ANALYSIS OF ANT COLONY CLUSTERING (ACC)

J.M. Lakshmi*

G. Raju**

ABSTRACT

Clustering, one of the fundamental tasks in Datamining, is also challenging field of research. Clustering can be considered as an optimization problem as it minimizes inter cluster similarity and maximize intra cluster similarity. Clustering problem has been solved by different algorithms ranging from simple K-means to Bio-inspired Evolutionary algorithms. Swarm intelligence, an emerging Evolutionary computational intelligent technique that has been derived from Bio- inspired behavior, has been applied to solve various real time implication problems of combinatorial optimization problem & computational problems. Swarm Intelligent encompasses the implementation of "Collective Intelligence" by groups of simple agents. It is based on the behavior of real world insect Swarms, as a problem solving tool. One such Swarm Behavioral model, Ant colony optimization(ACO) witnesses its application in the field of Datamining & knowledge discovery schemes in recent years. Ant colony optimization method has its ideology taken from the behavior of natural Ants. Ant colony based heuristics are widely studied in the field of computer science and its applications. This paper analyzes the application of Ant colony approach on clustering, by implementing it on some of the predominant data sets and also to compare the method with the existing method. The paper discusses the efficiency of the method based on the results.

Keywords: Ant colony Optimization, Clustering, Data Mining, Evolutionary computation, Swarm Intelligence.

*Research scholar , Manonmaniam Sundaranar University, Tirunelveli, TamilNadu .

**Head Of the Department, Department of Computer Applications, Kannur,Kerala

1. INTRODUCTION

Data Mining has attracted a great deal of attention in the information industry, Business organizations and other research area as the need for converting data into useful information and knowledge has emerged [9]. It is an Inter disciplinary field which has taken its core features from disciplines like Machine learning, Database systems, Neural networks and Statistics. Data mining field is in its infant stage which provides more scope for research. Datamining technology can generate new business opportunities like Market segmentation, Customer churn ,Fraud detection, Direct marketing, Interactive marketing, Market basket analysis, Trend analysis, etc. This can be achieved through automating prediction of trends and behavior and automating the discovery of previously unknown patterns. Datamining discover the patterns and relationships by which it can provide predictive information, through which knowledge can be gained and applied for decision making process. The Datamining methods fall into two groups called **discovery** and **predictive** techniques. Predictive mining methods include supervised learning technique. Discovery mining methods include unsupervised learning techniques [4].

The techniques of Datamining include Associations, Sequential patterns, Analysis of time series, Classification, and Cluster analysis. Clustering, one of the important Datamining techniques falls into the category of an unsupervised learning method. The next session follows with the discussion of clustering technique. Session three explains about the ideology and importance of Swarm intelligence (SI) and mention about various methods of it. Session four explains (ACO), one such method of (SI). Session five explains the application of ACO in the field of clustering, including the pseudocode algorithm of Ant Colony Clustering (ACC). Session six consists of Experimental results of ACC on selected sample database. Session seven analyze the Algorithm based on the findings of the experimental results. Session eight compare ACC with the traditional K –means algorithm followed by the conclusion in the last session.

2. CLUSTERING

Clustering of objects is an ancient concept as the human needs for describing the salient characteristics of men & objects based on similarity started. Clustering is the process of organizing data objects into groups (cluster) whose members are similar based on one or more features of the data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [11]. It can be defined mathematically as from the data object set X , X = {x_1, x_2,..., x_n} its partition into m parts (clusters) C_1,..., C_m, such that

None of the clusters is empty; $C_i \neq Ø$

Every data object belongs to a cluster

Every sample belongs to a single cluster (crisp clustering); $C_i \cap C_j = Ø$, $i \neq j$

Clustering is equivalent to breaking the graph into connected components, one for each cluster [15].The goal of clustering is descriptive. It‘s a challenging field of research in which its potential applications pose their own special requirements. The idea of clustering usage is evident in many brain functions including Pattern Matching. Clustering can be performed on different representations of data objects [9][14]

i) **Numerical**: Consists of numerical integer and float data types

ii) **Ratio scaled**: Positive measurement of non-linear scale

iii) **Interval scaled data type**: Continuous measurement of a roughly linear scale

iv) **Boolean**: also called as binary data type and has only two states zero or one.

v) Categorical: Non-numeric data attributes.

vi) Nominal: Is a generalization of binary data type variable, in the sense it can take more than two states defined in advance.

vii) **Ordinal:** A discrete ordinal data resembles a nominal data type, except that the ordinal values are ordered in a meaningful sequence.

