A Novel Approach for Enhancing Clustering Technique using Knowledge-based to Plan the Social Infrastructure Services

Hesham A. Salman[1]
[1] Department of Information Systems
Faculty of Computing and Information Technology King Abdulaziz University

Lamiaa Fattouh Ibrahim[2,3]
[2]Department of Information Technology
Faculty of Computing and Information Technology King Abdulaziz University
[3]Institute of Statistical Studies and Research, Cairo University

Zaki Taha[4]
[4]Department of Compute Science
Faculty of Computer and Information Sciences, Ain Shams University

B.P. 42808 Zip Code 21551- Girl Section, Jeddah, Saudi Arabia

ABSTRACT
This paper deals with social infrastructure planning problems to determine the location of the facilities of social infrastructure network and the layout. Each user must be assigned to the closest facility to be economically viable. The objective is how to make the accessibility to facilities maximum (i.e., to minimize the distance which the users traveled to reach the facilities). In this paper, we study the problem of clustering in the presence of obstacles to locate the public service facility. In this article we present a new algorithm in data mining in the presence of obstacles. Minimum pre-specified level of demand must served by each facility. The objective is to maximize the accessibility of the facilities this means also to minimize the distance travelled by users to reach the facilities. CSPOD-DBSC algorithm (Clustering with short path Obstructed Distance - Density-Based Spatial Clustering) is developed. Obstructed short path distance calculated in this algorithm by using Density-based clustering algorithm and Dijkstra algorithm. A case study involving the location of schools in districts of Mecca in Saudi Arabia is used to illustrate the application of this algorithm.

General Terms
Data mining, network planning.

Keywords
Clustering algorithm; infrastructure city planning; Spatial Clustering algorithm; Urban Planning; public service facility

1. INTRODUCTION
Civil engineers often play an essential role within the planning processes to determine the infrastructure location (or layout) and its capacity [1].

The planning problems of the social infrastructure faced by public authorities consist of determining where the facilities should be located and what the capacity of these facilities is?

Very often, the number of possible solutions of planning social infrastructure problems is very large and it is advantageous to handle them through a type of optimization model called location (or location-allocation) models. These models are classified as discrete or continuous depending on whether the facilities can be located only in some pre-specified points of the plane or anywhere on the plane. In real-world applications discrete location models are used more often than continuous location. Since the early 1960s, for this reason they have been extensively studied, and there is a vast body of literature describing solution and models methodologies.

To determine where the facilities of some infrastructure network should be located and what should be the capacity of these facilities and layout, clustering technique will be used. In data mining process clustering is one of the most useful tasks. To solve the problem of clustering large number of objects there are many algorithms that deal with this problem. These different algorithms can be classified regarding different aspects. These algorithms can be categorized into partitioning algorithms [2-4] hierarchical algorithms [2, 5, 6] density based algorithms [7-10] grid based algorithms [11- 13] and model based algorithms [14]. The clustering method consists of separating a set of objects into different groups according to some measures of goodness that vary according to the application. In spatial databases the applications of clustering present important characteristics. Spatial databases usually contain very large numbers of points. Thus, in spatial databases, clustering algorithms do not assume that the entire database can be held in main memory. So, their scalability to the size of the database is of the same importance as the good quality of clustering [15]. In spatial databases, objects are characterized by their position in the Euclidean space and, naturally, dissimilarity between two objects is defined by their Euclidean distance [16].

In many real applications the use of direct Euclidean distance has its weaknesses [16]. The Direct Euclidean distance ignores the presence of obstacles and streets paths that must be taken into consideration during clustering process.

A clustering–based solution is presented In this paper depending on using the obstructed short path distance and density-Based Clustering techniques.

A good typical application in the real-world is school network planning. In this case, the model would aim to determine the

locations and capacities of schools to minimize the distance traveled by students and taking into account that the students must assigned to the closest school of his residence. This paper is an extension version of papers [17] and [18]. In section 2 Motivation is discuss. In section 3, we introduced the CKB-WSP algorithm. In section 4 a case study was presented. Related work discussed in section 5. Conclusion was presented in section 6.

