International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:3, No:1, 2009

Optimizing Spatial Trend Detection By Artificial Immune Systems

M. Derakhshanfar, B. Minaei-Bidgoli

***Abstract*—**Spatial trends are one of the valuable patterns in geo databases. They play an important role in data analysis and knowledge discovery from spatial data. A spatial trend is a regular change of one or more non spatial attributes when spatially moving away from a start object. Spatial trend detection is a graph search problem therefore heuristic methods can be good solution. Artificial immune system (AIS) is a special method for searching and optimizing. AIS is a novel evolutionary paradigm inspired by the biological immune system. The models based on immune system principles, such as the clonal selection theory, the immune network model or the negative selection algorithm, have been finding increasing applications in fields of science and engineering.

In this paper, we develop a novel immunological algorithm based on clonal selection algorithm (CSA) for spatial trend detection. We are created neighborhood graph and neighborhood path, then select spatial trends that their affinity is high for antibody. In an evolutionary process with artificial immune algorithm, affinity of low trends is increased with mutation until stop condition is satisfied.

***Keywords*—**Spatial Data Mining, Spatial Trend Detection, Heuristic Methods, Artificial Immune System, Clonal Selection Algorithm (CSA)

I. Introduction

MANY organization have collected large mounts of spatially referenced data in various application areas such as geographic information system (GIS), banking, retailing and census. These are valuable mines of knowledge vital for strategic decision making and motivate the highly demanding field of spatial data mining i.e., discovery of interesting, implicit knowledge from large amount of spatial data [1].

So far many data mining tasks have been investigated to be applied on spatial databases. In [2] spatial association rules are defined and an algorithm is proposed to efficiently exploit the concept hierarchy of spatial predicates for better performance. Many improved spatial association rule mining algorithms are also recently proposed [2], [3]. Spatial classification models that predict some spatial phenomena are also studied in many research works [4]. Shekhar et. al [5] further improve spatial classification by considering the spatial autocorrelation concept. Also many spatial clustering algorithms have been developed e.g. [6]. One of the most valuable and interesting patterns potentially found in spatial databases are spatial trends [7], [8], [10].

In spatial trend analysis, patterns of change of some non-spatial attributes in the neighborhood of an object are explored [7], [8]; e.g. moving towards north-west from the city center, the average income of the population decreases (confidence 81%).

Ester et al. studied the task of spatial trend discovery by proposing an algorithm which applies a general clustering method [10]. This algorithm was further improved in [8] by exploiting the database primitives for spatial data mining introduced in [9].

In this later algorithm, first a specified start object *o* is given by the user. Then it examines every possible path in the neighborhood graph beginning from *o* to check its regression confidence. But in this approach the search space soon becomes tremendously huge by increasing the size of neighborhood graph and makes it impossible to do a full search.

Recently, many solutions for NP-Complete search and optimization problems have been developed inspired by the heuristic methods [11] and biological studies have always constituted a large pool of inspiration for the design of engineering systems. These last decades, two biological systems have provided a remarkable source of inspiration for the development of new types of algorithms: they are neural networks and evolutionary algorithms. In recent years, another biological inspired system has attracted the attention of researchers, the natural immune system and its powerful information processing capabilities. In particular, it performs many complex computations in a highly parallel and distributed fashion. The key features of the immune system, which provide several important aspects to the field of information processing, are: recognition, feature extraction, diversity, learning, memory, self-regulation, distributed detection, probabilistic detection, adaptability, specificity, etc. [12].

It is to be noted that the mechanisms of the immune system are remarkably complex and poorly understood, even by immunologists. Several theories and mathematical models have been proposed to explain the immunological phenomena. There are also a growing number of computer models called Artificial Immune System (AIS) to simulate various

M. Derakhshanfar is with the Computer Engineering Department of Iran University of Science and Technology,Tehran, Iran (corresponding author to provide phone: 0098-021-77510691; fax: 0098-021-77507401; e-mail: Derakhshanfar@Comp.iust.ac.ir).

