ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Advertisement Suggestions Based on Sentiment Analysis

Nikhil Vatwani, Vijay Kataria, Richard Joseph, Sujata Khedkar

Student, Dept. of Computer Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Chembur, India

Student, Dept. of Computer Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Chembur, India

Professor, Dept. of Computer Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Chembur, India

Professor, Dept. of Computer Engineering, Vivekanand Education Society's Institute of Technology (VESIT), Chembur, India

ABSTRACT: Due to the increasing interest of people in online purchasing there has been an inclination towards the digital marketing. So to make more and more purchasing possible from people the advertisement should be related to the people's interest. If the advertisement is not in interest with the people, it will not attract them. This paper attempts to study the person's interest and show them the relevant data using Sentiment Analysis with the help of Density Based Clustering Algorithms (DBSCAN). Density Based Clustering Algorithms need the Epsilon(Eps) value and Minimum points value(MinPts) to create the clusters. The proposed method accepts the domain knowledge about the data set as an input and calculation of Eps and MinPts is automated which helps to make the data certain to some extent. We first create the grids for the dataset related to domain then it derives the default Eps and MinPts which are input for DBSCALE[Density-Based Clustering Algorithm for Larger Datasets] Algorithm. Experimental results indicate that the proposed method gives the most relevant data which matches people's interest.

KEYWORDS: DBSCAN, DBSCALE, Density – Based Clustering, Epsilon, Domain Knowledge, Grid.

I. INTRODUCTION

As stated by Cheng-Fa Tsai and Chun-Yi Sung, Data Mining is widely employed in business management and engineering. The major objective of data mining is to discover helpful and accurate information among a vast amount of data, providing a reference basis for decision makers. Data Clustering is currently a very popular and frequently applied and analytical method in data mining. Research in data clustering focuses mainly on increasing the accuracy and reducing the clustering time cost [3]. Data mining techniques are broadly classified into two categories: Predictive and Descriptive. Predictive analytics turns data into valuable, actionable information. Predictive analytics uses data to determine the probable future outcome of an event or a likelihood of a situation occurring. Descriptive analytics looks at past performance and understand that performance by mining historical data to look for the reasons behind past success or failure. Clustering Algorithm comes under Descriptive analytics.

Clustering analysis is an active area in data mining and other research fields. Clustering algorithm can be categorized into three categories:

1: Directly Density Reachable: A point 'p' is directly density reachable from a point 'q' w.r.t. Eps, MinPts if

- P belongs to $N_{Eps}(q)$
- Core point condition: $|N_{Eps}(q)| >= MinPts$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

2:Density Reachable: A point 'p' is density reachable from a point 'q' w.r.t. Eps, MinPts if there is a chain of points $p_1,p_2,\ldots,p_n$, $p_1 = q,p_n = p$ such thatr p_{i+1} is directly density-rechable from p_i.

3: Density-connected: Apoint p is density-connected to a point q w.r.t Eps, MinPts if there is a point 'o' such that both 'p' and 'q' are density-rechable from 'o' w.r.t. Eps and MinPts.

The number of outliers generated is an important parameter as it is used for the stop condition of both the algorithm. The percentage of outliers generated in the DBSCALE algorithm is less than the DBSCAN algorithm. This is because the DBSCALE helps in reducing the outliers, but we can also see that the value of the percentage of outliers in the proposed algorithm is more than the value of outliers in DBSCALE.

II. RELATED WORK

Density-based spatial clustering of applications with noise (DBSCAN) is a data clustering algorithm. It gives a set of points in some space, it groups together points that are closely packed together (points with many nearby neighbours), marking as outliers points that lie alone in low-density regions (whose nearest neighbours are too far away).The DBSCAN algorithm also lacks in many aspects. DBSCAN is not entirely deterministic: border points that are reachable from more than one cluster can be part of either cluster, depending on the order the data is processed. Fortunately, this situation does not arise often, and has little impact on the clustering result: both on core points and noise points, DBSCAN is deterministic. It is a variation that treats border points as noise, and this way achieves a fully deterministic result as well as a more consistent statistical interpretation of density-connected components. The quality of DBSCAN depends on the distance measure used in the function RegionQuery(P,ε). The most common distance metric used is Euclidean distance. Especially for high-dimensional data, this metric can be rendered almost useless due to the so-called "Curse of dimensionality", making it difficult to find an appropriate value for ε. This effect, however, is also present in any other algorithm based on Euclidean distance. DBSCAN cannot cluster data sets well with large differences in densities, since the MinPts-ε combination cannot then be chosen appropriately for all clusters. If the data and scale are not well understood, choosing a meaningful distance threshold ε can be difficult [2].

III. PROPOSED ALGORITHM

The proposed algorithm works in two phases.

