Applying Data Mining for finding patterns from Spatio-Temporal Text Query Data

A DISSERTATION

submitted in partial fulfillment of the requirements

for the award of the degree of

Master of Technology

in

COMPUTER SCIENCE & ENGINEERING

by

Ravi Javiya

17535022

Under the supervision of

Dr. Durga Toshniwal

Professor, Department of Computer Science & Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

Indian Institute of Technology, Roorkee

Roorkee – 247667

May, 2019

AUTHOR'S DECLARATION

I declare that the work presented in this dissertation with title **"Applying Data Mining for finding patterns from Spatio-Temporal Text Query Data"** towards fulfilment of the requirement for the award of the degree of **Master of Technology in Computer Science and Engineering, Indian Institute of Technology Roorkee , India** is an authentic record of my own work carried out during the period of **July 2018 to May 2019** under the supervision of **Dr. Durga Toshniwal, Professor, Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, India.** The content of this dissertation has not been submitted by me for the award of any other degree of this or any other institute.

Date: ..... **Ravi Javiya**

Place: Roorkee,India **17535022**

CERTIFICATE

This is to certify that Thesis Report entitled **"Applying Data Mining for finding patterns from Spatio-Temporal Text Query Data"** which is submitted by **Ravi Javiya (17535022)**, towards the fulfilment of the requirements for the award of the degree of **Master of Technology in Computer Science and Engineering**, submitted in the **Department of Computer Science and Engineering, Indian Institute of Technology Roorkee, India** is carried out by him under my esteemed supervision and the statement made by the candidate in declaration is correct to the best of my knowledge and belief.

Date: ..... **Dr. Durga Toshniwal**

Place: Roorkee,India **Professor, Department of CSE**

Indian Institute of Technology, Roorkee

ABSTRACT

As the use of sensor network, GPS, and telecommunication technology increased, the availability of spatio-temporal data is also increasing exponentially. The use of these technologies also increased in the field of agriculture, which motivates to do data mining analysis of this data. For this work, we have selected farmers query data. In this work, we have experimented various data mining techniques to generate valuable insights from the data which can help in making policies. First, we have clustered the states of India. In the subsequent experiments we have successfully clustered the blocks and districts of India based on feature vector generated by combination of Crop and Query type. In the starting we have clustered based on month-wise data. But as in Agriculture similarity we measure tine in season, so in the next experiment we have used the season wise data for two season of India i.e. Kharif & Rabi. And in the end we found the co-occurring problems for various states in various months.

ACKNOWLEDGEMENT

First of all, I express my gratitude to the Almighty, who blessed me with the zeal and enthusiasm to complete this research work successfully. I am extremely thankful to my supervisor Dr. Durga Toshniwal, Professor, Computer Science & Engineering Department, Indian Institute of Technology, Roorkee for their motivation and tireless efforts to help me to get deep knowledge of the research area and supporting me throughout the life cycle of my M. Tech. dissertation work. Especially, the extensive comments, healthy discussions, and fruitful interactions with the supervisors had a direct impact on the final form and quality of M. Tech. dissertation work.

This thesis would not have been possible without the hearty support of my friends. My deepest regards to my Parents and my sisters for their blessings, affection and continuous support. Also, Last but not the least, I thank to the GOD, the almighty for giving me the inner willingness, strength and wisdom to carry out this research work successfully.

(Ravi Javiya)

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION & MOTIVATION

1.1 Introduction

Spatial-temporal data is one of the fastest growing types of data due to the rapid development of remote sensors, sensor networks, and telecommunication technology and devices. It provides a great opportunity to make such location-based data actionable and insightful with proper analysis techniques. Moreover, such geographic analyses enable a wide range of services such as location-based recommender systems.

As the use of technology in the field of agriculture increased, huge amount of data has been generated by it. By applying various data mining techniques on this data, we can generate many valuable insight and knowledge which can in turn be used to improve the farming practice and increase the production.

Data mining technique plays a vital role in the analysis of data. Data mining is the computing process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database system. Unsupervised (clustering) and supervised (classifications) are two different

types of learning methods in the data mining. Clustering is the process of examining a collection of "data points," and grouping the data points into "clusters" according to some distance measure. The goal is that data points in the same cluster have a small distance from one another, while data points in different clusters are at a large distance from one another. This type of clustering methods used to find the similar regions. Predictive data mining technique is used to predict future crop, weather forecasting, pesticides and fertilizers to be used, revenue to be generated and so on. For this work, we have used Agricultural data of farmer queries from different Indian states. By applying various data mining technique, we can get many valuable insights.

1.2 Motivation

Mining information from spatio-temporal data is an important problem. In most cases, these databases changes with time, it is therefore important to capture the evolutionary behaviour of the spatial data points with respect to time as these give insights for predicting future occurrences of events such as hurricanes, crime, insurance claim, etc. As agricultural data mining is novice field, for this work we have selected to work on farmer queries data of Indian States.

1.3 Problem Statement

Identifying clustering based patterns from spatio-temporal text query data using data mining techniques. To find the patterns in data, the following are the sub-problems:

- Finding Unique Query to represent all similar queries
- Finding Similarity between states, districts and blocks by using different clustering methods
- Finding co-occurring problems using Association rule mining

1.4 Thesis Organization

The report consists of five chapters. The current chapter gives the brief introduction about the spatio-temporal data mining. The remainder of the report is structured as follows.

Chapter 2 describes related work which has been done in the field of Spatio-Temporal Data Mining and also discusses research gaps. Chapter 3 describes Spatio-temporal Analysis work flow. Chapter 4 describes the dataset used. It also discusses experiments done. Chapter 5 describes the future work and concludes the report.

Chapter 2

Related Work

2.1 Literature Review

objective of clustering Spatio-Temporal points involves finding clusters that have an unusually high density of Spatio-Temporal points, also termed as hot-spots. This can be used for finding outbreaks of diseases or social movements, where there is a dense conglomeration of events both in space and time. This problem is also referred to as 'event detection' in the mining social media data [1]. In this work, they have used the twitter data stream having geo tag. They have proposed a framework for Hierarchical spatio-temporal hashtag clustering. For this they have used divide and conquer approach. They have proposed an efficient hashtag clustering algorithm, which can handle content-evolving hashtags in real-time.