2. MOTIVATION

There are four categories in spatial clustering algorithms. They are partition based, the hierarchical based, the density based and the grid based. Since our objective is to discover good locations that are hidden in the data we found that the most suitable clustering method is, partitioning based and density algorithms which include two major categories, k-means and k- medoids. To reduce the cost function these two methods are to randomly partitioning the database into k subsets and refine the cluster centers repeatedly. The cost function in the spatial domain is the sum of distance error E from all data objects to their assigned centers.

The non center data points are assigned to the centers that they are nearest to it. The k-means algorithm is easy to understand and implement, and also known for its quick termination. The k-means algorithm defines the cluster centers to be the gravity center of all the data points in the same cluster. In regular planar space, the cluster gravity center guarantees the minimum sum of distances between the cluster members and itself. However, the research proof [15] that the characteristic in obstacle planner space does not behave the same as the gravity center. The k-medoids algorithm chooses an actual object in the cluster as the clusters representative (medoid) instead of representing the clusters by their gravity centers. Using the real object decreases the k-medoids sensitivity to outliers. This technique also guarantees that the center is accessible by all data objects within the same cluster.

By comparing CLARANS and CLARA with PAM, CLARA first draws random samples of the data set and then apply PAM on those samples. Unlike CLARA, CLARANS draws a random sample from all the neighbor nodes of the current node in the searching graph. Efficiency depends on the sample size and a good clustering based on samples will not necessarily represent a good clustering of the whole data. The PAM (Partioning Around Medoids) algorithm which called also the K-medoids algorithm, represents a cluster by a medoid [16]. Initially, the number of desired clusters is input and a random set of k items is taken to be the set of medoids. Then at each iteration, all items from the input dataset which are not currently medoids are examined one by one to see if they available to be medoids. That is, the algorithm determines if there is an item that may replace one of the existing medoids. By looking at all pairs of medoids, non-medoids objects, the algorithm choose the pair that improves the overall quality of the clustering the best and exchanges them. Quality here is measured by the sum of all distances from a non-medoid object to the medoid of the cluster that they are in.

The total impact to quality by a medoid change TCih is given by:

$$TC_{ih} = \sum_{h=1}^{k} \sum_{n_i \in C_h} dis(n_h, n_i) \qquad (1)$$

An item is assigned to the cluster represented by the medoid to which it is closest (minimum distance or direct Euclidean distance between the customers and the center of the cluster they belong to).

PAM is not suitable for our problem because the Euclidian distance did not represent the reality in the presence of obstacle. The second reason, the number of facilities is not known before work. The last reason we need for each facility service a predefine number of population (density population).

The DBSCAN algorithm is a well-known algorithm of type Density-Based algorithm which is used when a cluster is a high dense region of points, separated by low-density regions. This algorithm makes clusters with different shapes. This algorithm also is not suitable for our problem where it has not this property. We search a clustering algorithm which construct cluster with a density within a given range and also a point in this cluster which represents this cluster such that the cost is minimum.

3. CKB-WSP ALGORITHM

The existing of the natural obstacles are affecting on distributing the service facility on the regions. Very often, the possible solutions for social infrastructure planning problems are extremely large and it is advantageous to handle them through a type of optimization model. These models are classified as discrete or continuous depending on whether the facilities can be located only in some pre-specified points of the plane or anywhere on the plane. In real-world applications discrete location models are used more often than continuous location.

In a certain city, we determine the number of public service facility requirements and define their boundaries to satisfy shortest path between facilities and users. We must take into account that each facility must serve a minimum level of demands to be economically viable and that each user must assign to the closest work facility.

The solution we proposed in this paper has the following features:

-The objective is to minimize the demand-weighted total distance (travel time - travel cost).

- A facility can only be found if it serves a minimum level of demand. Thus the capacity of each facility must exceed that given minimum to be economically viable.