Phase 1:
Initially we will form a grid structure whose object space is quantized to a particular no. of cells. This grid structure resembles a graph. Basic operation is to form clusters considering the plotted points, which comes from database. Database is a set of data points; the Point represents the core point. Begin scanning all data points within the entire database, these data points collected from database resembles the co-ordinates for graph which are plotted on the grid. These plotted data points are called "Facts". For example, considering an E-Commerce shopping website, each domain like Electronics, Furniture, etc. and each sub-domain like mobiles in electronics and beds in furniture, etc. are assigned an unique number/id. Suppose electronics is assigned id 1, Furniture is assigned id 2, Mobiles is assigned id 3, Beds is assigned id 4. Domains are plotted along Y-axis and sub-domains are potted along X-axis. Once X-axis and Y-axis are ready then "Facts" are plotted. Facts resembles co-ordinates which are combination of unique ids assigned to domain and sub-domain like a fact for person buying a mobile will be (3, 1). Here "3" is the id assigned to sub-domain Mobiles which is plotted along X-axis & "1" is the id assigned to domain Electronics which is plotted along Y-axis. Once the grid is ready and Facts are plotted we enter phase 2 of proposed algorithm.

Phase 2 [3]:
The parameters are set when implementing DBSCALE:(1) Radius (Eps), (2) Minimum number of included points (MinPts). The DBSCALE clustering algorithm can be described as follows:

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

Input: Datasets, Eps, MinPts
output: Clusters

```
DBSCALE (Datasets, Eps, MinPts)
 Intialization;
 ClusterID := NextID(First);
 FOR i FROM 1 TO Datasets.Size DO
  Point := Datasets.Get(i);
  IF Point.CID = UNCLASSIFIED THEN
   IF ExpandCluster(Datasets, Point, ClusterID, Eps, MinPts) THEN

    ClusterID := NextID(ClusterID);
    UnclassfiedData.Adjust();
   END IF
  END IF
 END FOR
END;

ExpandCluster(Datasets, Point, CID, Eps, MinPts) : Boolean;
 Neighbors := UnclassfiedData.RegionQuery(point, Eps); IF Neighbors.size < MinPts THEN
 Datasets.ChangeCIDs(point, NOISE);
  RETURN False;
   ELSE
 Datasets.ChangeCIDs(point,CID);
 Neighbors.AddExpansionSeedsO;
 FOR i FROM 1 TO Neighbors.size DO neighborPoint := Neighbors.Get(i);
  IF neighborPoint.CID = UNCLASSIFIED II neighborPoint.CID = NOISE THEN
   Datasets.ChangeCID(neighborPoint,CID); END IF;
 END FOR;
 WHILE Seeds <> Empty DO seedPoint := Seeds.FirstO;
  Seeds.Delete(seedPoint);
  Neighbors := UnclassfiedData.RegionQuery(seedPoint, Eps);

  IF Neighbors.size >= MinPts THEN Neighbors.AddExpansionSeedsO;
   FOR i FROM 1 TO Neighbors.size DO neighborPoint := Neighbors.Get(i);
    IF neighborPoint.CID = UNCLASSIFIED II neighborPoint.CID  NOISE
    THEN
     Datasets.ChangeCID(neighborPoint,CID);
    END IF;
   END FOR;
  END IF;
 END WHILE;
 RETURN True;
END IF END;
```

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

IV. PSEUDO CODE

The implementation steps for the DBSCALE algorithm are described as follows:

Step 1: Initialize all parameters, and define a new ClusterID. Step 2: Begin scanning all data points within the entire database. For data points belonging to the ClusterID of those unclassified data, implement the ExpandCluster processing procedure. The database is the set of data points; the Point represents the core point; the ClusterID denotes the current cluster ID;

e indicates the radius, and MinPts represents the minimum number of included points.

Step 3: If the data point returned by the expansion procedure function is a noise data point, then go directly to Step 2, until the Datasets database has been fully scanned. If an expansion data point is returned, then update the new ClusterID, and alter the index array of unclassified data, and then go to Step 2.

Step 4: End the algorithm when all data points have been processed.

The implementation steps for the ExpandCluster processing procedure are as follows.

Step 1: Search for neighborhood data within the range of radius e in the unclassified cluster index. If the number of neighborhood data is less than MinPts, then leave the procedure, and return the core point as the noise data point. Otherwise, go to Step 2.

Step 2: Set the core point as the current ClusterID.

Step 3: If the seed data points is empty then end the expansion processing procedure, otherwise go to Step 4.

Step 4: Search for the marking boundary point within neighborhood data, and add in the expansion seeds.

Step 5: Set all unclassified data points and neighborhood data points that are noise data as the current ClusterID.

Step 6: Extract the first seed from the expansion seeds; define it as the core point, and then delete it.

Step 7: In the unclassified data index, search for neighborhood data within the range of radius e of the core point. If the number of neighborhood data is greater than MinPts, then go to Step 3.