Several researches have been done on agricultural sector using data mining techniques. A recent study [2] in India analyzed agricultural data of Karnataka State by using different clustering methods like DBSCAN, PAM and CLARA. In this study, they have included the crop yield data, soil , temperature and rainfall data. First, they have clustered the districts on karnataka based on soil ,temperature and rainfall data. In the end, they applied a regression model to find out optimal parameters like optimal temperature, worst temperature and rainfall for a particular crop production.

They have also proposed Modified DBSCAN. In this they proposed to find the

minimum points and Epsilon (radius value) automatically. KNN plot is used to find out the epsilon value where input to the KNN plot (K value) is user defined.

In [3],they have proposed a "geographic, hierarchical self-organizing map (Geo-H-SOM)" to analyze geospatial, temporal and semantic characteristics of tweets which are geo-referenced. Their study includes three main tasks, which are twitter data retrieval & pre-processing, finding similarity of all information and computing an Self-Organizing-Map(SOM) in an unsupervised approach.

In [4], they have studied various classification, clustering, regression and association rule mining techniques which are applied on agriculture data. They have studied various application of data mining in agriculture data like using KNN simulate daily precipitation and other weather variables, Use of clustering and decision tree algorithm to find the patterns & prediction of class of soil in the data of soil.

In [5], they have proposed a system for government organizations and farmers to take decisions based on the various environmental factors. the system will suggest farmers by which crop they can have high production on their land by entering the location details and soil data. whereas based on the temperature, rainfall and previous crop production data, they gave the estimate of crop production. for this work, they have used clustering algorithms and decision tree based approach.

In [6], they have discussed various issues of spatio -temporal data mining. They have studied various clustering experiments done on the spatio-temporal data. The various application of spatio-temporal data mining in the fields of crop science, biology, Meteorology, Geophysics, etc. are also studied in this study.

Another study focuses on the analysis to predict Bangladesh's four most yielding crops; wheat, jute, T-Aman and mustard. [7] To carry out the whole experiment, they have analysed soil properties of medium high land and high land from different sub districts of Bangladesh and also their respective climatic data and crop production of the last 6 years. they have applied different data mining techniques such as K-means, PAM, CLARA and DBSCAN for clustering and four linear regression methods to predict crop yields.

2.2 Research Gaps

Many of the existing techniques are either based only on spatial analysis or only on temporal analysis of data. The work done in the field of spatio-temporal data is for the data having trajectories. The data has some unique record identifier to map various instances of the data. For Example, the data of taxi trajectories. There are no implementation on data mining techniques on the spatio-temporal text data which does not have any Unique Record identifier. i.e. all the instances of data are independent of each other. There are many existing spatial data mining techniques and temporal data mining techniques which are not expanded to work on spatio-temporal data. Like in clustering CLARANS is used for spatial data, while k-means and other simpler algorithms are important due to the speed of clustering.

The Standard Association Rule Mining algorithms like Apriori, FP Growth are not applicable in spatio-temporal data directly. There are many proposals to use it in temporal data as Time based association rule mining. To implement it in Spatio-Temporal Data is still a naïve field.

Chapter 3

Proposed Work

There are various stages in spatio-temporal data analysis workflow. The brief explanation for all the stages is given below.

1. Input Data

   In this stage, we collect the dataset to be used in the research. This data will be spatio-temporal in nature. i.e. the value of various attributes can be represented as function of time and space.

   For this work, we have collected data of farmers queries of various state of India from Jan,2013 to Sept,2017.

2. Pre-Processing

   In the pre-processing stage, we fill the missing value in data using various technique. We also normalize the data if required.

   For this work, first we have Geo-coded the data. Then we have applied word2vec based approach to find out similar queries.As the data is generated by more than one person, the sentence structure and choice of words are different. To find the similar query text, we have applied a word2vec based approach. In this we will find the vectors for all the words using word2vec model, if applicable. The Query Score for all the queries is generated by taking the sum of vector for the words in the query for which the vector is available.

$$QueryScore(Query) = \sum Vector(word) \ for \ word \ in \ Query \tag{3.1}$$

3. Dimensionality Reduction & Visualization

   As described earlier, the data will be a function of time & space. Now, if we create the vector based on the attributes data is having, it will be a multi-dimensional vector and important is it will be a sparse matrix. So, to remove this sparse nature in the vector we apply various dimensionality reduction technique like PCA to reduce the dimension of feature vector.

   As now the dimension is reduced, we can visualize the data get the initial idea of data that which are the methods we can apply, what can be the parameters for different methods, etc.

   For this work, we have used the crop and query type combination to generate the feature vector. First, we have found all the combination of crop and query type which are present in the data. And then filled the feature vector with the percentage of that type of combination of crop and query type in that block. Table 3.1 shows the schema overview of feature vector generation.

   Table 3.1: Schema of Feature Vector

	Crop QueryType Combination				
Block Number	Cotton, Plant Protection	Bengal Gram, Nutrient Management	Cotton, Weed Management	Chillies, Disease Control	...
1	0.10	0.09	0.11	0	
2	0	0.21	0	0	
3	0.11	0.07	0.31	0.08	
...					

   Now, as this leads to a high dimensional feature vector (approx. 7k), which are sparse in nature. To reduce the sparsity in the feature vector we have applied the dimensionality reduction on it.

4. Spatio-Temporal Analysis

   In this stage, we generate the useful insight of data using various data mining technique like clustering, association rule mining, etc.

5. Post-Processing & visualization

   As we have applied various algorithms, we can integrate the result of this analysis and if is satisfactory than we end the process or we continue applying the data mining techniques with different parameters to gain the knowledge.

6. Useful Knowledge

   As a result of this analysis, we end up having useful knowledge in Human understandable form.

Chapter 4

Experiments & Discussion

4.1 Dataset Used

For this work, we have used the farmer query data. The data includes queries from 20 Indian states for the month of Jan,2013 to Sept,2017.

The data has 11 different fields like Season, Sector, Category, Crop, Query Type, State, District, time, etc. There are total of 1,13,12,911 queries in the data. The summery of data is given in below Table 4.1.

Table 4.1: Summary of Data

Year	**2013**	**2014**	**2015**	**2016**	**2017**
No. of Queries	20,31,648	22,15,293	24,05,112	27,11,566	19,49,292

4.2 Experiments

We have Geo-coded the data based on the location information provided in form of block, district and state. We have added two new attributes latitude and longitude to be able to plot in map.

