International Journal of Applied Mathematics and Computing Science. Volume 3, Issue 4, (2014) 23-30
http://www.cayley-nielson.com/ijamcs

Accurate Trade Area Clustering by Using Micro-Blog POI Data

Xu Qiang
School Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China
E-mail: *xuqiang@bupt.edu.cn*

Abstract

The research of trade area division and classification always fascinating scientists, unfortunately, most reasonable division for values are based on the traditional market economy theory at present. With human's effort, clustering algorithm with more accurate and reasonable geographic location information take greater advantages when compared with the traditional research method of area dividing and positioning. In this paper, we will use micro-blog POI geographic information data and the current popular algorithm for finding high density clusters based on DBSCAN (S-DBSCAN) to perform clustering and precise positioning of division of work values. At last, with the micro-blog data, we check out our analysis and design by simulations, which show that the method is effective and practical.

Keywords: S-DBSCAN algorithm, trade area clustering, micro-blog POI data

1. Introduction

Trade area, also known as "business circle", refers to a radiation range, from the center of store location to extending areas along a certain direction and distance, to attract customers. To put It simply, it is the area limits where the customers live. In fact, there remains important significance as to whether the trade area layout is reasonable or not for the enterprise and the development of the business. If the layout is not reasonable, it will lead to low investment efficiency, low resource allocation efficiency and low economic benefit, resulting in a condition which is unable to meet the needs of consumers. Therefore, it is particularly important to make reasonable and accurate classification and clustering of the business circle.

As the core method of trade area research, clustering theory has more advantages than the market saturation theory [1] and purchasing power index method and so on. But we can find that the construction of trade areas has always been based on relevant definition on economics and city planning requirements [2]. Meanwhile, with small amount of data, we cannot acquire accurate position information of the trade area, leading to more difficulties to get the true trade area ranges. Today, with the advent of the information ages, micro-blog users grow so quickly with each passing day that the business potential cannot be ignored. When it comes to that how to use the micro-blog POI (point of interest) data for accurate location of the trade area to bring more economic benefits, there remains more important value and significance.

This paper creatively uses city POI micro-blog information data provided by the company, and does accurate clustering research of trade area with the help of S-DBCAN clustering algorithm (an algorithm for finding high density clusters based on DBSCAN).At last, we will take parts of Beijing (Xi Dan Area) as an example for simulation test to illustrate the validity of the method.