- The number of facilities to be opened is the output of the model.

- Users must be assigned to the closest open facility. If the travel distance is the same for two or more different facilities, users should assigned to one and only one of those facilities.

The problem statement:-

- Inputs:

-A set T data points {t_1, t_2… t_n} in two diminutions map.
- Surface of area to be planed.
- Obstacles location.
- MinPTS = minimum population services for this public services facility.
- MaxPTS= maximum population that can serviced by this public services facility.
- Candidate locations of services facilities.

- Objectives:

- Partitioning the city into k clusters C1, .., Ck that satisfying cluster constraints (minimum and maximum services population) such that the cost function is minimum.

Min TC= ∑ obstacle distance (i, j)* wi
Where:

TC is the cost function to be minimize

Obstacle distance (i, j) = min path obstacle distance between node i and facility j of the cluster which is calculate by Dijkstra algorithm

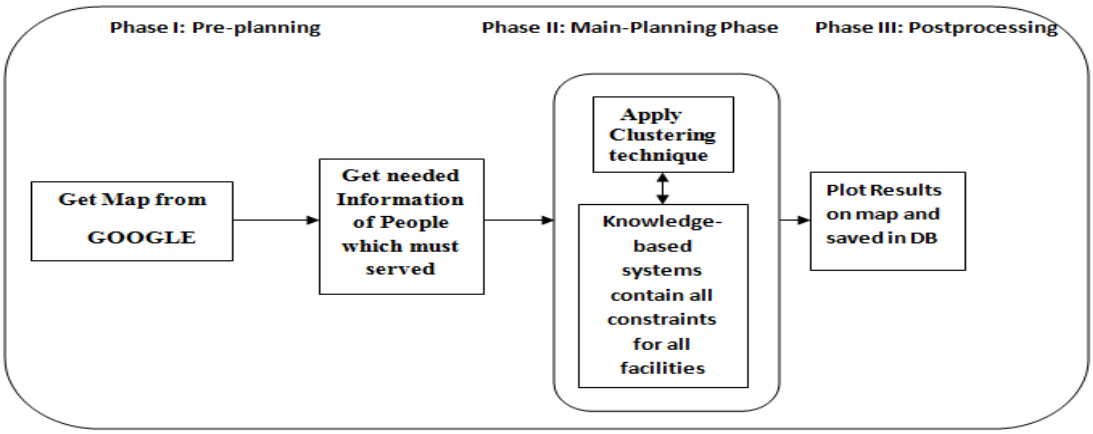
wi = weight of node i = population of node i

- Output:

- Optimal number of clusters which satisfy the required objectives.

- Locations of public services facility boundaries of each cluster.

The proposed algorithm contains three phases. Figure 1 shows the block diagram of the system overall. The following sections describe the three phases.



"Figure 1": Block diagram of CKB-WSP system

3.1 Phase I : Pre-Planning

The maps used for planning are scanned images obtained by the user from GOOGLE map. It needs some preprocessing operations before it used as digital maps, we draw the streets and intersection nodes on the raster maps, the beginning and ending of each street are transformed into data nodes, defined by their coordinates. The streets themselves are the links between data nodes. The populations are considered to be the weights for each node.

3.2 Phase II : Main-Planning Phase

CKB-WSP is divided into two steps:

1- Step 1: Preprocessing.
2- Step 2: CKB-WSP algorithm.

3.2.1 Preprocessing

During clustering the CKB-WSP often needs to compute the shortest obstructed path distance between a point and a temporary cluster center. The aim of pre-processing is to manipulate the information which will facilitate such computation.

