Mining Spatial Trends by a Colony of Cooperative Ant Agents

Ashkan Zarnani† Masoud Rahgozar†

Abstract

Large amounts of spatially referenced data has been aggregated in various application domains such as geographic information systems (GIS), environmental studies, banking and retailing, which motivates the highly demanding field of spatial data mining. So far many optimization problems have been better solved inspired by the foraging behavior of ant colonies. In this paper we propose a novel algorithm for the discovery of spatial trends as one of the most valuable and comprehensive patterns potentially found in a spatial database. Our algorithm applies the emergent intelligent behavior of ant colonies to handle the huge search space encountered in the discovery of this knowledge. We apply an effective greedy heuristic combined with the trail intensity being laid by ants using a spatial path. The experimental results on a real banking spatial database show that our method has higher efficiency in performance of the discovery process and in the quality of trend patterns discovered compared to other existing approaches using non-intelligent heuristics.

1 Introduction

Many organizations have collected large amounts of spatially referenced data in various application areas such as geographic information systems (GIS), banking and retailing. 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 amounts of spatial data [11].

So far many data mining tasks have been investigated to be applied on spatial databases. In [11] spatial association rules are defined and an algorithm is proposed to efficiently exploit the concept hierarchy of spatial predicates for better performance. In [8] and [10] algorithms are designed for the classification of spatial data. Shekhar *et al.* further improved spatial classification in [12] and also introduced algorithms to mine co-location patterns [9]. Spatial trends are one of the most valuable and comprehensive patterns potentially found in a spatial database. In spatial trend analysis, patterns of change of some non-spatial attributes in the neighborhood of an object are explored [6] e.g. moving towards north-east from the city center the average income of the population increases (confidence 82%).

Ester *et al.* studied this task proposing a general clustering algorithm and its application in trend detection [7] and further improved it in [6] exploiting the database primitives for spatial data mining introduced in [8]. Having constructed the neighborhood graph the algorithm proposed gets a specified start object *o* from the user. Then it has to examine every possible path in the graph beginning from *o*. For each path it performs a regression analysis on non-spatial values of the path vertices and their distance from *o*. But the search space soon becomes tremendously huge by increasing the size of neighborhood graph and makes it impossible to do a full search. In order to prune the search space it assumes that a desired trend will never have its regression confidence below a user given threshold. As we incrementally construct a possible path, we would have to resign from further extending it when the regression confidence of the current path becomes bellow the threshold. But this assumption is a restricting one, as it may mislead us by forcing a trend to stop from growth that would get much higher confidence if not blocked.

Many solutions for NP-Complete search and optimization problems have been developed based on the cooperative foraging behavior of ant colonies [2]. However less attention has been given to apply this powerful inspiration from nature in the tasks of spatial data mining. In this research we introduce a new spatial trend detection algorithm that uses the phenomenon of *stigmergy* i.e. indirect communication of simple agents by means of their surrounding environment, observed in real ant colonies [1]. It also combines this behavior with a new guiding heuristic that is shown to be effective. We succeeded to handle the non-polynomial growth of the search space, and at the same time retain the discovery power of the algorithm, by letting each ant agent to cooperatively exploit the colonies valuable experience. Also in contrast with the algorithm proposed in [6] our algorithm is not dependent on the user. It doesn't get a specified start object from the user nor needs it to input a pruning threshold. This brings ease of use and wider applicability to our method as its efficiency and performance is independent from the user. We have conducted some experiments on a real banking spatial database to compare the proposed method with the algorithm proposed in [6] which is being widely accepted and used. The results show that the proposed algorithm has higher efficiency in performance of the discovery process and in the quality of trend patterns discovered.

2 Spatial Trend Detection

Some spatial relations (called neighborhood relations) like direction, metric and topological relations between the objects are formally defined to be used in spatial data mining [8]. Based on these relations the notions of neighborhood graph and neighborhood path are defined as follows:

†Database Research Group, Electrical and Computer Engineering Department, University of Tehran

DEFINITION 1**:** let *neighbor* be a neighborhood relation and *DB* be a database of spatial objects. *Neighborhood graph G* = (N, E) is a graph with nodes $N = DB$ and edges $E \subseteq N \times N$ where an edge e = $(n1, n2)$ exists iff *neighbor*$(n1, n2)$ holds.

DEFINITION 2**:** A *neighborhood path* of length *k* is defined as a sequence of nodes [*n1, n2,…, nk*], where *neighbor*$(ni, ni+1)$ holds for all $ni \in N$, $1 \leq i < k$.