B. Minaei Bidgoli is with the Computer Engineering Department of Iran University of Science and Technology, Tehran, Iran. (e-mail: b_minaei@iust.ac.ir).

components of the immune system and the overall behavior from the biological point of view [13]. The models based on immune system principles, such as the clonal selection theory [14], the immune network model [15], [16], [17] or the negative selection algorithm [18], have been finding increasing applications in fields of science and engineering [19] such as: computer security, virus detection, process monitoring, fault diagnosis, pattern recognition, etc. Although the number of specific applications confirms the interest and the capabilities of these principles, the lack of a general purpose algorithm for solving problems based on them contrasts with the major achievements in that are for other biologically inspired models, and the presented algorithms only simulates a little of principles of the immune system. Since the immune system is a complex biologic system, which includes many principles, which can offer elicitations for engineering application.

In this paper, a novel immune algorithm for mining spatial trend patterns is introduced and customized cloning and mutation operators so that they conform to the requirements in the problem of spatial trend discovery. The effectiveness of these operators and the affinity analysis used is approved by the results obtained form the experiments conducted on a real-life large spatial database of a census data. Consequently, in contrast with the previous algorithm [7], [8], in our approach the user given confidence threshold does not affect the search process. In addition, the search for trends with different start objects are integrated and run cooperatively in parallel. Therefore, there will be no need to ask the start object from the user. In this way, the property of *user-independence* which is considered as an important advantage in knowledge discovery algorithms are gained.

II. Spatial Trend Detection

In order to model the mutual influence between the spatial objects some spatial relations between objects (called neighborhood relations) are formally defined [7]. These include direction, metric and topological relations. Based on these spatial relations the notions of neighborhood graph and neighborhood path are defined as follows [7], [9]:

Definition 1: let neighbor be a neighborhood relation and DB be a database of spatial objects.

A *neighborhood graph* G = (N, E) is a graph with nodes $N = DB$ and edges $E \subseteq N \cdot N$ where an edge $e = (n_1, n_2)$ exists iff $neighbor(n_1, n_2)$ holds.

Definition 2: A *neighborhood path* of length k is defined as a sequence of nodes $[n_1, n_2, \ldots, n_k]$, where $neighbor(n_i, n_{i+1})$ holds for all $n_i \in N$, $1 \leq i < k$.

As we have the location dimension in a spatial database, one useful pattern could be the change of a non-spatial attribute with respect to its distance form a reference object. E.g. beginning form a trade center in the city and moving on a specific highway towards the west, the unemployment rate grows (confidence 72%). Having available the desired neighborhood graph, the notion of spatial trends can be defined as follows [8]:

Definition 3: A *spatial trend* is a path on the neighborhood graph with a length k that the confidence of regression on its nodes data values based on their distance from the start node is above a user-given threshold figure 1.

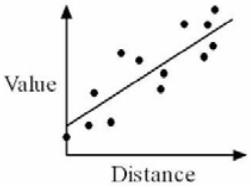

Fig.1 Regression line for a trend

The example spatial database contains the census Data and their various (non-spatial) cost of living data like the food, education, traveling, housing and etc. A map of these points and a sample trend are provided in figure 2. As an example we may need to find trends of "the cost of bread or other food usage per person in different cities" in the cities. Having discovered such trends, we can try to explain their existence by some spatial attributes [7], [8]. As an example a trend may approximately match with a road or a highway. We can also check if there are any matching trends on the same path but in other thematic layers such as demographic or land use layers. A trend can predict the non-spatial attribute value of a new point on its path using the regression equation. The reliability of this prediction is equal to the regression confidence.

Fig. 2 Cities spatial points and sample spatial trend

A desired informative spatial trend pattern would not be crossing the space in an arbitrary manner [7], [8]. So a direction filter is applied when forming the path of a candidate trend. Ester et al. used some direction filters in [7] like the starlike filter depicted in figure 3.a. and variable starlike filter in figure 3.b.

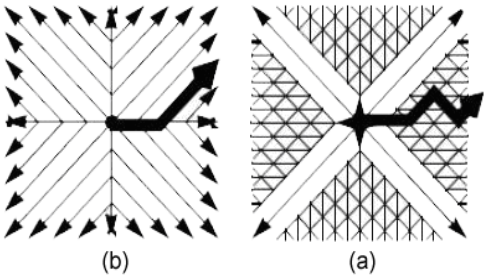

Fig. 3 (a) Variable starlike filter. (b) Starlike filter

We have used filter that proposed in [20]. In this filter that shown in figure 4, direction of the first edge of the path as the *main direction* of a candidate trend. This filter, accepts new directions to be the same as the main direction or rotated one step clockwise or counterclockwise.

