Expert Systems with Applications 36 (2009) 3923–3936

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A hybrid spatial data clustering method for site selection: The data driven approach of GIS mining

Bo Fan

School of International and Public Affairs, Shanghai Jiaotong University, Shanghai 200030, China

ARTICLE INFO

Keywords:
Spatial clustering
k-Means algorithm
GIS
Site selection

ABSTRACT

This article applies customer service to be the research background. Spatial data mining method is proposed to solve site selection of the service center. Firstly, a new data model for recording all the information of customer management is given, which transforms the traditional model-driven strategy to data-oriented method. Secondly, a hybrid spatial clustering method named OETTC–MEANS–CLASA algorithm is proposed. It has the advantages of applying *k*-means algorithm to reduce the result space and using simulated annealing method (CLASA) as result-searching strategy to find more qualified solutions. On the basis of GIS functions, we design deeper analytical function to take spatial obstacle factors, spatial environmental factors, spatial terrain factors, spatial traffic factors and cost factors into account. The result of the experiment declares that the algorithm does better at the both aspects of perform efficiency and result quality.

1. Introduction

The site selection of the service center is the key factor that decides the efficiency of service response. The current research of location problem usually applies the linear programming to reach the optimization of time response and cost of the models (Aboolian, 2007; Berman & Krass, 2002; Jain & Vazirani, 2001; Zhang, 2007), etc. But the above researches have two common shortages: (1) It should involve many spatial restriction factors, such as obstacle factors, environmental factors, traffic factors, terrain factors, etc. which require tremendous amount of variables and restrictions if it is solved by linear programming models. So it is quite difficult to build such a huge linear model for site selection. (2) The essence of the location problem is to find the global optimal *k*-centers or *k* distributions which is well known to be NP-hard (Domínguez et al., 2008; Owen & Daskin, 1998), etc. the linear model is not good enough to get the more qualitative results of *k*-centers/distributions.

In this article, we choose customer management as the application field for illustration our method of the site selection of service centers, proposes a data-driven method instead of the model-oriented method, and explore a process for finding service sites of the enterprise from the huge amount of customer locations which are recorded as two-dimension points in GIS (geographical information system) (Koret & Koret, 1997). The huge amount of spatial points are regarded as service receivers which might be the current and potential customers of an enterprise.

1.1. The application of GIS in location problem

Geographical information systems (GIS) have occupied the attention of many researches involving a number of academic fields including geography, civil engineering, computer science, land use planning, and environmental sciences (Domínguez et al., 2008). GIS can support a wide range of spatial queries that can be used to support location studies. Model application and model development are the major impact of GIS on the field of location science. These systems are designed to store, retrieve, manipulate, analyze, and map geographical data. GIS can serve as the source of input data for a location model. The applications of GIS in location problem only limit to the basic functions of visualization, querying, and preliminary analytical functions of overlapping, buffering, network analysis (Cheng, Li, & Yu, 2007; Koret & Koret, 1997; Nikolakaki, 2004). Location–allocation model can also be integrated in GIS software to realize the resulting presentation in a real map (Cheng & Chang, 2001; Nathanail, 1998; Vlachopoulou et al., 2001; Yeh & Chow, 1996). In all, the current researches focus on the basic functions of GIS, and the deeply analytical functions of spatial data have not been sufficiently developed.

1.2. The application of spatial clustering in location problem

Clustering is the organization of a data set into homogenous and/or well separated groups with respect to a distance or, equivalently, a similarity measure (Tran et al., 2003). Spatial clustering, which groups data for finding all distribution patterns and interesting correlation among geographical data set, has numerous appli-

E-mail address: fanbo411@163.com

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.02.056

cations in pattern recognition (Ayala, Epifaniob, Simó, & Zapatera, 2006; Domingo, Ayala, & Dı́az, 2002), spatial data analysis (Demir et al., 2007; Hu & Sung, 2006; Lin, 2004), image processing (Chu, Roddick, & Pan, 2001), market research, etc. (Ester & Kriegel, 1998; Han, Kamber, & Tung, 2000; Han & Kamber, 2000). Spatial data clustering is an important component of spatial data mining and further exploration of GIS functions (Ester & Kriegel, 1998). Spatial clustering method can be classified into four categories: partitioning method, hierarchical method, density-based method and grid-based method (Ester & Kriegel, 1998; Zhang & Rushton, 2008). Since our research is to ensure the minimization of the overall travel distance of all the customers in the city. The clustering algorithms of hierarchical method, density-based method and grid-based method mainly focus on finding natural clusters which do not guarantee the minimization of distances to cluster centers. The partitioning algorithm is a good choice as a solution, the two typical types of partitioning algorithm are k-means and k-mediod (Han et al., 2000; Han & Kamber, 2000).

