A MULTILAYER ALGORITHM FOR DISCOVERING CLUSTERS IN LARGE SPATIAL DATABASES AND ITS APPLICATION

Le Si Quang[1] , Huynh Thi Thanh Binh[2], Luong Chi Mai[3]

[1]Faculty of Technology, Hanoi National University
Email: lsquang@vnu.edu.vn

[2] Faculty of Information Technology, Hanoi University of Technology
Email: binhht@it-hut.edu.vn

[3] Institute of Information Technology, Hanoi

Email: lcmai@ioit.ncst.ac.vn

ABSTRACT

Spatial Data Mining (SDM), i.e., mining knowledge from large amounts of spatial data, is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. Spatial data mining is extraction of interesting spatial pattern and features, general relationships between spatial and non-spatial data, and other general data characteristics not explicitly stored in spatial database. In this paper we present a multilayer algorithm for discovering clusters in large spatial database. This algorithm is a modification of DBSCAN clustering algorithm for spatial database, relying on a density-based notion of clusters and a criteria which decide whether clusters are similar or dissimilar. Based on proposed algorithm, the paper also presents some experimental result on investigating of a level of health care for a populated area based on the map of Vietnam. The map have to be clustered in to areas with the goal: 1/ two objects (unit area) in one cluster will have same a health care lever and 2/ the defined clusters are maximum in area.

Keywords: Spatial Data Mining, Spatial Database, Clustering Algorithm,

1 INTRODUCTION

Knowledge discovery is the process using data mining methods to extract or identify valid, novel, potentially and ultimately understandable patterns or knowledge in databases. Knowledge discovery and data mining (KDD) on spatial databases faces the problem of not only large sizes of spatial databases but also particular attributes and relationships of spatial data objects [1, 2, 3]. Thus, normal knowledge discovery and data mining methods are not efficient in both time and space. In addition, the methods or algorithms for spatial databases have many limitations. One area of the new rapidly growing interdisciplinary field of knowledge discovery and data mining is clustering in spatial databases (CSD). Spatial data mining differs from regular data mining by differences between non-spatial data and spatial data. Spatial data and their associated nonspatial information are linked to each other in different architecture and there are different representation for spatial data. In many application, spatial information is stored as thematic maps. Each map contains specific features of spatial objects, e.g., forest type and coverage. There are two representations of thematic map: raster and vector.

Main focus in this paper is on identification of pilot Spatial Knowledge Discovery and Data mining (SKDD) problems, on development of clustering techniques that well adapted to spatial database. Based on proposed technique, the paper presents some experimental result on investigating of a level of health care for a populated area based on the map of Vietnam. The objective is to discovering and defining the level of health care for populated areas is crucial in formulating a development strategy for hospitals. In order to define the level of health care for a populated area, we need first to identify the level of health care by each hospital, then sum up the levels of health care by all hospitals in the area.

The level of health care by a hospital in an area is a function of the following:

- the prestige of the hospital (numbers of doctor, competence of doctor, medical facilities...)
- travel from the area to the hospital.

Holding the variable of hospital prestige constant, we at first look at the level of health care as a function of the travel to the hospital. The travel convenience from the area to the hospital is the most convenient path from the area to the hospital, the shortest path between the area and the hospital.

This paper is organized as follow. In the next Section, we present our approach to develop an clustering algorithm, that we called multilayer algorithm for discovering clusters in large spatial database. This algorithm is a modification of DBSCAN clustering algorithm for spatial database, relying on a density-based notion of clusters and a criteria which decide whether clusters are similar or dissimilar. The related techniques, such as quadtrees and algorithm for defining shortest path between center of one area and other point (hospital) also presented in this section. Section 3 presents an experimental evaluation on some test data. Finally, in Section 4, we will summarize the results and our future work.

2 TECHNIQUES FOR SOLVING PROBLEM

2.1 Multilayer DBSCAN

There are some problems must be solved in clustering techniques applied in KDD. Clustering is a process of grouping the objects which has great degree of compatibly into meaningful cluster. Clustering algorithm in large database require following conditions: 1/ require as smaller as better the number of input parameters, 2/ clusters with arbitrary shapes can be clustered eliminating noise, 3/ the algorithm should be effective working with large spatial database. The existing algorithm such as k-means, DBSCAN [4, 5],...are work well for some kind of data but ineffective with the databases which has following properties: 1/ large difference between minimum and maximum distances of data points, 2/ one cluster is crossed with another. The reason of ineffective of theses algorithms with this kind of data is they apply the same parameters to all data points or they consider all the points belong to unique layer of data. We will call these algorithms as monolayer-clustering algorithm. In [6], we first tried to apply different parameters in the clustering algorithm and we develop that idea on this paper.