The query type frequency counts for month of Jan,2016 in the state of M.P. is shown in below figure 4.1. From the figure 4.1 we can see that Plant protection,

weather and government scheme are the most frequent type of queries.

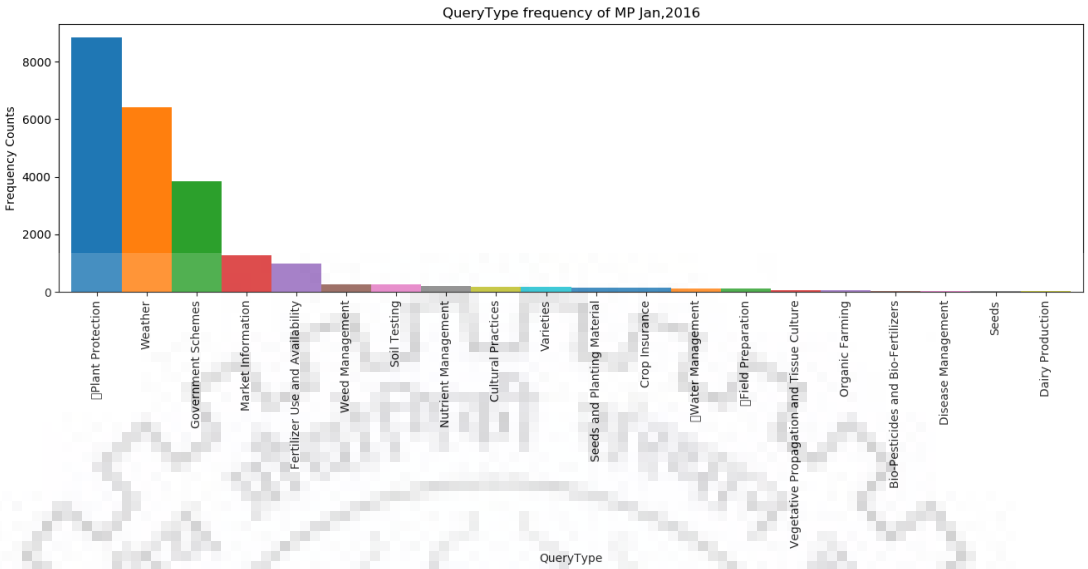


Figure 4.1: Query Type Frequency for MP Jan,2016

Table 4.2: List of Similar Queries with their Query Score

Similar Query List	**Query Score**
”flower drop mango”, ”flower drop of mango”, “flower drop” ”mango flower drop problem.....?”, ”mango flower drop”, “flower drop”, “problem of Flower dropping”	-18.775717
”yellow mosaic disease in black”, ”yellow mosaic disease in black garam.....?”, ”yellow mosaic disease in black gramp”, ”yellow mosaic disease in black gram...”, ”yellow mosaic disease in black gram..?”	-13.050985

By calculating the QueryScore as per Equation 3.1, we found out Unique Queries from the data. Unique queries are the Queries which have different QueryScore. To find Unique Query text is necessary for finding support of particular Query. As result of this, we successfully mapped the similar queries which are differing from word orders, having crop names as part of query or having different word selection for same meaning.

As shown in Table 4.2, the query score for the queries which are having words out of order or misspelled words are having same Query Score.

4.2.1 Clustering States based on Queries Similarity

To start with, we have first taken considered states as a spatial parameter to cluster. In this clustering experiment, we have clustered the states based on similarity of top 3 queries set. we have found out the set of Top 3 queries which represent the states. We have data for 20 states for the month of Jan,2016. The states are clustered in 10 clusters having same queries in the Top 3 query set. The result of the clustering is summarised in following Table 4.3. The Table 4.3 shows the cluster name along with the state list in particular cluster and respective Top 3 queries set. By looking at the result we can say that "Plant Protection" is the important query type as it was there in Top 3 queries set for all the states.

Same way, we have applied clustering for the month of April,2016 on the data of M.P. And the resultant clusters are shown in figure 4.2 on the map of India. The states which are not coloured or colored as grey shows the state for which we do not have data. As we can see in the figure 4.2 that all the neighbouring states are in same clusters that shows that queries has high effect of Environmental condition/Monsoon and type of land.

There are many cluster of size 1, which shows that states as clustering parameter is not appropriate. As Queries alone can not represent the location as multiple crops can have same queries. because of these limitations in further work we have considered crop and Query Type combination to generate the feature vector and considered blocks as spatial parameter to cluster.

Table 4.3: Clustering Result of States based on Top 3 Queries for month Jan,2016

Cluster No.	**States in Cluster**	**Top 3 Queries**
1	Bihar, Madhya Pradesh, Chhattisgarh, Jharkhand, Uttarakhand	'Plant Protection', 'Government Schemes', 'Weather'
2	Haryana	'Plant Protection', 'Bio-Pesticides and Bio-Fertilizers', 'Weather'
3	Jammu & Kashmir, Orissa, Punjab, Tripura	'plant Protection', 'Cultural Practices', 'Weather'
4	Gujarat	'Plant Protection', 'Cultural Practices', 'Government Schemes'
5	Himachal Pradesh, West Bengal	'Plant Protection', 'Fertilizer Use and Availability', 'Weather'
6	Andhra Pradesh, Karnataka, Maharashtra	'Plant Protection', 'Market Information', 'Weather'
7	Tamil Nadu	'Plant Protection', 'Fertilizer Use and Availability', 'Nutrient Management'
8	Rajasthan	'Plant Protection', 'Nutrient Management', 'Weather'
9	Uttar Pradesh	'Plant Protection', 'Weather', 'Weed Management'
10	Assam	'Plant Protection', 'Cultural Practices', 'Nutrient Management'

