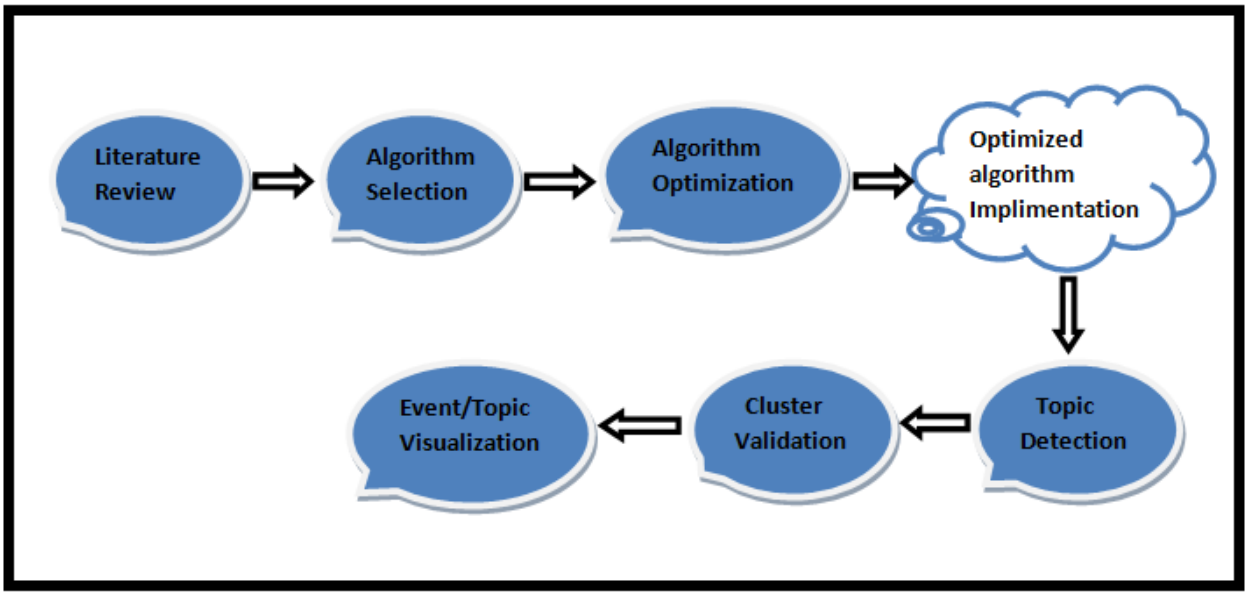


Figure 1 Workflow of master thesis

- The fifth step involves the cross cluster validation. This is accomplished via the WEKA machine learning tool whose input is the cluster result generated by the proposed clustering algorithm.

- The last step involves the visualization of the detected events.

At the beginning of the master thesis, the following research questions were identified:

• Which clustering methods are available and which ones are the most suitable for event clustering?

• Can a suitable algorithm be identified for extracting local hot topics by reviewing scientific research literature?

• How can the main local hot topics be extracted from spatial temporal data?

• How to evaluate the clustering result?

• How to visualize spatio-temporal event clusters in a way that will make them easy to understand for the user?

1.4 Thesis Outline

This introduction is followed by **Chapter 2 Social Media**, which gives a general overview of popular social media platforms.

Chapter 3 Overview of Clustering Methods, provides an overview of main clustering algorithms which play an important role to see inside the social media dataset. This section also includes the density based clustering algorithm which will provide the information about the existing extensions of DBSCAN based algorithms through charts.

Chapter 4 Spatial temporal, Event detection and Social media, provides a broad insight of the literature review and lists scientific papers which were selected for the literature review as well as the motivation to choose the "research paper by Tamura et al. (2013).

Chapter 5 (ϵ, τ) Density Based Spatial Temporal Clustering (DSC), provides the details of the Tamura et al. (2013) research paper.

Chapter 6 Proposed (ϵ, k, t)-density-based spatiotemporal clustering algorithm, provides the details, the workflow and implementation aspects of the proposed clustering algorithm. At the end of this chapter, the validation result of the proposed algorithm is included along with a comparison to the DBSCAN algorithm's results.

Chapter 7 Cluster result discussion and visualization, provides the details of local hot topic clustering results and visualization.

Finally, **Chapter 8 Conclusions and future work,** concludes the thesis with a summary, a description of the encountered problems during algorithm implementation and suggestions for the future research work.

2 Social Media

Social media is as one of the most important phenomena in 21 century provides an opportunity of data sharing freely for public in their daily lives. It is getting popularity day by day with the easy access of internet on our laptops, desktops, and mobile devices. Social media is a platform on which people are sharing their thoughts, messages, videos, Images through a social network. Main advantages of social media platforms are that people can enhance the latest political, social news, and updates about their friends, relatives.

According to (Ahlqvist et al., 2010) Social media is the interaction among people in which they create, share or exchange information and ideas in virtual communities and networks. (Kaplan et al., 2010) define social media as "a group of Internet-based applications that build on the ideological and technological foundations of Web 2.0, and that allow the creation and exchange of user-generated content ." Social media is an important platform to enhance the growth of business, to promote the new product among the people and also to know the review of product from the people, to make the better relation between government & public and for recruiters who are searching talented people according to job requirement. As shown in below figure 2 also, Facebook, Twitter, Flickr, YouTube, Instagram, Foursquare, Google+ and LinkedIn are popular example of social media.

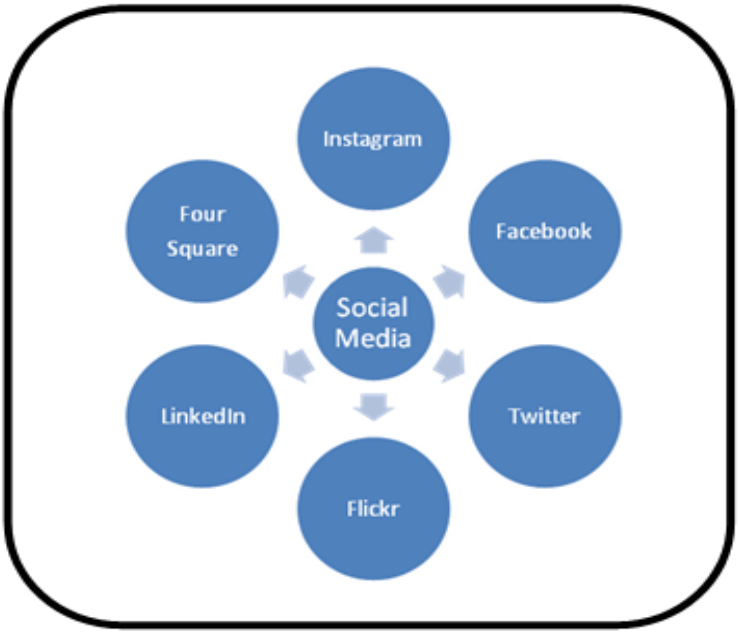


Figure 2 Famous Social media sites

2.1 Twitter

Twitter  is a social networking website launched in 2006, and it has been continuously gaining popularity ever since (Lee et al., 2011). Twitter allows its registered users to post short messages also called “Tweets Messages.” Tweets are constrained to 140 characters and can only include text. However, files like photos or videos can be also added as a URL and usually to save space, URLs are made tiny URLs. Each user creates its network of contacts by following the accounts s/he is interested in (might be friends, family, public persons or institutions that have a Twitter account). Basically following means subscribing to another user's Tweets or updates. By default, Twitter accounts are public and for this reason anyone, not necessarily only followers, can see a user's post. However, some users choose to set their posts as private, and only their followers can see their posts, in this case (Twitter, 2013b[1]). Posts from private accounts will also not be considered in the search result list. Users can reply to other users' posts by using the reply function or by beginning the Tweet with the user's id preceded by "@". In this way, conversations can be carried out on Twitter. Users also have the option to send each other direct messages that are private and unsearchable. However, in order to receive or send a direct message to another user, both users have to follow each other. Direct messages are also restricted to 140 characters. According to statisticbrain[2], Twitter has 645,750,000 total number of active registered Twitter users. This site is available in more than 35 languages.

2.2 Flickr

Flickr  is a platform on which user can share their videos and photos. Flickr platform was created in 2004 and later in 2005 it was acquired by Yahoo. In Flickr, people can upload and download the photos by following some license regulations. According to 2013 report on The Verge[3], Flickr had total 87 million registered users. This site is available in 10 different languages. According to Flickr website[4], Flickr offers three types of accounts: 1) Free 2) Ad Free and 3) Doublr account. By using the free account, a user has a limit to store their video and photos up to 1 terabyte and they can upload photos up to 200 MB. In a free account, video playback has only 3 minutes limit. Main advantage of the free account is the user has unlimited monthly bandwidth. If the User has Ad free account in that case s/he has to pay $49.99 per year or $5.99 per month and main advantage of this account is that your browsing will be ad-free. For Doublr account user has to pay $499 per year but they are getting more space for storage and all the benefits of the free account.

2.3 Instagram

Instagram  is another social media platform that was established by Kevin Systrom and Mike Krieger in 2010. In 2012, it was taken over by Facebook. By using Instagram, users can take the picture and share it in different social media platform. It is supported by Windows, IOS and Android. According to press news of Instagram[5] till 6 Sep 2013, Instagram had 150 million users. It is available in 25 different languages. It is freely available from Google play and Apple App Store. Users can share their photo either for public that means anybody can see the shared photos or privately by using private option, which means only people who follow you on Instagram will be able to see your photos.

2.4 Facebook

Facebook  is one of the famous social networking sites all over the world. It was established by Mark Zuckerberg and his four friends in 2004. Initially, the primary aim of Facebook was to connect the students of the University but after some time it allows to make an account to a people who are more than 13 years old. In Facebook, we can share photos, messages, videos. Through video call option users can connect with the other people from all over the world. Like button, Comment and share are main three options to express your view to other user's message. Like option is allowing an opportunity to user to express their appreciation on updates like photos, messages, and advertisement. Privacy setting is one of the important features of Facebook where people can select their own setting. For example, who can see their private stuff like photos or who will be able to see the friend list and who can write the text message and post the message on timeline. We can also block the particular Facebook accounts user if we do not want to share anything with them. This will restrict them from contacting us on Facebook. According to Scribd [6] as of the first quarter 2014, Facebook had 1.28 billion monthly active users. Facebook is available in more than 70 languages.

2.5 Linkedin

LinkedIn  is a professional social networking site that was launched in 2003. It is a platform where people can post their resume, recommend their friends, connect with people from same professional qualification and also search a job according to their professional qualification. LinkedIn has more than 300 million members from more than 200 countries. This site is available in 20 languages.

2.6 Four squares

Four squares  according to wikipedia[7] four squares is a mobile app that provides help to the user for personal searching. It was launched in 2009 by Dennis Crowley and Naveen Selvadurai. It is getting popularity day by day. Primary aim of Foursquare is to provide highly personalized recommendations of the best places to go around a user's current location. Since July 2014 user can also share their locations with friends using social networking layer. It is available in 12 different languages.

3 Overview of Clustering Methods

Clustering is a process in which we are trying to make groups of similar objects and these groups are representing a meaningful data. Grouping of an unknown dataset is the most challenging task in data mining as compared to classification. Clustering of any dataset needs profound domain knowledge and patience to get the good result. Now a day, clustering is the current theme for research area of Statistics, Business analysis, data mining and machine learning. Clustering is defined as "the task of assemble a set of data objects into numerous clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters" (Han et al., 2012.). There are various clustering algorithms available in the literature that can be used for very different purposes. Even similar algorithms with different configurations act very differently. According to (Mourya et al., 2013) clustering is an "iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure."

3.1 Partitioning Method

Partitioning algorithms are used to bifurcate the data into disjoint clusters. The most famous partitioning based algorithms are k -mean, k-medoid, k-mode and k –prototype. According to (Berkhin, 2006) there are two approaches to partition the data.

- Conceptual approach
- Objective function.

In conceptual point of view, clusters are identified with the help of predefined model and in objective function based partitioning approach either the pairwise computation of cluster or similarity-based relation between the clusters of dataset is considered. Main advantages and disadvantages of partitioning methods are discussed below

Advantages

- It is suitable for the dataset that includes the well separated compressed spherical clusters.
- It is a simple method.

Disadvantages

- User has to define number of clusters in advance.

- It is unable to deal with non-convex clusters with different sizes and density. It is very sensitive to noise and outliner.

3.1.1 K-mean

K-mean was proposed by **(**Macqueen. 1967**)**. It is the simplest unsupervised clustering algorithm. In k-mean, we assume the number of clusters (K) prior before partioning the data. It requires the user defined parameters: Number of clusters (K), Distance metrics and cluster initialization. Basic algorithm has some simple steps. First, we have to choose number of clusters as initial centroid, afterword it generates the number of clusters as a cluster center. In next step, it allocates each point to its nearby cluster center and again recomputed the center of each new cluster. This process will continue until some convergence criteria are met, in other words, until the centroid do not change. Fuzzy c mean (Dunn 1973), X-mean (Pelleg et al., 2000) and Kernel K-means (Schölkopf et al., 1998), K-prototype (HUANG, 1998) are some extension of k mean.

Advantages

- It is easy to understand.
- It gives good result when data is well separated.

Disadvantages

- We have to define number of clusters prior.
- It chooses center of the cluster randomly, which might not give positive results.
- It is applicable when mean value is defined.
- It is not a good choice for noisy data.
- It is Sensitive to the outliner.
- Final result always depends on the initial partition.

3.1.2 K-medoid

It is another important clustering algorithm based on partitioning. It was introduced by (Kaufman et al., 1987). In k-medoid algorithm, each cluster is represented by the most centric object (medoid) in the cluster. Medoids are more inflexible to noise and outliers as compared to centroids. The K-medoids algorithm has steps as follow: It starts with a random selection of objects as medoids for every k clusters then it assign each point to a cluster that is associated with cluster medoids. Afterward, it recalculates the k-medoids position. This process will continue until medoid becomes fixed. PAM, CLARA and CLARANS are main extensions of K-medoid.

Advantages

- Easy to implement and understand.
- It is less sensitive to the outliner as compared to k-mean.

Disadvantages

- It needs a prior knowledge about the number of cluster parameter.
- Final result and run time always depend on the initial partition.

3.2 Hierarchical Method

Primary aim of the hierarchical method is to demonstrate the cluster similarity into tree pattern that is also called dendrogram. The nested clusters in the dendrogram represent the clusters that are related to each other in dataset. There are mainly two types of algorithm of the hierarchical method: 1. Agglomerative method 2. Divisive method. Dendrogram can demonstrate both methods. Hierarchical clustering approach uses different restraint to decide locally which cluster should be merged at every step.

3.2.1 Agglomerative algorithms

According to (Jain et al., 1988), it is also known as bottom-up method. Agglomerative method considers each point as cluster, and it merges the point until we do not get the final desired cluster. Rock, BIRCH, Cure, CFT, Chameleon are main extension of agglomerative algorithm.

3.2.2 Divisive algorithms (top to down)

According to (Kaufman et al., 1990), it is opposite to agglomerative algorithm. In this method, all the points or objects are considered as part of only one cluster but further points are subdivided into a small cluster until we get the final desired result.

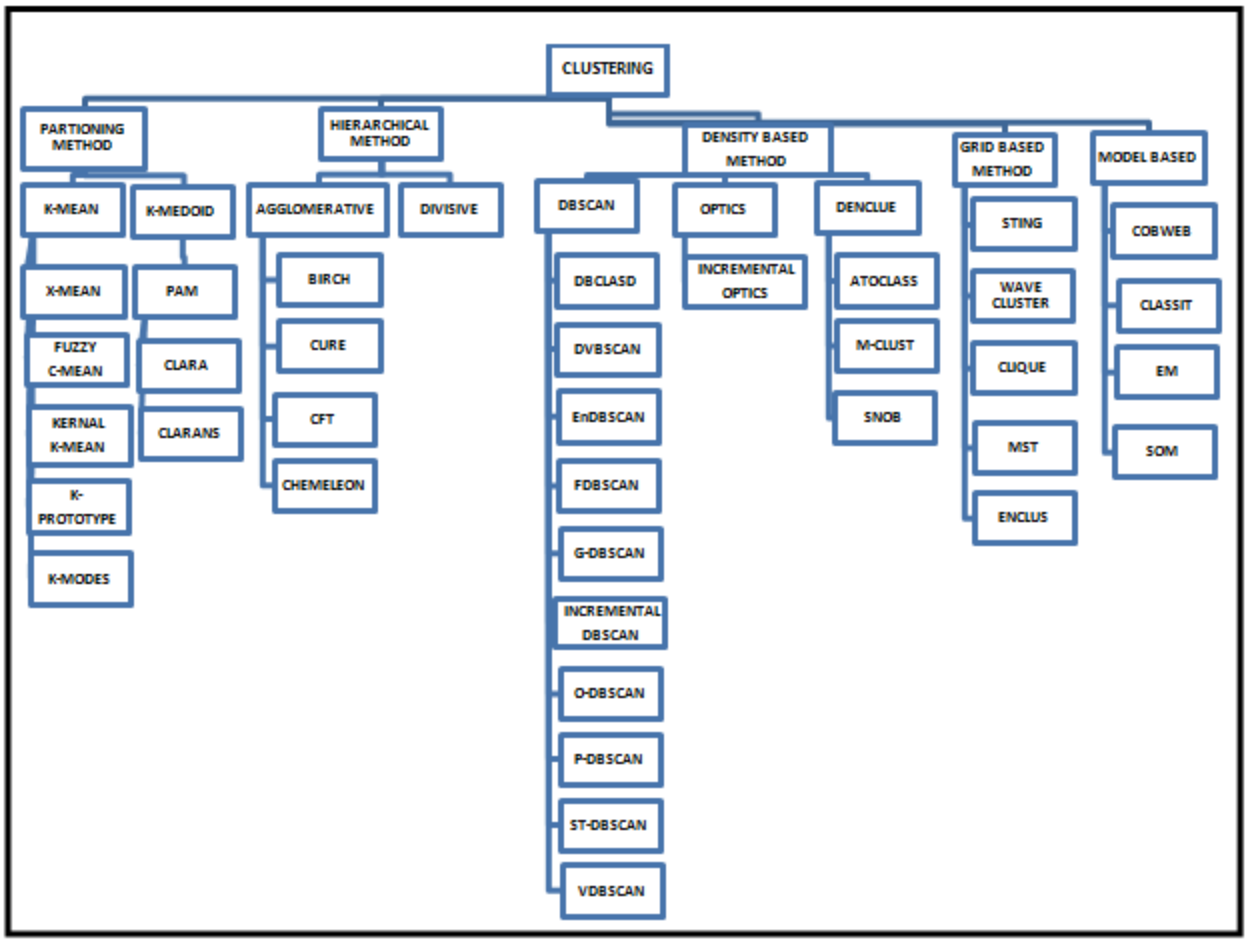


Figure 3 Overview of clustering methods

Advantages

- No prior information about the number of clusters required.
- Easy to implement and gives the best result in text categorization.
- It can find the cluster when data is in nested form.
- It is suitable for point linkage related problems.

Disadvantages

- It is hard to understand the number of clusters through dendrogram in case of complex dataset.
- It is not good choice for beginners.

3.3 Density Based Methods

Density based method finds clusters based on density in dataset. Central idea of density-based clustering is that clusters are surrounded by low dense cluster. The objects which are part of a cluster are placed close to each other within a certain range. In other words, in density based approach for each point of the cluster the neighborhood of given radius has to carry at least minimum number of points called minimum points. We do not need a prior knowledge of number of clusters in density-based algorithm. It can find arbitrary shape cluster in dataset. It can deal with noise or outliner of dataset. It is most appropriate method to find clusters in real time data. DBSCAN, OPTICS and DENCLUE are the main algorithms based on density.

Advantages

- It can recognize the arbitrary shaped clusters with different sizes.
- It can handle noisy dataset.

Disadvantages

- It is highly sensitive to the input parameter.
- It is not suitable for many dimensions dataset because of the curse of dimensionality phenomenon.

3.3.1 DBSCAN

DBSCAN (Ester et al., 1996) is one of the appropriate clustering algorithms that can cluster the data with noise. According to the (Ester et al., 1996) DBSCAN divides the data mainly in three classes. 1. Core point that is an internal part of the cluster. 2. Border points that are not core point. 3. Noise which is not consider as core and border point. Central idea of DBSCAN is that the densest objects within a certain range are viewed as a cluster and small dense area are called as noise of data. A group of central (core) objects with its overlapped neighborhood define the structure of cluster and non-core objects or nodes that are part of neighborhood of core objects represent the boundaries of the cluster and the rest are noise. DBSCAN requires two parameters Epsilon and Minimum Points. DVDBSCAN, VDBSCAN, ODBSCAN, EnDBSCAN, LDBSCAN, FDBSCAN, PDBSCAN and STDBSCAN are some main extensions of DBSCAN.

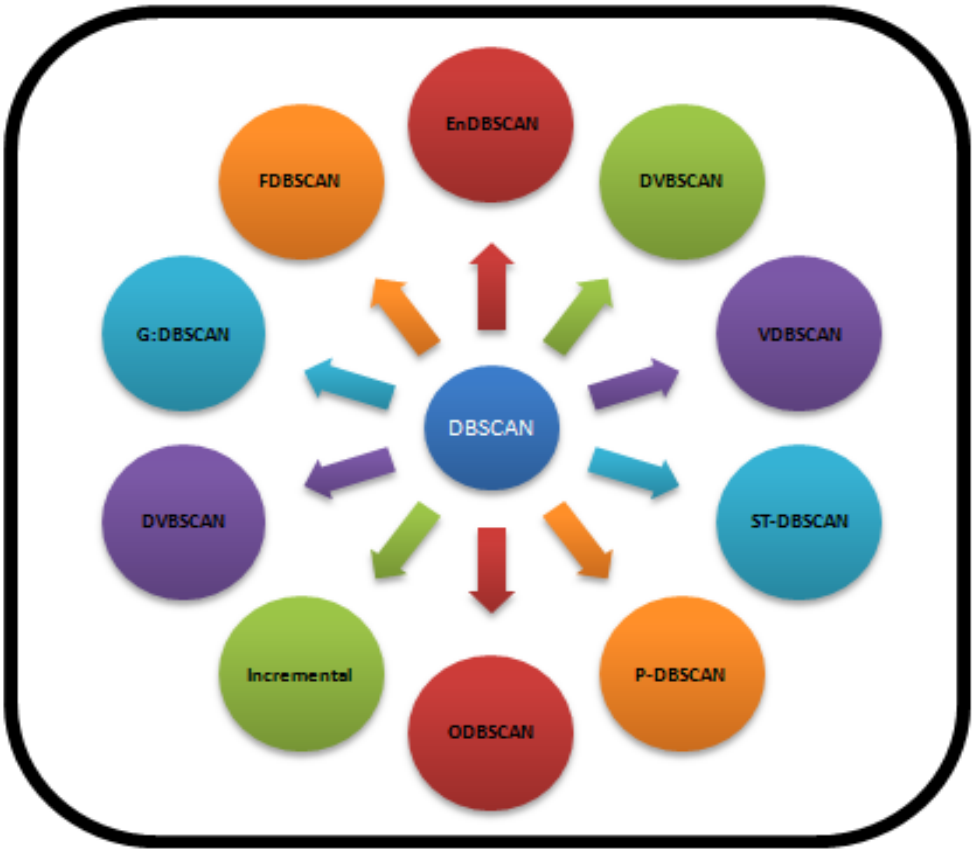


Figure 4 Extension of DBSCAN

Advantages

- DBSCAN can find arbitrary shape of clusters.
- It can handle dataset with noise.
- It requires two users' defined parameters.
- It does not require prior knowledge of number of clusters.