V. SIMULATION RESULTS

The proposed algorithm takes Eps and MinPts and datasets as an input. The calculation of Eps and MinPts are automated. Initially, for calculation of Eps(Radius), the diagonal end points of a particular cell from the grid are considered. Through these 2 points we get a distance by applying distance formula. By taking half of the distance, we obtain the radius of particular cell which is required initial Eps for input in proposed algorithm. For MinPts, each cell of the grid is scanned and the cel with least number of points is recorded. This value is assigned to MinPts. After assigning initial Eps and MinPts the clusters are formed in the phase 2 of the proposed algorithm. After each iteration in phase 2 of proposed algorithm “Eps” value is reduced and “MinPts” value is increased by a particular amount and hence denser clusters are formed. Once “Eps” and “MinPts” reaches a particular value, we get the dense clusters where all the points are taken into consideration and least number of noise points are left. The clusters formed gives us the most relevant data i.e. advertisement that will be most likely in the people’s interest.

We made a dummy e-commerce website and collected datasets i.e. person’s interest from it. These datasets were processed by “DBSCAN” and then by “DBSCALE” algorithms separately. The implemented algorithms give the following results:

Algorithm	Number of clusters formed	Number of noise points
DBSCAN	10	6
DBSCALE	15	2

Thus from the above table it is clear that the proposed DBSCALE algorithm performs better than DBSCAN and gives the best advertisement suggestions which will be most likely in the interest of people.

The explanation of below mentioned figure is given in phase 1 of proposed algorithm.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 2, February 2017

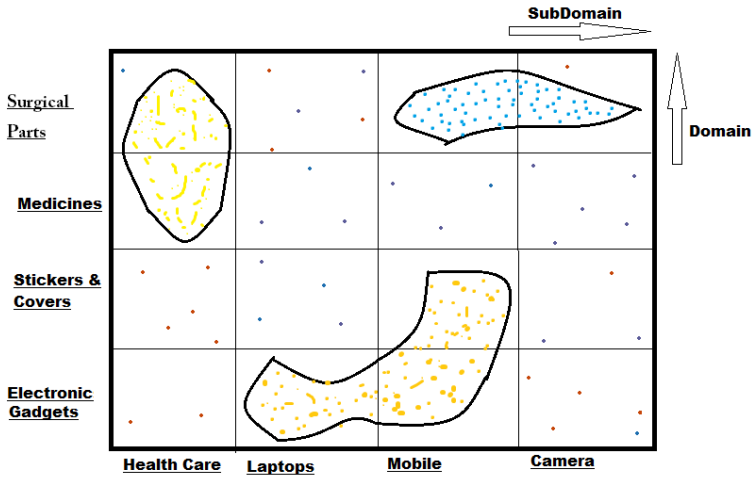

Fig.1. Suggestion of advertisement based on formation of clusters.

VI. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better than the other data mining algorithms like DBSCAN algorithm and shows the advertisement suggestions that are most relevant to person's purchase history and interest. In the proposed algorithm value of Eps and MinPts are automated. The disadvantage of the system will be providing the domain knowledge. This system will not be able to provide much help in optimizing the process of clustering to the user who cannot provide the domain knowledge and will work as any other clustering algorithm works but reduces the time considerably in selecting the Eps and MinPt which otherwise would be a daunting task for the user. In future we plan to show the advertisement on the area of webpage where user's interaction is most in addition to the relevant advertisement.

REFERENCES

1. JiaWei Han, Micheline Kamber. Data Mining Concepts and Techniques, Beijing: Higher Education Press, pp. 37-39, 2001.
2. Neethu Antony, Arti Deshpande, Domain Driven Density Based Clustering Algorithm, Springer Publications, pp.3-4, 2014.
3. Cheng-Fa Tsai, Chun-Yi Sung, An Efficient Density Based Clustering Algorithm for Data Mining in Large Databases, Second Pacific-Asia Conference on Circuits, Communications and System (PACCS), pp.6, 2010.
4. G. H. Shah, C. K. Bhensdadia, A. P. Ganatra, "An Empirical Evaluation of Density Based Clustering Techniques", Vol. 1, Issue 2, pp.8, March 2012.
5. Ms K. Santhisree, Dr. A. Damodaram, SSM-DBSCAN and SSM-OPTICS: Incorporating new similarity measure for Density based clustering of Web usage data, International Journal on Computer Sciences and Engineering, pp.4, August 2011.
6. M.Ester, H.-P Kriegel, J Sander, Xiaowei Xu. "A density based algorithm for discovering clusters in large spatial databases with noise". In: Proc. of Knowledge Discovery and Data Mining, Portland, AAAI Press, pp.226-231, 1996.
7. Guba S, Rastogi R, Shim K. "CURE: an efficient clustering algorithm for large databases". In: Haas LM, Tiwary A, eds. Proceeding of the ACM SIGMOD International Conference on Management of Data. Seattle: ACM Press, pp.73-84, 1998.
8. W.Wang, J.yang, R.Muntz. "Sting: a statistical information grid approach to spatial to spatial data mining". In: Proc. Of VLDB'1997, pp.186-195, 1997.
9. Hongmei Wang, Yingying Wang, Shitao Wan, A Density-based Clustering Algorithm For Uncertain Data, International Conference on Computer Science and Electronics Engineering, pp.6-7, 2012