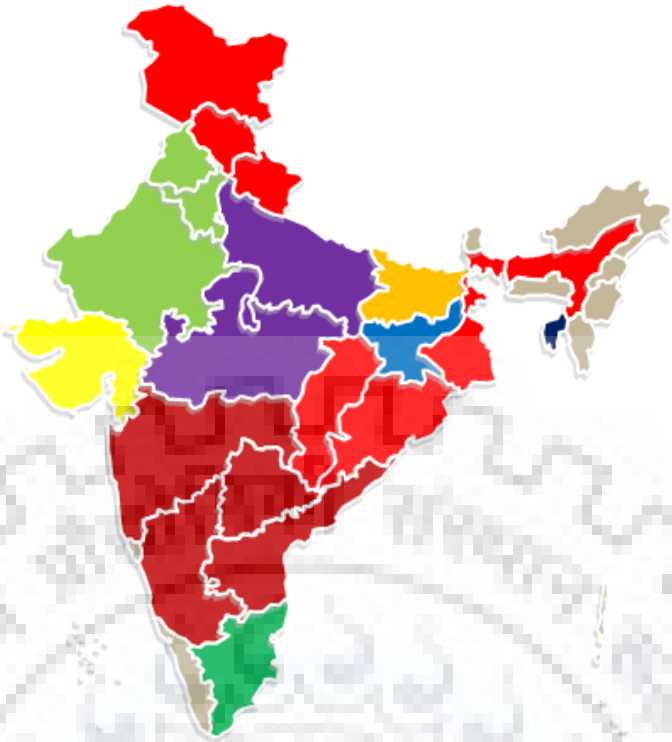


Figure 4.2: Clustering Result for States based on Top 3 Queries for the month of April,2016

Clusters Achieved By Top 3 Queries for the month of April,2016.

Cluster: 1 Haryana, Punjab, Rajasthan

Cluster: 2 Assam, Himachal Pradesh, Jammu & Kashmir, Orissa, West Bengal, Chhattisgarh, Uttarakhand

Cluster: 3 Bihar

Cluster: 4 Tamil Nadu

Cluster: 5 Tripura

Cluster: 6 Andhra Pradesh, Karnataka, Maharashtra

Cluster: 7 Gujarat

Cluster: 8 Madhya Pradesh, Uttar Pradesh

Cluster: 9 Jharkhand

4.2.2 Clustering District based on Queries Similarity

In the next step, we have used the crop and query type combination to generate the feature vector. First, we have found all the combination of crop and query type which are present in the data. And then filled the feature vector which the percentage of that type of combination of crop and query type in that block. Table 4.4 shows the schema overview of feature vector generation.

Table 4.4: Schema of Feature Vector

Block Number	Crop QueryType Combination				
	Cotton, Plant Protection	Bengal Gram, Nutrient Management	Cotton, Weed Management	Chillies, Disease Control	...
1	0.10	0.09	0.11	0	
2	0	0.21	0	0	
3	0.11	0.07	0.31	0.08	
...					

Now, as this leads to a high dimensional feature vector (approx. 7k), which are sparse in nature. To reduce the sparsity in the feature vector we have applied the dimensionality reduction on it. For dimensionality reduction we have used the PCA method and then applied the Density based clustering algorithm DBSCAN to find the clusters with various values of Eps and MinPts. To validate the result, we have used silhouette width as parameter.

Here, we have applied the clustering algorithm on the data of M.P. for the month of Jan,2016. Here we have applied the clustering at block level data. i.e. the feature vector is calculated for all the blocks in the state of M.P. The highest cluster validation index is found for the value of Eps = 0.8 and MinPts= 2 which is 0.392797.

In the result Figure 4.3, there are 4 clusters formed having 244,2,2,2 points. This shows that as the data points are in the form of closer cluster it is not able to find the cluster effectively.

When we applied the partitioning-based clustering algorithm K-Means to the data, we found better results. We have used the cluster validating index to find

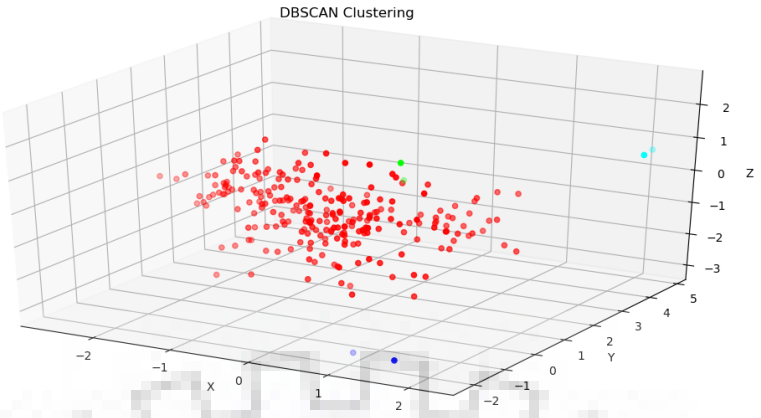
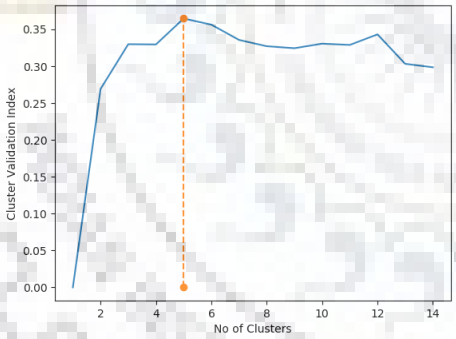
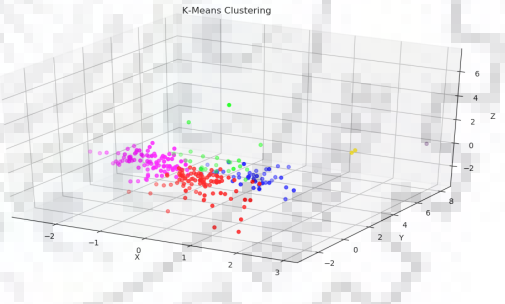


Figure 4.3: DBSCAN Clustering result on Block wise data of M.P. Jan,2016

the value of K – no. of clusters. Which shows the high index for the value of K=5 (Figure 4.4a). As you can see in the Figure 4.4b, we have found 5 cluster having data points 95,25,42,3,93 respectively. Which shows that K-means has found the cluster in data which are closer to each other and made an extra small cluster of 3 points which are also removed by DBSCAN as noise point.



(a) Cluster Validation Index vs. No. of clusters graph



(b) Clustering Result

Figure 4.4: K-Means Clustering result on Block wise data of M.P. Jan,2016

By doing clustering at block level, we can get the best result for the data of particular state only. Because when we apply the same on the data of India, there are many blocks for which data will not be available for all the months. So, for further work we have focused on clustering district having similar pattern in farmers queries. In this case also, the DBSCAN algorithm failed to find clusters in dense data points, where as the K-Means algorithm has performed better in finding clusters.