Disadvantages

- It is not efficient when we have different densities in same dataset.
- It requires many iterations to get a good result.
- Parameter selection needs profound knowledge of the subject.

3.3.2 OPTICS

OPTICS (Ankerst et al., 1999) algorithm finds the cluster in spatial dataset and it is a density-based approach. OPTICS uses different multiple parameters setting to discover clusters in various density level. In OPTICS, objects are arranged in linear form so points that are neighbor, can arrange next to each other in dataset with their spatial distance. Such type of arrangement of data helps to see the similar point in the cluster, which can represent by hierarchical method called dendrogram. Optics secures the quality of cluster by maintain the sequence in which data are processed for example it always gives high priority to most dense clusters as compared to less dense cluster. Optics always saves the two values of each processed point. One is reachability distance and another is core distance. Main advantage of OPTICS is that it is not limited to one parameter.

3.3.3 DENCLUE

DENCLUE (Hinneburg et al., 1998) is based on kernel density based estimation. According to (Hinneburg et al., 1998) on the basis of the mathematical function we can model each object of dataset and result can be obtained by the influence function. This method is based on following criteria: 1. by using a mathematical function each data point can be formally modeled 2. By calculating total sum of function which is applied to all data point, we can model the density of data space. 3. After that cluster can be identifying mathematically by recognize the density attractors.

Advantages

- It can scale arbitrarily shaped cluster.
- It can deal with noisy data.
- It is not sensitive to the data ordering.

Disadvantages

- It is highly sensitive to the input parameter.
- It allows a compact mathematical description of arbitrarily shaped clusters in high dimensional data sets.

3.4 Grid Based Method

Grid-based clustering algorithm is different from traditional clustering algorithms. According to(Han et al., 2012) ``It analysis's the object space into a defined number of cells that form a grid structure on which all of the processes for clustering are performed.'' Grid

Based clustering algorithm quantizes the clustering space into finite number of cell to make a skeleton of grid. Afterward, it performs all the operation on it. At last dense cells are connected to form the cluster that have more than certain number of cells. Sting, Wave cluster, Clinique, MST and ENCLUS are some example of grid-based algorithms.

Advantages

- It is a fast approach when we are concerned about processing time.
- It uses multi-resolution data structure.
- Easy to identify neighbor cluster.

Disadvantage

- Main disadvantage of grid-based algorithm is that it has limited shapes to make a structure of grid cells.

3.5 Model Based Methods

According to (Han et al., 2012) the primary aim of model-based clustering is endeavor to enhance the fitment into given data and some mathematical model. Such types of techniques are usually based on the presupposition that the data are produced by a combination of underlying probability distribution. Main advantage of the model-based clustering is that it can find characteristic of the group of objects. SOM, EM COBWEB and CLASSIT are examples of model-based algorithm.

3.5.1 Expectation Maximization (EM)

Expectation Maximization (EM) is famous k-mean based statistical based clustering algorithm. Central idea of EM is that it allocates each point to a cluster according to weight criteria, and new means are estimated on the basis of weight.

3.5.2 COBWEB

COBWEB is incremental clustering algorithm that is based on a hierarchical approach. (Fisher et al., 1987) Invented COBWEB. It does not require predefined number of clusters in advance, although it presumes that all the variables are independent. In COBWEB, clusters describe the probabilistically. Primary aim of COBWEB is to get a high prediction of nominal variable values within a given cluster. Main disadvantage of COBWEB is that it is not suitable for large dataset. CLASSIT is extension of COBWEB and it is incremental clustering algorithm. It can handle numeric and symbolic dataset.

3.5.3 SOM

SOM (Kohonen, 2001) Self Organizing Map is a model-based clustering method that uses neutral network approach. It is also known as SOFM and topological ordered map. SOM consist nodes with weight as input data vector. In SOM, neurons are arranged in two dimension lattices. It is useful to visualize high dimension dataset.

4 Spatial temporal, Event detection and Social Media

4.1 Literature review

Data mining is the technique to find concealed and fascinating pattern from dataset, which can be used in decision making and future prediction (Fayyad et al., 1999). Spatial temporal analysis of any social media platform mainly depends on the clustering method that is applied on the dataset to find a meaningful result. For literature review, the research papers were searched with relevant keywords like social media, clustering algorithm, real-time algorithm, algorithms for spatial temporal data, and event detection using a clustering algorithm. All papers were selected from well known publications and journals. Author has found many papers but only 60 were selected for the literature review, which were written in English language. Summary of the selected research papers are shown in below table 1 which is mainly focused on the algorithm used, its input parameters, dataset type on which the algorithm was applied.

1. **Ester et al. (1996)** proposed DBSCAN. In DBSCAN, most dense object within a certain range are considered as a cluster and low dense areas are considered as noise of the data. Radius (r) and minimum points (MinPts) are two parameters of DBSCAN algorithm, which are defined by the user. (Parimala et al., 2011) have defined the steps of DBSCAN in a very simple manner as follows (i) Choose a random point p (ii) Regain all points density-reachable from 'p' w.r.t. Eps and MinPts. (iii) If p is a central point, a cluster is formed. (iv) If p is a border point, no points are density reachable from p and DBSCAN visits the next point of the database. (v) Procedure continues until all the points are processed.

2. **Xu et al. (1998)** introduced DBCLASD (A distribution-based clustering algorithm for mining in large spatial databases) to overcome the problem of DBSCAN that user has to define two parameters, which in itself is a quite challenging task, when we have large spatial dataset. DBCLASD does not require any parameter. It is very effective in large dataset because it finds clusters of arbitrary shapes. A distribution-based clustering algorithm for mining in large spatial databases algorithm is based on the hypothesis that the points inside a cluster are uniformly distributed.

3. **Sander et al. (1998)** proposed generalized density-based algorithm (G-DBSCAN) which can cluster the core object based on spatial or non-spatial attribute of dataset. It requires two parameters. The experimental performance of G-DBSCAN is compared with BIRCH and CLARANS with visual inspection. Author has applied G-

DBSCAN to solve the real time problem related to biology, earth science, astronomy, and geography.

4. **Yang et al. (1998)** proposed an agglomerative clustering algorithm (GAC: augmented Group Average Clustering) to extract retrospective events from the news story to analyze the correlation between cluster quality and efficiency of computations, an iterative bucketing and re-clustering model was applied. Hierarchical and non-hierarchical document clustering algorithms were used to 15,836 stories to exploit their content and temporal information. The cluster hierarchies were the key to detecting unidentified events previously retrospectively, supporting both query free and query-driven retrievals. In addition, the temporal distribution of document clusters provided useful information to improve both retrospective detection and online detection of events.

5. **Shou et al. (2000)** invented F-DBSCAN to improve the performance and run time of DBSCAN. Fast DBSCAN is faster as compared to original DBSCAN. Runtime complexity and time calculation complexity are also far better than DBSCAN, which saves the expenditure of cost due to its fast performance, but main drawback is that it does not maintain the accuracy. According to the author F-DBSCAN, needs only a few objects representative of core object neighbor as starting point to enlarge the cluster.

6. **Hammouda et al. (2003)** proposed an incremental clustering using cluster similarity histogram. The key feature of this algorithm is pairwise document similarity. Main concept of SHC (similarity histogram-based clustering method) is to retain a high degree of consistency at any time. Experimental result shows that it requires less computation time to get a better cluster quality as compared to another clustering algorithm like Hierarchical Agglomerative Clustering, Single-Pass Clustering, and k-Nearest Neighbor Clustering.

7. **Roy et al. (2005)** introduced EnDBSCAN. It is a mixed approach of DBSCAN and Optics to overcome the stumbling block of DBSCAN and OPTICS. Algorithm can detect the nested or natural cluster structure in dataset with noise. En-DBSCAN has extended the idea of core distance of optics and initiates the idea of core neighborhood to solve the global parameter setting of density-based approach. Experimental result of EnDBSCAN showed that it can detect embedded and nested cluster, but it has same runtime complexity as DBSCAN and optics. According to

(Roy et al., 2005) it needs few Specifications and has fewer obstacles as compared to OPTICS.

8. **Singhal et al. (2005**) proposed methodology for clustering multivariate time series data. They have calculated similarity between multivariate time series data with batch fermentation algorithm that is based on two factors. First similarity factor is built on principal component analysis and the angles between the principal component subspaces and the second factor is constructed on Mahalanobis distance formula at dataset. The main advantage of this similarity factor with batch fermentation is effectual in clustering multivariate time series datasets and has preferable result to existing methodologies.

9. **Procopiuc et al. (2005)** proposed a new algorithm called local kernel that is based on KD- tree and can handle spatial data dream stream. They have used kernel estimator to compute the local statistics and in addition can handle the maintenance of local statistics.

10. **Zhong (2005)** explored an online Spherical K-mean algorithm for clustering high dimensional text data that applies the "Winner Take-All" competitive learning technique with the combination of annealing-type learning rate schedule. This algorithm requires a set of N unit-length vector data and number of predefined clusters. To achieve the speedy and modifying clustering result, they have combined online spherical k mean algorithm with an existing scalable clustering strategy. Experimental result shows that online spherical k-means algorithm can accomplish exceptionally better clustering results than the batch version.

11. **Viswanath et al. (2006)** proposed hybrid clustering method L-DBSCAN to overcome the complication of DBSCAN like run time when it applies on large dataset and also to get arbitrarily shaped clusters. According to the authors, they have used leaders clustering method to derived two levels of prototype. L-DBSCAN requires two user-defined parameters. Experimental result shows that when L-DBSCAN applied on dataset with suitable parameter it requires less time to find the cluster as compared to DBSCAN.

12. **Wang et al. (2006)** invented a new clustering approach which is combination of ST GRID (Spatial Temporal Grid) and ST DBSCAN (An algorithm for clustering spatial–temporal data that is based on DBSCAN) to find spatial temporal cluster from geo-database. Main advantage of this method is that we do not need to calculate the spatial and temporal distance. Experiment has done on seismic dataset.

13. **Cao et al. (2006)** proposed Den stream clustering algorithm that rely on density-based approach and find the arbitrary shape clusters in the data stream. They have defined clusters with random shape as dense or core micro cluster and also explained the difference between core and outliner cluster. Den stream is two-phase clustering algorithm. In online phase, it keeps the micro-clusters and in the offline mode it creates the cluster based on DBSCAN. Den stream requires four parameters that are defined by the user. The experimental results show the advantages and potency of Den Stream in searching clusters of arbitrary shape in data streams.

14. **Nanni et al. (2006)** proposed a density-based clustering approach for moving objects trajectories. According to authors, in this paper they tried to find out the answers to two central questions 1. What is the most suitable clustering algorithm for trajectories and how can we exploit the intrinsic semantics of the temporal dimension to improve the quality of trajectory clustering. For clustering, they have used density based T-OPTICS (Trajectory-OPTICS) and TF-OPTICS (Time Focused version of OPTICS) algorithms and to find out the solution of the second question they have considered time interval between trajectories. Experiment has been done on six different dataset to get the promising result.

15. **Liu et al. (2007) introduced** VDBSCAN (Varied Density-Based Spatial Clustering of Application with Noise) to overcome the problem of DBSCAN to find the relevant clusters if we have different densities in single dataset. User does not need to feed input parameter himself. Instead, VDBSCAN automatically chooses some values of the parameter for varied densities. According to the author, it is two steps algorithm: 1. Select parameter epsilon and 2. Cluster with diverse densities. VDBSCAN has same time complexity as DBSCAN.

16. **Birant et al. (2007)** proposed a spatial temporal density-based algorithm that is based on original DBSCAN. ST-DBSCAN can detect the cluster in both spatial and non-spatial attributes of dataset. It requires three user defined parameter to identify the cluster. Main advantage of ST-DBSCAN in contrast to DBSCAN is that it can detect noise in varied density by assign density factor to each cluster. This approach can be used in many applications such as geographic information systems, medical imaging and weather forecasting. Experimental result shows that ST-DBSCAN had very promising result when it was applied to spatial-temporal dataset to detect the cluster.

17. **Chen et al. (2007)** proposed density based approach call D- stream for clustering the DataStream. This framework is a combination of online and offline approach

that uses density-based algorithm to capture the dynamic data stream. Main advantage of D-Stream is that it is automatically and dynamically adjusts the clusters without requiring user specification of target time horizon and number of clusters. In addition, it can detect real-time clusters. Experimental results show that it is a fast and efficient algorithm to identify the clusters from real-time data stream.

18. **Duan et al. (2007)** invented new density-based algorithm called L-DBSCAN (Local-density based spatial clustering algorithm with noise) that is based on local density approach and also consider the advantages of LOF (Local Outlier Factor). It can find clusters with noise in the spatial database. It requires three user-defined parameters. Experimental results show that LDBSCAN can generate meaning clusters as compared to other clustering approaches.

19. **Palma et al. (2008)** proposed a new spatial-temporal clustering method for discovering interesting places in trajectories, which is based on speed rather than distance. It requires three parameters. This approach is a variation of DBSCAN, and in addition it is an implementation of SMoT (Stops and Moves of Trajectories) which is known as CB-SMoT (Clustering Based Stops and Moves of Trajectories). All the implementation had done on weka tool.

20. **Nosovskiy et al. (2008)** proposed a new cluster algorithm called ADACLUS (Adaptive Density-based Clustering algorithm) to find automatic arbitrary shaped clusters automatically, which is based on an introduced adaptive influence function. In some cases, it also gives an opportunity to define three parameters by users. The algorithm was applied to two-dimension dataset for evaluation and results were quite promising as compared to other clustering approaches.

21. **Mu et al. (2008)** proposed a parameterless density based approach that is based on nearest neighbor concept. It does not require predefined or user defined parameter via range scaling and the proportional criterion technique. Main advantage of the algorithm is that it can remove the noise around the clusters.

Serial No	Name of the Algorithm used in Paper	Input Number of Parameter	Varying Density	Geometry Shape	Data Type	References
1	DBSCAN	Two parameters (radius, minimum points)	No	Arbitrary shape	Spatial dataset with noise	Ester et al. (1996)
2	DBCLASD	Automatic	Yes	Arbitrary shape	Spatial dataset with equally distributed point	Xu et al. (1998)

3	G-DBSCAN	Two parameters	No	Arbitrary shape	Spatial Dataset	Sander et al. (1998)
4	Agglomerative and incremental cluster	Six parameters	Yes	Number of clusters shown in the histogram	TDT project dataset	Yang et al. (1998)
5	F-DBSCAN	Two parameters (radius and minimum points)	No	Arbitrary shape	Synthetic dataset and real data set	Shou et al.(2000)
6	Single pass incremental clustering algorithm	Automatic	No	Number of clusters	Two web document data sets	Hammouda et al. (2003)
7	En DBSCAN	Two parameters (radius and minimum points)	Yes	Arbitrary shape	Dataset with noise	Roy et al. (2005)
8	For clustering multivariate time-series data modified k- mean algorithm is used	Automatic	No	Number of clusters	The batch fermentation case study data	Singhal et al. (2005)
9	Kd-tree	Automatic	No	-	Synthetic dataset nm2 and real data	Procopiuc et al. (2005)
10	Online spherical k- mean for text clustering	Three parameters	No	Number of clusters	Twenty news groups dataset	Zhong (2005)
11	L-DBSCAN	Two parameters	No	Arbitrary shape	Synthetic and real dataset called Pen digits data from UCI machine learning repository with noise	Viswanath et al. (2006)
12	Density based spatial temporal clustering	Automatic	Yes	Number of clusters	Database of 'integrating seismic catalog in China	Wang et al. (2006)
13	Density based den stream algorithm	Four parameters	No	Arbitrary shape	Synthetic and real time both data set with noise	Cao et al. (2006)
14	Density-based T-optics and TF optics	Two parameters	Yes	Arbitrary shape	Six different real-time data set (four without noise, two with noise)	Nanni et al. (2006)
15	VDBSCAN	Automatic	Yes	Arbitrary shape	Spatial Dataset	Liu et al. (2007)
16	ST-DBSCAN	Three parameters	No	Arbitrary shape	Spatial temporal dataset	Birant et al. (2007)
17	D-Stream (density-based approach)	Automatic	No	Arbitrary shape	Synthetic dataset	Chen et al. (2007)
18	Density-based L-DBSCAN	Three parameters	No	Arbitrary shape	Dataset with 473 points with noise	Duan et al. (2007)
19	DBSCAN based CB-SMOT	Three parameters	No	Arbitrary shape	Trajectory data collected in the city of Amsterdam	Palma et al. (2008)

20	ADACLUS density based	Automatic	Yes	Arbitrary shape	Real world data set collected from European topic center on air and climate change	Nosovskiy et al. (2008)
21	Algorithm based on density model	Parameter free	No	Arbitrary shape	Synthetic datasets	Mu et al.(2008)
22	Single linkage (hierarchical based)	User defined (time, similarity)	Yes	Number of clusters	Mainichi news paper dataset	Sato et al. (2008)
23	Key graph algorithm	Two parameters	No	Graph	Dataset of live lab's social streams platform with noise	Sayyadi et al.(2009)
24	DBSCAN is used for clustering in text tag in Event Detection from Flickr Data through Wavelet-based Spatial Analysis	Two parameters	No	Number of clusters	Flickr photo data	Chen et al. (2009)
25	Single pass incremental clustering	Three parameters	No	Number of clusters	Dataset with Flickr photos	Becker et al. (2010)
26	An automatic topic detection method (TPIC) based on an incremental clustering algorithm	Automatic	No	Number of clusters	Standard corpora TDT-4 from the NIST TDT corpora	Zhang et al. (2010)
27	K-means clustering to classify the tweets	Three parameters	No	Number of clusters	Twitter data	Lee et al.(2010)
28	ODBSCAN	Radius, min pts and number of identical circle	No	Arbitrary shape	Two-dimensional synthetic dataset	Peter et al.(2010)
29	P-DBSACN	Two parameters	No	Arbitrary shape	Spatial dataset with noise	Kisilevich et al.(2010)
30	DVBSCAN	Two parameters	Yes	Arbitrary shape	Spatial Dataset with varied density	Ram et al. (2010)
31	K-mean algorithm with SVM	Five parameters	No	Number of clusters	Sina sport data	Li et al.(2010)
32	Probabilistic models based on Semantic	Automatic	No	Arbitrary shape	Spatial dataset with timestamp	Sakaki et al. (2010)
33	Incremental PRE-decon based on density based concept	Three parameters	Yes	Arbitrary shape	Real-time data set	Kriegel et al. (2011)
34	Incremental DBSCAN	Two parameters (radius and minimum points)	Yes	Arbitrary shape	Warehouse data (temporal)	Goyal et al. (2011)

35	Incremental clustering	Two parameters	No	Number of clusters	Flickr data	Wang et al. (2011)
36	Single linkage	User defined	Yes	Number of clusters	Dataset of Flickr photo	Reuter et al. (2011)
37	Framework used Incremental clustering algorithm with a threshold parameter	Two parameters	No	Arbitrary shape	Twitter database	Becker et al. (2011)
38	Incremental k-clique clustering algorithms	Automatic	Yes	Arbitrary shape	ENRON and DBLP datasets	Duan et al. (2012)
39	Density-based Incremental DBSCAN algorithm for mining microblogging text streams	Two parameters	Yes	Arbitrary shape	Twitter data with noise	Lee,(2012)
40	Multimodal clustering based on similarity (based on k-means)	User defined	No	Number of clusters	Media Eval social event dataset (numerical data)	Petkos et al. (2012)
41	Social stream clustering	Automatic	Yes	-	1.Twitter dataset 2. Enron email data set	Aggarwal et al. (2012)
42	Density-based spatial clustering algorithm (DBSC)	Automatic	Yes	Arbitrary shape	Data with noise	Liu et al. (2012)
43	Density-based clustering approach as event detection algorithm	-	-	Arbitrary shape	Twitter data	Lee et al. (2012)
44	k-means algorithm	Three parameters	No	Number of clusters	Flickr dataset of San Francisco	Cheng et al.(2012)
45	Single-pass incremental clustering algorithm	One parameter (similarity)	No	Arbitrary shape	Baidu news dataset	Xiaolin et al. (2013)
46	SED-RHOCC(Social event detection with robust high-order co-clustering)	Three parameters	No	Graph	Media Eval SED dataset 2012 (Flickr photos)	Bao et al. (2013)
47	Incremental DBSCAN	Two parameters	No	Arbitrary shape	Twitter and Flickr dataset with noise	Samangooei et al. (2013)
48	Agglomerative hierarchical clustering technique	Two parameters	Yes	Number of clusters	Two datasets 1. Rst dataset 2011 2. Tweets (from Jan 13 to Jan 19, 2013)	Parikh et al. (2013)
49	Robust clustering algorithm based on GPR	User defined	No	Number of clusters	Instagram data with noise	Xie et al. (2013)

50	Density-based Spatial temporal clustering algorithm	Three parameters	No	Arbitrary shape	Spatial Dataset	Tamura et al. (2013)
51	Adaptive K - mean similarity used for detection of geographical, social events	Automatic	No	Number of clusters	Spatial dataset with timestamp	Gao et al. (2013)
52	Dense K-mean algorithm used in rare framework	Two parameters	No	Number of clusters	Multiple dataset with noise	Székely et al. (2013)
53	Hierarchical algorithm used in LEED framework for clustering	Automatic	-	Graph	Twitter data	Unankard et al. (2013)
54	Graph-based Scan algorithm for event class detection	Automatic	No	Arbitrary shape	Flickr data of 2011 with noise	Nitta et al. (2014)
55	LTT graphical model based on Bayesian	Two parameters	No	Graph	Twitter data set	Zhou et al. (2014)
56	Multilayer event detection algorithm based on agglomerative	User defined	No	-	Twitter data	Tan et al. (2014)
57	Density-based approach	Automatic	Yes	Arbitrary shape	Four real-world benchmark datasets	Abulaish et al.(2014)
58	Extended DEN stream	Four parameters	Yes	Arbitrary shape	Twitter data with noise	Popovici et al.(2014)
59	Density based	Four parameters	-	Arbitrary shape	GeoSN datasets	Shi et al. (2014)
60	Constrained Incremental clustering via ranking	Automatic	No	Number of clusters	Flickr data and synthetic data	Sutanto et al. (2014)

Table 1 Main characteristics of different algorithms applied in scientific papers for the literature review

22. **Sato et al. (2008)** invented a trend based clustering algorithm to detect the topic and track them in dataset. This method is based on single linkage agglomerative clustering algorithm. The fundamental key of this algorithm is to provide a weight to each word according to their frequency in dataset that is based on gradient model. It requires user defined parameter. This clustering approach has consider mainly two steps 1)In the first step, distances between a new document and prior ones are calculated upon its arrival, and the nearest one is recorded in a nearest neighbor table. 2) In the second step, document clusters are generated based on the threshold given by the user. Experimental result of F- measure shows that algorithm is able to detect relevant documents.