As we have the location dimension in a spatial database one useful potential pattern could be the change of a non-spatial attribute on a neighborhood path with respect to its distance form a reference object. E.g. beginning form a trading center in the city and moving on a specific highway towards the west, the unemployment rate grows. Having available the desired neighborhood graph, the notion of spatial trends can be defined as follows [6]:

DEFINITION 3: A spatial trend is a path on the neighborhood graph with a length k of nodes that the confidence of regression on its nodes data values and their distance from start node (see figure 1) is above a minimum fraction.

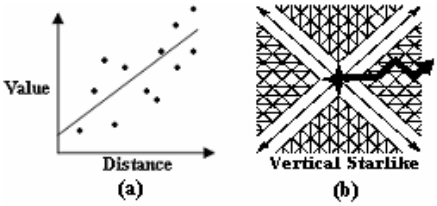

Figure 1: (a) Regression line for a trend, (b) A direction filter

Our example spatial database contains the agency locations of a national bank and their various financial data like the count and remainder of different kinds of accounts. A map of these point and city regions is shown in figure 2. A possible trend of 15 nodes is also depicted.

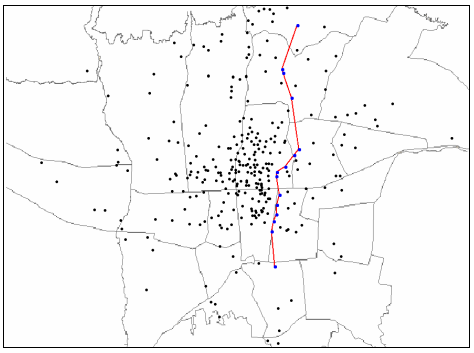

Figure 2: The city regions with bank points

As an example we may want to find trends of the number of long term accounts in bank agencies starting from arbitrary agency points. Having discovered such trends, we can try to explain them by some spatial attributes. For example a trend could be matching a road or a highway. We can also check if there are any matching trends on the same path but in other thematic layers like demographic or land use layers. A trend can also predict the data value of a new point on its path with a reliability fraction equal to the regression confidence. A desired informative spatial trend pattern would not be crossing the space in an arbitrary manner. So a direction filter (see figure 1.a) is applied when forming the path of the trend being examined [6, 7].

To discover the trends in a neighborhood graph by the algorithm proposed in [6] having *a* feasible nodes on average to extend a path, we would have to meet a^n paths to examine their regression confidence, where n is the maximum trend length. It's impossible to examine this amount of paths even for not much large values of n (e.g. n=20). This condition gets worse when the user has not any specific start object in mind, and wants the algorithm itself to check different start objects.

For efficiency, the algorithm allows a path to be extended further by the next set of feasible nodes if its current confidence is not bellow the threshold given by the user. This heuristic force the search space to become smaller but can easily miss a high confidence trend if its confidence is bellow the threshold somewhere in the middle of its path. In figure 3 a sample of this situation is shown in which the algorithm will stop path extension when it is in node *i* as the regression confidence of the path from the first node to *i* is bellow the threshold. However this path would have a confidence much higher than the threshold if not blocked and continued.

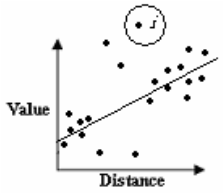

Figure 3: Missing a trend in node i

We have managed to remove this restricting assumption by using distributed cooperative ant agents to improve the performance of the search process.

3 Ant Colonies for Search and Optimization

Ant Colony Optimization (ACO) is a new meta-heuristic inspired by real ant colonies in nature. Ant colonies intelligently solve complex discrete problems like finding shortest path although its individuals are very simple and not intelligent enough to solve such problems on their own. The main underlying idea is to use a multi-agent parallel search on the different possible combinations of solution components. The decision to choose a component is based on a local problem data and a dynamic shared global memory of the colony that contains a history on the quality of previously obtained results [2, 5].

3.1 ACO Meta-Heuristic. To solve a combinatorial optimization problem, the ant agents concurrently move to

the next state selecting a component and forming a partial solution of the problem. This move is done by a stochastic decision policy directed (i) by the ant's private information (internal state, or memory) and (ii) by publicly available pheromone trail and a priori problem-specific local information, i.e. the two parameters of *trail intensity* and *attractiveness* [2]. For an ant k in state *i*, the probability to choose state *j* depends on two values:

- η_{ij}: The attractiveness of the next feasible component, which is computed by a problem dependent heuristic providing *a priori* desirability of the move.
- τ_{ij}: The trail intensity of the move that represents the quality of the previously evaluated solutions containing the component,, thus providing a posteriori desirability of the move.