The parameter for DBSCAN is : Eps – 0.8 Min. Pts. – 9 The two clusters are

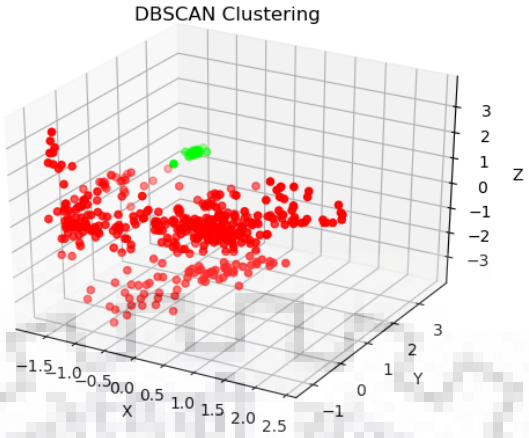


Figure 4.5: DBSCAN Clustering result on District wise data of July,2017

of size 442 & 23 as shown in figure 4.5.

For same data, the K-Means Algorithm Cluster validation index vs. K – no. of clusters graph is shown in figure 4.6. The result of K-Means clustering is shown in

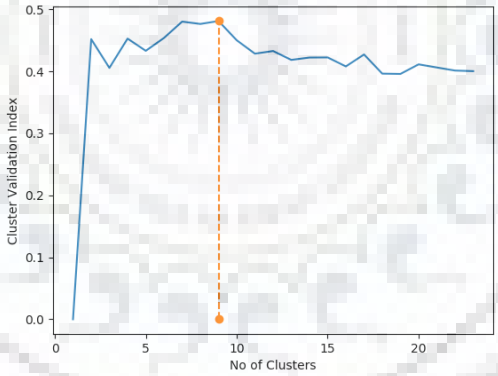
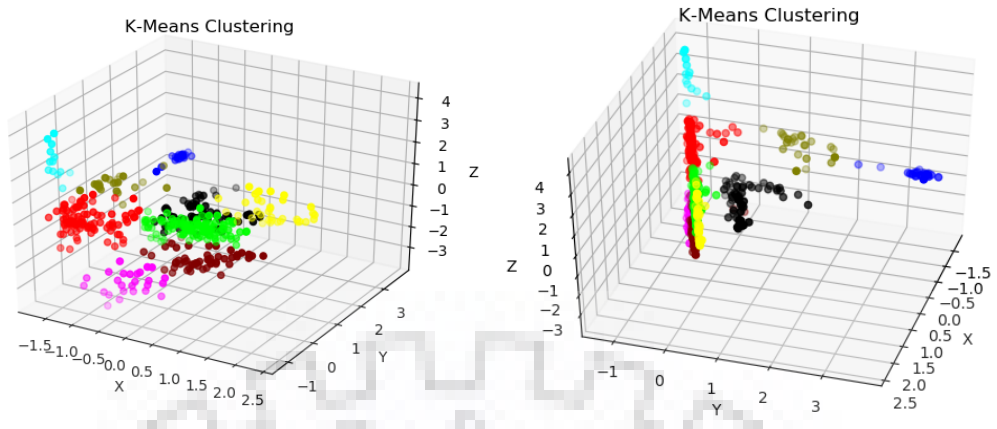


Figure 4.6: Cluster Validation Index vs. No. of clusters graph for K-means clustering for District wise data of July,2017

figure 4.7 having 9 clusters.

As we can see in figure 4.7, the clusters formed are clearly distinguishable and valid clusters. But the DBSCAN has only formed two clusters. One of them having majority of points. The clustering result of K-Means algorithm can be visualised on map of India as shown in figure 4.8. The district which are not coloured and shown as grey shows the districts for which data is not available. The summery of cluster result is also given in table 4.5.