By the reason that spatial data contains above described properties, we propose new effective algorithm, called *MultiLayer-Clustering Algorithms* that based on the following condition: a database could be make distinctions in different layers, each layer is a group of

clusters which have the same properties and be produced by a monolayer-clustering algorithm with the same parameters. For *MultiLayer-Clustering Algorithm*, we need the followings definitions:

Definition 1: Layer of clusters

Given a spatial database $\mathcal{D}$ and a way to cluster $\mathcal{D}$ into a set of clusters $\{c_1, c_2,...,c_k\}$. A layer of clusters L_i in $\mathcal{D}$ is defined as a subset of clusters $\{c_{k1}, c_{k2},...,c_{kn}\}$ that satisfied the following conditions:
a/ $L_i \cap L_j = \Phi$
b/ $L_1 \cup L_1 \cup ... \cup L_p = \mathcal{D}$

With this definition, we propose a multilayer-clustering algorithm to cluster whole database $\mathcal{D}$ with the following methods.

Definition 2: Epsilon make-layer

Given some parameters $\varepsilon \in [\alpha, \beta]$ and Δ; α, β, Δ are real numbers. Using the ε for the algorithm DBSCAN to cluster a database $\mathcal{D}$ into clusters $\{c_1, c_2,...,c_k\}$, (k>0). If increase eps to $\varepsilon+\Delta$ without changing the result of clusters, then eps is called epsilon make-layer.

Based on two described definitions, suppose that a parameter ε is epsilon make-layer for a database $\mathcal{D}$ and clusters $c_1, c_2,...,c_p$ are produced when apply DBSCAN to database $\mathcal{D}$ with parameter ε, so $\{c_1, c_2,...,c_p\}$ can be considered as a layer of clusters. The basic idea of Multilayer DBSCAN (MLDBSCAN) is that the difference between minimum and maximum distances between data points in the cluster is not too large.

Multilayer DBSCAN algorithm:

1. Define a suitable epsilon make-layer $\varepsilon \in [\alpha, \beta]$ by using some clustering technique (e.g. DBSCAN) with the parameter ε.
2. Let $\{c_1, c_2,...,c_k\}$ is a set of clusters given by the clustering technique using ε, set $\mathcal{D} = \mathcal{D} \setminus \{c_1, c_2, ...c_k\}$. If $\mathcal{D}$ contain only noisy data then goto Step 3, else goto Step 1.
3. Stop

2.2 Combined techniques

The problem of discovering and defining the level of health care for populated areas is based on a map of North Vietnam in vector format, that includes the following information: 1/ highways and provincial roads are presented in polylines, 2/ each hospital presented by a point with its coordinates. Holding the variable of hospital prestige constant, we at first look at the level of health care as a function of the travel to the hospital. The travel convenience from the area to the hospital is the most convenient path from the area to the hospital, we consider it as the shortest path between the area and the hospital. The problem then becomes defining the shortest path from two points between a hospital and a center of area. The level of health care in an area is the sum up the shortest paths from all the hospitals in an area to the center of it. The problem of defining an area and its center to be solved by applying quadtree technique to divide the map of North VietNam using some population threshold.

And then, based on levels of health care as a function of the travel to the hospital, we applied clustering technique MLDBSCAN to define clusters with the same level of health care.

Quadtree technique

Quadtree is the useful technique to divide spatial areas in to sub areas easily and flexible. Based on tree concept, quadtree is a tree that its node has four children. Decision of dividing process is based on some criterion. In our case, the number of population is chosen as criterion for dividing process. The number of population in an area should not too big or too small. An area should be divided if its population is more than a predefined threshold, otherwise stop dividing process. Other criterion is based on similar weights of different points (in our case, a point is a hospital) inside an area. It means that in the same area, weight of any points need to be approximate to average weight of the area. So, the division process will be continuing when a weight of points in an area is large from each other until meeting the above condition.