CKB-WSP algorithm calculate the shortest path from one source (public service facility) to all destinations by using Dijkstra algorithm to determine the suitable location and layout of public service facility and from one node to all public service facility to determine the nearest suitable public service facility that will serve this node. Figure 2 shows the pseudo code of the Dijkstra algorithm**.**

```
Function Dijkstra(G, w, s)
For each vertex v in V[G]//      Initializations
    d[v] := infinity
    previous[v] := undefined
    d[s] := 0
    S := empty set
    Q := set of all vertices
While Q is not an empty set // The algorithm itself
    u := Extract_Min(Q)                    S := S union {u}
  for each edge (u,v) outgoing from u
   if d[v] > d[u] + w(u,v)          // Relax (u,v)
      d[v] := d[u] + w(u,v)
      previous[v] := u
End Function
```

"Figure 2": pseudo code of the Dijkstra algorithm

3.2.2 CKB-WSP algorithm

Figure 3 shows implementation of pseudo code of CKB-WSP algorithm used**.** The algorithm begins with estimate number of clusters which is equal to sum of population of all points divided by MaxPTS, where MaxPTS is the maximum population this facility can served.

The user first inserts the location of candidate. Our package arbitrarily selects K points from candidates to be the initial location of services facilities. Then the package determines the boundaries of cluster by calculating the obstacle distance from each node to each facility. The algorithm iterate until chooses one which satisfies the conditions and minimize the cost function. Each type of facility has his constraints. These constraints differ from one to another. To open school, government determines two constraints MinPTS which is the minimum number of students that can open a school and MaxPTS the maximum number of students which determine from the surface of the school. The maximum distance which can the students wake to go to school Eps must be know. If the government needs to open mosque, any human to pray need .5m*1 m to know how many humans can pray in this mosque (MaxPTS), divide surface of mosque by .5 m^2. The maximum distance which can the human can wake to go to mosque, Eps, must be know. Knowledge-Based system is constructed and contain all constraints which is affected the plan of facilities services.

1.1 Phase III: Post processing

The output mined knowledge is presented graphically and as data in data base. The following section shows case study and demonstrates the output knowledge.

4. CASE STUDY

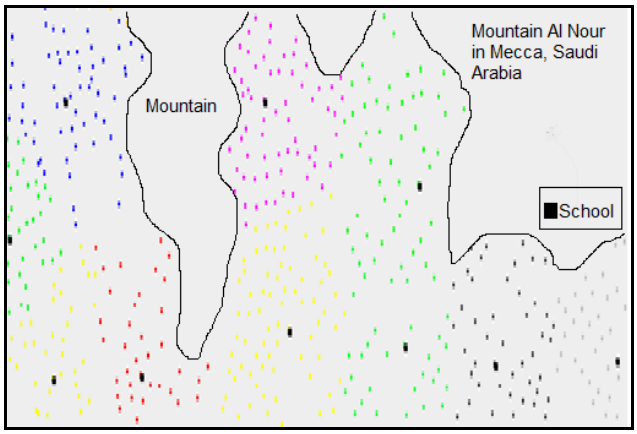
For real application, the proposed algorithm is applied on two maps representing a district in Mecca in Saudi Arabia. This area suffer of mountains the bigger one called Al Nour mountain. We scanned the actual map, then the beginning and the ending of each street are transformed into data points, defined by their coordinates; the streets themselves are transformed into linkages between data points; after that the population of each node is added.

Figure 4 shows this area; all dark areas represent mountains area. In this case study the facility is a school which must serves at least 100 students and maximum 700 students. Figure 5 shows the map after applying CKB-WSP algorithm, MaxPTS= 700. As shown in the figure the facilities are located in actual place not in the obstacle location. The layout of each facility is not cutting any obstacle.

Algorithm CKB-WSP
Input
D={t_1, t_2, t_3,……..t_n) / * set of elements
Surface of area to be planed
Obstacles location
Eps maximum distance between I and public service facility
MinPTS = minimum population services for this public service facility
MaxPTS= maximum population that can service by this public services facility
Output
A partition of the D objects into K cluster
Location of public service facility
Boundaries of each cluster
CKB-WSP Algorithm
Enter type of facility
Load constraints parameter from Knowledge-Based system
Estimate number of cluster= K= (∑ weight of node I / MaxPTS)
currentTC= big number
Label 1 Arbitrarily select K points from Candidate to be the location of services facilities
For (i=1 to candidate No.)
For (j=1 to number of node)
Calculate the short path obstacle distance from public service facility (i) to node(j)
If (path obstacle distance < Eps km)
Then current population (i)= current number of population (i) + population of node
End For
If (Current population (i) >= MaxPTS)
Then add one center of public service facility and go to label 1
If (Current population (j) < MinPTS)
Go to label 1
End for
For each node in the city select the best location service for it by calculating the obstacle distance between the nodes and each location service
Calculate the number of population of each core
Calculate TC If TC < currentTC than currentTC = TC go to label 1
Save solution