23. **Sayyadi et al. (2009)** invented a new event detection algorithm that is based on the keyword graph and applied community detection methods to find the event on social stream. They have used cosine similarity to detect the clusters for document. According to authors, they assumed all keywords in one community as keywords for the event and weight is assigned to keyword graph to discover the betweenness centrality score.

24. **Chen et al. (2009)** proposed an algorithm to detect event on Flickr data where the user-defined tags are analyzed on spatial temporal bases. Furthermore, wavelet transform approach is used to remove the noise from the data. According to the authors, they had determined two types of event from the tags Flickr photo 1) aperiodic events 2) periodic events. Experimental result shows that algorithm is most suitable to detect periodic events with high accuracy as compare aperiodic events, but still aperiodic events detection is much more effective as compare to existing approaches.

25. **Becker et al. (2010**) implemented a weighted clustering algorithm considering multiple features listed as title, description, tags, location, and time. They continued their work by presenting a framework in 2010 to achieve high quality clustering results. They examined ensemble based and classification-based techniques for combining a set of similarity metrics. This offers the possibility of finding similarity among detected events. Their experiments revealed that the similarity metric learning methods produce better performance.

26. **Zhang et al. (2010)** have proposed automatic topic detection algorithm based on incremental algorithm. According to the author algorithm is able to automatically discriminate the topic from other topics and furthermore it assign a weight to each topic. This algorithm has used Bayesian Information Criterion (BIC) to detect the topic automatically. Experiment results show that TPIC (Topic Detection method based on Incremental Clustering) method is taking less time during execution and it has good performance results as compare to other methods.

27. **Lee et al. (2010)** proposed a new framework to detect local event from geo tagged messages from twitter by focus on geographical regularities of local crowd behaviors. This method detects local events using spatial partitions. This approach is divided in three steps. 1) collection of geo tagged messages.2) Identify the region of interest and furthermore measuring geographical

regularities of crowd behaviors 3) Identify the events on the bases of comparison through regularities. According to the author for event detection, it is very important to detect region of interest. They had applied k-mean clustering algorithm to classify the geo tagged tweets and furthermore they have formed veronoi diagram using the center points (lat., long.) of the K-means results. Experimental result shows a promising result of this approach.

28. **Peter et al. (2010)** proposed ODBSCAN, which is combination of fast DBSCAN algorithm and ME-DBSCAN algorithm to improve the performance of DBSCAN. It requires number of identical circles; radius and minimum points as input parameter. According to the author, main function of F-DBSCAN in this algorithm is to choose representative point as seed point at the time of cluster development to reduce region query function call. As the region query retrieves the neighbor point that belongs to radius, Circle lemmas are given and which can be directly used in the region query optimization.

29. **Kisilevich et al. (2010)** introduced a clustering algorithm called P-DBSCAN. They have developed photo based DBSCAN for event detection through geo-tagged photograph. It requires dataset of points with coordinates and ownership attributes neighborhood radius, adaptive density and adaptive density drop threshold as input parameters. In P-DBSCAN author has considered a user as density threshold to detect the unique event rather than personal event.

30. **Ram et al. (2010)** introduced DVDBSCAN (Density Variation Based Spatial Clustering of Applications with Noise) which is based on DBSCAN but opposite to DBSCAN, it is effective to handle density variation that exists within the cluster. Minimum objects (μ), radius, threshold values (α, λ) are the main parameter of algorithm. According to the author basic idea of DVDBSCAN it that it calculate the density mean and density difference of any central object of growing cluster and If cluster density difference for a main object is less than or equal to a given threshold value and it is also satisfying the cluster similarity index, in that case object will expand. The experimental result of DVDBSCAN shows that it detects the cluster not only in sparse region but also in the area having different density differences.

31. **Li et al. (2010)** proposed a method to identify hotspot by using sentiment analysis and text mining on online forum. In this paper, a k-mean clustering algorithm has applied along with support vector machine to develop the text mining approach. According to authors, they had used five input parameters in k

mean module. Experimental results show that SVM forecasting achieves highly consistent results with K-means clustering.

32. **Sakaki et al. (2010)** proposed a real time event detection method by applying semantic analysis on twitter tweets. They had applied support vector machine algorithm to classify the tweets in two different domains positive and negative. In this study, authors have applied two filtering approaches Kalman and particle for the estimation of event location. Additionally to make this method better they had consider user as sensor. The final result shows that particle filter works better than other compared methods in estimating the centers of the earthquake which was considered as event to see the performance of the proposed model.

33. **Kriegel et al. (2011)** proposed density based subspace clustering approach called Pre-Decon for dynamic data. Pre-Decon require three parameters (the distance thresholdε, the neighborhood size thresholdμ and the dimensionality thresholdλ) as input for algorithm. According to author dimension plays pivotal role to cluster the object if neighborhood of object along this dimension has a small variance.

34. **Goyal et al. (2011)** proposed an incremental density based algorithm, it has ability to add many points in the present group of cluster in dataset. In contrast to DBSCAN, in this algorithm new clusters can be added with present cluster to come with modified set of cluster after adding data points by using DBSCAN. According to author, clusters are joining incrementally rather than adding points. Main merits of an efficient Density Based Incremental Clustering Algorithm is that we can see the clustering pattern of both new as well as existing data and also we can merged the clusters.

35. **Wang et al. (2011)** proposed a method to detect event on social media using clustering and filtering. They have applied single pass clustering method to classify the data. Furthermore, they have also applied similarity approach on different feature of dataset. Flickr dataset included the spatial, temporal, textual and visual feature. Experimental result shows that F1-measure (It is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score in statistical analysis of any data. Mainly used in machine learning) value is acceptactable but NMI (Normalized Mutual Information) value need still work on algorithm to improve the result.

36. **Reuter et al. (2011)** proposed a method to handle the problem of event identification on social media platform by applying record linkage approach that is based on state of the art. According to the authors, they have applied single linkage algorithm to get the scability and to avoid the extra computation work. In this approach author has used two dataset and cluster quality is evaluate by using F-measure (It is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score in statistical analysis of any data. It is mainly used in machine learning) and NMI (Normalized Mutual Information) on dataset.

37. **Becker et al. (2011)** analyses the twitter stream to classify the massages of twitter related to real and non-real world events. They have applied incremental clustering algorithm to find clusters in data stream. Main advantage of incremental clustering is that it does not require prior knowledge of number of clusters. Event results are statistically evaluated to see the performance of proposed algorithm.

38. **Duan et al. (2012)** proposed incremental k-clique clustering algorithm for dynamic social network. Main advantage of K-clique incremental method is that we do not need to define the parameter. ENRON (ENRON is natural gas pipeline company of Texas, United States) and DBLP (Data Base systems and Logic Programming) dataset are used to evaluate the algorithm and results shows that it is more efficient algorithm as compare to others.

39. **Lee et al. (2012)** have proposed an approach to detect the events on social media micro blogs by analyzing spatial temporal features and contents. This method is able to assign the rank for each event to evaluate the impact of events. According to the authors, they have applied a density-based approach as event detection algorithm. For experiment they have used twitter data of Jan - Sep 2011 to detect the Virginia earthquake as event to see the performance of approach .The result shows that the approach has ability to quickly find the related events. Furthermore, it executes the event analysis on their spatial-temporal affects.

40. **Petkos et al. (2012)** introduced a supervised multimodal clustering algorithm. The algorithm was tested on the challenge data of Media-Eval social event detection and is compared to an approach using multimodal spectral clustering and early fusion. Using the explicit supervisory signal, the algorithm is able to achieve higher clustering accuracy and at the same time it required the

specification of a much smaller number of parameters. The authors claim that their algorithm can be applied not only to the task of social event detection, but to a wider scope for other multimodal clustering problems as well.

41. **Aggarwal et al. (2012)** proposed a new social event detection approach which is directly related to clustering. In this paper, author has assumed that each text message is related with least pair of actor in data stream. According to the author, data stream can be used as important resource to detect the interesting events. Experimental result shows the usefulness of this method.

42. **Liu et al. (2012)** invented a new density based algorithm called Density based spatial clustering algorithm to detect spatial cluster using geometrical properties and attributes. DBSC requires two users defined input parameters. According to authors, they have used delaunay triangulation with edge length constraints to make model for spatial proximity connection between spatial points. Afterwards, they have applied modified density based clustering to find spatial cluster in dataset.

43. **Lee et al. (2012)** proposed novel spatial temporal topic detection framework to explore the micro blogging social network that are the main item for real time event. According to authors, they have applied density based clustering algorithm in data stream to detect the events by using spatial temporal feature. The proposed framework consist two steps module. First step is content and temporal analysis module, which is able to categories twitter DataStream in to thematic topics. Afterwards in second step spatial analysis has done to assign the topic to real time location.

44. **Cheng et al. (2012**) proposed a new novel based method to extract hot spot of any area by using social media dataset. This method has three steps; First of all k- mean clustering algorithm is applied on Flickr dataset to group the geo tagged photos which helps to extract the events from the scene. In next step, TF-IDF (Term Frequency–Inverse Document Frequency) method was applied to find top keywords from the clusters and in final step, photos with description on the bases of extracted keyword are visualized. Experiment was done on Flickr geo tagged photos taken from San Francisco area and results shows that the algorithm could effectively improve the original tag results.

45. **Xiaolin et al. (2013**) invented an improved single pass clustering algorithm, which calculates the similarity between new coming document and the category

seed documents. It requires similarity as input parameter. Main advantage of the algorithm is that it decreases the false detection and cost of false detection. Experimental result shows that it enhanced the speed and quality of clustering.

46. **Bao et al. (2013)** proposed a Social Event Detection with Robust High-Order Co-Clustering (SED-RHOCC) algorithm that is based on start structured k partite graph to overcome the challenge of processing the associated heterogeneous metadata, such as timestamp, location, visual content and textual content. SED-RHOCC algorithm is two-step processes. In first step algorithm detects the event from dataset and in second step, it refines the cluster result by using post processing. The experimental experiences on Mediaeval Social Event Detection Dataset showed the effectiveness of the proposed approach in social media datasets.

47. **Samangooei et al. (2013)** proposed a new method for event detection on social media platform called multi modal clustering approach. In this paper author has talked about how they had combined the feature, what was the relative importance of feature? Furthermore, he has discussed in detail about the event detection process on large dataset. Proposed algorithm requires two parameters and experimental result shows a promising result of the method.

48. **Parikh et al. (2013)** developed a scalable system, called ET (Events from Tweets), for detecting real world events from a set of micro blogs (tweets). It is automatically detect the event by investigating the textual and temporal components. The key feature of their system was clustering the related keywords based on content similarity and appearance similarity among keywords. ET used a hierarchical clustering process for determining the events. It was tested on two different datasets from two different domains. The results for both of these domains were precise.

49. **Xie et al. (2013**) proposed a method for real time hyper local event detection from social media data. They applied two step processes, first step is combination of time series predicting component and classifier, which made a model for time series and helps to identify the unusual signals in a small geographical area. In second step they used classifier to identify the real events. According to the author, this model can apply in any social media platform but for demonstration, they have used Instagram photos that were taken in New York and results shows that this approach is able to find different type of event as well as minor events.

50. **Tamura et al. (2013)** proposed a density based spatial temporal clustering algorithm for extracting bursty areas from Geo-referenced documents. Density-based spatiotemporal clustering algorithm is a natural extension of DBSCAN. Latitude, longitude, time interval and minimum document as threshold value are main input parameters for algorithm. The proposed clustering algorithm is able to recognize the temporally and spatially separated clusters. The clustering algorithm separated coordinate space from time space. In this paper authors claimed that they did not find any study on spatiotemporal clustering algorithms that can recognize clusters that are both temporally and spatially separated from other clusters for geo referenced documents before their work. Evaluation has done on real dataset of twitter, which has included 480,000 tweets of more than one year and has given significant result of the proposed algorithm.

51. **Gao et al. (2013)** proposed a novel approach to overcome problem of detecting geo tagged social event in micro blogging social media sites. According to the authors, they have applied k-mean algorithm to detect the clusters from geo tagged tweets and afterwards they perceived spatial social event by the tweets in cluster. The approach is applied on Sine Wiebo realistic dataset and experiments approved the benefit of their tool in location related social event detections.

52. **Székely et al. (2013)** proposed a novel two-stage framework to detect the group of separated outliner in dataset. Authors have introduced backward approach to outlying unusual event categories in large dataset. According to authors, first of all they detect the centers of compact areas and afterword they expand the area according to density based condition. Result shows that framework is able to detect rare event in large dataset.

53. **Unankard et al. (2013**) proposed a system to detect event location from micro blog messages with geo location. The main idea of this approach is to create correlation between user and event location to make event visible in dataset. According to author, he has used hierarchical based clustering algorithm to make text-based cluster and identify the events in dataset. Furthermore, he has applied sliding window manager to keep record of each message arrival in system. The experiments result shows that the proposed method is able to detect emerging events over the baselines.

54. **Nitta et al. (2014)** proposed method for real time event detection, which is based on latitude, longitude, time and text tag in Flickr images. It is two-step

event detection processes. Author has define a class to event by applying SCAN (Structure Clustering Algorithm for Networks) clustering algorithm which is based on graph and for event detection author has taken similarity of time and text tag of image. Experimental result shows that it is efficient approach to detect the event from small amount of images.

55. **Zhou et al. (2014**) proposed a framework for monitoring online social events from tweet streams for real applications such as crisis management. To represent the tweets, a graphical model called location-time constrained topic (LTT) was proposed to fuse various information such as social content, location and time of tweets. The similarity of messages was caught using a complementary distance, which considered the differences between two messages over four attributes; content, location, time, and link. To prove the effectiveness and efficiency of their proposed approach, they have conducted two experiments over long tweet streams during two occurred crisis in Australia.

56. **Tan et al. (2014)** proposed multilayer hot event detection algorithm to make a distinction between global and local events. This approach has four steps to detect the event, which gives us meaningful events as result. The evaluation result that is based on F- measure shows that algorithm has good performance as compare to traditional approaches.

57. **Abulaish et al. (2014)** proposed density based approach for detecting overlapping community structures in online social networks. Main advantage of this method that it does not require user defined parameter, which is tuff task for any algorithm, rather than it compute the neighborhood for every single node automatically. Experimental result shows a quiet positive result of this approach.

58. **Popovici et al. (2014)** proposed an online clustering approach called extended DEN stream (Density-Based Clustering over an Evolving Data Stream with Noise) which is based on density based approach and also an extension of Den stream. It requires four parameters and evaluation result shows us quiet positive result.

59. **Shi et al. (2014)** proposed an extension of DBSCAN to cluster places in geo social network which are visited by users. This method has considered socio spatial information between users who visit the clustered places. DCPGS (Density-based Clustering Places in Geo-Social Networks) requires five input

parameters. This method is applied on real dataset and result shows that DCPGS algorithm can cluster millions of places within a few seconds.

60. **Sutanto et al. (2014)** introduced ranking based constrained document clustering method to deal with large dataset problem. According to the authors, they have applied semi incremental procedure to make the algorithm faster and more efficient in terms of memory consumption. For document ranking, they have chosen the criteria of vicinity of best cluster instead of applying distance criteria between documents. Experimental result shows that this approach provides good accuracy and requires less memory.

4.2 Literature review conclusion

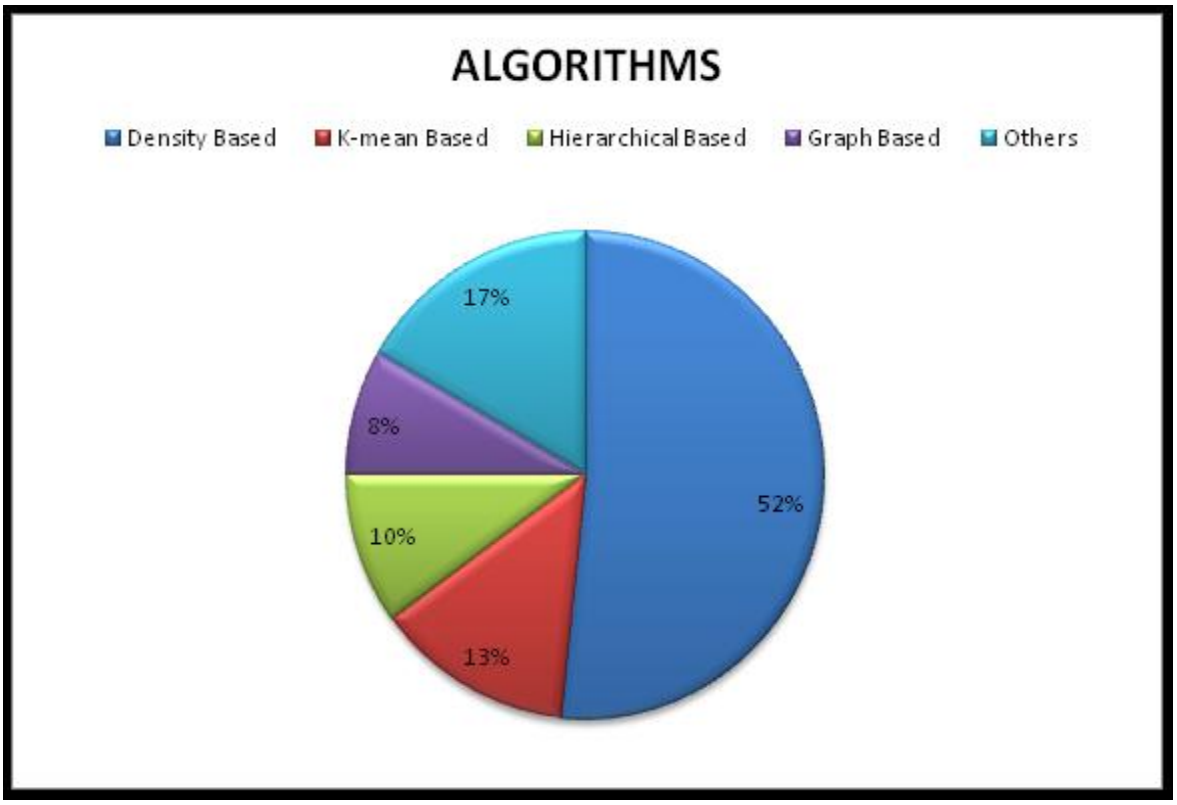


Figure 5 Percentage of different clustering algorithms used in reviewed literature

As a conclusion of review work of scientific research papers mentioned in table 1 which are taken from different time frame and different journals and magazine; it is found that density based algorithms specially DBSCAN has significant impact on the research in the similar domain.

Above Figure 5 shows a short summary of the percentage of different type of clustering algorithms used in reviewed literatures. As clearly visible 52% of papers applied density based algorithm, 13% have applied k-mean, 10% have applied hierarchical based clustering, 8% papers applied graph based approach and 17% papers had applied rest of clustering algorithm other than mentioned above. This in itself also shows the significant impact and importance of density-based algorithm in the scientific research community in the context of problem of the similar domain and dataset/data type. Based on the various algorithm's advantages, disadvantages (section 3.1 - 3.5) and of the type of dataset on which the algorithm is applied and final results of above said papers; I have finally selected density-based criteria and "Density-based Spatiotemporal Clustering Algorithm" is selected as a base research paper for this Thesis. This algorithm is natural extension of DBSCAN algorithm. In this paper author has taken spatial as well as temporal feature with text based threshold criteria to implement the clustering algorithm.

4.3 Motivations to choose DBSCAN algorithm variant

Following are the main reasons for choosing this variant of DBSCAN algorithm in this Thesis as a base algorithm for achieving the Thesis objective:

1. This algorithm is using one extra dimension of time (in comparison to base DBSCAN algorithm) apart from radius (distance) to find out the clusters of relevant topic, which are made considering these two parameters. This new parameter of "time" will give us more control over finding the clusters considering time via predefined threshold value and based on the dataset type this parameter shall be tuned. As a result we get refined clusters located in terms of time and location i.e. who all are discussing certain topics in which time frame and location area.

2. Text based spatial temporal dataset shall be used in this Thesis, which is also used in selected paper. This is another motivation to modify and extend "Density-based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Geo referenced Documents'' paper's algorithm in comparison to other reviewed papers.

   There are also some other factors which motivated me to opt density based clustering algorithm is that

3. It is an unsupervised clustering algorithm, which has the ability to detect arbitrary shape clusters even in noisy dataset.

4. It can be used for real world and real time data. We can use it locally as well as global level. This is main reason that is very popular as compared to other algorithms. It does not require prior knowledge of number of clusters as k-mean, which is quite tough job when we have large dataset.

5. As we know that geo spatial data usually vary in different form and normal clustering algorithms are not able to handle the variation in such type of data. DBSCAN is most appropriate approach to handle it. As above mention the characteristics of DBSCAN popularity is its simplicity, which motivates the scholars to work on this algorithm.

6. Last but not the least is the consideration of the time involved in the implementation of the whole framework to achieve the current Thesis's objective was also one of the criteria to choose this algorithm.

5 (ϵ, τ) Density Based Spatial Temporal Clustering (DSC)

(ϵ, τ) DSC algorithm was introduced by (Tamura et al. 2013) which is an extension of DBSCAN (Ester et al. 1996) clustering algorithm. Main difference of DSC algorithm from base DBSCAN algorithm is that it is able to recognize the temporally and spatially separated clusters.

The basic idea behind (ϵ, τ) DSC algorithm (Tamura et al. 2013) is described in below section.

5.1 Definitions of (ϵ, τ) Density Based Spatial Temporal Clustering (DSC)

Definition 1: ((ϵ, τ) neighborhood N(ϵ, τ)(dp))

The (ϵ, τ) neighborhood of a document dp, denoted by $N_{(\epsilon, \tau)}(dp)$ is defined as,

$N(\epsilon, \tau)(dp) = \{ dq \in D | dist(dp,dq) \leq \epsilon \text{ and } iat(dp,dq) \leq \tau \}$

Distance between documents dp and dq is return by the function *dist* and interarrival time between documents dp and dq is return by *iat* function.