Shortest path between center of an area and a hospital

As explained above, defining level of health care for a populated area problem become defining shortest path form two point: start point is center of populated area and destination point is a hospital. We convert data of highways and roads into undirected graph with its vertex, most of them are intersec with each other. In this problem, start point and destination point are not belong to vertex set of graph. So, we admit them to vertex set and new edges to edges set.

Problem of admittance of new vertex and edge to set of vertex and edge of graph to be solved as follows: in shortest path problem, it's better if the new edge is more short. So, we have to define the points nearest to a start point and a destination point and it will be a new vertex admited to vertex set. A nearest point to a start point and a destination point is defined according to following algorithm:

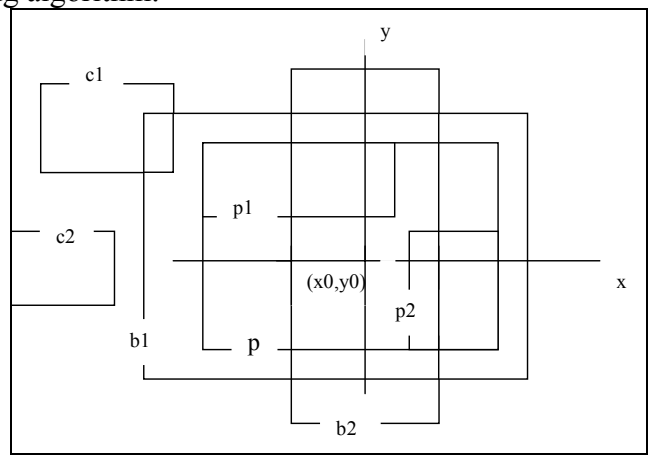

Figure 1. Minimum bounding rectangles are examined

1. B is set of minimum bounding rectangles (MBRs) of the polylines contain (x0, y0)
2. C is set of MBRs do not contain (x0, y0)
3. Define two MBRs p1, p2 in C which have minimum distance along x and y axis to (x0, y0).
4. P is MBR contain p1 and p2
5. Define set of MBRs D in C which overlap with P
6. Define nearest point to (x0, y0) in MBRs in B and D.

And then, using Dijkstra algorithm to find shortest path between start point (hospital) and destination point (center of defined area).

3 EXPERIMENTAL RESULT ON HEALTH CARE APPLICATION

Experiment result of evaluation of the level of health care is shown as below. All the data (highways, provincial roads, hospital) belong to the map of North VietNam. The total number of hospitals are 303. Experiment result was carrying out on Pentium III and takes 50s for defining shortest path from all the hospitals to defined area and sum up.

Classification of level of health care is based on two criterions: (a) level of health care less than some threshold ξ; (b) distribution of level of health care.