ISSN 2333-2425
December, 2015

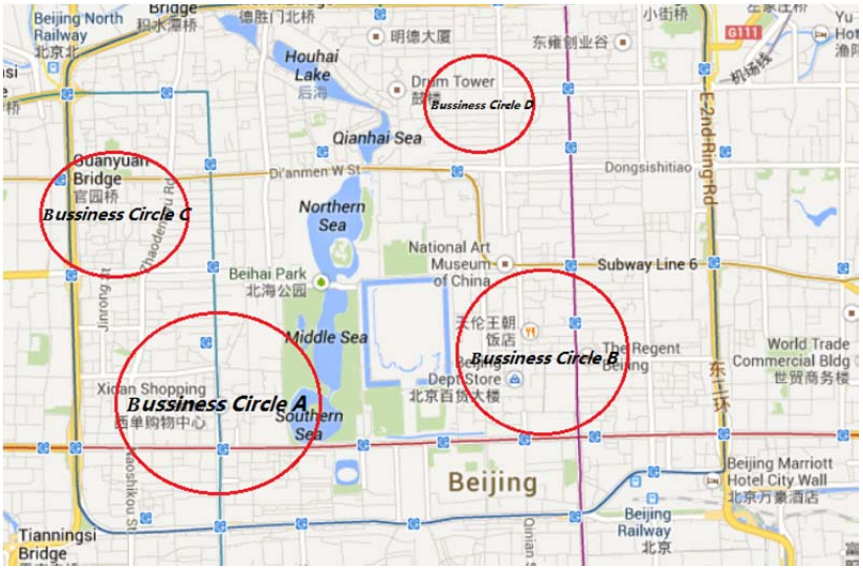

Figure 1. Traditional business circle Clustering

2. An algorithm for finding high density Clusters based on DBSCAN

2.1 The introduction of DBSCAN clustering algorithm

DBSCAN, density-based spatial clustering of applications with noise, is one of the classic algorithm in clustering data mining. The algorithm will divide enough high density regions into a class, and can find out the clusters of arbitrary shapes in spatial database with "noise". DBSCAN algorithm distinguishes a class based on certain density threshold which is composed of Eps and MinPts. Eps is the radius; MinPts said that there are at least a minimum number of points in an Eps-neighborhood of that point. Some concepts and terms to explain the DBSCAN algorithm [3] can be defined as follows.

Definition 1. *(Core object).* A core object *p* refers to such point that its neighborhood of a given radius (*Eps*) has to contain at least a minimum number (*MinPts*) of other points.

Definition 2. *(Border object).* An object *p* is a border object if it is not a core objects but density-reachable from another core object.

Definition 3. *(Directly density-reachable).* An object *p* is directly density-reachable from the object *q* if *p* is within the *Eps*-neighborhood of *q*, and *q* is a core object.

Definition 4. *(Density-reachable).* An object p is density-reachable from the object *q* with respect to *Eps* and *MinPts* if there is a chain of objects $p_1,p_2,\ldots,p_n$, $p_i=q$, $p_n=p$, such that p_{i+1} is directly density-reachable from p_i with respect to *Eps* and *MinPts*, for $p_i \in D, (1 \leq i \leq n)$.

The algorithm starts with the first point O in database, and retrieves all neighbors of point *O* within *Eps* and *MinPts* distance. If the total number of these neighbors is greater than *MinPts* - if *O* is a core object - a new cluster is created. The point O and its neighbors are assigned into this new cluster [4]. Then, it iteratively collects the neighbors within *Eps* distance from the core points. The process is repeated until all of the points have been processed.

But in an interactive data mining, such as mining in web, we often hope that by changing the parameters to obtain more meaningful results. At the same time, we do some periodic update instead of real-time update transaction on the data mining. So, to finish the task of trade area clustering based on micro-blog POI information, we chose the clustering algorithm proposed by Sun Peng, etc., algorithm for finding high density clusters based on DBSCAN (S-DBSCAN) [5], to make the results of the trade area clustering more efficient, more accurate.

For example, the following 18 points is handled by DBSCAN clustering, where *Eps*=2, *MinPts*=8.

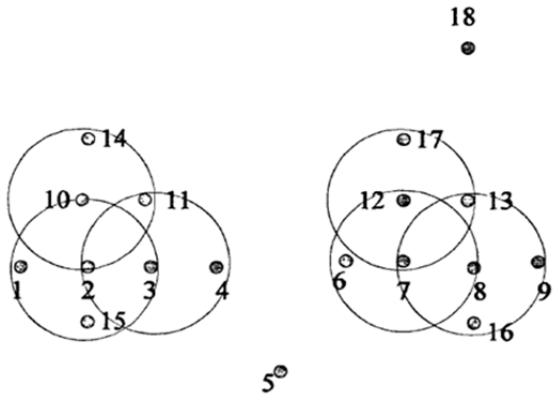

Figure 2. Cluster diagram of 18 data objects

2.2 The sign convention in S-DBSCAN algorithm

1) The parameter marker in DBSCAN algorithm: *Eps1*， *MinPts1*,

2) The parameter markers in S-DBSCAN algorithm: *Eps2*， *MinPts2*,

3) The markers of all data sets: *n*; the original number of noise quantity: *noise1*.

2.3 The description of S-DBSCAN algorithm

Input: Eps2, MinPts2, Eps1, MinPts1, U.clu_id2

Output: U.clu_id1

Proc

For all objects u **in** U **do**

U.clu_id = (u.clu_id2 ≡ NOISE) ? NOISE: Un_defined

End For

Cluster_id : = next_id(NOISE);

For all clusters c **in** U.clu_id2 **do**

For all objects u **in** c **do**

If u.clu_id1 ≡ Un_defined **then**

If ExpandCluster (c,u,Cluster_id,Eps1,MinPts1) **then**

Cluster_id: =next_id(Cluster_id)

End If

End If

End For

End For

End

The core idea of S-DBSCAN is just to calculate in the original cluster again according to DBSCAN algorithm, regardless of the original noise object. The algorithm firstly set the new *"cluster_id"* as *"Un_defined"*, and set the noise object as *NOISE*, including all of the object clusters, and selects any point for expanding. In addition, the *ExpandCluster* function and DBSCAN algorithm are same in the comments. What is different is that, the algorithm narrows the search scope of object in the cluster that is only for *Eps1* close to the query in the original cluster.