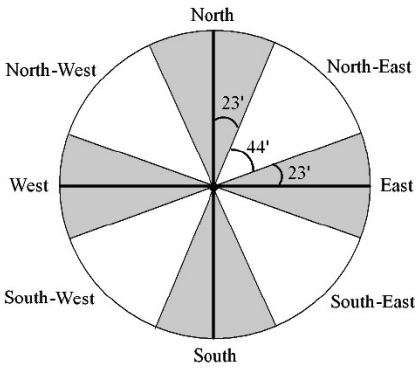

Fig. 4 Direction filter.

III. Evolutionary Mining of Spatial Trends

A. Introduction and Motivations

Many problems in data mining are search and optimization problems. In fact the database can be considered as a search space and the data mining algorithm as a search strategy [1]. The search strategy is used to efficiently walk and explore the space as non-exhaustive search becomes inevitable in large databases. In [20] proposed a genetic algorithm for efficient discovery of spatial trends and in [23] shown the Ant Colony Optimization (ACO) for spatial trend detection. ACO is a widely-used meta-heuristic inspired by the behavior of real ant colonies in the nature. Ant colonies can intelligently solve complex discrete problems (e.g. finding shortest path) although their individuals are so simple and never intelligent enough to solve such problems on their own [21], [22]. This approach was first applied by Dorigo *et al.* on the traveling salesman problem.

Recently, Immune algorithms have become a search strategy that use for search and optimization problems in many fields including data mining. Immune algorithms have been studied for the Traveling Salesman Problem (TSP) which is a typical NP-Hard graph search problem like spatial trend discovery.

There are many motivations to apply the evolutionary search strategy used in a immune algorithm, in the problem of discovering spatial trends.

1. the problem of mining spatial trends is NP-Complete and the full exploration of the search space is infeasible in the huge geo-spatial databases available today.
2. the graph representation of the problem suggests easy coding and assessment of the antibodies.
3. the nature of the problem matches well with the search mechanism used in artificial immune system.

Consequently, it generally seems that the evolutionary process can effectively and efficiently guide the search for spatial trends in the neighborhood graph.

B. Neighborhood Graph Construction

In the first step we need to construct a neighborhood graph to perform the spatial trend knowledge discovery on this graph. The definition of the neighborhood relations and edges are to be specified and tailored in a way to match with the knowledge discovery demands in the geo-spatial database and the business problem. In our sample we used the following procedure to get the neighborhood graph.

the city points P_i and P_j are connected with two directed edges E_{ij} and E_{ji} iff: Distance(i, j) < *Max-Distance* and there is no other point P_k in the Hallow-Area of P_i and P_j namely H_{ij}. H_{ij} is the area where for any point P_k in this area:

– The distance between P_k and the spatial line connecting P_i and P_j, is less than a given maximum value, namely *Hallow-Distance*.

– The angle between the lines of (P_k, P_i) and (P_k, P_j) with respect to the line of (P_i, P_j) is less than a given maximum value namely *Hallow-Angle*.

This means that if there is to be a spatial trend where P_j comes after P_i, P_k must be also present in that, making a spatial sequence of P_i, P_k and P_j. figure 5 graphically shows how the *Hallow-Area* is defined.

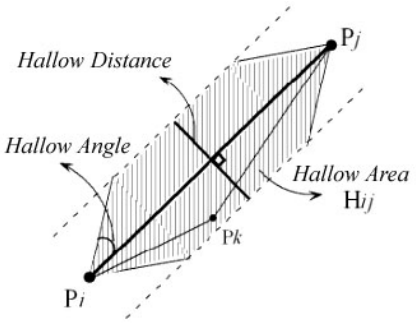

Fig. 5 Hallow Area

Finally a direction filter that shown in figure 4, is assigned to each spatial edge with respect to its angle with the

horizontal line connecting west to east.

C. Immune Algorithm For Spatial Trends (IA4ST)

Artificial immune systems (AIS) are adaptive systems, inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving [11]. Clonal selection theory, where an active B-Cell produces antibodies through a cloning process: the produced clones are also mutated. The clonal selection algorithms (CSA) are a special kind of Immune Algorithms using the clonal expansion and the affinity maturation as the main forces of the evolutionary process.