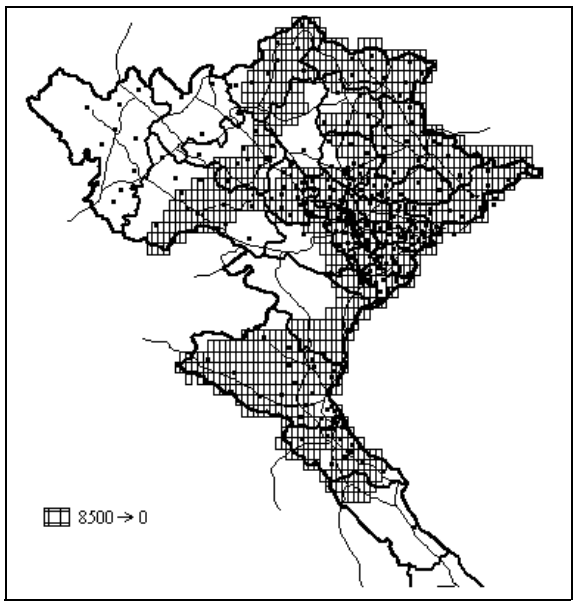

Figure 2. Level of health care less than $\xi = 8500$

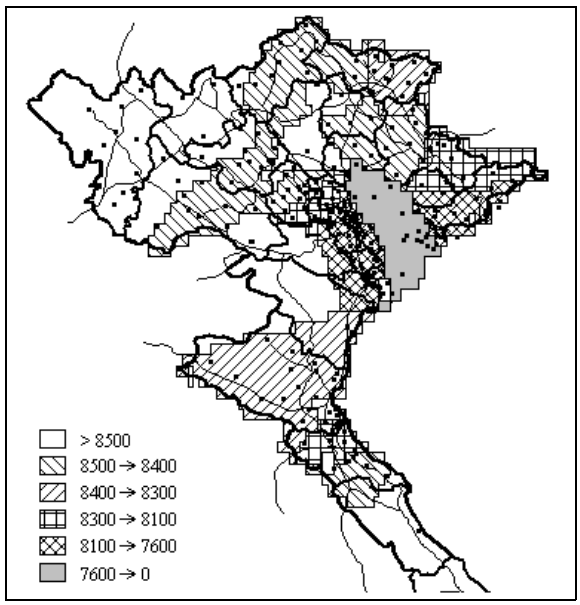

Figure 3. Multilayer level of health care

(a) Weigh per unit of highways is set to 1 and 1000 if there is no way from two points correspondingly. Using quad-tree technique in the North of Viet Nam with 7 depth degree.

ξ	0-8500	0-8400	0-8300	0-8200	0-8100	0-8000	0-7600
Number of hospitals	255	208	180	157	147	129	71

As shown in the Figure 2, in the white area with ξ>8500, the density of the hospital is small or transportation is less convenient than red area with ξ<8500. It means that level of health care in the defined area influenced by travel from defined area to the hospital around that area.

(b) Results of clustering on distribution of 1481 of levels of health care (corresponds to 1481 leaf nodes of quadtree) into 6 ranks (clusters) with ξ are: rank 1 ($\xi > 8500$), rank 2 ($\xi \in (8400 \div 8500)$), rank 3 ($\xi \in (8300 \div 8400)$), rank 4 ($\xi \in (8100 \div 8300)$), rank 5 ($\xi \in (8100 \div 7600)$), rank 6 ($\xi \in (0 \div 7600)$).

As shown in the Figure 3, if ξ is higher, level of health care is lower. And if ξ is lower, level of health care is higher. So, the region near Hanoi has ξ lowest value. We have tested a distribution of hospitals in corresponding ranks (regions) and found that with the low value of ξ, distribution of hospitals is high, it's proved that in these regions, hospitals are in locations near to highways and provincial roads.

4 CONCLUSION AND FUTURE RESEARCH

Our research presented in previous sections is, in one hand, as a pilot problem for combined different techniques to solve some aspect of Spatial Knowledge Discovery and Data mining. On the other hand, our research focused to develop clustering technique that well adapted to spatial database. Based on proposed technique, the paper also presented some experimental result on investigating of a level of health care for a populated area based on the map of North Vietnam. It's necessary to say that we have tried on some kind data, that is not up to date but we just want to show how to combine different techniques, including our proposed technique to give advices to solve a kind of societal problem, i.e. the problem for discovering and defining the level of health care for populated areas is crucial in formulating a development strategy for hospitals. The objective in the future work is to discover and evaluate factors that influence on/from population, economy, etc. to the distribution of hospitals and clinics. i.e. to use the factor with nonspatial attribute.

5 REFERENCES

[1] Usama M.Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, "*Advances in Knowledge Discovery and Data mining*",

[2] Martin Ester, Hans-Peter Kriegel, Jorg Sander and Xiaowei Xu, "*Density Based Algorithm for Discovering clusters in Large Spatial Database with Noise*", *Proceeding of 2nd International Conference on Knowledge Discovery and Data Mining* (KDD 96).

[3]. M. Hosheimer and A. Siebes, Data mining: the search for knowledge in Database, *Report CS-R9406*, CWI, Amsterdam, The Nertherland, 1994.

[5]. G. Piatetsky-Shapio and W.J. Frawley, *Knowledge Discovery in Databases*, AAAI/MIT Press, 1991.

[4] Martin Ester, Hans-Perter Kriegel, Jorg Sander, "*A Distribution-Based Clustering Algorithm for Mining in Large Spatial Database*" *Proceeding of 2nd International Conference on Data Engineering* (ICDE'98).

[5]. R.Ng and J.Han, *Efficient and effective clustering method for spatial data mining.* Proceeding of 1994 Int'l Conference on Very Large Databases (LVDB'94), September 1994.

[6]. Luong Chi Mai and Le Si Quang, *Clustering Algorithm for Large Spatial Databases.* International Forum cum Conference on Information Technology and Communication at the Dawn of the New Millennium, Thailand, Dec 2000.