(a) Front View (b) Side View

Figure 4.7: K-Means Clustering result on District wise data of July,2017

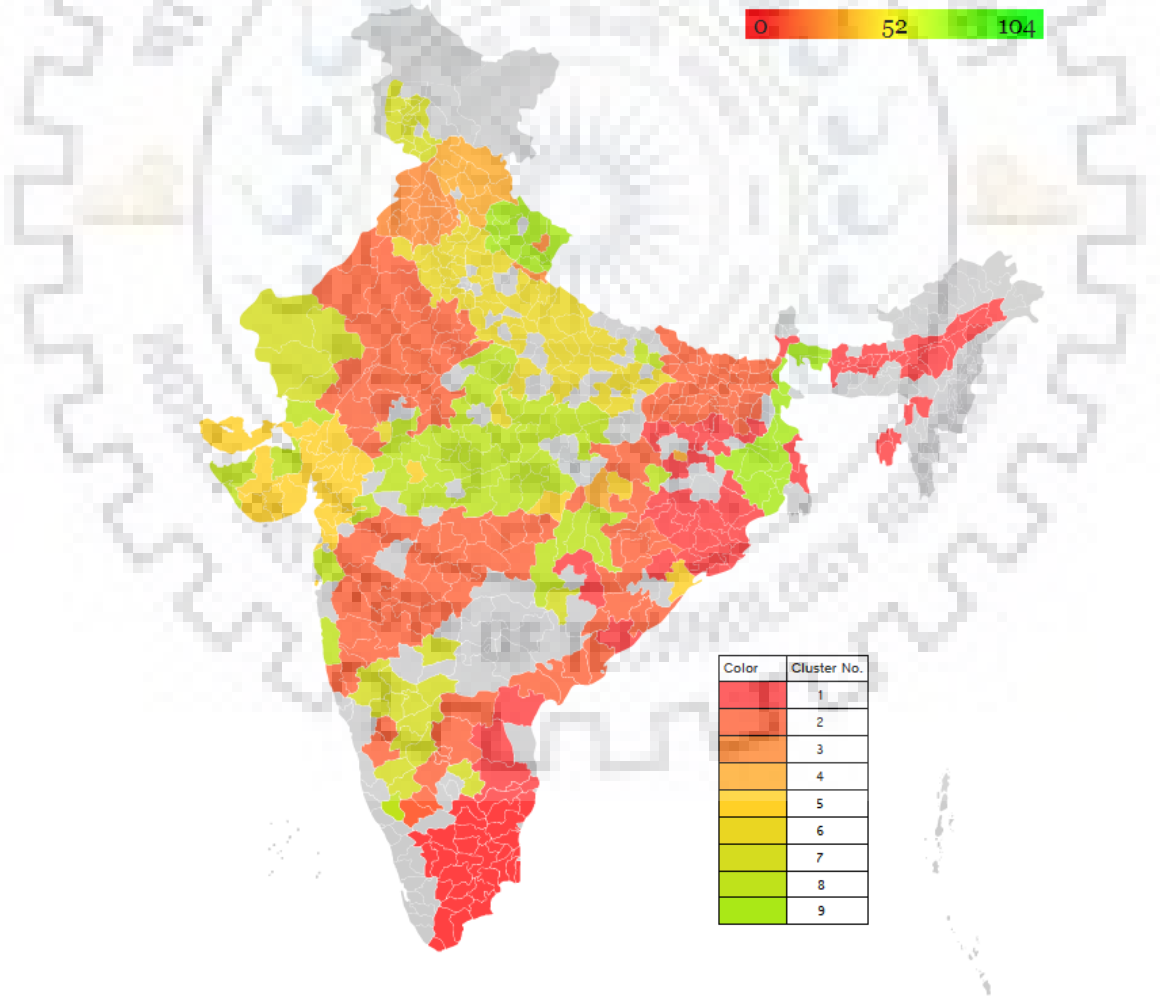


Figure 4.8: District wise clustering visualization for July,2017

Table 4.5: Cluster vs. No. of districts in cluster for District wise clustering of July,2017

Cluster No.	**No. of Districts**	**Some Important Districts**
1	97	Visakhapatnam (Andhra Pradesh) Dibrugarh (Assam), Puri (Odisha), Vellore (Tamilnadu), West Tripura (Tripura), Darjeeling (West Bengal), Dhanbad (Jharkhand)
2	128	Krishna (Andhra Pradesh), Gaya (Bihar), Mysore (Karnataka), Pune (Maharashtra), Sonepur (Odisha), Ajmer (Rajasthan), Bilaspur (Chhattisgarh)
3	24	Jalandhar (Punjab), Mandi (Himachal Pradesh)
4	13	Chamba (Himachal Pradesh), Kullu (Himachal Pradesh)
5	26	Valsad (Gujarat), Indore (Madhya Pradesh)
6	67	Rohtak (Haryana), Saharanpur (Uttar Pradesh)
7	36	Pulwama (Jammu & Kashmir), Bellary (Karnataka), Jodhpur (Rajasthan)
8	53	Jamnagar (Gujarat), Sagar (Madhya Pradesh), Durg (Chhattisgarh)
9	24	Pithoragarh (Uttarakhand), Hooghly (West Bengal)

4.2.3 Clustering Districts based on Season Data

In agriculture, the time duration is generally taken as a season. There are mainly two seasons in Indian Agriculture namely Kharif & Rabi. As most part of India suffers from water shortage in summer, The Zaid season is not taken as important as other two because it depends on water availability. Because of this, we have also focused only on the other two seasons. The distribution of Months in season is shown below in table 4.6.

Table 4.6: Distribution of Months into Season

Season	**Months**
Kharif	June, July, August, September, October
Rabi	November, December, January, February, March

As we have a field of Season in data, we matched the standard duration with the result from data. And the result is almost similar to standard duration Except October. As October generally considered in both seasons, but the data has a greater number of Kharif queries in month of October, we have considered it in Kharif season. We have applied both the density-based clustering approach and partition-based approach on the season wise data, and the result for the same is shown in the table 4.7 & table 4.8.

Here also Density based approach in most of the cases finds one big cluster and one small cluster having few data points, whereas the K-Means algorithm finds a greater number of clusters than density-based approach and the cluster validation index score is also better for K-Means algorithm in most of the cases.

Table 4.7: Summery of District wise Clustering results for Kharif Season

Year	**Clustering Parameters**					
	DBSCAN				***K-Means***	
	Min. Pts.	Eps.	No. of Clusters formed	Cluster Validation Index	No. of Clusters - K	Cluster Validation Index
2013	5	0.8	2	0.5534	7	0.38137
2014	4	0.8	2	0.5533	5	0.37914
2015	4	0.7	2	0.4043	9	0.42557
2016	5	0.8	2	0.45139	2	0.47081
2017	13	0.7	2	0.45639	2	0.5027

Table 4.8: Summery of District wise Clustering results for Rabi Season

Year	**Clustering Parameters**					
	DBSCAN				***K-Means***	
	Min. Pts.	Eps.	No. of Clusters formed	Cluster Validation Index	No. of Clusters - K	Cluster Validation Index
2013	9	0.8	2	0.4155	4	0.49544
2014	3	0.6	2	0.5758	2	0.66146
2015	7	0.5	7	0.3357	12	0.41511
2016	6	0.8	2	0.53647	5	0.50572
2017	2	0.5	3	0.4035	2	0.582

The result for the season Rabi 2016 is shown in the map of India in figure 4.9. As shown in figure 4.9, there are 5 cluster for the data of Rabi season for year 2016. Which clusters the most of coastal districts in one cluster.

Now, to find the closely related districts, we have found out the set of districts which are in same clusters for all 5 years. For this we have only shown the set of