The k-means algorithm uses the mean value of objects in a cluster as the cluster center. The objective criterion used in algorithm is squared-error function defined as

$$E = \sum_{i=1}^{k} \sum_{x \in C_i} |x - m_i|^2.$$

In the above formula, x is the point in space representing the given object, and m_i is the mean value of cluster C_i. The method can be described as: Step 1: k-means algorithm arbitrarily choose k-centers/distributions as initial solutions. Step 2: k-means algorithm assigns each object to its nearest center forming a new set of cluster. Step 3: all the centers of those new clusters are then computed by taking the mean of all the objects in each cluster. The steps 2–3 are repeated until the criterion of function E does not change after an iteration. The k-means algorithm is relatively scalable and efficient in processing large data set because the computational complexity of the algorithm is $O(n * k * t)$, where n is the total number of objects, k is the number of clusters, and t is the number of iterations. Normally, $k \ll n, t \ll n$. The method often terminates at local optimum (Han et al., 2000; Han & Kamber, 2000). Except for that, k-means algorithm is also very sensitive to noise and outlier data points since a small of such data can substantially influence the mean value.

k-Mediod method uses the most centrally located object(mediod) in a cluster to be the cluster center instead of taking the mean value of the objects in a cluster (Tung, Hou, & Han, 2001). Because of this, k-mediod is less sensitive to noise and outlier data, but it results in a higher running time. The typical k-mediod algorithm is called CLARANS (clustering large application based on randomized search) (Han et al., 2000; Han & Kamber, 2000; Ng & Han, 2002; Chu, Roddick, & Pan, 2002). When searching for a better center in step 3 of k-means algorithm, CLARANS tries to find a better solution by randomly picking one of the k-centers and replacing it with anther randomly chosen object from other $(n - k)$ objects, and if no better solution is found after a certain number of attempts, the local optimum is assumed to be reached. Though CLARANS is more effective than other k-mediod algorithm, its computational cost is as high as $O(n * n)$, where n is the number of objects.

1.3. Section arrangement

On the basis of GIS function and the research of spatial data clustering method, we intent to improve the current method for better solving the problem of site selection. In Section 2, a data model will be proposed to comprehensively organize the information of the management process of customer service which is the key aspect to realize our idea of data-driven method. The section includes the modeling method and operating method of the spatial data cube for introducing the spatial analytical function to the solution of site selection. In Section 3, the computation method of spatial distance between service center and service receiver is proposed, it takes the obstacle factor and the location weight into account. In Section 4, a hybrid spatial clustering method named OETTC–MEANS–CLASA algorithm is designed to realize the mining oriented method for location selection of customer service center. Obstacle factor, environmental factor, traffic factor, terrain factor and cost factor are taken into consideration in the process of the algorithm. Finally a series of experiments will be carried out to test the efficiency and the quality of our algorithm.

2. Data model for spatial clustering

2.1. Spatial data model of customer location

The distribution of spatial objects in urban district is vital for the location selection of customer service center, the principle of which is to ensure the minimization of overall travel distance of all the customers. Data model of customer management has the function of spatially enabled analysis and can organize huge amount of customer attributes as multidimensional model which is described in the right part of Fig. 1. Spatial dimensions and non-spatial dimensions of the model are joined by SDE (spatial data engineer) which can integrate the non-spatial attributes of a customer in transaction database and spatial attributes of the same customer in GIS (Nebojsa, 1997; Papadias et al., 2001). The model can not only provides fundamental analytical function of GIS and OLAP(On-line analytical processing), but also presents an interface for developing deeply analytical functions to realize the purpose of scientific site selection.

2.1.1. The model of spatial data cube

Spatial data cube model can be defined as: $MD = (D,H,M,\rho,\boldsymbol{\Gamma})$.The spatial data cube for recording the customers' fact is shown as Fig. 1.

Definition 1 (*Dimension of spatial data cube*).
D represents the dimensional attribute set of a customer O. D is made up of M non-spatial dimension named SD and N spatial dimensions named ND, $ND \subseteq D$, $SD \subseteq D$. Where $ND = \{ND_1,ND_2,\ldots,ND_m\}$, $SD = \{SD_1,SD_2,\ldots, SD_n\}$. For example, "*district dimension*" and "*residential area dimension*" belong to spatial dimensions, "*time dimension*", and "*commodity dimension*" belong to non-spatial dimensions. The dimensional value D_i of O is denoted as $O(D_i)$.

$D_i(H) = \{H_1,H_2,\ldots,H_n\}$, where $D_i(H)$ is the value set of hierarchies for dimension D_i. n is the number of concept hierarchies for D_i, each dimension of data cube can be generalized into different concept level. For example, if D_i = "*Time dimension*", then $D_i(H) = H_1 \leftarrow H_2 \leftarrow H_3$ (*year* ← *month* ← *day*).

$D_i(H_j) = (\beta_{ij,1},\beta_{ij,x},\ldots,\beta_{ij,t})$, where $\beta_{ij,x}$ is the detailed value of $D_i(H_j)$, t is the number of the value of $D_i(H_j)$. For example $D_i(H_j)$ = "*year*", $\beta_{ij,x}$ = "*the year of 1999*".

Definition 2 (Measure of spatial data cube).
$M = \{NM,SM\}$ – M represents the measure of spatial data cube, and includes numerical measure NM and spatial measure SM.

$NM = (NM_1,NM_2,\ldots,NM_n)$ – NM is the numerical measure of spatial data cube, and is a set of numerical values, e.g. "the *expenditure* of a customer", etc. in Fig. 1.

$SM = (SM_1,SM_2,\ldots,SM_n)$ – SM is the spatial measure of spatial data cube, and is a huge set of *object pointer* (Stefanovie, Han, & Koperski, 2000), e.g. "*object pointer* points to the location of customer's residential-area/office-area of customer" in Fig. 1.

Definition 3 (*Functions of the measures*).
$\rho = (f_1(NM),\ldots,f_r(NM))$ – ρ is the function operations of numerical measure. For example, "*the total expenditure in a certain month of a customer*" can be computed by the functions of $f_i(NM) = Sum()$.