We Proposed algorithm depend on CSA to discover trend patterns in geo spatial databases. This algorithm is described as follows:

1) Generate initial antibodies (each antibody represents a solution that represents a path in neighborhood graph).
2) Compute the fitness of each antibody. The used fitness function computes the average log probability over training data.
3) Select antibodies from population which will be used to generate new antibodies (the selection can be random or according to the fitness rank). The antibodies with highest fitness are selected such that they are different enough as described later.
4) For each antibody, generate clones and mutate each clone according to fitness.
5) Delete antibodies with lower fitness form the population, then add to the population the new antibodies.
6) Repeat the steps from 2- 5 until stop criterion is met. The number of iterations can be used as the stop criterion.

1) Representation

Each antibody represents a candidate solution**.** Each member (antibody) of the generation represents the sequence of spatial nodes like figure 6.

Fig. 6. The antibody that represent the candidate trends by coding their node sequence.

2) Affinity and selection

Selection in clonal Selection algorithm depends on the affinity values for each antibody; the antibodies with the highest affinity are selected such that they are different enough.

The affinity value for each antibody is computed as follows:

$$r = \frac{\frac{\sum xy}{n} - \frac{\sum x \sum y}{n^2}}{\sqrt{\frac{\sum_{i=1}^{n} x^2}{n} - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n^2} \times \frac{\sum_{i=1}^{n} y^2}{n} - \frac{\left(\sum_{i=1}^{n} y\right)^2}{n^2}}} \quad (1)$$

The quality of an antibody which is a candidate spatial trend is measured with an evaluation function. To evaluate a spatial trend we consider the main criterion that is the regression model. The coefficient of determination (R^2) in the linear regression is used to assign an affinity value to an individual. This value is a fraction between 0.0 and 1.0, and has no units; regardless of the trend is increasing or decreasing i.e. the slope of the regression line is positive or negative. The valid trends having higher confidence are given to the output.

3) Operations

a) Cloning

The cloning operator copy a selected antibody to mating pool (cloning rate).

b) Mutation

The mutation operator changes a node in the path sequence with a probability equal to the mutation rate. To replace a node in position *i* in the path sequence, the new node must be connected to the nodes in position *i-1* and *i+1* with edges that satisfy the direction filter.

c) Crossover

After using the roulette-wheel to select one of the parents, we search for the group of potential mating partners. The local neighborhood of an individual is defined as the set of individuals that:

- Have the same main direction as the other partner.
- Have the same sign in the slope of regression line (either increasing or decreasing).
- Have at least one node shared on their sequence in the same position.

After selecting the second partner from the set of neighbor individuals (roulette-wheel), recombine the two parents to get two new offspring candidate spatial trends.

d) Insertion

A few antibody (insertion percentage) randomly generated in per generation and add to population. The use of this method was proven to be effective in the experiments.

D. Hybrid Genetic -Immune System Method

The proposed hybrid method depends on genetic algorithms

and immune system. The main forces of the evolutionary process for the genetic algorithm are crossover and the mutation operators.

For the Clonal selection algorithm the main force of the evolutionary process is the idea of clone selection in which new clones are generated. These new clones are then mutated and the best of these clones are added to the population plus adding new generated members to the population. The hybrid method takes the main force of the evolutionary process for the two systems.

The hybrid method is described as follow:

1) Generate the initial population (candidate solutions).
2) Select the (N) best items from the population.
3) For each selected item generate a number of clones (N_c) and mutate each item from (N_c).
4) Select the best mutated item from each group (N_c) and add it to the population.
5) Select from the population the items on which the crossover will be applied. We select them randomly in our system but any selection method can be used..
6) After selection make a crossover and add the new items (items after crossover) to the population by replacing the low fitness items with the new ones.
7) Add to the population a group of new generated random items.
8) Repeat step 2- 7 according to meeting the stopping criterion.

IV. Experimental Study

We studied the performance of our proposed immune algorithm on the spatial database introduced in section II and compared it with the algorithm proposed in [7], [8]. Based on our census application problem we used the neighborhood graph construction parameters in Table I.

Spatial trends of "the cost of bread or other food usage per person among the cities" starting from arbitrary points were to discovered.

TABLE I
NEIGHBORHOOD GRAPH PARAMETERS

Node	Candidate Edge	Max Distance	Hallow Distance	Hallow Angle	Obtain Edges
303	39458	550000	30000	70	1646

To discover trends with trend length equal to 10, we initially created one antibody for each possible start node. The best results were gained by setting the parameter that shown in table II.