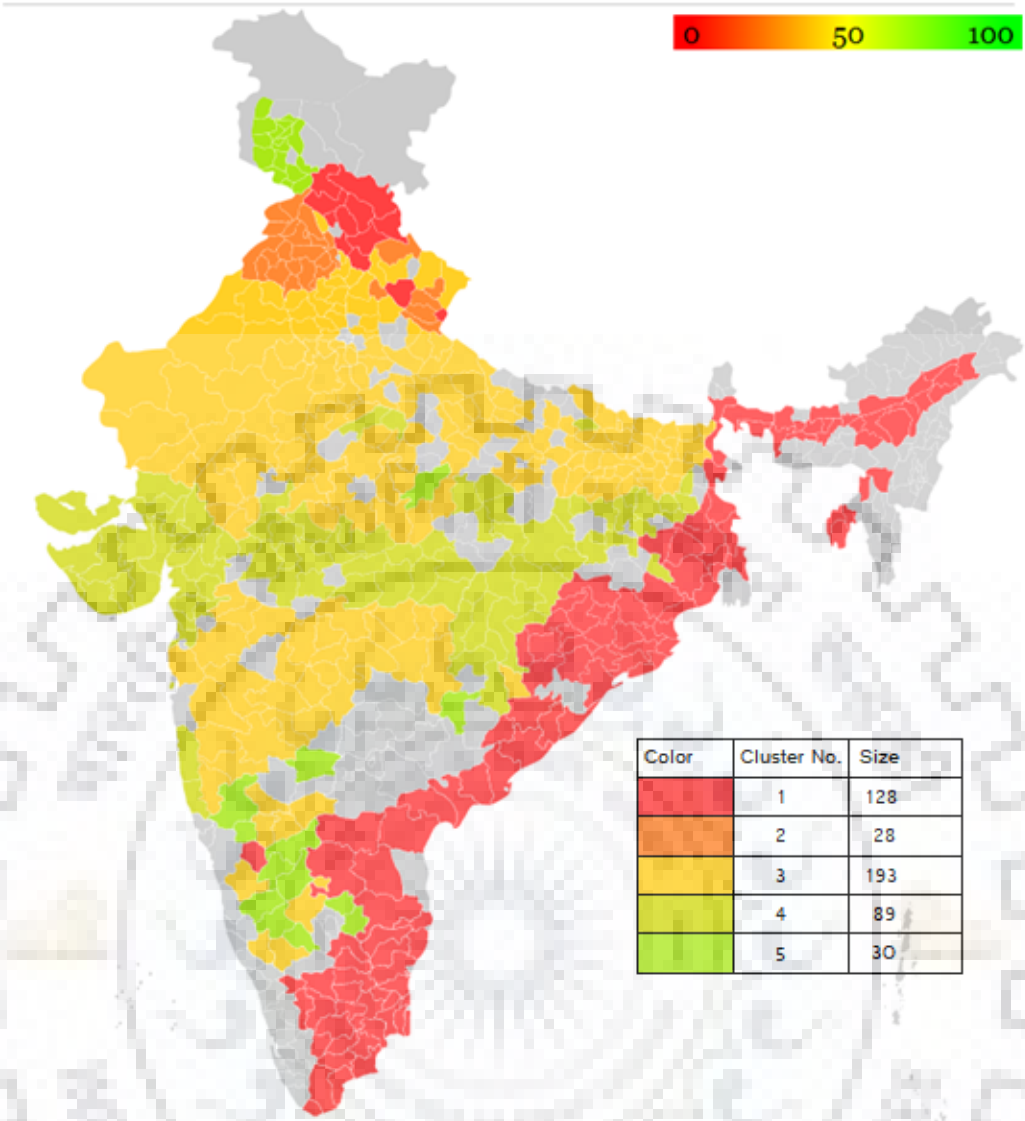


Figure 4.9: District wise clustering visualization for the Rabi season,2016

districts which are having size greater than 20. Figure 4.10 shows the set of districts which are in same cluster for all 5 years in Kharif season whereas figure 4.11 shows the set of districts which are in same cluster for all 5 years in Rabi season. The number of districts which are in particular set is mentioned as the size in map.

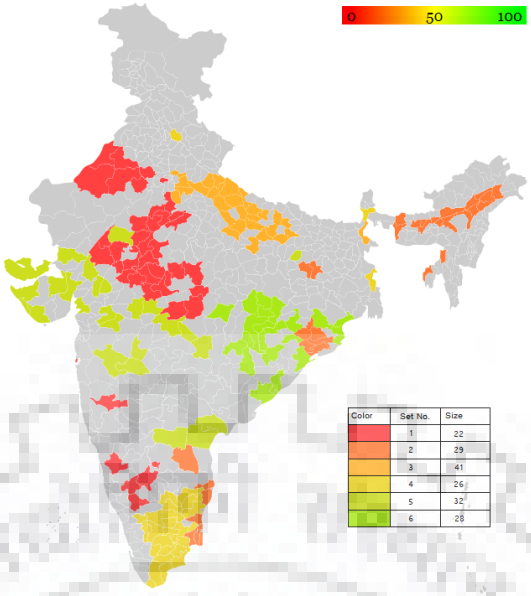


Figure 4.10: Similar District Set visualization for Kharif Season

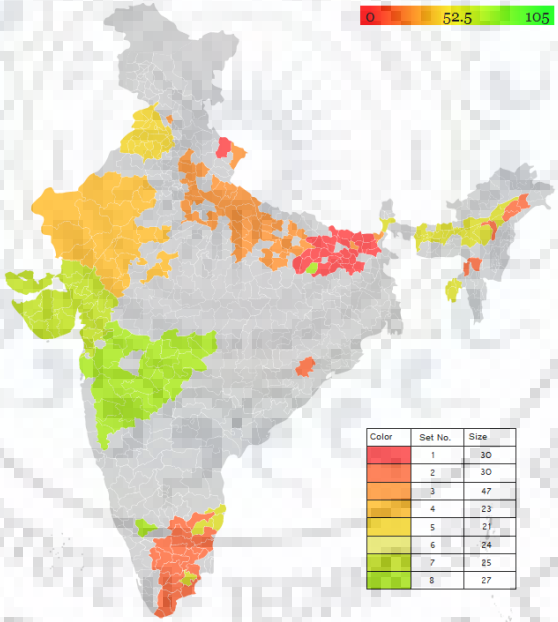


Figure 4.11: Similar District Set visualization for Rabi Season

In table 4.9, some of important districts which are in particular set is shown. The cluster number and size of cluster is also shown in table 4.9.

Table 4.9: Overview of the Similar District Set

Season	**Cluster No.**	**Some Important Districts**
Rabi	1	Gaya (Bihar), Patna (Bihar)
	2	Coimbatore (Tamilnadu), Tirupur (Tamilnadu), Dibrugarh (Assam)
	3	Varanasi (Uttar Pradesh), Etawah (Uttar Pradesh)
	4	Udaipur (Rajasthan), Ajmer (Rajasthan)
	5	Mansa (Punjab), Gurdaspur (Punjab)
	6	Morigaon (Assam), Thiruvallur (Tamilnadu), West Tripura (Tripura)
	7	Navsari (Gujarat), Anand (Gujarat)
	8	Amravati (Maharashtra), Mysore (Karnataka)
Kharif	1	Bikaner (Rajasthan), Ramgarh (Madhya Pradesh), Sangli (Maharashtra)
	2	Thanjavur (Tamilnadu), Morigaon (Assam), Hazaribagh (Jharkhand), Cuttack (Odisha)
	3	Fatehpur (Uttar Pradesh), Bulandshahar (Uttar Pradesh)
	4	Tirupur (Tamilnadu), Bilaspur (Himachal Pradesh)
	5	Nandurbar (Maharashtra), Porbandar (Gujarat)
	6	Kalahandi (Odisha), Korba (Chhattisgarh), Sambalpur (Odisha)

4.2.4 Association Rule Mining

To find the frequent problems which are occurring together gives you a valuable insight about how to overcome it. As the data has may types of queries, for this part we have selected the Queries with type "Plant Protection". This type of queries include queries about fertilizer, pesticides, and other problems farmers are facing.