3. Trade area clustering and simulation test using micro-blog POI data

3.1 Introduction of micro-blog POI data

With the maturity and popularization of web technology, LBS (Location based service) has become one of the fastest developing technological applications [6]. LBS is a service based on the known position of the platform, and this platform needs the support of digital map. POI is the “Point of Interest”, which can be called “information points”. As the name suggests it is attractive point of interest on a digital map. Each POI contains four aspects of information, name, category, the longitude, latitude and other information. POI is the most crucial information on the digital map, because the purpose of using map is to find the interested target location and its corresponding properties [6].

With the development of the times, the popularity of mobile terminal makes communication become more and more convenient, and also promotes the rapid rise of the micro-blog as a kind of means of communication. Because of its large number of micro-blog users, there will be a massive data with the value of data mining. So, using micro-blog POI location information to do district mining and clustering has great market prospect and economic value, and it is also a hot spot in relevant mining research.

Micro-blog location check-in data uploaded by the user through the mobile terminal with GPS positioning, with the features that it has a large amount of data, the social attribute, etc. a kind of potential available data source which can make POI have high quality and efficiency. Micro-blog location check-in data covers all necessary information to update POI, fast and accurate.

3.2 Process of district clustering

The flow chart of district clustering analysis with the micro-blog POI data is as follows

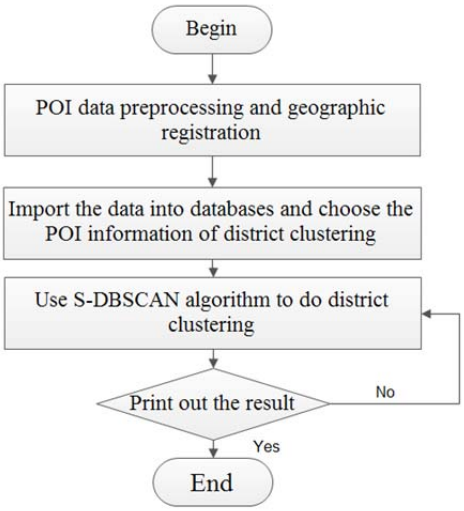

Figure 3. Process of district clustering

3.2.1 POI data preprocessing and geographic matching

Micro-blog location check-in data is uploaded voluntarily, so that there remain problems which include low precision, data redundancy and incorrect format etc. So, we must do data preprocessing at first to eliminate some data which is meaningless, lack in concerns or has few information points and combine the points with lots of repetitions. Also we must set registration with existing POI data to improve the data accuracy, reduce the data redundancy and meet the POI requirement.

Among them, the data preprocessing includes the following:

1) Set in the threshold of check-in times and the number of people, in order to remove some meaningless data.

2) Check the attribute information of data is complete or not. For the missing information, we need to establish a standard format according to which to make modification of the data which need to preserve.

3) Combine the repeated attendance data. This operation can use the POI data dictionary to compare with the micro-blog location check-in data, so that we can combine different alias with the standard name corresponding to the same geographic target.

In addition, due to the existence of certain error in localization of mobile intelligent terminal, micro-blog location check-in data may set certain deviation with existing POI data in the space, so in the use of data, we need to make geographical registration of micro-blog location check-in data at first.

According to the data preprocessing and geographical registration process mentioned above, we now show the example of POI data after processing the data in our simulation test (POI data in Beijing region of Xi Dan fast food restaurants).