The trail update is usually done when all of the ants have finished their incremental solution construction. The amount of pheromone to be laid over the components used in a solution depends on its quality, defined differently in various problems. The whole procedure is iteratively repeated in a loop until a stopping criteria is satisfied. Also a mechanism of *trail evaporation* is applied which lets the colony to avoid unlimited accumulation of trials over some component [4].

3.2 ACO for Spatial Trend Discovery. ACO has been recently used in some data mining tasks, e.g. classification rule discovery [13]. However considering the challenges faced in the problem of spatial trend detection (see section 2) we can see that ACO can suggest efficient properties in these aspects. Firstly as the definition of the problem suggests, the ant agents can search for the trend starting from their own start point in a completely distributed manner. This omits the need to get the start point node from the user. Secondly to guide the stochastic search of the ants, the pheromone trails can help the ants to exploit the trend detection experience of the colony. This guides the search process to converge to a better subspace potentially containing more and better trend patterns. Finally some measures of attractiveness can be defined for selecting a feasible spatial object from the neighborhood graph which can effectively guide the trend detection process of an ant.

4 ACO Algorithm for Trend Detection

In our approach for trend discovery different ant agents will start form different points of the graph searching for increasing spatial trends. Note that each decreasing trend is also an increasing one.

4.1 Pheromone Trails. The quality of an ACO application depends very much on the definition assigned to the pheromone trail. As previously mentioned the ants will search for increasing trends only, adding the direction of an edge to the properties of the pheromone laid. A pheromone trail value is considered for each directed edge of the neighborhood graph, making the pheromone matrix asymmetric. When an ant is in node P*i* and selects P*j* as the next node, its pheromone is laid on the edge E*ij* (and not E*ji*). Thus $\tau_{i,j}$, the amount of pheromone of E*ij* encodes the favorability of selecting node P*j* when in node P*i*, to form a high confidence increasing spatial trend.

4.2 Heuristics for Spatial Trend Detection. Another important feature of an ACO application is the choice of a good heuristic to incrementally build possible solutions, which will be used in combination with the pheromone information. We applied two heuristics for guiding the discovery of spatial trends. First closer nodes are preferable to the ones far from the current node, as they are more likely to be correlated. Second we would like the value of the next node to match better with an increasing linear regression model. This means that we would prefer the nodes with a value higher than the last node of the current increasing path. To apply this second heuristic we used equation 3.2. The heuristic value for selecting node P*j* from node P*i* will be calculated by the following formulas:

$$D_{i,j} = 1/Distance(i,j) \tag{4.1}$$

$$S_{i,j} = 1 - \left| \frac{Slope(i,j) - 45}{135} \right| \tag{4.2}$$

$$\eta_{i,j} = {D_{i,j}}^{\beta} \cdot {S_{i,j}}^{\gamma} \tag{4.3}$$

Where β and γ are the relative importance of distance heuristic and slope heuristic, respectively and $\eta_{i,j}$ is the attractiveness of the node P*j* when in node P*i*. *Slope(i,j)* is the slope(in degrees) of the line from P*i* to P*j* (x coordinate is the distance form P*i* and y coordinate is non-spatial values).

4.3 Building Candidate Trends. The pheromone trail and the heuristic information defined above will now be used by the ants to build feasible solutions. Ant agents start from different nodes and form a candidate trend by adding nodes iteratively. Each ant k has a memory M^k to store the information of the current path including the nodes, their values, distances and also the main direction of the candidate trend. The selection of the next node is stochastically done by assigning a probability $P_{i,j}^k$ for ant k to select node P*j* when the last node of the trend is P*i* by the following formula:

$$P_{i,j}^k = \begin{cases} \dfrac{\left[\tau_{i,j}\right]^{\alpha} \cdot \left[\eta_{i,j}\right]}{\sum_{l \in allowed_k} \left[\tau_{i,l}\right]^{\alpha} \cdot \left[\eta_{i,l}\right]} & \text{if j} \in allowed_k \\ 0 & \text{otherwise} \end{cases} \tag{4.4}$$

Where α is the relative importance of pheromone trail and $allowed_k$ is the set of nodes that are connected with an

edge to node P*i* and have also passed the direction filter. The direction of the trend of ant k is the direction of its second node with respect to its first node. The filter we have used here accepts new directions to be the same as the trend direction or rotated one step either clockwise or counterclockwise. As an example if the direction of the trend is South-East then the nodes with direction South-East, South or East will be in $allowed_k$.