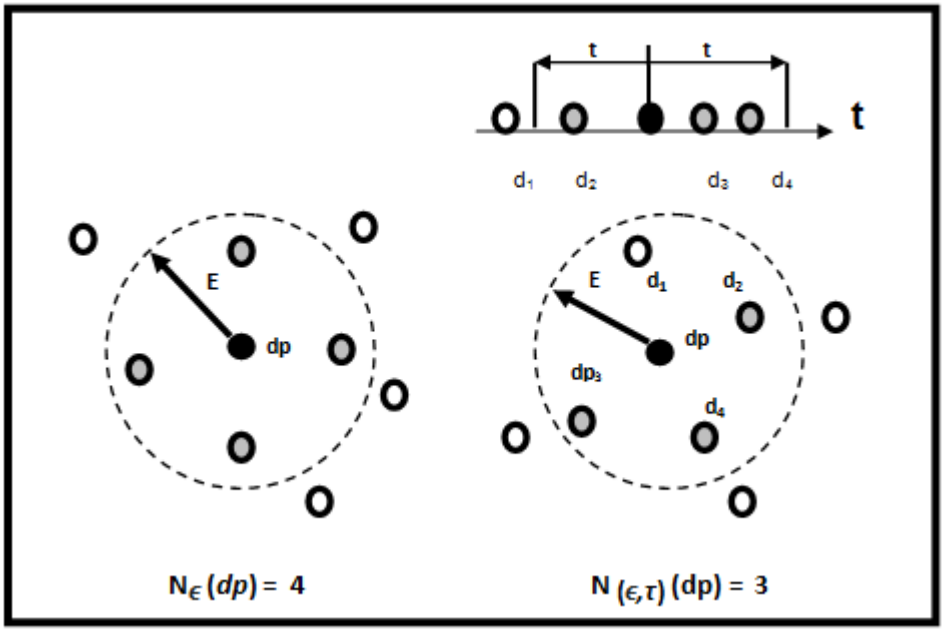


Figure 6 Explanation of definition 1 (DSC algorithm)

As shown in above figure 6 (left side figure), there are 4 documents in ϵ neighborhood of document dp and in the same figure on the right side when the author had applied (ϵ, τ) neighborhood of document dp, only 3 documents (d2, d3, d4) are present because in right above side figure the (ϵ, τ) neighborhood of document dp is group of documents which lies within the radius of document dp and each document in (ϵ, τ) neighborhood is posted before or after the posted time of dp.

Definition 2: (Core document, Border document)

A document dp who satisfies below conditions are accordingly said core/border document.

If ($N_{(\epsilon, \tau)}(dp) \geq$ Min Doc) in this case dp called core document and

If ($N_{(\epsilon, \tau)}(dp) \leq$ Min Doc) in this case dp called border document.

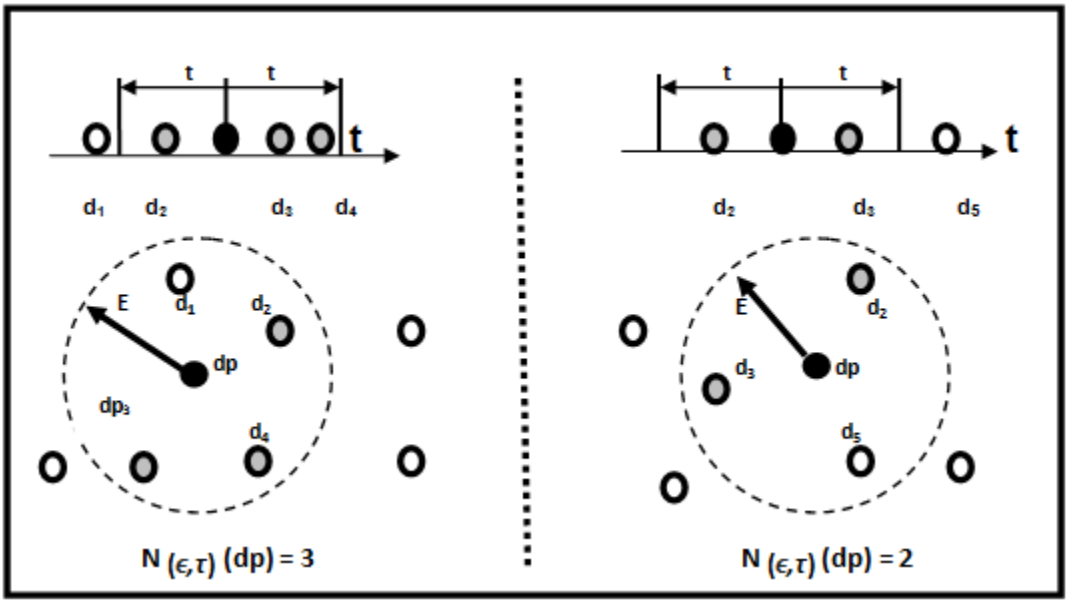


Figure 7 Example of definition 2 and 3 (DSC algorithm)

In above example in figure 7, if user defined minimum document value is set to 3. In that case, left part of diagram dp is core document because it is satisfying the minimum document condition and in right side of figure dp is border document because the number of documents in $N_{(\epsilon, \tau)}$(dp) is less than minimum document.

Definition 3: ((ϵ, τ)-density based directly reachable)

Suppose that a document dq is the (ϵ, τ)-neighborhood of a document dp. If the number of documents in the (ϵ, τ)-neighborhood of document dp is greater than or equal to MinDoc,

i.e. $N_{(\epsilon,\ \tau)}(dp) \geq$ MinDoc, document dq is (ϵ, τ)-density-based directly reachable from document dp. In other words, documents in the (ϵ, τ)-neighborhood of a core document are (ϵ, τ)-density-based directly reachable from the core document.

Definition 4: ((ϵ, τ)-density based reachable)

If there is a document sequence like (dp_1, dp_2,⋯, dp_n) and *i+1* th doc dp $_{i+1}$ is density-based directly reachable from the i-th document dp_i. Document dp_n is (ϵ, τ)-density-based reachable from document dp_1.

Definition5: ((ϵ, τ)-density based connected)

According to definition suppose that document dp and dq are (ϵ, τ) density based reachable from document do and if $N_{(\epsilon,\ \tau)}(do) \geq$ MinDoc, in this case document dp is (ϵ, τ) density based connected to document dq.

(ϵ, τ)-Density-based Spatiotemporal Cluster

A (ϵ, τ)-density-based spatiotemporal cluster consists of two types of document: core documents, which are mutually (ϵ, τ)-density-based reachable; and border documents, which are (ϵ, τ)-density-based directly reachable from the core documents. A (ϵ, τ)-density-based spatiotemporal cluster is defined as follows.

Definition 6: ((ϵ, τ)-density based spatial temporal cluster)

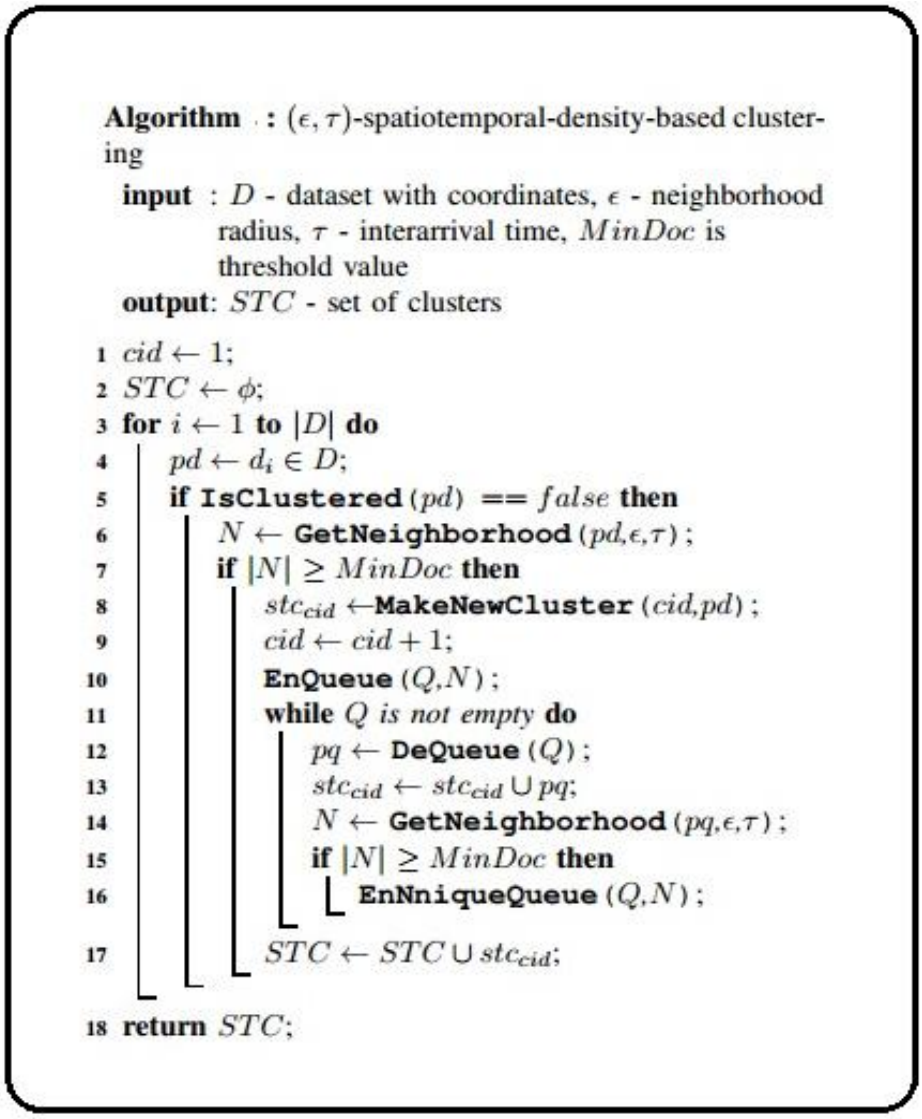
An (ϵ, τ)-density-based spatiotemporal cluster (DSC) in a document set D satisfies the following restrictions:

1. ∀dp, dq ∈ D, if and only if dp ∈ DSC and dq is (ϵ, τ)-density-based reachable from document dp, document dq is also in DSC.
2. ∀dp, dq ∈ DSC, document dp is (ϵ, τ)-density based connected to document dq.

Even if dp and dq are border documents, dp and dq are in a same (ϵ, τ)-density-based spatiotemporal cluster if dp is (ϵ, τ)-density-based connected to document dq.

5.2 Description of (ϵ, τ)-Density Based Spatial Temporal Clustering

The steps involved in the (Tamura et al. 2013) algorithm are shown in below figure 8,



Algorithm : (ϵ, τ)-spatiotemporal-density-based clustering

input : D - dataset with coordinates, ϵ - neighborhood radius, τ - interarrival time, $MinDoc$ is threshold value

output: STC - set of clusters

```
1  cid ← 1;
2  STC ← φ;
3  for i ← 1 to |D| do
4      pd ← d_i ∈ D;
5      if IsClustered(pd) == false then
6          N ← GetNeighborhood(pd,ε,τ);
7          if |N| ≥ MinDoc then
8              stc_cid ← MakeNewCluster(cid,pd);
9              cid ← cid + 1;
10             EnQueue(Q,N);
11             while Q is not empty do
12                 pq ← DeQueue(Q);
13                 stc_cid ← stc_cid ∪ pq;
14                 N ← GetNeighborhood(pq,ε,τ);
15                 if |N| ≥ MinDoc then
16                     EnNniqueQueue(Q,N);
17             STC ← STC ∪ stc_cid;
18 return STC;
```

Figure 8 DSC Algorithm (Tamura et al. 2013)

Step1: In first step, for each document dp in D, the function Is Clustered checks whether document dp is already assigned to a cluster or not.

Step 2: Then the (ϵ, τ) density based neighborhood of document dp is obtained using the function Get Neighborhood.

Step 3: If document dp is core document according to definition 2, it is assigned to a new cluster, and all the neighbors are queued in to Q for further processing. The processing and assignment of documents to the current cluster continue until the queue is empty, if that document dp is core document then it is assigned to new cluster and other neighbor will be

added in the queue for further processing and this process and allocation of documents in current cluster will continue until the queue is not finished.

Step 4: In last step, next document is dequeued from queue Q and if the dequeued document is not already assigned to current cluster, it is so assigned to the current cluster. Then if the dequeued document is a core document, documents in the (ϵ, τ) density-based neighborhood of the dequeued document are queued to queue Q using the function EnNniqueQueue, which puts input document into queue Q if they are not already in queue Q.

5.3 Definition of Cosine Similarity

Similarity measure plays a significant role to group the clusters in dataset. A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Euclidean distance and the Jaccard correlation coefficient (Huang et al, 2008 and Korenius et al, 2007). Cosine similarity is one of the famous techniques for text matching between two different georeferenced documents that is widely used measure to find out the similarity between text documents.

Let dt_i denote all words in $text_i$ of i-th georeferenced document:

$dt_i = \{w_{i,1}, w_{i,2}, \cdots, wi,nw(i)\}$,

Where wi,j ∈ W , W is a set of all words including in {text1, text2, · · ·, textn}.

The Cosine distance between *u* and *v*, is defined as below in terms of vector dot products.

$$1-(u.v)/(|u|.|v|)$$

Where u.v is the dot product of u and v.

In other words, the cosine similarity between two documents Di and Dl is calculated as

$$sim(D_i, D_j) = \cos(D_i, D_j) = \frac{\sum_{j=1}^{m} W_{ij} W_{li}}{\sqrt{\sum_{j=1}^{m} W_{ij}^2 . \sum_{j=1}^{m} W^2_{lj}}}, \text{ i, l =1, ..., n}$$

The resulting similarity ranges from −1 meaning exactly opposite, to 1 meaning exactly the same, 0 usually indicating independence, and in-between values indicating intermediate similarity or dissimilarity.

For text matching, the attribute vectors u and *v* are usually the term frequency vectors of the documents. The cosine similarity can be seen as a method of normalizing document length during comparison.

In the case of information retrieval, the cosine similarity of two documents will range from 0 to 1, since the term frequencies cannot be negative.

6 Proposed (ϵ, k, t) density-based spatiotemporal clustering algorithm

This chapter contains the details about the implementation aspects of proposed density based spatial temporal clustering algorithm which is referred in this Thesis as (ϵ, k, t)-DBSCAN. This algorithm is a natural extension of selected scientific paper "Density-based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Geo referenced Documents'' [Tamura et al. 2013], and implicitly it is also an extension of "A density-based algorithm for discovering clusters in large spatial databases with noise." (Ester et al., 1996).

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

The (ϵ, τ)-density-based spatiotemporal clustering algorithm extracts (ϵ, τ)-density-based spatiotemporal cluster, which are both temporally separated and spatially separated from other cluster; however it does not take in to account the semantic similarity criteria of the spatiotemporal documents. In this study we define the (ϵ, k, t) neighborhood of geo-referenced document to extract semantically similar spatial and temporal clusters. Main advantage of the proposed algorithm is to extract the semantically dense areas, which will allow us to identify the local hot topics under discussion among different social media users. The algorithm uses the cosine similarity concept between two spatiotemporal documents to achieve the semantically similar clusters.

6.2 (ϵ, k, t)-DBSCAN algorithm

DBSCAN algorithm was first introduced by Ester et al., 1996 which is based on a density based characteristic of clusters, which was the base of Tamura et al. 2013 research paper "Density-based Spatiotemporal Clustering Algorithm for Extracting Bursty Areas from Geo referenced Documents''. The proposed algorithm, (ϵ, k, t)-DBSCAN (where 'ϵ' is the Radius, 'k' is similarity rate constant, 't' stands for time when a document is posted) is an extension of (ϵ, τ)-spatiotemporal-density based clustering algorithm (Tamura et al. 2013) which implicitly means, this is also an extension of DBSCAN (Ester et al., 1996).

In DBSCAN the key concept is of "data point density", where the clusters are defined as maximal group of dense object within a certain radius ϵ and low dense areas are called noise. The beauty of DBSCAN algorithm is that it can discover arbitrary shape clusters too in the dataset apart from being able to handle outliers (noise) of the dataset. Outliers

(noise) is defined as those data points which do not belong to any generated clusters. Due to such inherent wonderful properties of DBSCAN, it is quite well used in spatial temporal databases. In another words, a cluster is made by set of data points which has certain density which is defined by parameters "Minimum number of points" denoted by MinPts and "radius" criteria and both of these are predefined threshold values configured by the user.

In the following pages, wherever I mentioned "document" it refers to "spatial-temporal document". In the following section, the definitions 1 to 6 are the extension of (Tamura et al. 2013) research paper to accommodate the newly added input parameter 'k' (cosine similarity rate constant) and also the modification of the meaning of "MinPts" of base DBSCAN (Ester et al.) algorithm's input.

Definition 1 ((ϵ, k, t) - neighborhood N (ϵ, k, t) (dp))

The (ϵ, k, t) neighborhood of a document dp, denoted by N (ϵ, k, t) (dp), is defined as

$$N(\epsilon, k, t)(dp) = \{dq \in D \mid dist(dp, dq) \leq \epsilon \text{ and } iat(dp, dq) \leq t \text{ and } sim(dp, dq) \geq k\},$$

Where,

The function dist returns the distance between document dp and document dq, and

the function iat returns the inter arrival time between document dp and document dq,

and the function sim returns the similarity between document dp and document dq.

An example of **ϵ**-neighborhood is shown on the left side of below Fig. 9. The **ϵ**-neighborhood of the document dp is a set of documents that exist within **ϵ** from document dp. In this example, there are five documents in the **ϵ**- neighborhood of document dp.

An example of an **(ϵ,k,t)** - neighborhood of document dp is shown on the right side of Fig. 9 As indicated by the example, the **(ϵ,k,t)** - neighborhood of document dp is a set of documents which fulfill the following three main criteria as said above i.e.

- **(ϵ,k,t)** - neighborhood of document dp is a set of documents that exist within **ϵ** distance from document dp;
- each document in the **(ϵ,k,t)** - neighborhood is posted in 't' before or after the posted time of document dp.

- each document in the (ϵ, k, t) - neighborhood has similarity with document dp which is $\geq k$ (user defined threshold).

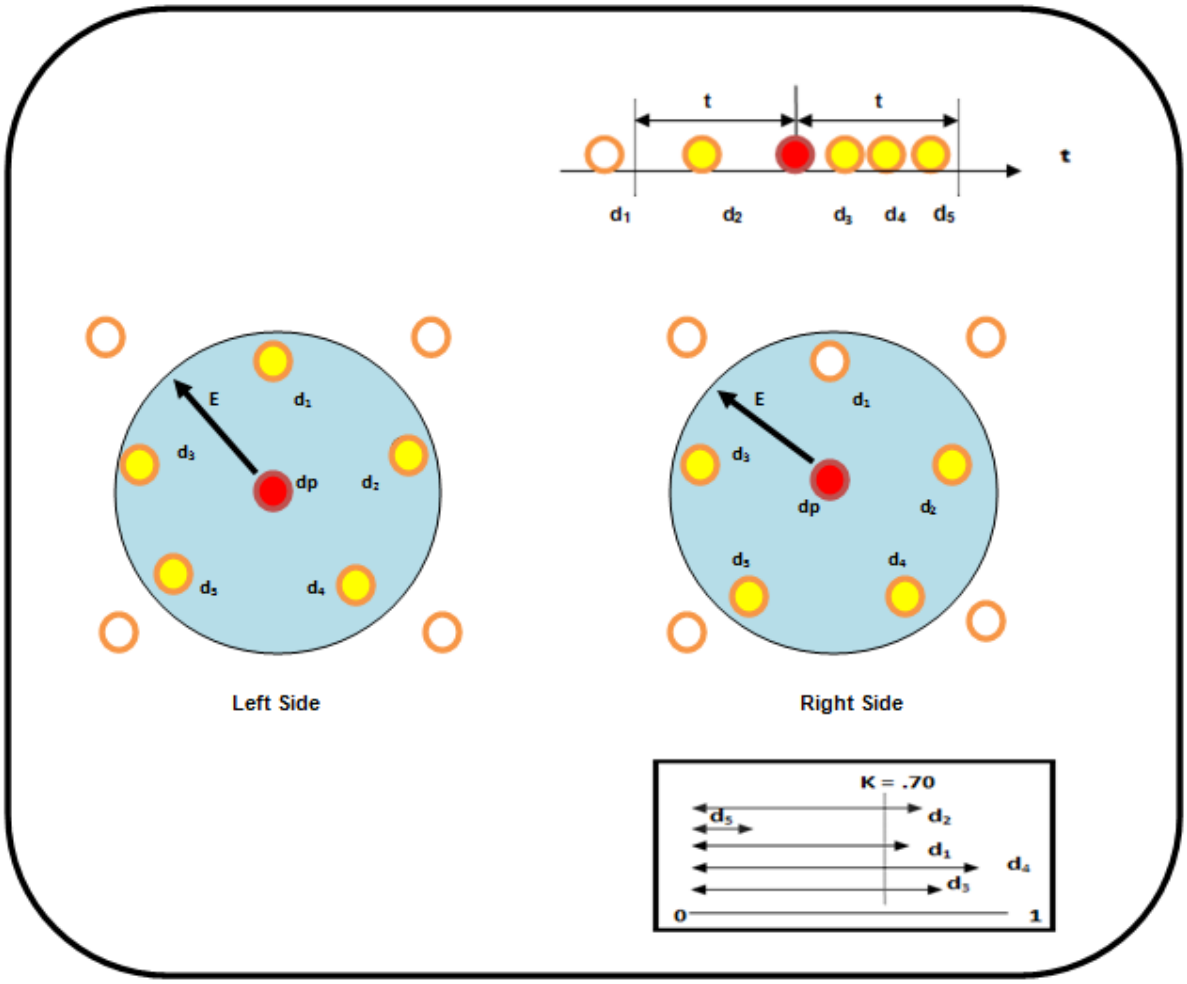


Figure 9 Example of Definition 1 ((ϵ, k, t)-DBSCAN)

In above figure 9 (right side) example, there are three documents, $N_{(\epsilon,k,t)}(dp) = \{d_2, d_3, d_4, \}$. Document d1 is within ϵ distance from document dp; however, it is not in $N_{(\epsilon,k,t)}$ because it is not posted in '**t**' before or after the posted time of document dp.

Document d5 is within ϵ distance from document dp and also posted within the time '**t**' however it is not meeting the similarity criteria and hence it is not in $N_{(\epsilon,k,t)}$.

Definition 2 (Core Document, Border Document)

A document dp is called a *core document,* if there are at least a minimum number of

documents, MinDoc, in the (ϵ, k, t)-neighborhood $N_{(\epsilon,k,t)}(dp)$ of that document ($N_{(\epsilon,k,t)}(dp) \geq MinDoc_{DifferentUsers}$).