"Figure 3": Implementation of CKB-WSP algorithm

"Figure 4" Area in Macca City in Saudi Arabia



"Figure 5" Using CKB-WSP algorithm considering the location of obstruct MinPTS= 100 and MaxPTS=700

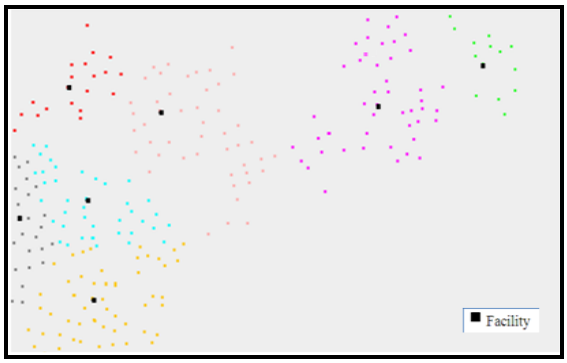
Figure 6 shows the second area; all dark areas represent mountains area. Figure 7 shows the map after applying CKB-WSP algorithm which divides the map into 7 clusters when Eps= 50 and Minpnt= 4000. Figure 6 shows the map after applying CKB-WSP algorithm which divides the map into 3 clusters when Eps= 70 and Minpnt= 6000. From figure 5 and 6, the location of facilities is not showed in the center of cluster but move towards heavy nodes (population) due to the weight parameter whish is introduced in the cost function.

In the proposed algorithm you can enter in each run the values of the minimum number of population (MinPNT), maximum number of population (MaxPNT) the facility can served and the radius of region (Eps) which the facility will be serve. This makes the algorithm more flexible to plan any facility (e.g. preschool, hospital, telecommunication service……)

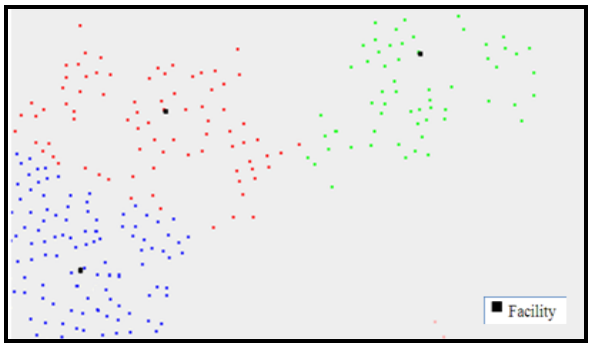
"Figure 6": Area in Macca City in Saudi Arabia

5. RELATED WORK

In [19], the most common objective is the minimization of cost. However, the authors' experience, tells that this objective in participatory social infrastructure planning processes often poses problems. The main reason is that users are avers to accept a cost minimization objective, especially when the matter is the location of facilities as important as schools or hospitals. Another relevant reason is that cost information is often rare and poor, and cost values can be difficult to estimate (specially the value of fixed costs).



"Figure 7": using CKB-WSP algorithm considering the location of obstruct Eps= 50 and Minpnt= 4000



"Figure 8": using CKB-WSP algorithm considering the location of obstruct Eps= 70 and Minpnt= 6000

On the contrary, the maximization of accessibility to facilities tends to be a more consensual objective among the different stakeholders. This objective is usually represented in location models by the minimization of the demand-weighted total (or average) distance traveled to obtain the service. One classic model that considers this objective is the *p*-median model [20] originally; the *p*-median model is based upon two important assumptions: we know a priori how many facilities should be opened and the capacity of facilities does not have to satisfy maximum and/or minimum limits. However, in a social infrastructure planning problem, the number of facilities to locate is typically one of the desired outcomes (rather than a parameter) and the capacity of facilities must be within certain limits.