The addition of a node is done by all of the colony's ants independently and in parallel. This process repeats until there is no node in the set of allowed nodes for ant k or the number of nodes in the trend reaches to *TrendLength*, which is an input integer parameter of the algorithm.

4.4 Updating the Pheromone Trail. For updating the pheromone trail, we chose the quality of the trend. This quality is evaluated by the r^2 value of the linear regression. This value is a fraction between 0.0 and 1.0, and has no units. The better you can predict Y from X, the nearer is this value to 1.0. If we call an iteration of the algorithm: the possible addition of nodes by m ants, then in every *TrendLength* iterations of the algorithm (called a cycle) each ant completes its candidate trend. At this time the trail intensity is updated by the following formulas:

$$\tau_{i,j}(t+1) = \rho.\tau_{i,j}(t) + \Delta\tau_{i,j} \quad (4.5)$$

$$\Delta\tau_{i,j} = \sum_{k=1}^{m} \Delta\tau_{i,j}^{k} \quad (4.6)$$

$$\Delta\tau_{i,j}^{k} = \begin{cases} Q \times confidence(k) & \text{if ant k uses edge(i,j) in its trend} \\ 0 & \text{otherwise} \end{cases} \quad (4.7)$$

Where ρ is a coefficient such that (1-ρ) is the *evaporation* of trail between cycle t and t+1. The *confidence(k)* is the value of r^2 for the regression line of the nodes present in the trend detected by ant k and Q is a constant.

5. Experimental Results

In this section we provide the experimental results to study the properties of our algorithm and to compare it with the algorithm proposed in[6].

In Table 1 (a) the properties of the neighborhood graph are given and in (b) some statistics for the values we were searching their spatial trends (i.e. the number of long term accounts in each bank agency point) are provided.

Table 1 (a): Node values

Minimum	Maximum	Average	Std. Deviation
1	39025	4623	15631

(b) Neighborhood graph properties

Edges	Min. Degree	Max. Degree	Avg. Degree	Min Distance	Max Distance	Avg. Distance
1535	2	30	10.23	48	5988	2582

To find the trends of length 10 with our method, we gained the best results by putting two ants in each node and the values of α, β, γ and ρ being respectively equal to 1, 4, 0.2 and 0.5. These values were used for the all of the experiments. We considered a path with its confidence (r^2) over 75% as a valid trend. Figure 4 shows the number of trends that the two algorithms found when a certain number of paths in the neighborhood graph have been examined by the algorithms.

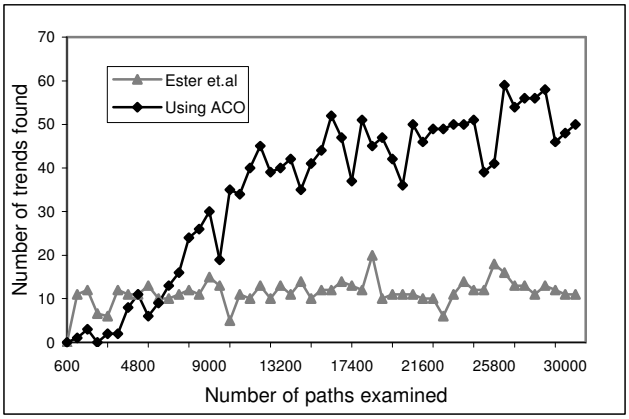

Figure 4: Comparison between the two algorithms

As can be seen the algorithm proposed by ester does not use any dynamic guiding heuristic and its performance will not improve as examined paths increase. However, our proposed method will improve its trend discovery power as the colony aggregates its population's experience gained form the previous cycles and soon outperforms the other algorithm drastically. Figure 5 shows how the algorithm improves its discovery power in search for trends of length 15. The regression line of the graph confirms a smooth increase of the number of trends discovered in each cycle.

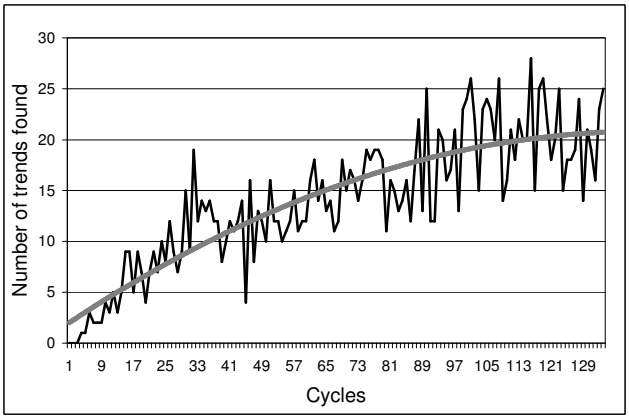

Figure 5: Improvement in Trend Discovery of Trend-Length=15

There is also an evolution observed in the average confidence of paths examined in every cycle shown in figure 6. This also confirms the improvement of the

discovery process in such a way that the confidence values of candidate paths increase, although they are not considered as valid trends.