TABLE II
BEST PARAMETERS OF ALGORITHMS

Algorithm	Insertion %	Mutation Rate	Cloning %	Crossover %	Selection power
CSA	40	70	60	-	0.4
HYBRID	20	70	60	20	0.4
ESTER	-	-	-	-	-

In figure 7 the number of discovered trends by the three algorithms when a certain number of paths have been examined is shown. The results are averaged over 3 independent runs of the algorithm. As can be seen our proposed algorithms improves its performance in subsequent generations as the search experience is exploited.

Also in figure 8 the evolution in the average confidence of the population can be observed. This confirms the effectiveness of the affinity measure applied in the selection operator, which will soon improve the average confidence of the population and maintain this property in subsequent generations. figure 9 shown that the hybrid algorithm is better than only CSA and both of them are better than Ester.

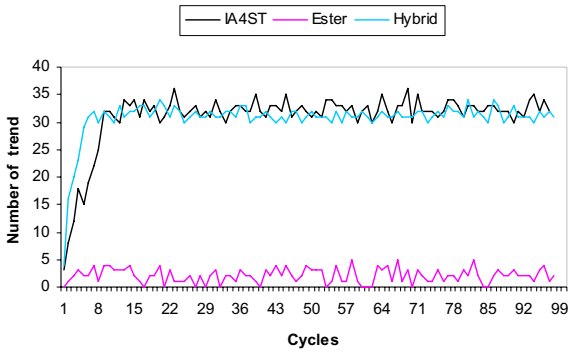

Fig. 7. number of discovered trends by the three algorithms

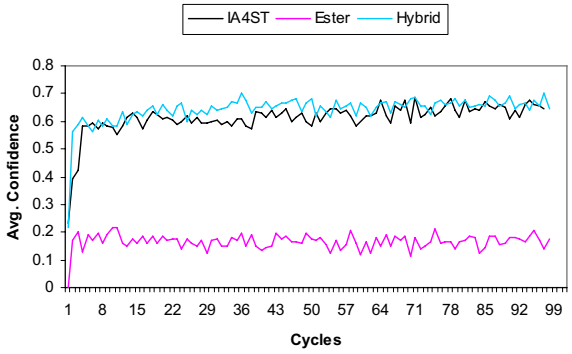

Fig. 8. Comparison in avg. confidence

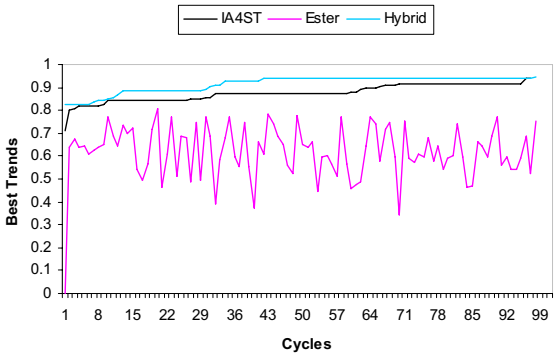

Fig. 9. Comparison in avg. best trends

V. CONCLUSION

In this paper we proposed a novel immune algorithm for efficient discovery of trend patterns in geo-spatial databases. Some customized operators were designed that match well with the nature and requirements of the problem. Also in our algorithm the user-given confidence threshold does not affect the search process and does not force us to miss the valid trends. Noticeable improvement in the performance of the discovery process was observed in the experimental study on a census spatial database. The future research directions include a detail study on the affects of different parameters in the immune algorithm and the integration of search for trends of different lengths with improved representation and operators.

REFERENCES

[1] K. Koperski and J. Han, "Discovery of spatial association rules in geographic information databases," *in Proc. 4th Int. Symp. on Large Spatial Databases* , 1995, pp.47-66.