Although there is an abundant literature on incapacitated *p*-median models, capacitated versions have been less studied. Moreover, most of the existing models only take into account maximum capacity constraints (recent examples include [21], [22], [23]. However, minimum capacity constraints are important because they model the minimum level of demand that facilities must satisfy to be economically viable. Above this level, possible economies of scale have already been made, and unit facility costs can be considered to be constant.

In [24], the model solve small-size instances exactly using exact method. For large-size instances, it uses heuristic methods. First, Using a classic local search heuristic (Add + Interchange) and classic population heuristic (GA) failed to identify optimum or near-optimum solutions but the solutions provided by the (Add + Interchange) heuristic were better than those given by the GA. Second, Using, Tuba Search TS and Specialized Local Search Heuristic SLSH in which the neighborhood structure was improved to better represent the specific features of the model. SLSH generally provides better solutions than TS, though requiring a larger computing effort.

Table 1 described different comparison between the proposed method and other methods in facilities planning. There are two methods that are frequently used here: Tabu search [24] and Genetic Algorithms [24].

6. CONCLUSION

Clustering analysis is one of the major tasks in various research areas. The clustering aims are identify and extracting significant groups in underlying data. Based on certain clustering criteria; the data are grouped so that the data points in a cluster are more similar to each other than points in different clusters. In this paper, we introduced a clustering solution to the problem of locate public Service facility in the presence of physical obstacles, the CKB-WSP algorithm. This algorithm is density-based clustering algorithm using distances which are weighted shortest obstacle path distance (not Euclidian distance) and satisfying facilities constraints due to use of knowledge-based system. The result is a realistic solution representing the population demand with minimum costs due to modify in cost function.

The CKB-WSP algorithm application was illustrated by a case study in a district in Mecca in Saudi Arabia. Experimental and analysis results indicate that the CKB-WSP algorithm is effective to satisfy population demands with facility constructed in an area where population is non-homogeneous due to the presence of obstacles.

The existence of Knowledge-Based System helps us to plan any new facility serves after define the constraints of this facility in the Knowledge-based.

7. REFERENCES

[1] Bigotte, J. F., Antunes, A. P. 2007. Social Infrastructure Planning: A Location Model and Solution Methods, Computer-Aided Civil and Infrastructure Engineering 22 (2007) 570–583.

[2] Kaufman L., and Rousseeuw, P, 1990. Finding groups in Data: an Introduction to cluster, John Wiley & Sons.

[3] Han, J., Kamber, M., and Tung, A. 2001. Spatial Clustering Methods in data mining: A Survey, Geographic Data Mining and Knowledge Discovery.

[4] Bradly, P., Fayyad, U., and Reina, C. 1998. Scaling clustering algorithms to large databases. In proc. 1998 Int. Conf. Knoweldge Discovery and Data mining.

[5] Zhang, T., Ramakrishnan, R., and Livny, M. 1996. BIRCH: an efficient data clustering method for very large.

[6] Guha, S., Rastogi, R., and Shim, K. 1998. Cure : An efficient clustering algorithm for large databases. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98).

[7] Ester, M., Kriegel, H., Sander, J. and Xu, X. 1996. A density based algorithm for discovering clusters in large spatial databases. In Proc. 1996 Inc. Conf. Knowledge discovery and Data mining (KDD'96).

[8] Ankerst, M., Breunig, M., kriegel, H. and Sander, J., 1999. OPTICS: Ordering points to identify the clustering structure. In Proc. 1999 ACM-SIGMOD Int. Conf. Management of data (SIGMOD'96).