Otherwise, ($N_{(\epsilon,k,t)}(dp) \leq MinDoc_{DifferentUsers}$), document dp is called a *border document*.

In this algorithm, I have also extended the definition of minimum number of documents (MinDoc), which refers here as the minimum number of documents of *different users* ($MinDoc_{DifferentUsers}$).

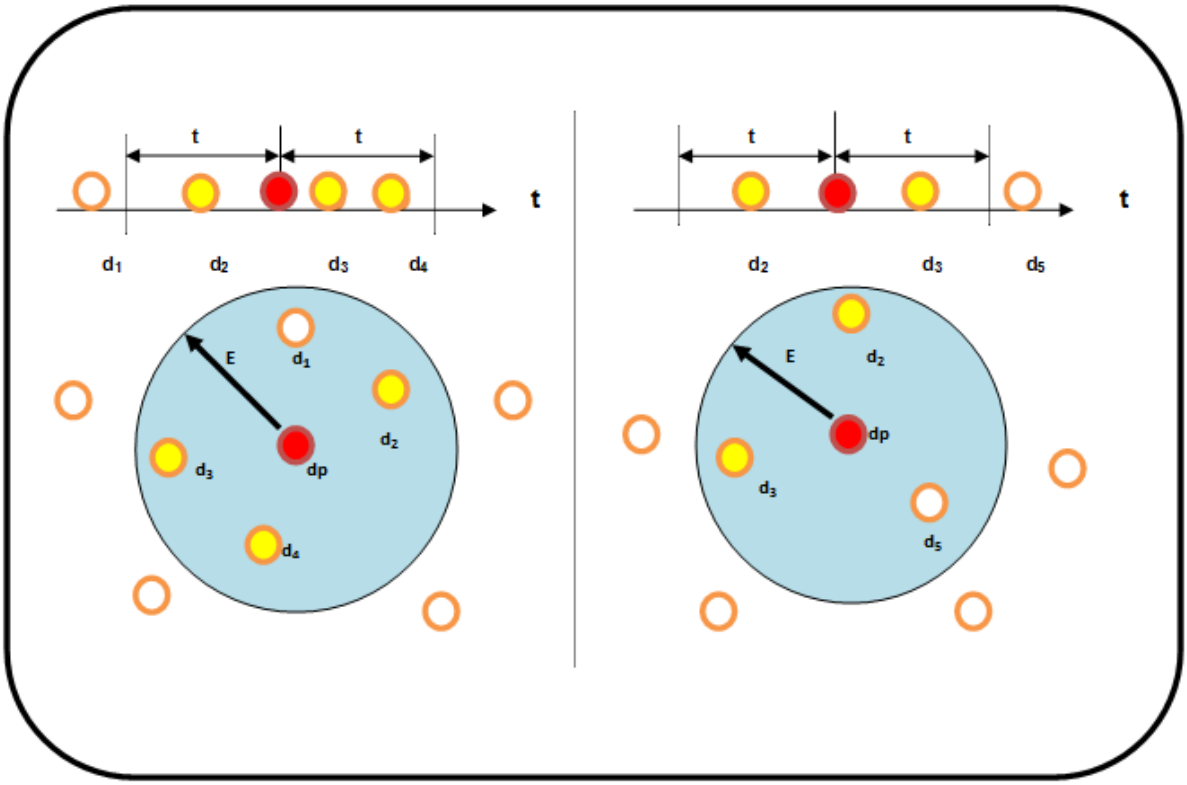


Figure 10 Example of definition 2 and 3 ((ϵ, k, t)-DBSCAN)

Suppose that $MinDoc_{DifferentUsers}$ is set to three. On the left side of figure 10, document dp is a core document, because the requirement is met i.e. $MinDoc_{DifferentUsers} = 3$, but on right side of figure10, the document dp is a border document because the number of documents in $N_{(\epsilon,k,t)}(dp)$ is less than $MinDoc_{DifferentUsers}$ i.e. three. Note here in the right side of the figure, the document d5's owner is same as d3 and hence the $MinDoc_{DifferentUsers} = 2$ only because the documents d5 does not full fill minimum document with different users criteria.

Definition 3 ((ϵ, k, t)-density based directly reachable)

Suppose that a document dq is the $(\boldsymbol{\epsilon,k,t})$-neighborhood of a document dp. If the number of documents in the $(\boldsymbol{\epsilon,k,t})$-neighborhood of document dp is greater than or equal to $\text{MinDoc}_{\textbf{DifferentUsers}}$, i.e.,

$N_{(\epsilon,k,t)}(dp) \geq \text{MinDoc}_{\textbf{DifferentUsers}}$, document dq is $(\boldsymbol{\epsilon,k,t})$-density-based directly reachable from document dp. In other words, documents in the $(\boldsymbol{\epsilon,k,t})$-neighborhood of a core document are $(\boldsymbol{\epsilon,k,t})$-density-based directly reachable from the core document.

On the left side of figure 10, document dp is a core document, because

$N_{(\epsilon,k,t)}(dp) = \text{MinDoc}_{\textbf{DifferentUsers}}$.

and on the right side of figure 10, Documents d2 and d4 are in the $(\boldsymbol{\epsilon,k,t})$-neighborhood of document dp. These two documents are not $(\boldsymbol{\epsilon,k,t})$-density-based directly reachable from document dp considering $\text{MinDoc}_{\textbf{DifferentUsers}}$ is set to three.

Definition 4 ((ϵ, k, t)-density-based reachable)

Suppose that there is a document sequence $(dp_1, dp_2, \cdots, dp_n)$ and the $\mathbf{i+1}^{th}$ document dp_{i+1} is $(\boldsymbol{\epsilon,k,t})$-density-based directly reachable from the $\mathbf{i^{th}}$ document dp_i. Document dp_n is $(\boldsymbol{\epsilon,k,t})$-density-based reachable from document dp_1.

Definition 5 ((ϵ, k, t)-density-based connected)

Suppose that document dp and document dq are $\boldsymbol{(\epsilon,k,t)}$-density-based reachable from document dx.

If $N_{(\epsilon,k,t)}(dx) \geq \text{MinDoc}_{\textbf{DifferentUsers}}$, we denote that document dp is $\boldsymbol{(\epsilon,k,t)}$-density-based connected to document dq.

(ϵ,k,t)-Density-based Spatiotemporal Cluster

A (ϵ,k,t)-density-based spatiotemporal cluster consists of two types of document: core documents, which are mutually (ϵ,k,t)-density-based reachable; and border documents, which are (ϵ,k,t)-density-based directly reachable from the core documents. A (ϵ, k, t)-density-based spatiotemporal cluster is defined as follows.

Definition 6 ((ϵ, k, t)-density-based spatiotemporal cluster)

An (ϵ, k, t)-density-based spatiotemporal cluster in a document set D satisfies the following restrictions:

1) $\forall dp, dq \in D$, if and only if $dp \in DSC$ and dq is (ϵ, k, t)-density-based reachable from document dp, document dq is also in DSC.
2) $\forall dp, dq \in DSC$, document dp is (ϵ, k, t)-density-based connected to document dq

Even if dp and dq are border documents, dp and dq are in a same (ϵ, k, t)-density-based spatiotemporal cluster if dp is (ϵ, k, t)-density-based connected to document dq.

6.3 Data Model of (ϵ, k, t)-Density Based Spatial Temporal Clustering Algorithm

In this thesis, the terms like "document" or "georeferenced document" or "spatiotemporal document" are used interchangeably and they mean the same thing which is explained in below section.

6.3.1 Definition of Spatiotemporal Document

A document is nothing but the single instance/record of any social media user's publicly posted information e.g. status etc. which consists of primarily 3 main parts, which are position, time and text. Where

Text is nothing but the actual textual content of the document e.g. Twitter's tweet string of certain tweet which is posted by certain Twitter user.

Time is nothing but the actual time of posting the document e.g. e.g. Twitter's tweet precise posting time, and

Position is the combination of latitude and longitude of the posted document e.g. Twitter's tweet posting precise actual location at the time of posting; It primarily just specifies the precise location of some document in the map/co-ordinates.

Following figure 11 is the visual representation of the concept of document/spatiotemporal document/geo-referenced document, which are referred here as sd1 … sd5 which were posted in timeline from time1 to time5 at locations referred as Position1 to Position5 respectively.

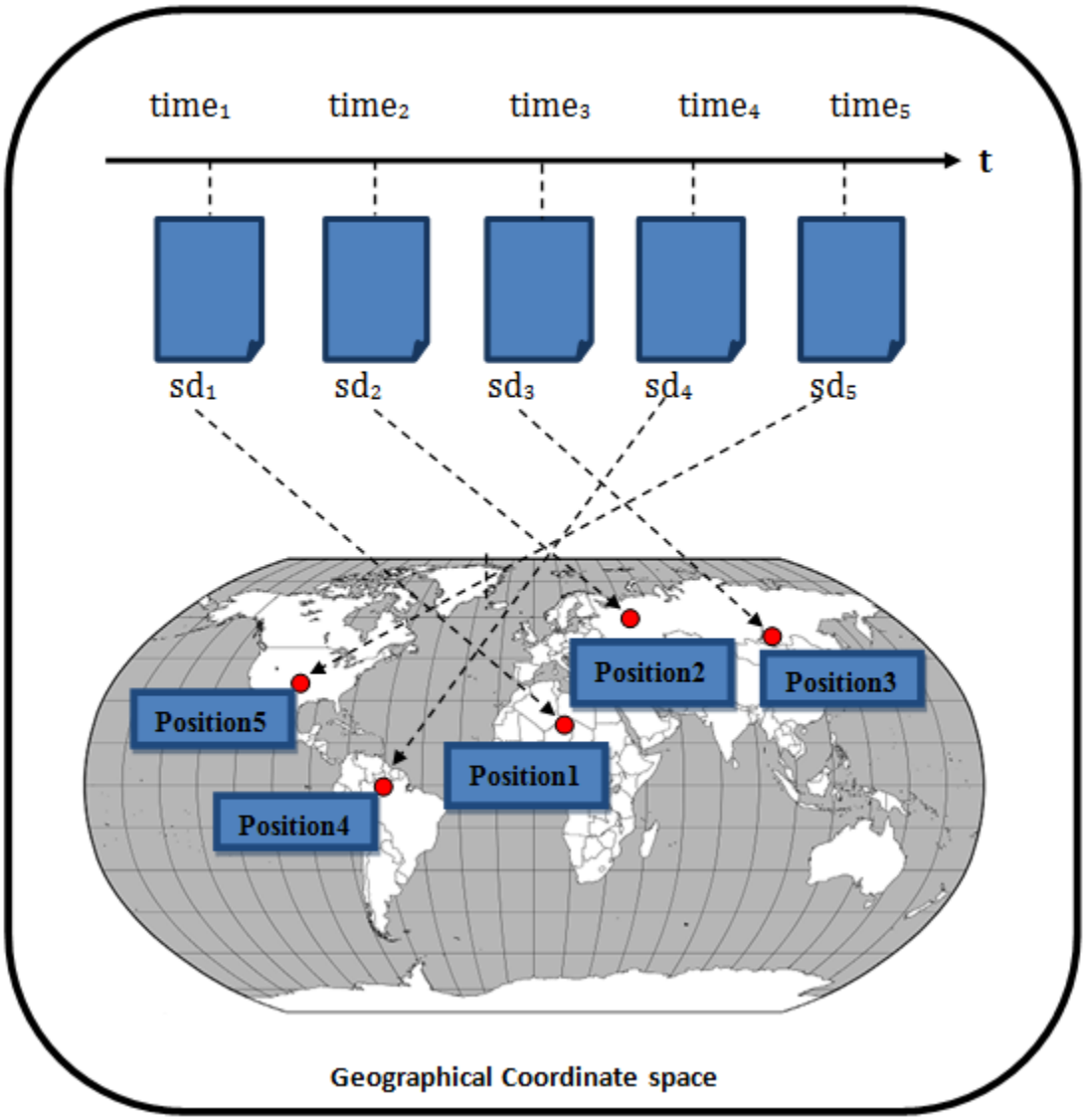


Figure 11 Data model of (Є, k, t)-Density-Based Spatial Temporal Clustering Algorithm (Tamura et al. 2013).

6.4 Description of (ϵ, k, t)-DBSCAN Algorithm

Below pseudo code steps describes the (ϵ, k, t)-density-based spatiotemporal clustering algorithm.

The logic and processing of the algorithm is clear in itself by seeing the pseudo code shown below however for more clarity the steps are elaborated as below:

Step 1: For each document 'pd' in D, first we check whether this document is already assigned to some existing cluster or not and this job is done by ***IsClustered*** function. Next step proceeds further depending upon whether this document is already assigned to existing cluster or not and if it not assigned then the control goes further to next step 2 otherwise next document 'pd' is taken from D again and the same step 1 starts again. This explains the step 3 – step 5 of the below said algorithm's pseudo code.

Step 2: Next step involves the finding of all the (ϵ, k, t)-density-based neighborhood in the dataset of the selected document 'pd'. This job is done by GetNeighborhood function, which returns all the neighbors of the document 'pd'. If the document 'pd' is a core document as per Definition 2 i.e. .e. the count of the found neighbors of this document 'pd' is greater than predefined threshold value of minimum number of different users, it is assigned to a new cluster, and all the neighbors are queued to Q for further processing. This explains step 6 – step 10 of the below said algorithm's pseudo code.

Step 3: Now the document 'pq' is fetched from the queue Q. If the fetched document is not already assigned to the current cluster, it is so assigned to the current cluster. Then, if the fetched document 'pq' is a core document as per Definition 2 i.e. the count of the found neighbors of this document 'pq' is greater than predefined threshold value of minimum number of different users, then all the neighbors are added to queue Q only if they are not already in queue Q. This job of adding the members is done by EnNniqueQueue function. This explains the step 11 – step 17 of the below said algorithm's pseudo code.

Algorithm: (ϵ, k, t) spatiotemporal-density-based clustering

Input: Data - dataset with coordinates, ϵ - neighborhood radius predefined threshold value,

k - cosine document similarity constant predefined threshold value, t – inter arrival time predefined threshold value, MinDocDifferentUsers – minimum number of different users predefined threshold value

Output:

STC - set of clusters

```
1   cid ← 1;
2   STC ← ϕ;
3   for i ← 1 to |D| do
4    pd ← di ∈  D;
5    if IsClustered (pd) == false then
6     N ← GetNeighborhood (pd, ϵ, k, t);
7     if |N| ≥ MinDoc_DifferentUsers then
8       stccid ← MakeNewCluster (cid, pd);
9       cid ← cid + 1;
10       EnQueue (Q, N);
11       while Q is not empty do
12         pq ← DeQueue (Q);
13         stccid ← stccid ∪  pq;
14         N ← GetNeighborhood (pq, ϵ, k, t);
15         if |N| ≥ MinDoc_DifferentUsers then
16           EnNniqueQueue (Q, N);
          end
17       STC ← STC ∪  stccid;
```

```
    end
  end
end
18 return STC;
```

Pseudo code (ϵ, k, t)-Density-based spatiotemporal clustering algorithm.

6.5 Workflow of (ϵ, k, t)-DBSCAN

As shown in the figure 12, following is the workflow of the program:

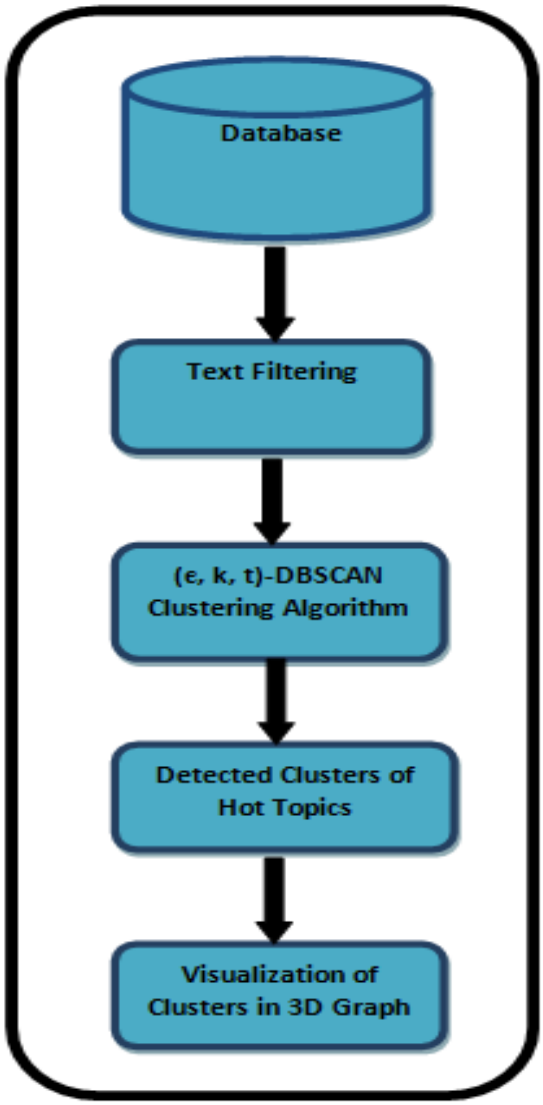


Figure 12 Workflow of (ϵ, k, t)-DBSCAN

Step 1. Twitter data is downloaded by external tool referred as SMM and saved as Twitter.sql file and this dataset shall be used by the algorithm in the next steps.

Step 2. Text filtering is done to clean up the noise from tweets and this involves many things which is further described in section 6.6.2

Step 3. Final data set is given to the algorithm and we get the set of clusters as a *.csv (comma separated value) file which can be used for cluster validation by external weka the data mining tool.

Step 4. Clustering results are visualized by using t-SNE in Python 3D scatter diagram.

6.6 Experiment

6.6.1 Dataset

In this study, twitter micro blogging post were collected from external tool named as SMM which is developed by Polous et al, 2014 at TUM Munich. From Twitter API one can only download the data of last 7 days. So to collect the larger data, Twitter data was weekly downloaded during the period 6.09.2014 - 08.11.2014(9 weeks) for Munich, Germany and finally it is manually merged in single data file. This dataset include URL, Id, owner, time stamp, text fields, latitude and longitude. An example of twitter record from dataset is shown below in table 2.

Twitter Tweet Example - https://twitter.com/Modnizzaksyu/status/506938112757747712 "506938112757747712" "105206531" "Опубликовано фото @ New Town Hall, Munich http://t.co/aL42vaU2pj :)" "1409698439" "48.1373478" "11.57560918"		
Query String Parameter	**Example of each parameter of tweet**	**Description**
Url	"https://twitter.com/Modnizzaksyu /status/506938112757747712"	This is a full link which is combination of Id column and user id column.
Id	"506938112757747712"	This column is unique in nature generated by Twitter system. Effectively this is a primary key of Tweet record when any user posts something in Twitter.
Owner	"105206531"	This is the user id of registered

		Twitter user.
Text	"Опубликовано фото @ New Town Hall, Munich http://t.co/aL42vaU2pj :)"	This is the textual content which is posted by Twitter user online from his/her id. This is the key field which gives us the text to find out what different users are discussing online in Twitter i.e. source of local topic under discussion among Twitter users.
Timestamp	"1409698439" (Time shown as since epoch i.e. 01.01.1970 00:00) 2/9/14 23:53 (Timestamp in human readable format)	This contains the time when any users posted some text on the twitter system online.
Geo-latitude and longitude	"48.1373478" "11.57560918"	Together these 2 values are the location of user when he/she posted something online in twitter from his/her Twitter account.

Table 2 Twitter Tweet Example

6.6.2 Text Preprocessing

Words which contain too many replicates or not includes the valid character are deleted. Weka stop word list is also used to clean the dataset for removing unwanted words. The text value shown in above **Example 1** is as below:

"Опубликовано фото @ New Town Hall, Munich http://t.co/aL42vaU2pj"

In order to get the meaningful results by the algorithm, we have to do preprocessing to filter out many things e.g.

- url's e.g. as shown above http://t.co/aL42vaU2pj. Such url's are of no interest for getting some meaningful local hot topics discussion related information, so they are filtered out.

- Certain special symbols e.g. as shown above :). For handling such special characters; regular expressions are used in this framework.
- Non-English characters are also removed. For this also, regular expressions are used in this framework to filter out such unwanted unicode characters.
- After above processing, WEKA stop words are also used to filter out the words which does not contain any specific and meaningful real world events e.g. A, An, the, on, at, becoming, but, despite ……the list goes on.

At the end of such special preprocessing, we get the preprocessed text result as "**New Town Hall, Munich**" (in above **Example 1**) which is the input to the algorithm for further extraction of local hot topics. This is the final text on which the cosine similarity is applied between two different documents of the data set of respective field i.e. text field during algorithm processing.

6.6.3 Cosine Similarity

As said in previous section 5.3, the technique used for text matching between 2 different georeferenced documents is cosine similarity. This measure computes the cosine of the angle between two feature vectors and is used frequently in text mining.

For example:

Text 1 "I'm enjoying Oktoberfest at Munich Germany"

Text 2 "I'm staying at Hilton-hotel Munich for Oktoberfest Germany"

I'm	1 1
enjoying	1 0
Oktoberfest	1 1
at	1 1
Munich	1 1
Germany	1 1
staying	0 1
Hilton-hotel	0 1
for	0 1

In above example we are not interested in the words itself, instead we are more interested in those two vertical vectors of counts of each word. We are going to decide how close these two texts are to each other by calculating the cosine distance between these 2 vectors, namely the cosine of the angle between them.

The two vectors corresponding to above text are:

a: [1, 1, 1, 1, 1, 1, 0, 0, 0]

b: [1, 0, 1, 1, 1, 1, 1, 1, 1]

and the Cosine distance/similarity between them is : 0.717137165601

These vectors (a and b above) are 9-dimensional.

6.6.4 Most Frequent Words

Once the proposed algorithm generates the cluster results, it also shows the most frequently discussed words for each generated cluster. This feature is configurable and user can configure it to show top 5 or top 3 or top 10 most discussed words among social media users in the given dataset. Default configuration is to show top 5 words.

6.6.5 Python t-SNE 3D scatter diagram

The clusters identified by the algorithm are visualized in 3D scatter plot using t-SNE (Please see figure 15 and 16 for more detail.). t-distributed stochastic neighbor embedding (Van der Maaten et al, 2008) is a machine learning algorithm for dimension reproduction. It was invented by Van der Maaten and Geoffrey Hinton. It is a well known technique to visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map, which can be visualized in a scatter diagram. According to the Van der Maaten, it converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

6.6.6 Parameter selection for proposed (ϵ, k, t)-DBSCAN algorithm

To experiment the algorithm's efficiency, Following values of different parameters are considered as shown in the below table 3. These values are considered after multiple runs of the program during development and seeing the output and manual verification from the dataset. During literature also review I got the information that such parameter values may be different for different type of dataset for example many things depends upon how much users are active and posting the information online in the social media; this means the values of different parameters for this algorithm which is good for Twitter might not be

suitable for Instagram or vice versa; which in itself indicates that this algorithm is sensitive to the input parameter values, which is well fact of DBSCAN algorithm.