	Name	Lon	Lat	poi_id
1	YIPINSANXIAO(Xi Dan Branch)	116.37458	39.90958	382804
2	McDonald's (Xi Dan Branch)	116.37642	39.9102	363503
3	Pizza Li (Xi Dan Branch)	116.373	39.9181	447595
4	Carlos Pizza (Xi Dan Branch)	116.374137	39.908843	469179
5	KFC (Xi Dan Branch)	116.376	39.9107	456031

Figure 4. Example of POI data

3.2.2 Process the POI data

According to the results of POI data preprocessing, we need to import the POI data we got into database, in order to facilitate further processing. We use MATLAB to program S-DBSCAN algorithm. Firstly, we need to find the POI location information and mark them on the map. Then, make a two-dimensional points figure and import the data into MATLAB program, we would get the clustering results. Finally, we would draw the Trade Area we got on the Google map, in order to show our results more clearly.

3.3 Simulation test

To show our new method effectively, we use the micro-blog POI data, provided by Sina, to make a simulation test. The goal is to find the trade area of fast food restaurants in Xi Dan.

3.3.1 Test environment

Operation System: *Window* 7

Computer: 2*G RAM*, *CPU* 2.20*GHz*

Simulation tools: *MySQL* 2008, *MATLAB*

Data from: *Sina micro-blog Company*

Step 1

We need to import the micro-blog POI data of all the fast food restaurants in Xi Dan into MySQL database, and with the help of Google map, mapped the POI data to the map and draw out two-dimensional coordinate scatter diagram as the data clustering source. After data preprocessing,

geographical registration and classification, we get the following data.

```
'B2094757D66FA5FB499C',' Nan Zhou Beimian',116.37425,39.91156,68);
'B2094757D66FA0FC4299','Ma La Tang',116.3787,39.91413,68);
'B2094757D66EA2F8489A','YIPINSANXIAO(Xi Dan Branch)',116.37458,39.90958,68);
'B2094654D06FA7F94498','King Banboo',116.37595,39.9105,68);
'B2094757D668A1F54598','JUICE TIME',116.37348,39.91228,68);
'B2094757D668A1FD489D','KFC',116.37452,39.9123,68);
'B2094757D16DA2F8489E','KFC',116.477943,39.919437,68);
'B2094757D16DA3F8439F',' Kungfu Xidan Shopping Centre Shop',116.37423,39.91286,68);
'B2094757D064AAFE479C','Xidan winged Cool (Yongding Road)',116.26467,39.90282,68);
'B2094757D16AA5FD479F',' Cooking potato king Xidan Pearl Shop',116.37591,39.9091,68);
'B2094757D16FAAF94592','Half an acre of garden Xidan',116.37087,39.91016,68);
'B2094757D669A3F84192','Willow home Dousha Bao',116.373107,39.907272,68);
'B2094757D668AAFC479E','The Taste of Health Restaurant edge',116.371145,39.909135,68);
'B2094757D669A3FA4693','Must taste duck neck Xidan',116.3739,39.91285,68);
'B2094757D668A5FA4999','Gokokuji snack Xidan',116.37339,39.91892,68);
'B2094757D668A5F9409C','Kung Fu Bao Zipu Xidan steam',116.3778861,39.9089695,68);
'B2094757D668A6F9469E','Malan Noodle Xidan',116.36999,39.9091,68);
'B2094757D668A5F44799','Honey cake Xidan',116.37521,39.9082,68);
'B2094757D668A5F4459E','Motoki December Baiji steamed meat juice Store',116.37689,39.90698,68);
'B2094757D06FA4FC4398','McDonald's',116.3751,39.90913,68);;
'B2094654D168A2FC4293','Chu taste duck neck Xidan',116.37615,39.91047,68);
'B2094654D069A7F4459F','SUBWAY Xidan',116.37557,39.90586,68);
'B2094654D065A1FE449A','Heung Yuen Bridge bridge noodle (Xidan)',116.374801,39.908847,68);
'B2094654D069A4FD409B','Chen Po (Friends of Xidan Store)',116.375,39.9094,68);
```

Figure 5. POI data(Before test)

Two-dimensional scatterplots after mapping are as follows (on the Google map).