As we already applied the word2vec based approach for finding the similar queries, we know the frequency counts of each query. We had applied Apriori algorithm for finding the Association rules for all months. few are shown in table 4.10.

Table 4.10: Result of Month wise Association Rule Mining

State, Month, Year	**Association Rule in the form A→B**		**Support**	**Confidence**
	A	B		
Rajasthan, March, 2016	"Lemon control flower dropping"	"Onion disease control"	0.33333	0.71428
	"Bengal Gram insect control"	"Lemon control flower dropping"	0.33333	0.76923
	"Onion nutrient management"	"Citrus control flower dropping"	0.33333	0.90909
	"Lemon control flower dropping"	"Citrus control flower dropping"	0.43333	0.92857
Haryana, April, 2014	"Tomato control early blight"	"Berseem control mundi"	0.36842	0.5
	"Berseem control mundi berseem"	"Tomato control early blight"	0.36842	1.0
	"Tomato control fruit borer"	"Tomato control early blight"	0.31578	0.6666
	"Bottle Gourd control Lali pest vegetables"	"Tomato control early blight"	0.42105	1.0

As mentioned in Article [8] Tomato seeds can be planted in any season, but preferred months are March,June & November. From Article [9] we know that the problem of Early Blight happens around 1 month from planting i.e. in April, which support the truthness of our data as the problem is listed in the month of April.

Chapter 5

Conclusion & Future Scope

5.1 Conclusion

In the first part of Chapter 4, we have found out the Unique Queries from the data using the "QueryScore" approach based on word2vec. Which helps us in finding support of particular Query in data. In Section 4.2.1, we have used the Top3 Queries to represent all states of India, and clustered them based on Similarity of Top3 Queries. As State is a bigger location, we can not say about the border locations that they also shows the same pattern as rest of part of state. Because of this, in Section 4.2.2, we have focused on clustering Blocks(Tehsil - Smallest Region available in data). based on Crop,Query Type Combination. In previous part, we only used the queries to cluster. but the Queries only are not sufficient as it can be of any crop. So, we have Used a new Feature Vector generated from combination of Crop, QueryType. As the number of Crop, QueryType combination is more than 7k, the data became too sparse in nature. To remove the sparsity in data, we have applied dimensionality reduction on the data before clustering.

As blocks are very small region, there is many instances where no queries are there from particular block in particular months. Due to this, In further work, we have focused on clustering districts instead of blocks. The various experiments done on Month wise data of all districts of India in section 4.2.2. As Agricultural season is good measure rather than months for finding agricultural similarity of districts, we

have taken season wise data for clustering in section 4.2.3. we have also found out the districts which are in same clusters in all the years. In section 4.2.4, by applying the Association Rule mining on the Queries of "Plant Protection", we have found out the co-occurring problems.

From the experiments above, we can conclude that in the spatio-temporal data when we have very close cluster then density-based approach is not effective in finding clusters. While the partitioning-based algorithms can find these types of clusters. And the states located near to each other or having similar geographical conditions are having similar type of queries generally. Result of association rule mining shows the problems which comes together. Based on result of this one can also predict the which of the fertilizer or pesticides required in near future.

5.2 Future Work

- Clustering Spatio-temporal events based on Attributes:
  - For the Agriculture Query data, we can think of Finding Similar Region with Soil, Climate and similar queries. i.e. finding location and time of all events having similar attributes.
  - we can also apply some other clustering algorithms and neural network approach to find the similarity between districts.
  - In association rule mining, we can use different approaches to find the frequent itemset and association rule.

Bibliography

[1] W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, and J. Huang, "Streamcube: Hierarchical spatio-temporal hashtag clustering for event exploration over the twitter stream," in *2015 IEEE 31st International Conference on Data Engineering*, pp. 1561–1572, April 2015.

[2] J. Majumdar, S. Naraseeyappa, and S. Ankalaki, "Analysis of agriculture data using data mining techniques: application of big data," *Journal of Big Data*, vol. 4, p. 20, Jul 2017.

[3] E. Steiger, B. Resch, and A. Zipf, "Exploration of spatiotemporal and semantic clusters of twitter data using unsupervised neural networks," *International Journal of Geographical Information Science*, vol. 30, no. 9, pp. 1694–1716, 2016.

[4] H. Patel and D. Patel, "Article: A brief survey of data mining techniques applied to agricultural data," *International Journal of Computer Applications*, vol. 95, pp. 6–8, June 2014.

[5] R. Shirsath, N. Khadke, D. More, P. Patil, and H. Patil, "Agriculture decision support system using data mining," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, pp. 1–5, June 2017.

[6] A. K.Venkateswara Rao and K. Rao, "Spatiotemporal data mining:issues, tasks and applications," *International Journal of Computer Science & Engineering Survey (IJCSES)*, vol. 3, p. 14, Feb 2012.

[7] S. Afrin, A. T. Khan, M. Mahia, R. Ahsan, M. R. Mishal, W. Ahmed, and R. M. Rahman, "Analysis of soil properties and climatic data to predict crop yields and cluster different agricultural regions of bangladesh," in *2018 IEEE/ACIS 17th*

International Conference on Computer and Information Science (ICIS), pp. 80–85, June 2018.

[8] "Agrifarming." `https://www.agrifarming.in`. Agricultural Guide for beginners.

[9] D. Tewari, "Agropedia." `http://agropedia.iitk.ac.in`. An initiative of IIT Kanpur & Indian Council for Agricultural Research(ICAR).

Publications

[1] Ravi Javiya and Durga Toshniwal, "Spatio-Temporal analysis of agricultural problems" in *ICDM 2019: 19th IEEE International Conference on Data Mining, Beijing, China, 8-11 November 2019.*(ERA A, Qualis A1) **(Communicated)**