Input Parameter Name	**Various Values Considered**					
Radius (km)	0.7	1	2	3		
Similarity Rate (Cosine Distance)	0.55	0.60	0.65	0.70	0.75	0.80
Time (Hours)	24	48	72			
Minimum Number Of Different Documents/Users	6	7	10			

Table 3 Different parameter values of (ϵ, k, t)-Density-based spatiotemporal clustering algorithm.

216 different parameters combinations was prepared based on above said values and checked all the clustering results with different parameters through cross validation in weka tool and best fit combination selected for algorithm. Final parameter for (ϵ, k, t)-Density-based spatiotemporal clustering algorithm were radius (ϵ = 2 km), t = 48 hrs (172800 in seconds), Cosine similarity (k) =0.70 and minimum number of documents of different users = 10.

Main motivation to consider the cosine similarity rate between 0.55 and 0.80 is because if cosine similarity is taken below 0.55, in that case algorithm had meaningless clusters and if it is greater than 0.80 in such case it is observed after seeing the clustering results that algorithm missed some very important local hot topics clusters which were discussed among the Twitter users.

Similarly, the radius is also taken in the range of 0.7 km – 3 km. The main motivation to keep this parameter value as short/long in the range depends on the kind of dataset, location of dataset, user's habit in the social media, where we want to run this algorithm. For current dataset in this experiment, these values are very good for Munich location to get the good results. As said previously also this radius can be different for different country and the user's habit of posting information in the social media. In this small range we get very refined clusters of local hot topic under discussion among Twitter users of Munich location, which is visible in the experiment results and also in manual verification.

Hardware used in this experiment setup was as below:

Processor: Intel Core i7 (4^{th} generation processor) 4210U and

RAM: 4GB and

Operating system: Linux Fedora 20.

Language: Python version 2.7.8

6.7 Cluster validation

Cluster validation is one of the important parts to evaluate the cluster results produced by (ϵ, k, t)-density-based spatial temporal clustering algorithm. There are many methods to explore the clustering results and researchers are using

- either 2D or 3D visualization of result to verify the validity of results or
- statistical/quantitative approach to check i.e., how well the clustering algorithm discovered the clusters from the data set (Halkidi et al., 2001) i.e. the goodness of cluster results.
- manual result verification.

In order to evaluate the clustering results, weka the data mining tool is used (Hall et al., 2009). WEKA version used for this purpose was "3.6.11".

Figure 13 details the workflow of cluster validation process, which contains mainly four steps.

First step – The output/result of (ϵ, k, t) Density based spatial temporal clustering algorithm (i.e. clustering result) which is in the form of comma separated value file (.csv file) of all the different combinations of algorithm input parameters are used as an input to the weka tool. To provide this Cluster-Result-Output.csv file as an input to weka tool is a manual step; however the Cluster-Result-Output .csv file is automatically generated by the (ϵ, k, t)-Density based spatial temporal clustering algorithm at the end.

Second step- In this step "clustering via classification" option is used in weka tool to see the correctly and incorrectly classified instances of clusters generated by (ϵ, k, t) - density based spatial temporal clustering algorithm as a result. The result produced by weka tool's "clustering via classification" option also includes F-measure, Recall and Precision values. The meanings of these terms are explained in below last step section.

Third step - 10 fold cross validation option (approach) is used on cluster result. According to professor Witten (Witten et al., 2011) "In cross-validation, data is divided in to 10 pieces. 9 pieces of data is used for training, and the last piece is taken for testing. Then, with the same division, another 9 pieces are use for training and the held-out piece for testing. We do the whole thing 10 times, using a different segment for testing each time. In other words, we divide the dataset into 10 pieces, and then we hold out each of these pieces in turn for testing, train on the rest, do the testing and average the 10 results. That would be '10-fold cross-validation'."

In last step, Result is generated by weka tool which includes correctly classifies instances, incorrectly classifies instances, F- measure, Recall and Precision values. The explanations of these terms are as below:

- **Correctly classified instances** (/points) shows the accuracy of precisely classified data points.
- **Incorrectly classified instances** (/points) shows inaccurate or wrong classified data points (instances).
- **Precision** - Precision is the fraction of correctly retrieved instances.
- **Recall** - Recall is the fraction of correctly retrieved instances out of all matching instances.
- **F- measure** considers both the precision and the recall of the test to compute the score in statistical analysis of any data. F-measure is weighted average of precision and recall, which is also known as harmonic mean of precision and recall. The value of F-measure is always between 0-1, where 0 value shows negative value and 1 shows the positivity of result, this means the higher the F-measure value tending towards 1 shows a better result and lower the F-measure value tending towards zero is bad result.

$$\text{F-measure} = 2^* (\text{Precision}^* \text{Recall}) / \text{Precision} + \text{Recall}$$

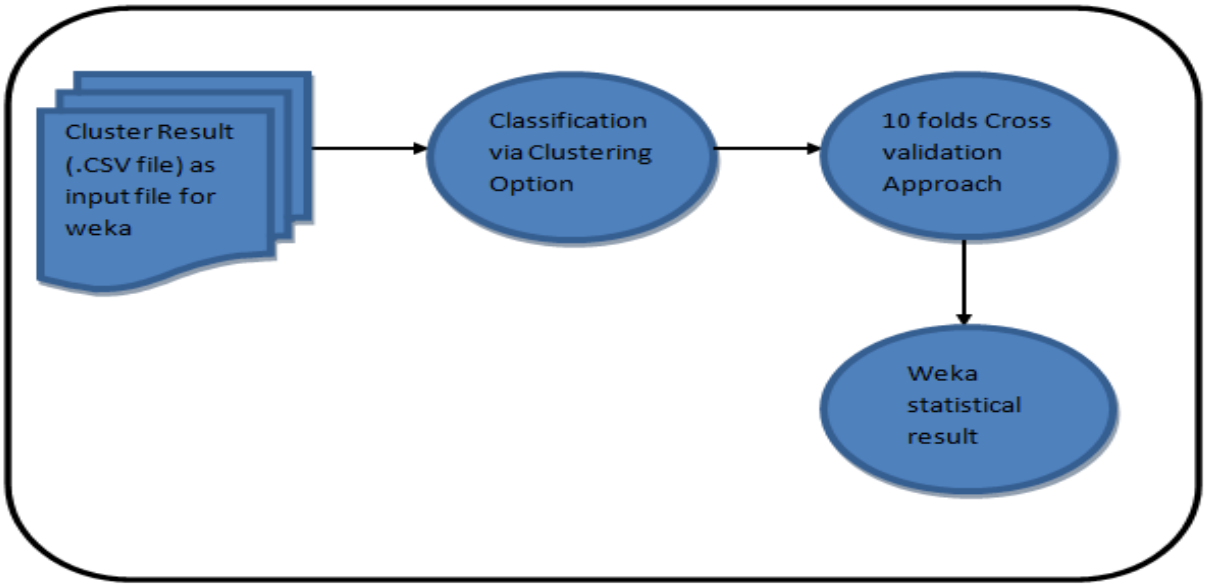


Figure 13 Workflow of statistical analysis

As mentioned previously that 216 combinations of different (ϵ, k, t) Density based spatial temporal clustering algorithm's input parameters are used to and generated out file is given to weka tool's one by one as an input to see the result about above mentioned parameters (i.e. correctly classifies instances, incorrectly classifies instances, F-1 measure, Recall and Precision values). Table 4 shows only best 5 results for different parameter values which are shown in the table itself. As said earlier the weka tool's result include the correctly classified instances, incorrectly instances, F measure, precision and recall values. Best result considered not only incorrectly classified points, but also taken in to account the result of precision, recall, F- measure to select the best input parameter combination of the (ϵ, k, t) Density based spatial temporal clustering algorithm. Apart from above; the manual verification of all the clustering results were also done and on the basis of all the results, it is found that time = 2 days, radius = 2 km, minimum number of different users =10 and document similarity rate constant k= 0.70 has best result and it is also able to reveal the local hot topic discussed among the Twitter users of the dataset used in the experiment.

	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	82.96	17.04	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

Table 4 Weka result comparison of different input parameter values

6.8 Comparison with DBSCAN algorithm

The proposed (ϵ, k, t)-Density-based spatiotemporal clustering algorithm's result is compared with DBSCAN to see how well the clustering algorithm discovers the desired (Thesis objective) clusters from the given Twitter data set. The parameter of (ϵ, k, t)-Density-based spatiotemporal clustering algorithm were set to radius (ϵ) = 2 km, t = 48 hrs (172800 in seconds), Cosine similarity (k) =0.70 and minimum number of different users = 10. The parameter of DBSCAN were set to radius (ϵ = 2 km), minimum number of documents (MinDoc) = 10. Table 5 shows the summary of the both the above said

algorithm's results, which are generated from WEKA machine learning tool of the respective algorithm's output i.e. clustering result of the respective algorithm.

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

Table 5 Comparison of (ϵ, k, t)-DBSCAN and DBSCAN results

As mentioned above the (ϵ, k, t)-Density-based spatiotemporal clustering algorithm has only 17.17 % incorrectly classified instances of the dataset whereas DBSCAN has 58.3% incorrectly classified instances. Other result i.e. Precision, Recall and F-measure also have significant difference between results of (ϵ, k, t)-Density-based spatiotemporal clustering and DBSCAN algorithm. All the cluster results are also manually verified and it is found that (ϵ, k, t)-Density-based spatiotemporal clustering algorithm has very meaningful clusters which reflects the real world event whereas DBSCAN have many clusters but results are not very promising for this aspect i.e. discovering the local hot locals among Twitter/social media users.

In addition, 3D scatter diagram using t-SNE is used to visualize the cluster result to verify the algorithm's results to see how well the proposed algorithm discovers the clusters of local hot topics from the data set. In figure 14, (ϵ, k, t)-DBSCAN 3D scatter diagram has 5 clusters shown in different colors. We can see them very clearly, however sometime we need to zoom and rotate the plot to see few of the points, which are hidden behind some front documents. It is clearly visible that each cluster is showing one unique hot topic discussed among the social media users of the dataset. Whereas figure 15 is 3D scatter diagram of DBSCAN algorithm, which is displaying 14 clusters however, it is not clearly visible from the diagram that it contains 14 clusters. The clusters itself are not showing the local hot topics instead they are just showing the clusters which are based on the spatial radius and location (longitude and latitude). It is not easy to interpret the result of DBSCAN from the 3D visualization.

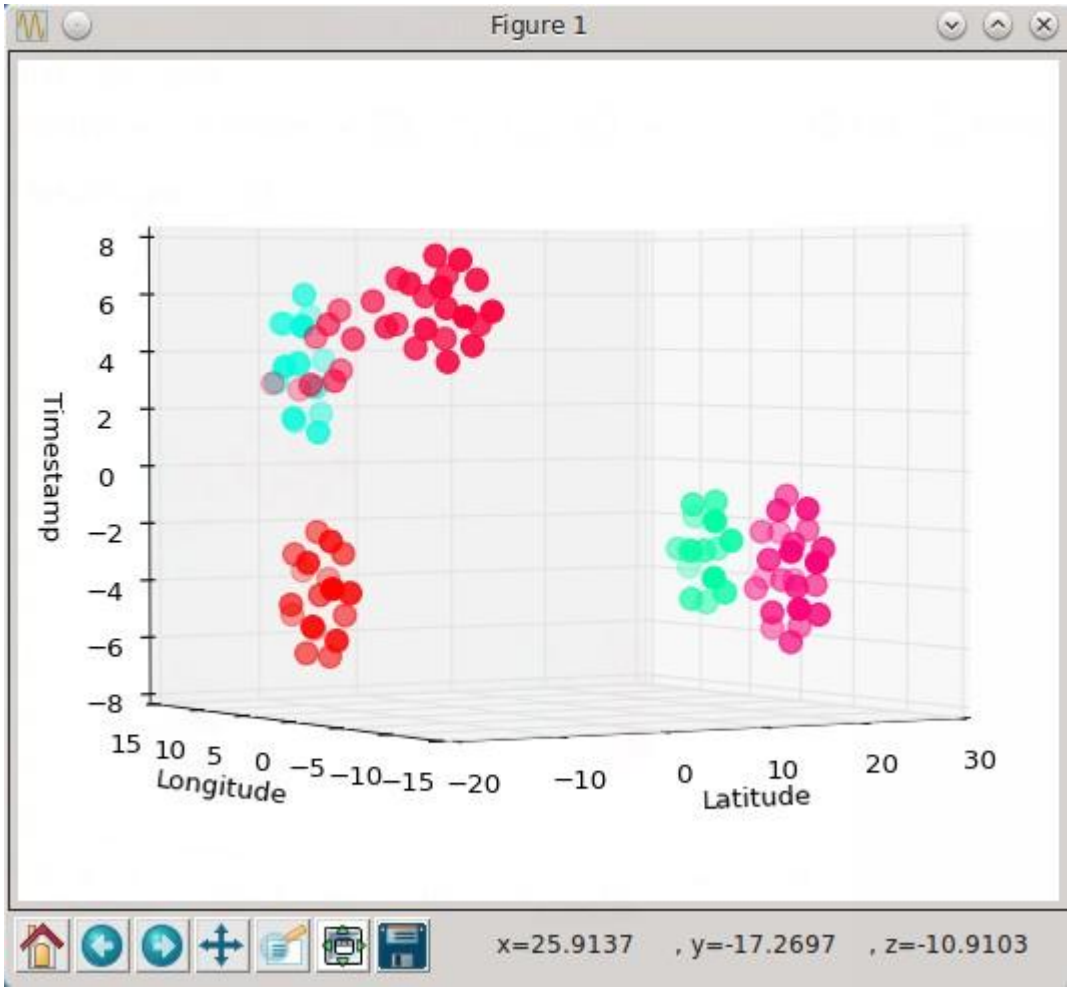


Figure 14 (ϵ, k, t)-DBSCAN cluster visualization in 3D scatter graph

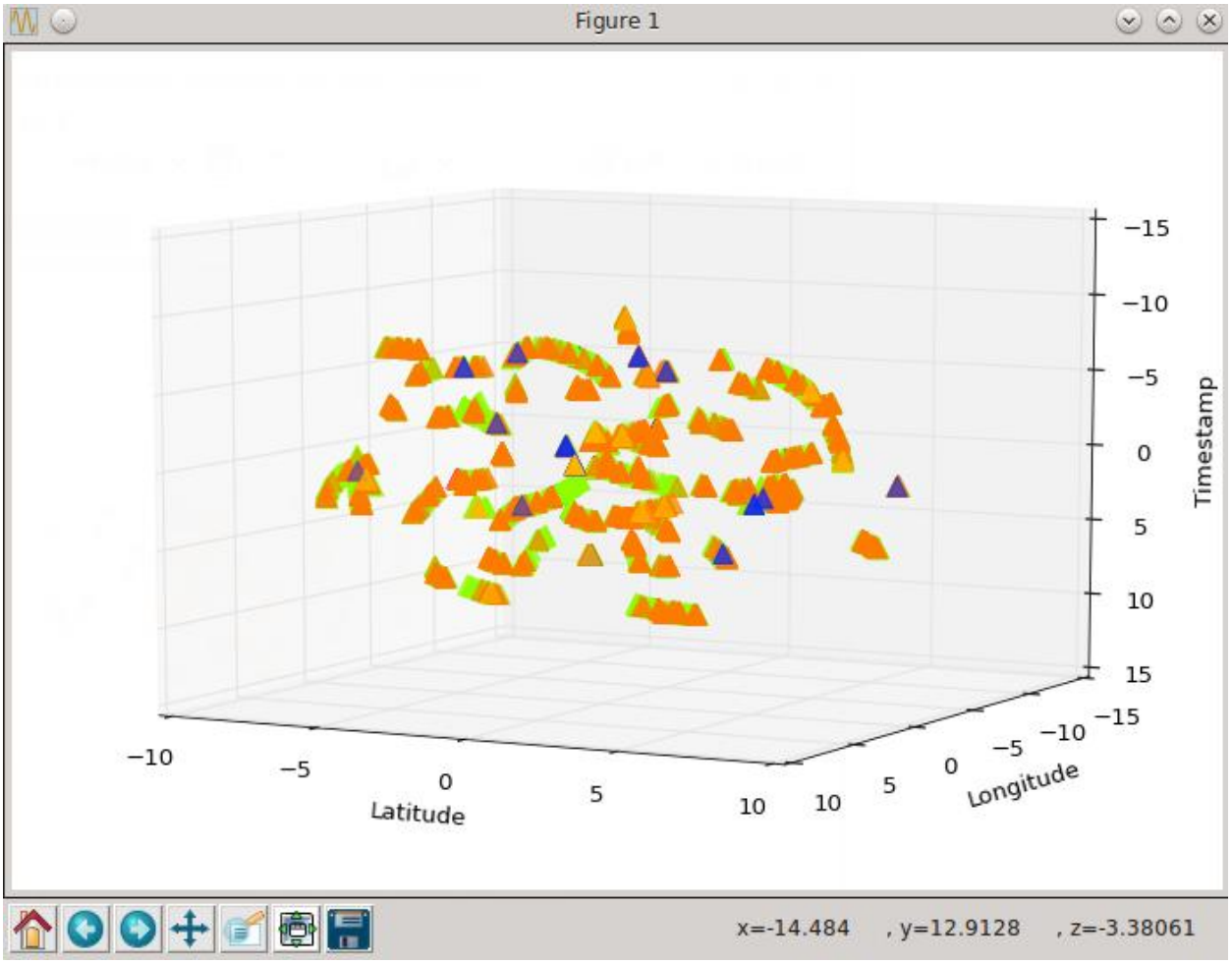


Figure 15 DBSCAN cluster visualization in 3D scatter graph

7 Cluster result discussion and visualization

This chapter contains the clustering result of (ϵ, k, t)-Density-based spatiotemporal clustering algorithm and DBSCAN algorithm and also the clustering result visualization of the documents that are part of clusters visualized in geographical maps etc via online visualization tools. For text visualization, Voyant online tool[8] and for cluster result visualization, Google map using Google fusion table[9] and Carto DB[10] tool are used.

7.1 (ϵ, k, t)-DBSCAN and DBSCAN cluster result discussion

Table 6 shows the characteristics of spatial temporal clusters extracted using (ϵ, k, t)-Density-based spatiotemporal clustering algorithm. This table includes rank, cluster number, the range of latitude and longitude, time stamp, most frequent words and local hot topic discussed among the users. (ϵ, k, t)-Density-based spatiotemporal clustering algorithm was able to detect 5 clusters from the dataset. Rank is assigned to each extracted cluster based on the number of tweets. Furthermore, we are able to find local hot topic or real world event based on the result of most frequent words discussed in each cluster. Rank 1 is assigned to cluster 3 because of the highest number of tweets in cluster 3 and on the basis of top 5 most frequent words discussed among the users.

Rank	Cluster Number	No of Tweets	Range of latitude	Range of longitude	Time	Most 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

Table 6 (ϵ, k, t)-DBSCAN cluster results

Manual verification was also done on the cluster results to assign event names to the cluster. In cluster 3, users were talking about Oktoberfest, which is one of the famous

Germany's festivals. It happens in Munich for 16 days in an area named Theresienwiese. It starts at the mid of September until first week of October every year.

Second rank assigned to cluster 1 which has 23 tweets in which people were discussing about Bayern München Vs Manchester match at Allianz arena stadium which was held on 17 September 2014. Rank 3 was assigned to cluster 5, rank 4 to cluster 4 and rank 5 to cluster 2. In cluster 5 people were discussing about Bayern München Vs Roma match, in cluster 4 people were discussing about Bayern München Vs Hannover match and in cluster 2 people were discussed about Bayern München Vs Paderborn match. These football matches were held at Allianz area, München. Based on cosine text similarity, proposed (ϵ, k, t)-Density-based spatiotemporal clustering algorithm was able to detect all real world events discussed as local hot topics among the Twitter users in the dataset. During manual verification of all the above said events on the internet, it is found that all the cluster result have promising results.

Rank	Cluster Number	No of Tweets	Range of latitude	Range of longitude	Time (duration)	Most Frequent words	Hot topic description (real world event)
1	Cluster 1	3044	48.0745388 - 48.22814874	11.42242258-11.70642367	2014-08-30 15:00:47 - 2014-11-08 00:54:00	Germany, Bavaria, oktoberfest, münchen, bayern	Oktoberfest
2	Cluster 2	1296	48.1229698 - 48.15201	11.55690142-11.60542916	2014-08-30 17:59:33 - 2014-11-07 20:44:16	vorhersage, wetter, qfe, qff, leiser	Not a valid event
3	Cluster 6	165	48.12830293 -48.15848837	11.54425979-11.56879072	2014-09-09 16:46:43 - 2014-11-07 15:01:12	germany, oktoberfest, Bavaria, sighting, landesleitung	Not a valid event
4	Cluster 9	62	48.10980211 - 48.132158	11.59502826-11.6330728	2014-09-17 06:30:02 - 2014-11-07 06:26:22	blog, bier, oktoberfest, musings, live,	Oktoberfest
5	Cluster 5	33	48.13081076 - 48.1872116	11.50598867 - 11.5704359	2014-09-02 13:18:28 - 2014-11-07 15:19:08	germany, münchen, Deutschland, bayern, bmw	Not a valid event

6	**Cluster 7**	**28**	**48.11607293 - 48.12414412**	**11.53559452-11.56824478**	**2014-09-10 23:50:32 - 2014-10-30 19:49:02**	**Rachel, flo, eigenbroduction, zf42, münchen**	**Not a valid event**
7	**Cluster 11**	**22**	**48.14082147 - 48.16382642**	**11.49973675 - 11.526408**	**2014-09-24 16:10:06 - 2014-11-06 11:20:39**	**Sieber, partner garden, münchen, botanic**	**Not a valid event**
8	**Cluster 3**	**20**	**48.04585266 - 48.07893911**	**11.59688934-11.62888033**	**2014-09-02 06:31:23 - 2014-11-06 09:35:31**	**Unterhaching, marienplatz, bayern, oktoberfest, love**	**Not a valid event**
9	**Cluster 8**	**16**	**48.15097687 - 48.16396334**	**11.6157685 - 11.62550725**	**2014-09-11 17:03:59 - 2014-11-03 11:26:22**	**westin , sheraton, arabellapark, münchen, grand**	**Not a valid event**
10	**Cluster 4**	**15**	**48.18403708 - 48.22086359**	**11.55294366-11.61093085**	**2014-09-02 07:11:51 - 2014-11-04 11:56:03**	**Germany, münchen, bayern, bavaria, bier**	**Not a valid event**
11	**Cluster 14**	**14**	**48.10485375 - 48.1273655**	**11.4296243 - 11.47847881**	**2014-09-01 08:04:56 - 2014-11-04 16:31:23**	**alberthammond, münchen, beer, munich, germany**	**Not a valid event**
12	**Cluster 12**	**13**	**48.10153094 - 48.11655702**	**11.5704559 - 11.58565004**	**2014-09-26 21:53:11 - 2014-11-05 07:57:41**	**Germany, bayern, digitals, oktoberbest, digital**	**Not a valid event**
12	**Cluster 13**	**13**	**48.13206954 - 48.13234716**	**11.69104099-11.69129305**	**2014-10-09 13:07:25 - 2014-11-07 13:10:21**	**messe, knig, ludwig, pastabar, mammas**	**Not a valid event**
13	**Cluster 10**	**10**	**48.12574983 -48.15150857**	**11.66920847-11.7016964**	**2014-09-22 23:03:33 2014-11-04 20:16:15**	**messe, ereal, münchen, investment, european**	**Not a valid event**

Table 7 DBSCAN cluster results

Table 7 shows the details of extracted spatial clusters using DBSCAN. This table includes rank, cluster number, the range of latitude and longitude, time stamp, most frequent words and local hot topic discussed among the users.