Figure 6. Micro-blog POI mapping map

Step 2

According to the S-DBSCAN algorithm described above, with MATLAB code, we write out the processing of two-dimensional POI function program, and import the POI information into the MATLAB database to do clustering, finally draw out the scatter diagram after two-dimensional S-DBSCAN clustering. After clustering, the scatter diagram is as follows

From the figure we can clearly see that, after the clustering with S-DBSCAN algorithm, the results of POI points showed that it roughly comes to three clusters, which means three clustering trade areas which we get.

Step 3

According to the simulation results of MATLAB, we can get clustering trade areas from the Google

map. Finally we can get three fast food business circles which are located in the Beijing Xi Dan area, and the specific trade areas of the clustering are as follows.

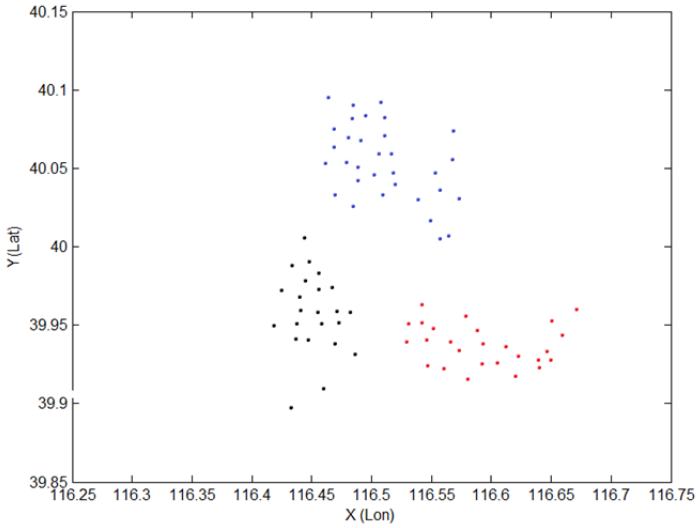

Figure 7. The MATLAB clustering simulation graph

Figure 8. Trade Area A

Figure 9. Trade Area B and C

3.3.2 Clustering results

From the above we can see that, in the Xi Dan area there are POI points of many fast food restaurants, and through cluster analysis, we can divide the Xi Dan area of fast food shops into three business circle. We can foresee, if business circles have been set by clustering, there will be more important significance and greater commercial value for the future of Xi Dan fast-food business when we do research on reasonable layout and related economic problems.

4. Conclusion

Nowadays, the “Trade Area” is a hot topic. Being different from the traditional method of analyzing business district, this paper presents a new method of analyzing, clustering and districting business circle precisely. This method uses the micro-blog POI location information data with the S-DBSCAN clustering algorithm to achieve the goal of clustering business circle. Among them, the micro-blog POI data is provided by the user through the sign location with GPS positing mobile intelligent terminal uploads. This huge amount of data, which are highly realistic with social attributes and other characteristics, is a potential source of available data to update POI, which is more accurate than traditional position data. S-DBSCAN algorithm is very suitable for business circle clustering applications. In the end of the paper, by using simulation test, we can find that the new method is reasonable, effective and practical, with reliability and great commercial value.

References

[1] Reilly, William John. Methods for the study of retail relationships. University of Texas, Bureau of Business Research, 1959.

[2] Hertog, Pim den. "Knowledge-intensive business services as co-producers of innovation." International Journal of Innovation Management 4.04 (2000): 491-528.

[3] Zhong, Ning, and Setsuo Ohsuga. "Discovering concept clusters by decomposing databases." Data & Knowledge Engineering 12.2 (1994): 223-244.

[4] Birant, Derya, and Alp Kut. "ST-DBSCAN: An algorithm for clustering spatial–temporal data." Data & Knowledge Engineering 60.1 (2007): 208-221.

[5] Zhang, Hua-Ping, Qiang Mo, and He-yang Huang. "Structured Poi data extraction from Internet news." Universal Communication Symposium (IUCS), 2010 4th International. IEEE, 2010.

[6] Goodchild, Michael F., and J. Alan Glennon. "Crowdsourcing geographic information for disaster response: a research frontier." International Journal of Digital Earth 3.3 (2010): 231-241.