[2] Y. Huang, S. Shekhar, H. Xiong, "Discovering Spatial Co-location Patterns from Spatial Datasets: A General Approach.", *IEEE Transactions on Knowledge and Data Eng.* vol.17, no.12 (2004) 1472-1485

[3] L. Wang, K. Xie, T. Chen, X. Ma, "Efficient Discovery of Multilevel Spatial Association Rules Using Partitions", *Information and Software Technology,* Vol. 47, no. 13 (2005) 829- 840

[4] K. Koperski, J. Han, N. Stefanovic, "An Efficient Two-step Method for Classification of Spatial Data", *Proc. International Symp. On Spatial Data Handling* (1998) 320-328

[5] S. Shekhar, P. Schrater, W. R. Vatsavai, W. Wu , S. Chawla, "Spatial Contextual Classification and Prediction Models for Mining Geospatial Data.", *IEEE Transactions on Multmedia*, vol. 2, no.4 (2002) 174-188

[6] R. Ng, J. Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining." *IEEE Transactions on Knowledge and Data Engineering*, vol. 14, no. 5 (2005) 1003–1017

[7] M. Ester, A. Frommelt, H.P. Kriegel and J. Sander, "Spatial data mining: database primitives, Algorithms and efficient DBMS support," *Data Mining and Knowledge Discovery,* vol. 4, no.2/3, pp. 193-217, 2000.

[8] M. Ester, A. Frommelt, H.P. Kriegel and J. Sander, "Algorithms for characterization and trend detection in spatial databases," *in Proc. 4th International Conf. on Knowledge Discovery and Data Mining,* 1998 pp. 44-50.

[9] M. Ester, H. P. Kriegel, J. Sander, "Spatial Data Mining: A Database Approach." *Proc. 5th Int. Symp. On Large Spatial Databases*. (1997) 320-328

[10] M. Ester, H. P. Kriegel, J. Sander, X. Xu, "Density-Connected Sets and Their Application for Trend Detection in Spatial Databases." *Proc. 3rd Int. Conf. on Knowledge Discovery and Data Mining*. (1997) 44-50

[11] A.P. Engelbrecht, "Computational Intelligence: An Introduction Second Edition" Wiley, 2007, pp. 431-435.

[12] C. Guangzhu, L. Zhishu, Y. Daohua, Nimazhaxi and Zhai yusheng , "An Immune Algorithm based on the Complement Activation Pathway", *IJCSNS International Journal of Computer Science and Network Security,* VOL.6 No.1A, January 2006

[13] L.N. De Castro, F. J. Von Zuben, "Artificial Immune Systems: Part I – Basic Theory and Applications," *Technical Report, RT–DCA 01/99,* December 1999.

[14] L.N. De Castro, F. J. Von Zuben, "the Clonal Selection Algorithm with Engineering Applications," *In Proceedings of GECCO'00 Las Vegas, Nevada, USA*, 2000.

[15] N. K. Jwenw, "Towards a Network Theory of the Immune System," *Annual Immunology,* vol.125c, 1974.

[16] J.Timmis, M.Neal, J.Hunt, "An Artificial Immune System for Data Analysis," *Biosystems,* vol.55 (1/3), 2000.

[17] L.N. De Castro, F. J. Von Zuben, "AiNet: an Artificial Immune Network for Data Analysis," *International Journal of Computation Intelligence and Application (IJCIA),* vol.1 (3). 2001.

[18] P.D'haeseleer, S.Forrest, "An Immunological Approach to Change Detection: Algorithm, Analysis and Implication," *In Proc. of IEEE Symposium on Research in Security and Privacy, Oakland, CA,* 1996.

[19] L.N. De Castro, F. J. Von Zuben, " Artificial Immune Systems: Part II – A Survey of Applications", *TechnicalReport, DCA–RT,021/00,February*, 2000.

[20] A. Zarnani, M. Rahgozar, "Efficient Discovery of knowledge from larg Geo- Spatial Databases: An Evolutionary Approach" 2006.

[21] M. Dorigo, V. Maniezzo, A. Colorni, "The Ant System: Optimization by a Colony of Cooperating Agents." *IEEE Transactions on Systems, Man, and Cybernetics Part B: Cybernetics,* vol. 26, no.1 (1996) 29–41

[22] M. Dorigo, T. Stützle , "The Ant Colony Optimization Meta-Heuristic: Algorithms, Applications and Advances. " *In: Glover F., Kochenberger G.: Handbook of Meta-heuristics. Kluwer Academic Publishers* (2002)

[23] A. Zarnani and M. Rahgozar, "Mining spatial trends by a colony of cooperative ant agents," in *Proc. SIAM Conf. on Data Mining'06 Workshop on Spatial Data Mining* [Online]. 2006, Available: http://www.siam.org/meetings/sdm06/workproceed/Spatial%20Data%20Mining/index.html