DBSCAN was able to detect 14 spatial clusters from the dataset. Table 7 shows that DBSCAN is able to detect only one event Oktoberfest in two different local areas. Only cluster 1 and cluster 9 were able to detect local hot topic as real world event although in manual examination of cluster results, it is found that they included some other local hot topic in cluster furthermore they included many combination of several local hot topics in single cluster. Due to many local hot topics included in single cluster, we are not able to detect any real world event in cluster results except cluster 1 and cluster 9.

It is clear that, in contrast to DBSCAN, the (ϵ, k, t)-Density-based spatiotemporal clustering algorithm was able to recognize most relevant spatial temporal clusters and further more it is able to detect real world events as local hot topic discussed among users. Figure 16 shows that (ϵ, k, t)-Density-based spatiotemporal clustering algorithm was able to detect 100% real world event whereas DBSCAN was able to detect only 14.29% relevant result.

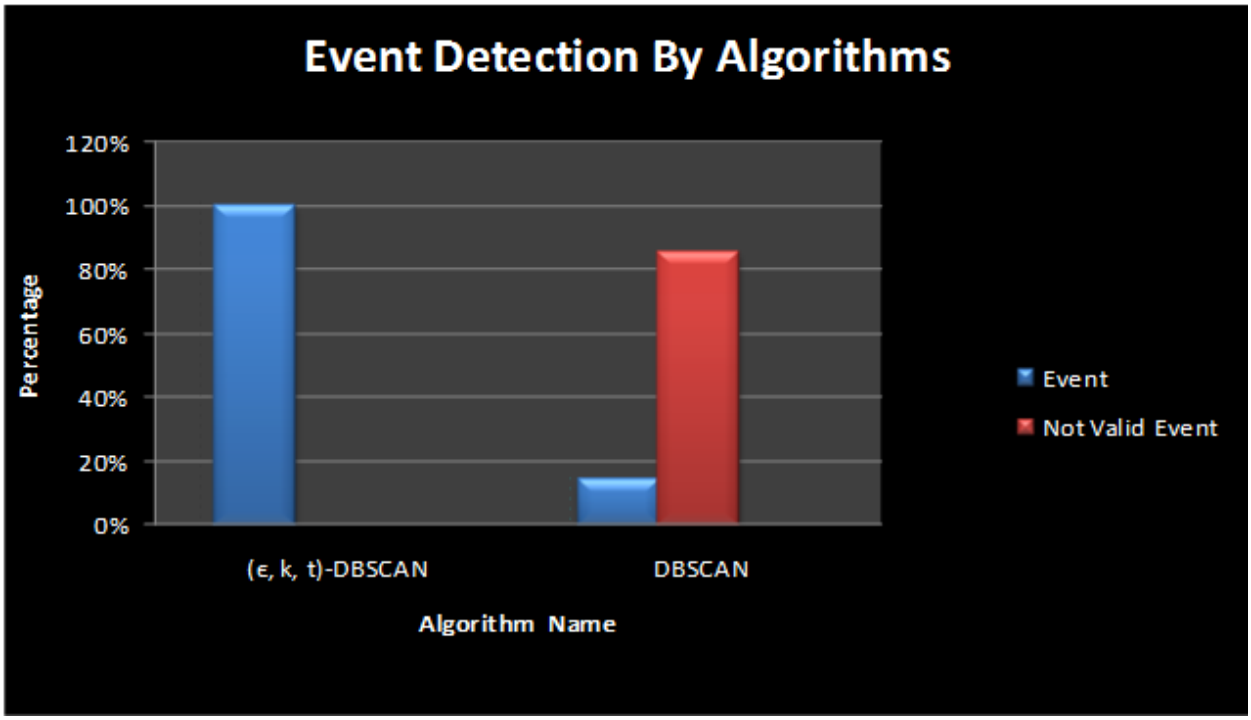


Figure 16 Statistics of event detection by two algorithms from the dataset

7.2 Visualization

Visualization plays an influential role in data analysis and is used in maps, scientific drawings, and data plots for over a thousand years. Visualization technique, facilitate the interpretation of a big amount of data to understand the meaning of the data, which is always easy to comprehend.

7.2.1 Text Visualization

There are many online text visualization tools available. For text visualization of (ϵ, k, t)-DBSCAN and DBSCAN algorithm, voyant online text visualization tool is applied on filtered text. Main advantage of this tool is that user does not need any programming skills before applying it on data. It is user-friendly web based text analysis tool. It accepts different file format i.e. text, MS word, Pdf, HTML and XML. User can also directly paste the text on text box. After getting input file, it visualizes the most frequent words and it also provides details about frequent words used in the file. Moreover, user can also apply stop word on dataset. User can export the result as URL of the tool with current tool or bibliographic citation or with HTML button for this tool and current data option. Figure 17 and 18 are the screenshot of (ϵ, k, t)-DBSCAN filter text and DBSCAN filter text respectively. Figure 18 shows the text visualization with detail of most frequent words used in dataset including their number of counts. In this figure 17, we can see that München, Bayern, Allianz, Arena and Oktoberfest Manchester, Roma, Hannover and Paderborn are the main word detected from the tool, which are actually the local hot topics itself and this reveals that all the events are captured here in the form of local hot topics.

Figure 17 Text visualization of (ϵ, k, t)-DBSCAN with number of counts

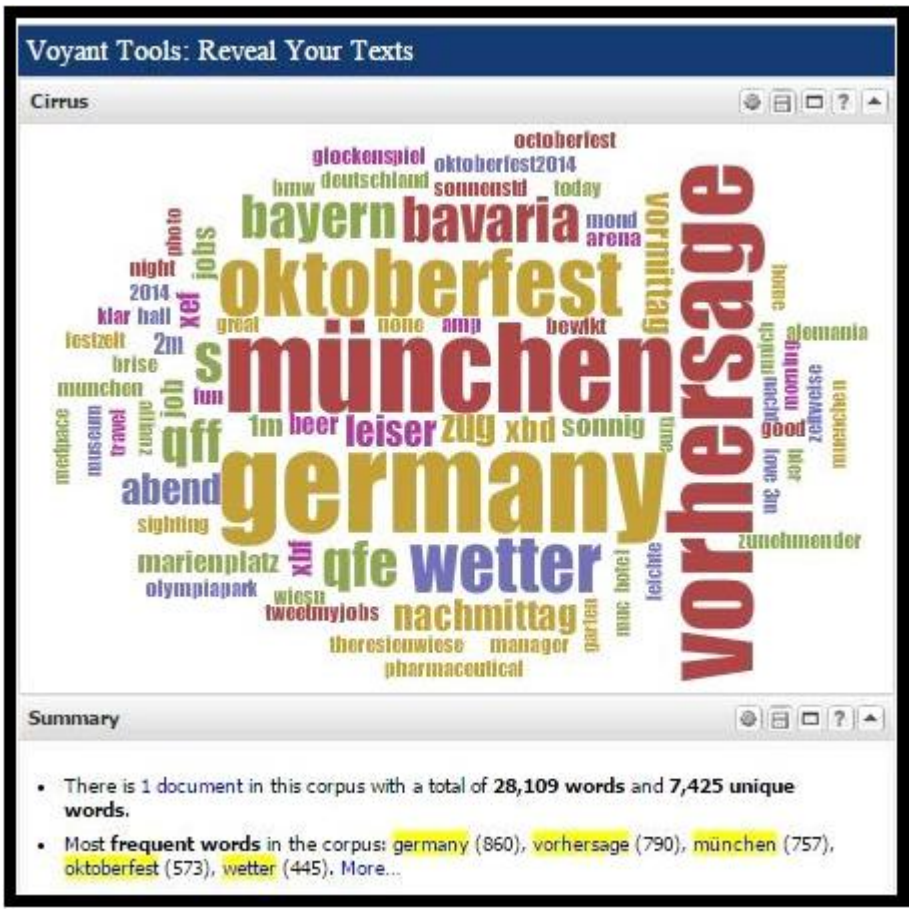


Figure 18 Text visualization of DBSCAN with number of counts

Figure 18 is a screen shot of DBSCAN algorithm's filter text and in this figure Germany, Bayern, Vorhesage, München, Oktoberfest, Bavaria, wetter qfe etc. are the main words detected by the tool. These words are detected as local hot topics from the DBSCAN algorithm but they are not the events except oktoberfest word, which means DBSCAN is able to reveal only one event i.e. Oktoberfest as a result.

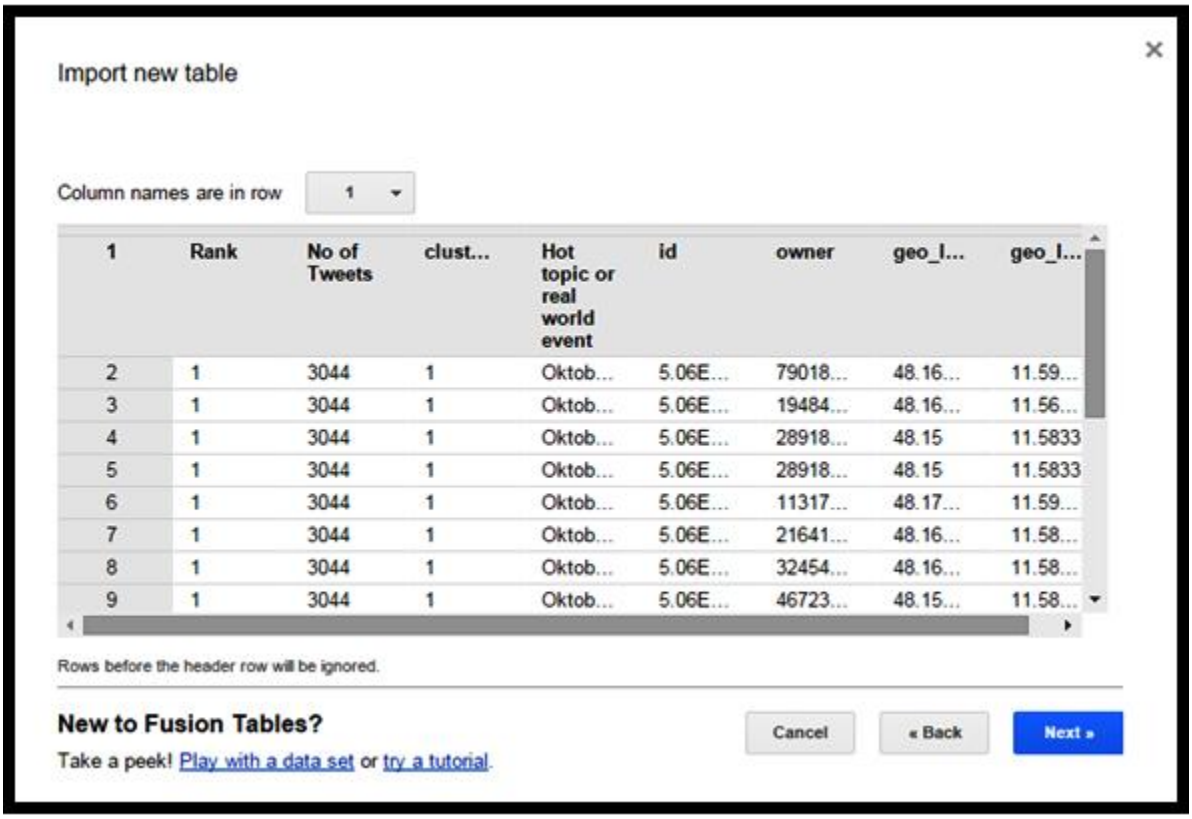
7.2.2 Online Cluster Visualization

Google Maps using the Google Fusion Table and for the online cluster visualization CartoDB online tool is used to visualize the cluster results.

7.2.2.1 Google Fusion Table

The Google Fusion Table is applied to visualize the cluster results. The Google Fusion Table is a web service provided by Google for data management. The user can visualize the data in many forms for example as a bar chart, pie chart, scatter plot or line plot on a geographical map. The user only needs a google account to make a map by Fusion Table on

Google Maps. The user can upload the data from an excel file, CSV file, KML file or a spreadsheet. After importing the data file, our data is shown in figure 19. Here user has to select the location column to visualize the data location on map. User can choose either geo latitude and longitude or address of the location.



1	Rank	No of Tweets	clust...	Hot topic or real world event	id	owner	geo_l...	geo_l...
2	1	3044	1	Oktob...	5.06E...	79018...	48.16...	11.59...
3	1	3044	1	Oktob...	5.06E...	19484...	48.16...	11.56...
4	1	3044	1	Oktob...	5.06E...	28918...	48.15	11.5833
5	1	3044	1	Oktob...	5.06E...	28918...	48.15	11.5833
6	1	3044	1	Oktob...	5.06E...	11317...	48.17...	11.59...
7	1	3044	1	Oktob...	5.06E...	21641...	48.16...	11.58...
8	1	3044	1	Oktob...	5.06E...	32454...	48.16...	11.58...
9	1	3044	1	Oktob...	5.06E...	46723...	48.15...	11.58...

Figure 19 Data in fusion table

After loading the data on map, user can change map style and cluster marker colors. If many clusters have same location, in that case all the clusters will be overlapped to each other. In this situation, if we want to see all the clusters, user can select the find option to search the cluster and by clicking on the cluster, user can see the more details about the cluster.

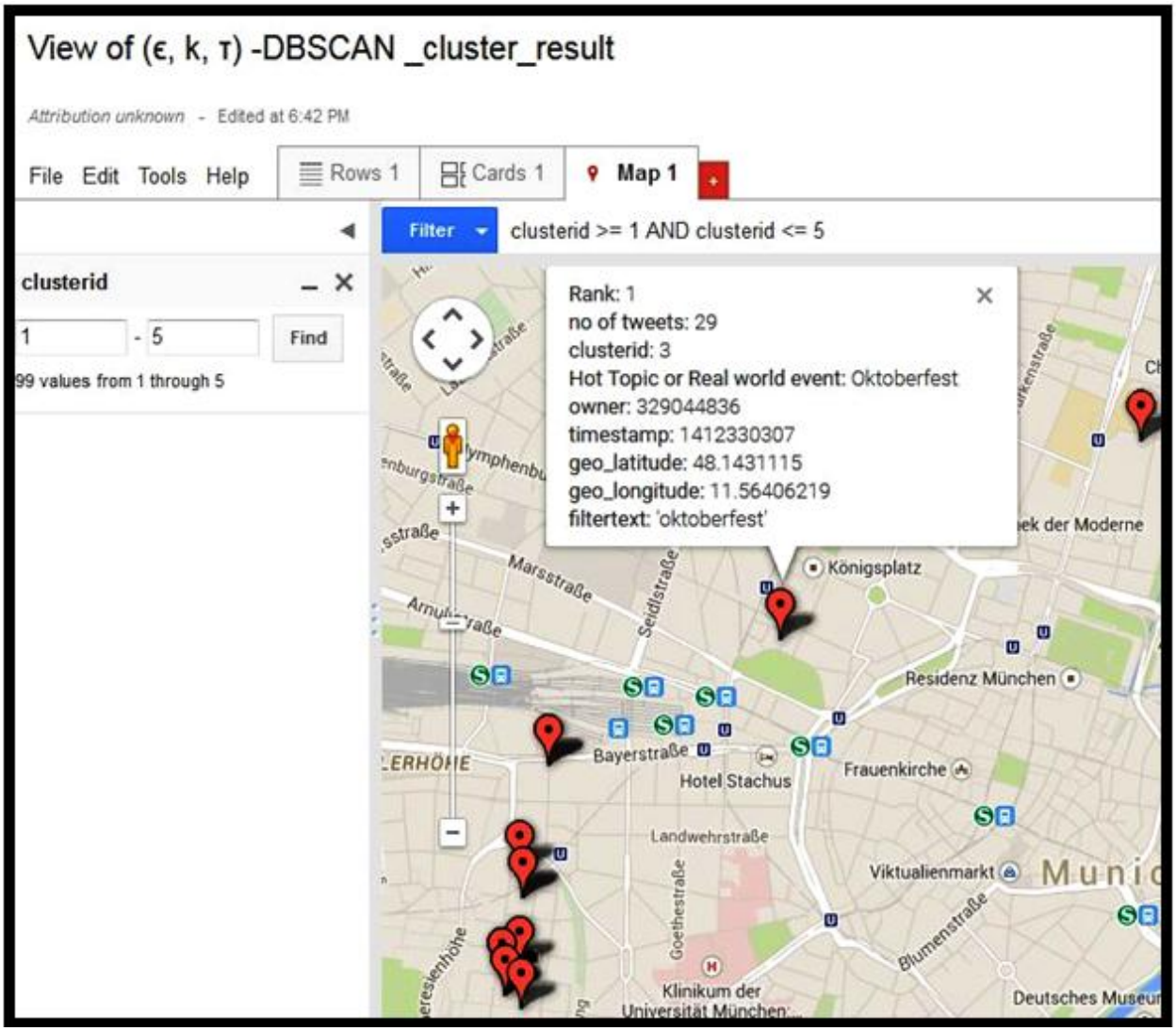


Figure 20 Screen shot of (ϵ, k, t)-DBSCAN result on google map

Figure 20 is the screen shot of (ϵ, k, t)-DBSCAN cluster results. In this screen shot all the clusters are shown in different colors but out of five clusters four are overlapping on to each other because they are four different real world events which happened at same place but at different times and the place is the football stadium at Allianz arena, Munich. The red color documents shown in above figure 20 belong to cluster id 3, which is the Oktoberfest event. The different information associated with this document is shown for only one document in above figure 20, however if we click on other documents the similar information can be seen. For all of the above red color documents the longitude, latitude, owner, timestamp will be different however rank, number of tweets, cluster id, filter text will remain same for all other documents.

7.2.2.2 CartoDB

CartoDB[10] is an open source tool that allows storage and visualization of geospatial data on the web. According to CartoDB editor[12], CartoDB accepts data in different formats (Excel, CSV, XML, SHP, and GeoJSON) and from various sources. User needs a CartoDB account and internet connection to create a map on CartoDB. User can share the map publically or private and in addition can share the map by providing the link/url. Map author can also embed the map on the website. CartoDB's, Isarithmic map and animated maps are shown in below section.

7.2.2.2.1 Isarithmic Map

The maps that represent data sets that have a "continuous distribution and smooth change in value'' (Kraak et al., 2003). After the Choropleth map, the isarithmic map is the most widely used mapping technique. Main advantage of isarithmic map is that the map shows the distribution of a spatial phenomenon that varies spatially. In isarithmic map, data type must be quantitative. In CartoDB tool, Intensity map option is used to create isarithmic map. Intensity map is very useful to represent the density of data points. It measures the density of data points by darkening the areas with many points in contrast to those with fewer points. Intensity map visualization can be also be defined as combining all the points on the map and showing it with more intense color. Multiple points (i.e. dense area) are represented by one single intense color. In figure 21, the yellow color represents the area with fewer points (low dense areas) whereas, the areas with many points (high density area) are shown in red color. Each intense color point also shows other information for example the latitude, the longitude, the cluster id.

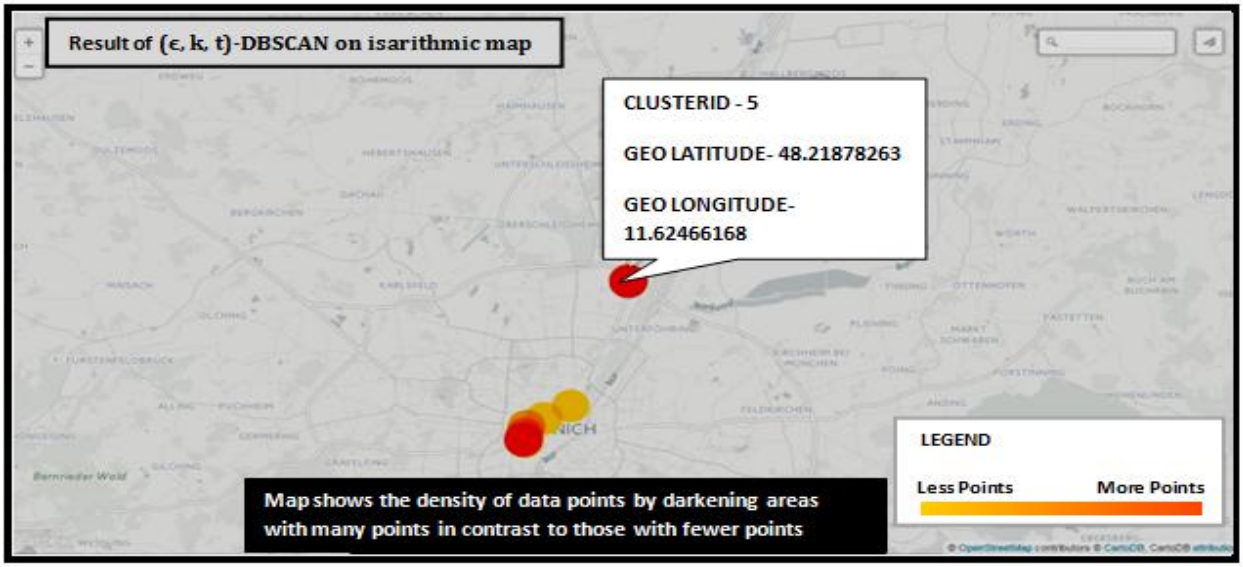


Figure 21 Screenshot of (ϵ, k, t)-DBSCAN Isarithmic map on CartoDB

7.2.2.2.2 Animated Map

Animated Map continuously shows different points (points refer here a cluster) on timeline during viewing the map. According to (DiBiase.1992) animated maps can emphasize the existence of an occurrence at a location, emphasize an attribute of an occurrence or representing change in the position or attributes of an occurrence. CartoDB Torque Cat option is used to describe the (ϵ, k, t)-DBSCAN result as animated map. Torque Cat is ideal for displaying the spatial-temporal data. This tool animates a progression of points based on a table column containing the time stamp[10]. The map shown in figure 22 visualizes the (ϵ, k, t)-DBSCAN result as it changes over time. This map represents five clusters with unique local hot topic, which happened at a specific period and represented by different colors. As the time slider moves in below figure, the different clusters appear on the map.