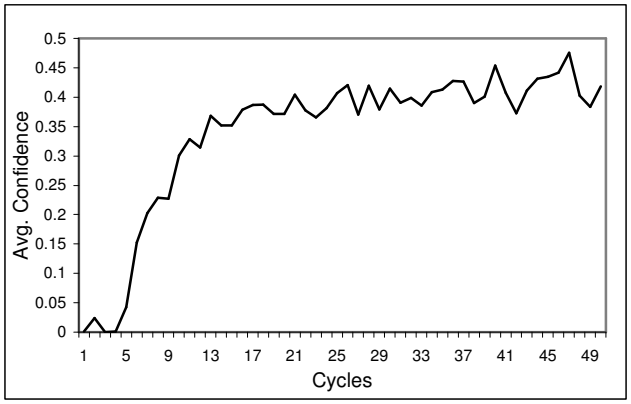

Figure 6: Evolution of average trend confidence

6. Conclusion

In this paper we proposed a new application of ant colony optimization for the task of trend detection in spatial data mining. This algorithm applies the emergent intelligent behavior of ant colonies to handle the huge search space encountered in the discovery of this invaluable knowledge. We proposed two heuristics for edge selection by ants in spatial trend discovery shown to be much more effective than the previous ones proposed in the literature [6]. Our algorithm is also independent from the user in a way that it does not need a start node and applies no threshold in the discovery process. The experiments run on a real banking spatial database show that the proposed method outperforms the current approaches in performance and discovery power. In our future research we will further improve the heuristics used and investigate the use of some modified versions of ACO like MMACO in spatial trend detection. Currently a limitation of our algorithm is that it searches for trends of a certain length. We plan to tackle this problem by integrating the search process for trends of different length.

Acknowledgments

We would like to thank the Mellat Bank Research Center for their financial support and also for providing us the banking spatial database of Tehran. We also thank Dr C. Lucas and Dr. N. Memariani for their useful comments on this work.

References

[1] M. Dorigo, E. Bonabeau & G. Theraulaz, *Ant algorithms and stigmergy,* Journal of Future Generation Computer Systems, 16(8), 2000, pp.851–871,

[2] M. Dorigo & G. Di Caro, *Ant algorithms for discrete optimization,* International Journal Artificial life, 5(2), 1999, pp.137-172.

[3] M. Dorigo, G. Di Caro, *Ant colony optimization: A new meta-heuristic,* Proc. 1999 Congress on Evolutionary Computation, 1999, pp.1470-1477

[4] 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, 26(1), 1996, pp.29–41.

[5] M. Dorigo, T. Stützle. *The ant colony optimization meta-heuristic: Algorithms, applications and advances,* In F. Glover and G. Kochenberger, *Handbook of Metaheuristics* (Kluwer Academic Publishers, 2002).

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

[7] M. Ester, H.P. Kriegel, J. Sander, X. Xu, *Density-connected sets and their application for trend detection in spatial databases,* Proc. 3rd International Conf. on Knowledge Discovery and Data Mining, New York City, NY ,1997, pp.44-50.

[8] M. Ester, H.P. Kriegel, J. Sander, *Spatial data mining: A database approach,* Proc. 5th International Symp. On Large Spatial Databases, Berlin, Germany, 1997, pp.320-328.

[9] Y. Huang, S. Shekhar & H. Xiong, *Discovering Spatial Co-location Patterns from Spatial Datasets: A General Approach,* IEEE Transactions on Knowledge and Data Eng., 16(12), 2004, pp.1472-1485.

[10] K. Koperski, J. Han, N. Stefanovic, *An efficient two-step method for classification of spatial data,* Proc. International Symp. On Spatial Data Handling, Vancouver, Canada, 1998, pp.320-328.

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

[12] S. Shekhar, P. Schrater, W. R. Vatsavai, W. Wu & S. Chawla, *Spatial Contextual Classification and Prediction Models for Mining Geospatial Data,* IEEE Transactions on Multimedia, 2(4), 2002, pp.174-188.

[13] R.S. Parpinelli, H.S. Lopes & A.A. Freitas, *Data mining with an ant colony optimization algorithm,* IEEE Transactions on Evolutionary Computation, 6(4), 2002, pp.321-332.