[9] Hinneburg, A. and Keim, A. 1998. An efficient approach to clustering in large multimedia databases with noise. In Proc. 1998 Int. Conf. . Knowledge discovery and Data mining (KDD'98).

[10] Ibrahim, L. F. 2011. Enhancing Clustering Network Planning Algorithm in the Presence of Obstacles, KDIR International Conference on Knowledge Discovery and Information Retrieval, KDIR is part of IC3K, the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, Paris, France 26- 29 October 2011.

[11] Wang, W., Yang, J., and Muntz, R. 1997. *STING: A statistical information* grid approach to spatial data mining. In Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97).

[12] Sheikholeslami, G., Chatterjee, S. and Zhang, A. 1998. Wave Cluster : A multi- resolution clustering approach for very large spatial databases. In Proc. 1997 Int. Conf. Very Large Data Bases (VLDB'97).

[13] Agrawal, R., Gehrke, J., Gunopulos, D., Raghavan,P. 1998. Automatic subspace clustering of high dimensional data for data mining application. In Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIGMOD'98).

[14] Kohonen, T. 1982. Self organized formation of topologically correct feature map. Biological Cybernetics.

[15] Nanopoulosl , A., Theodoridis , Y., Manolopoulos, Y. 2001. C2P: Clustering based on Closest Pairs. Proceedings of the 27th International Conference on Very Large Data Bases, p.331-340, September 11-14.

[16] Tan, P., Steinback, M., and Kumar, V. 2006. Introduction to Data Mining. Addison Wesley.

[17] Salman, H. A., Ibrahim L. F., Fayed Z. 2013. Enhancing Clustering Technique Using Knowledge-Based to Plan The Social Infrastructure Services. 5th International Conference on Agents and Artificial Intelligence (ICAART 2013), Bacelona, Spain, 15-18 February 2013.

[18] Salman, H. A., Ibrahim L. F., Fayed Z. 2013. Enhancing Clustering Technique to Plan Social Infrastructure Services. ISMS2013 Fourth International conference on Intelligent Systems, Modelling and Simulation, Bangkok (Thailand) 29-31 January 2013, IEEE Xplore Press.

[19] Cornuejols, G., Nemhauser, G. L. &Wolsey, L. A. 1990. The uncapacitated facility location problem, in P.B. Mirchandani and R. L. Francis (eds.), Discrete Location Theory, JohnWiley & Sons, New York, pp. 119–71.

[20] Mirchandani, P. B. 1990. The p-median problem and generalizations, in P. B. Mirchandani and R. L. Francis, (eds.), Discrete Location Theory, John Wiley & Sons, New York, pp. 55–117.

[21] Lorena, L. A. N., Senne, E. L. F. 2004. A column generation approach to capacitated p-median problems, Computers and Operations Research, 31(6), 863–76.

[22] Ceselli, A. ,Righini,G. 2005. A branch-and-price algorithm for the capacitated p-median problem, Networks, 45(3), 125–42.

[23] Diaz, J. A., Fernandez, E. 2006. Hybrid scatter search and path relinking for the capacitated p-median problem, European Journal of Operational Research, 169(2), 570– 85.

[24] Bigotte, J. F., Antunes, A. P. 2007. Social Infrastructure Planning: A Location Model and Solution Methods, Computer-Aided Civil and Infrastructure Engineering 22 (2007) 570–583.

"Table 1": Relative Works

Algorithm	Input Parameters	Results	Location of facility	Constraints	Type of distance	consider obstacles
Genetic	- Data points - Population - Initial probability - Mutation probability - Crossover probability - Number of iteration - Selection pressure	# of facilities	Optimal placement	Fitness Function	Euclidean distance	NO
Tabu Search	- Data points - Init probability - Generation probability - Recency factor - Frequency factor - Number of iteration - Number of neighbors	# of facilities	Optimal placement	Tabu List	Euclidean distance	NO
CKB-WSP	- Data points -Eps - MinPNT - MaxPNT	- Core points - # of facilities	Core point	MinPts, Eps and obstacles	Short path obstacles distance	YES