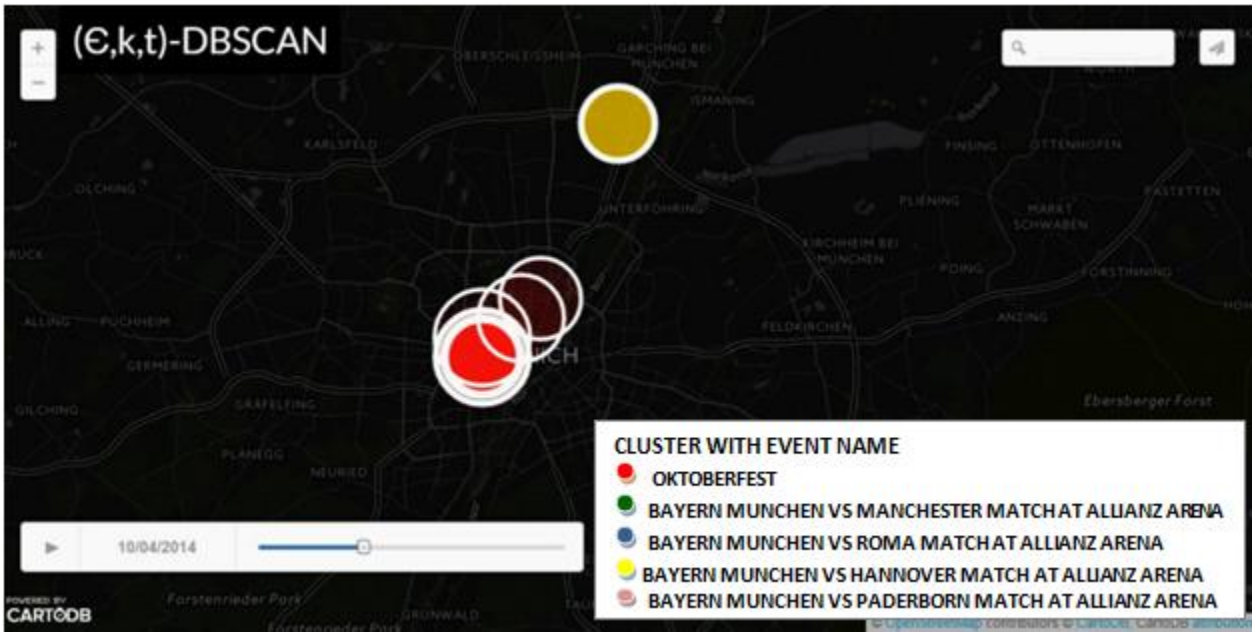


Figure 22 Screenshot of (ϵ, k, t)-DBSCAN result on Animated map.

8 Conclusions and future work

The primary objective of this thesis was to get insight into spatial temporal social media data to detect any kind of significant changes named as event. The event has been defined as any anomalous user activity, which happened at a specific time or within a particular period at a particular location.

To accomplish the above goal, the DSC algorithm from Tamura et al. (2013) was identified as the base algorithm after the literature review of many scientific research papers in the similar domain. The next challenge was optimization and modification of the selected algorithm in order to get the thesis's objective and desired results. This optimized version of the algorithm is proposed as the (ϵ, k, t)-Density-based spatial temporal clustering algorithm, which is an extension of the chosen DSC algorithm. In this (ϵ, k, t)-Density-based spatial temporal clustering algorithm, another new dimension is added as an input parameter to the algorithm which is referred to as "similarity rate constant (parameter k)". Furthermore, the parameter MinDoc is changed to the minimum number of documents of different users (parameter $MinDoc_{DifferentUsers}$). This is done in order to receive better and refined results and to fulfill the thesis's objective. Better and refined results means hereby cluster results, which are able to reveal the events from the social media dataset based on the user defined algorithm input parameters.

The basic DBSCAN algorithm is sensitive to its input parameters. During experimentation, 216 input parameter combinations were executed with the algorithm and the results were further cross-verified with the clustering via the classification option in the weka tool for all 216 combinations. I have manually verified all the cluster results because I believe involvement of human analysts is mandatory to reflect and rate the information transported by the Twitter tweets. Following are the most optimal parameters for the given dataset for (ϵ, k, t)-Density-based spatiotemporal clustering algorithm. Spatial radius ϵ = 2 km, timestamp t = 48 hrs (172800 in seconds), Cosine similarity constant k = 0.70 and minimum number of documents of different users = 10.

In the experiment (section 7.1), it is evident that the changes done in this algorithm are capturing all the real world events in the form of 5 clusters from the Twitter dataset used (study period of 9 weeks covering Munich). Whereas DBSCAN captures 14 clusters and 12 clusters as noise. All the clusters, which are captured by the proposed algorithm are shown in section 7.1 of (ϵ, k, t)-DBSCAN algorithm's execution result which captured the following five real world events:

- The Oktoberfest

- The football match Bayern München Vs Manchester City at the Allianz arena
- The football match Bayern München Vs Roma at the Allianz arena
- The football match Bayern München Vs Hannover at the Allianz arena
- The football match Bayern München Vs Paderborn at the Allianz arena

This gives a positive indication of the accuracy of the optimized (ϵ, k, t)-Density-based spatial temporal clustering algorithm.

In this study we define the (ϵ, k, t) neighborhood of geo-referenced documents to extract semantically similar spatial and temporal clusters.

The main advantages of this algorithm are:

- The proposed algorithm is able to reveal all the events from the datasets based on the user defined algorithm input parameters. The input parameters have a decisive impact on the cluster result.
- The proposed algorithm can extract spatial, temporal and semantic based clusters, which allow users to identify local hot topics under discussion among social media users.
- It is suitable for any text based social media dataset to reveal the local hot topics and further revealing the events. Certain extra preprocessing might be required for some datasets except Twitter and Instagram to remove the noise.

During the Implementation of the (ϵ, k, τ)-Density-based spatial temporal clustering algorithm many problems occurred. For example, the parsing of smileys and other special characters from the Twitter dataset's text fields remains an obstacle. Text pre-processing, i.e. to filter out the noise (for example non-English text handling) from the tweet sentence, took lot of time. This was iterative work during algorithm implementation. It was solved by the use of regular expressions in python along with WEKA stop words usage. To receive the correct report generation from the clustering result was also an issue that had to be solved, i.e. range of minimum and maximum longitude/latitude and timestamp in each cluster. The solution was to correctly update the right variables during report generation, which was a little tedious job.

Future Work

The objective of this thesis has been achieved but there are still few aspects which can be further optimized. For example, a real time data downloading capability should be added to the framework in order to minimize the human involvement for data set preparation work. There is also further research scope of enhancing the speed of the algorithm execution by using other nearest neighbors learning algorithms in this framework for example Ball Tree or KDTree.

9 Bibliography

9.1 Books, Journals, articles and conference proceedings

- Abulaish, M., & Bhat, S. Y. (2014). A Density-Based Approach to Detect Community Evolutionary Events in Online Social Networks. In State of the Art Applications of Social Network Analysis (pp. 189-208). Springer International Publishing.

- Aggarwal, C. C., & Subbian, K. (2012). Event Detection in Social Streams. In SDM (Vol. 12, pp. 624-635).

- Ahlqvist, T., Bäck, A., Heinonen, S., & Halonen, M. (2010). Road-mapping the societal transformation potential of social media. foresight, 12(5), 3-26.

- Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. In ACM Sigmoid Record (Vol. 28, No. 2, pp. 49-60). ACM.

- Bao, B. K., Min, W., Lu, K., & Xu, C. (2013). Social event detection with robust high-order co-clustering. In Proceedings of the 3rd ACM conference on International conference on multimedia retrieval (pp. 135-142). ACM.

- Becker, H., Naaman, M., & Gravano, L. (2010). Learning similarity metrics for event identification in social media. In Proceedings of the third ACM international conference on Web search and data mining (pp. 291-300). ACM.

- Becker, H., Naaman, M., & Gravano, L. (2011). Beyond Trending Topics: Real-World Event Identification on Twitter. ICWSM, 11, 438-441.

- Berkhin, P. (2006). A survey of clustering data mining techniques. In grouping multidimensional data (pp. 25-71). Springer Berlin Heidelberg.

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

- Cao, F., Ester, M., Qian, W., & Zhou, A. (2006). Density-Based Clustering over an Evolving Data Stream with Noise. In SDM (Vol. 6, pp. 326-337).

- 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.

- Chen, Y., & Tu, L. (2007). Density-based clustering for real-time stream data. In Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 133-142). ACM.

- Cheng, W., Li, A., & Gao, B. (2012). Scenic discover and representation based on social media. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on (pp. 1013-1017). IEEE.
- DiBiase, D. (1992). Animation and the role of Map Design. Cartography and Geographic Information Systems 19(4): 201–214, 165–266.
- Duan, D., Li, Y., Li, R., & Lu, Z. (2012). Incremental K-clique clustering in dynamic social networks. Artificial Intelligence Review, 38(2), 129-147.
- Duan, L., Xu, L., Guo, F., Lee, J., & Yan, B. (2007). A local-density based spatial clustering algorithm with noise. Information Systems, 32(7), 978-986.
- Dunn, J. C. (1973). A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.
- Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Kdd (Vol. 96, pp. 226-231).
- Fayyad, U. & Uthurusamy, R. (1999)"Data mining and knowledge discovery in databases: Introduction to the special issue," Communications of the ACM, 39(11).
- Fisher, D. H. (1987). Knowledge acquisition via incremental conceptual clustering. Machine learning, 2(2), 139-172.
- Gao, X., Cao, J., He, Q., & Li, J. (2013). A novel method for geographical social event detection in social media. In Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service (pp. 305-308). ACM.
- Goyal, N., Goyal, P., Venkatramaiah, K., Deepak, P. C., & Sannop, P. (2011). An efficient density based incremental clustering algorithm in data warehousing environment. In 2009 International Conference on Computer Engineering and Applications, IPCSIT (Vol. 2).
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. Journal of Intelligent Information Systems, 17(2-3), 107-145.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: an update. ACM SIGKDD explorations newsletter, 11(1), 10-18.
- Hammouda, K. M., & Kamel, M. S. (2003). Incremental document clustering using cluster similarity histograms. In Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on (pp. 597-601). IEEE.
- Han, J., & Kamber, M. (2012). Data Mining: Concepts and Techniques. Morgan Kaufmann.
- Hinneburg, A., & Keim, D. A. (1998). An efficient approach to clustering in large multimedia databases with noise. In KDD (Vol. 98, pp. 58-65).

- Huang, A. (2008). Similarity measures for text document clustering. In Proceedings of the sixth New Zealand computer science research student conference (NZCSRSC2008), Christchurch, New Zealand (pp. 49-56).

- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. Data mining and knowledge discovery, 2(3), 283-304.

- Jain, A. K. & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

- Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. Business horizons, 53(1), 59-68.

- Kaufman, L. & Rousseeuw, P.J. (1987). Clustering by Means of Medoids, In Dodge, Y. (ed.), Statistical Data Analysis, based on the L1 Norm and Related methods, Elsevier/North Holland, Amsterdam. pp. 405-416,

- Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

- 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.

- Kohonen, T. (2001). Self-organizing maps. Springer Series in Information Sciences.

- Korenius, T., Laurikkala, J., & Juhola, M. (2007). On principal component analysis, cosine and Euclidean measures in information retrieval. Information Sciences, 177(22), 4893-4905.

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

- Kriegel, H. P., Kröger, P., Ntoutsi, I., & Zimek, A. (2011). Density based subspace clustering over dynamic data. In Scientific and Statistical Database Management (pp. 387-404). Springer Berlin Heidelberg.

- Lee, C. H. (2012). Mining spatio-temporal information on micro blogging streams using a density-based online clustering method. Expert Systems with Applications, 39(10), 9623-9641.

- Lee, C. H., Yang, H. C., Wen, W. S., & Weng, C. H. (2012). Learning to explore spatio-temporal impacts for event evaluation on social media. In Advances in Neural Networks–ISNN 2012 (pp. 316-325). Springer Berlin Heidelberg.

- Lee, K., Palsetia, D., Narayanan, R., Patwary, M. M. A., Agrawal, A., & Choudhary, A. (2011). Twitter trending topic classification. In Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on (pp. 251-258). IEEE.

- Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for Twitter-based geo-social event detection. In Proceedings of the 2nd ACM SIGSPATIAL international workshop on location based social networks (pp. 1-10). ACM.

- Li, N., & Wu, D. D. (2010). Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems, 48(2), 354-368.

- Liu, P., Zhou, D., & Wu, N.(2007). "Varied Density Based Spatial Clustering of Application with Noise" In proceedings of IEEE Conference ICSSSM, pg 528-531.

- Liu, Q., Deng, M., Shi, Y., & Wang, J. (2012). A density-based spatial clustering algorithm considering both spatial proximity and attribute similarity. Computers & Geosciences, 46, 296-309.

- Macqueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

- Mourya, M., & Prasad, P. (2013). An Effective Execution of Diabetes Dataset Using WEKA. (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (5), 2013, 681-682

- Mu, J., Fei, H., & Dong, X. (2008). A Parameter-Free Clustering Algorithm Based on Density Model. In Young Computer Scientists, 2008. ICYCS 2008. The 9th International Conference for (pp. 1825-1831). IEEE.

- Nanni, M., & Pedreschi, D. (2006). Time-focused clustering of trajectories of moving objects. Journal of Intelligent Information Systems, 27(3), 267-289.

- Nitta, N., Kumihashi, Y., Kato, T., & Babaguchi, N. (2014, January). Real-World Event Detection Using Flickr Images. In Multimedia Modeling (pp. 307-314). Springer International Publishing.

- Nosovskiy, G. V., Liu, D., & Sourina, O. (2008). Automatic clustering and boundary detection algorithm based on adaptive influence function. Pattern Recognition, 41(9), 2757-2776.

- Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. In Proceedings of the 2008 ACM symposium on Applied computing (pp. 863-868). ACM.

- Parikh, R., & Karlapalem, K. (2013). Et: events from tweets. In Proceedings of the 22nd international conference on World Wide Web companion (pp. 613-620). International World Wide Web Conferences Steering Committee.

- Parimala, M., Lopez, D., & Senthilkumar, N. C. (2011). A survey on density based clustering algorithms for mining large spatial databases. International Journal of Advanced Science and Technology, 31(1).

- Pelleg, D., & Moore, A. W. (2000). X-means: Extending K-means with Efficient Estimation of the Number of Clusters. In ICML (pp. 727-734).

- Peter, J.H., Antonysamy, A. (2010) ``An Optimised Density Based Clustering Algorithm" International Journal of Computer Applications (0975 – 8887) Volume 6– No.9.

- Petkos, G., Papadopoulos, S., & Kompatsiaris, Y. (2012). Social event detection using multimodal clustering and integrating supervisory signals. In Proceedings of the 2nd ACM International Conference on Multimedia Retrieval (p. 23). ACM.

- Polous, K., Mooney, P., Krisp, J. M., & Meng, L. (2013). Mining Event-Related Knowledge from OpenStreetMap. In Progress in Location-Based Services (pp. 109-124). Springer Berlin Heidelberg.

- Polous, K., Freitag, A., Krisp, J., Meng, L., Singh, S. (2014). A General Framework for Event Detection from Social Media in; Joint International Conference on Geospatial Theory, Processing, Modeling and Applications, Toronto, Canada. October 6-8, 2014

- Popovici, R., Weiler, A., & Grossniklaus, M. (2014). On-line Clustering for Real-Time Topic Detection in Social Media Streaming Data. In SNOW-DC@ WWW (pp. 57-63).

- Procopiuc, C. M., & Procopiuc, O. (2005). Density estimation for spatial data streams. In Advances in Spatial and Temporal Databases (pp. 109-126). Springer Berlin Heidelberg.

- Ram, A., Jalal, S., Jalal, A.S., & kumar, M. (2010). "A density Based Algorithm for Discovery Density Varied cluster in Large spatial Databases", International Journal of Computer Application Volume 3, No.6.

- 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.

- Roy, S., & Bhattacharyya, D. K. (2005). An approach to find embedded clusters using density based techniques. In Distributed Computing and Internet Technology (pp. 523-535). Springer Berlin Heidelberg.

- Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World Wide Web (pp. 851-860). ACM.

- Samangooei, S., Hare, J., Dupplaw, D., Niranjan, M., Gibbins, N., Lewis, P. H., & Preston, J. (2013). Social Event Detection via sparse multi-modal feature selection and incremental density based clustering.

- Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (1998). Density-based clustering in spatial databases: The algorithm gdbscan and its applications. Data mining and knowledge discovery, 2(2), 169-194.

- Sato, Y., Kawashima, H., Okuda, H., & Oku, M. (2008). Trend-based Document Clustering for Sensitive and Stable Topic Detection. In PACLIC (pp. 331-340).

- Sayyadi, H., Hurst, M., & Maykov, A. (2009). Event Detection and Tracking in Social Streams. In ICWSM.

- Schölkopf, B., Smola, A., & Müller, K. R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. Neural computation, 10(5), 1299-1319.

- Shi, J., Mamoulis, N., Wu, D., & Cheung, D. W. (2014). Density-based Place Clustering in Geo-Social Networks. In Proceedings of the 2014 ACM SIGMOD international conference on Management of data (SIGMOD '14). ACM, New York.

- Shou, S., Zhou, A., Fan, Y., & Qian, W. (2000). "A Fast DBSCAN Algorithm" Journal of Software: 735-744.

- Singhal, A., & Seborg, D. E. (2005). Clustering multivariate time-series data. Journal of chemometrics, 19(8), 427-438.

- Sutanto, T., & Nayak, R. (2014). The Ranking Based Constrained Document Clustering Method and Its Application to Social Event Detection. In Database Systems for Advanced Applications (pp. 47-60). Springer International Publishing.

- Székely, E., Poncelet, P., Masseglia, F., Teisseire, M., & Cezar, R. (2013). A density-based backward approach to isolate rare events in large-scale applications. In Discovery Science (pp. 249-264). Springer Berlin Heidelberg.

- 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.

- Tan, Z., Zhang, P., Tan, J., & Guo, L. (2014). A Multi-layer Event Detection Algorithm for Detecting Global and Local Hot Events in Social Networks. Procedia Computer Science, 29, 2080-2089.

- Unankard, S., Li, X., & Sharaf, M. A. (2013). Location-based emerging event detection in social networks. In Web Technologies and Applications (pp. 280-291). Springer Berlin Heidelberg.

- Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.

- Viswanath, P., & Pinkesh, R. (2006). L-dbscan: A fast hybrid density based clustering method. In Pattern Recognition, 2006. ICPR 2006. 18th International Conference on (Vol. 1, pp. 912-915). IEEE.

- Wang, M., Wang, A., & Li, A. (2006). Mining spatial-temporal clusters from geo-databases. In Advanced Data Mining and Applications (pp. 263-270). Springer Berlin Heidelberg.

- Wang, Y., Xie, L., & Sundaram, H. (2011). Social Event Detection with clustering and filtering. In Mediaeval.

- Witten, I. H., & Frank, E., & Hall, M. (2011). "Data Mining: Practical machine learning tools and techniques, 3rd Edition". Morgan Kaufmann, San Francisco. Retrieved 2011-01-19.

- Xiaolin, Y., Xiao, Z., Nan, K., & Fengchao, Z. (2013). An improved Single-Pass clustering algorithm internet-oriented network topic detection. In Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on (pp. 560-564). IEEE.

- Xie, K., Xia, C., Grinberg, N., Schwartz, R., & Naaman, M. (2013). Robust detection of hyper-local events from geotagged social media data. In Proceedings of the Thirteenth International Workshop on Multimedia Data Mining (p. 2). ACM.

- Xu, X., Ester, M., Kriegel, H. P., & Sander, J. (1998). A distribution-based clustering algorithm for mining in large spatial databases. In Data Engineering, 1998. In Proceedings of 14th International Conference on (pp. 324-331). IEEE.

- Yang, Y., Pierce, T., & Carbonell, J. (1998). A study of retrospective and on-line event detection. In Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 28-36). ACM.

- Zhang, X., & Li, Z. (2010). Automatic topic detection with an incremental clustering algorithm. In Web Information Systems and Mining (pp. 344-351). Springer Berlin Heidelberg.

- Zhong, S. (2005). Efficient online spherical k-means clustering. In Neural Networks, 2005. IJCNN'05. Proceedings. 2005 IEEE International Joint Conference on (Vol. 5, pp. 3180-3185). IEEE.

- Zhou, X., & Chen, L. (2014). Event detection over twitter social media streams. The VLDB Journal—the International Journal on Very Large Data Bases, 23(3), 381-400.

9.2 Online Resources

- Twitter, 2013b[1]. About public and protected Tweets.
  [Online] Available at: https://support.twitter.com/articles/14016-about-public-and-protected-tweets
  [Last Accessed 18 December 2014].

- statisticbrain[2]. Twitter Company Statistics.
  [Online] Available at: http://www.statisticbrain.com/twitter-statistics/
  [Last Accessed 7 December 2014].
- Verge[3] Report "The man behind Flickr on making the service 'awesome again'"
  [Online] Available at: http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer
  [Last Accessed 17 December 2014].

- Flickr website[4] Detail about different type of Flickr accounts and help
  [Online] Available at: https://www.flickr.com/help/limits/#150470666
  [Last Accessed 19 December 2014].

- Press news of Instagram [5]
  [Online] Available at: http://instagram.com/press/
  [Last Accessed 27 December 2014].

- Scribd [6] Statistical Report about Facebook
  [Online] Available at http://www.scribd.com/doc/229617868/Facebook
  [Last Accessed 17 December 2014].

- Detail about Four Square available on Wikipedia[7]
  [Online] Available at: http://en.wikipedia.org/wiki/Foursquare
  [Last Accessed 17 December 2014].

- Voyant [8] text Visualization tool
  [Online] Available at: http://voyant-tools.org/
  [Last Accessed 19 December 2014].

- Google map via google fusion tables[9] help and details are available at
  https://support.google.com/fusiontables/answer/2571232
  [Last Accessed 29 December 2014].

- CartoDB[10] "CartoDB Documentation."
  [Online] Available at: https://cartodb.com/docs
  [Last Accessed 9 February 2015].

- Definition of Visualization on Wikipedia[11]
  [Online] Available at:
  http://en.wikipedia.org/wiki/Visualization_%28computer_graphics%29
  [Last Accessed 14 February 2015].

- CartoDB-Editor[12]
  [Online] Available at: http://docs.cartodb.com/cartodb-editor.html
  [Last Accessed 9 February 2015].