There exist different categories of clustering methods for cluster formation such as

i) **Partitioning algorithms**: Construct various partitions and then evaluate them by some criterion. (K- means, K-medoids)

ii) **Hierarchy algorithms**: Create a hierarchical decomposition of the set of data (or objects) using some criterion, has the classification of Agglomerative, Divisive.

iii) **Density-based**: Based on connectivity and density functions.

iv) **Grid-based**: Based on a multiple-level granularity structure.

v**) Model-based:** A model is hypothesized for each of the clusters and the idea is to find the best fit of that model to each other.

Similarity computation for clustering can apply any of the distance measures like Euclidean distance formulae, Manhattan distance formulae, or Mahalanobis distance methods.

The choice of Distance or Similarity methods and Clustering methods depends on the problem domain, nature of the data object set, predefined number of clusters, level of Accuracy and users convenience. This paper discuss and analyze one such emerging method for clustering, Ant Colony Clustering (ACO) [20][1] based on Swarm Intelligence.

2. SWARM INTELLIGENCE

Swarm Intelligent has been technically defined as, a decentralized group of elements having similar characteristics with a similar aim. In a colony of social insects, such as ants, bees, wasps and termites, each insect usually performs its own tasks independently from other members of the colony. Swarm Intelligence derives it base from biological natures of bird flocking, ant foraging and fish schooling, animal herding and bacterial growth mechanism. The reason why researchers were so fascinated by these biological computations was because it promises self-organization, cooperative work, division of work load and collective sorting and clustering. However, the tasks performed by different insects are related to each other in such a way that the colony, as a whole, is capable of solving complex problems through cooperation.

Important, survival-related problems such as selecting and picking up materials, finding and storing food, which require sophisticated planning, are solved by insect colonies without any kind of supervisor or centralization. Research in using the social insect metaphor for solving problems is still in its infancy. The systems developed using swarm intelligence principles emphasize distributiveness, direct or indirect interactions among relatively simple agents, flexibility and robustness [17].

There are some problems that look rigid enough for an individual to solve. Swarm Intelligence gives flexibility to these complications. Some of the well-known strategies in this area are Particle swarm optimization (PSO), Ant Colony Optimization (ACO), Artificial Bee Colony (ABC), Bacterial Foraging Algorithm (BFA), FireFly Algorithm (FFA), Artificial Immune Systems (AIS), Fish Swarm Algorithm (FSA), Intelligent Water Drops (IWD), Shuffled Frog Leaping Algorithm (SFLA), Glowworm swarm optimization (GSO) [2].

Based on the advantages of Swarm technology, many optimization algorithms have been designed to simulate such swarm intelligence and these algorithms have been successfully applied to the areas of functional optimization & combinatorial optimization [8]. Some work have highlighted clustering approaches based on Ant Colony Optimization (ACO) which is a branch developed from Swarm Intelligence [3]. Among the Nature's social insect's behaviors, the most widely recognized is the Ant's ability to work as a group in order to finish a task that cannot be finished by single Ant. It has the effect of synergy and seems to indicate unity is strength. In early nineties a algorithm called Ant system was proposed as a novel heuristic approach for the solution of combinatorial optimization problems (dorigo et .al.,) [13]. Ant Meta heuristics has been applied to well -known classic problems like travelling salesman, knapsack.

4. ANT COLONY OPTIMIZATION (ACO)

Ant colony optimization method has its ideology taken from the behavior of natural ants. An ant colony has many characteristics that are considered useful as it is composed of many agents which, although simple, individually, can perform complex tasks as a group, without central coordination. Existing research areas analyzed ant colony behavior under 3 models

Brood sorting behavior

Foraging model

Piling model

Ant based clustering algorithms are based upon the brood sorting behavior of ants-Deneubourg.[12]

Ant based clustering algorithms can be considered non-hierarchical hard, agglomerative clustering methods [5]. Non hierarchical means that there is no parent –child relationship between the objects or the clusters formed. Hard means that each object is assigned to only 1 cluster (hard clustering). Agglomerative means that the clusters are formed bottom-up in other words, isolated objects are progressively put together to form bigger cluster. Ant- based Clustering algorithms are an appropriate alternative to traditional clustering algorithms [2].

5. ANT COLONY CLUSTERING

ACO algorithm for Clustering is a promising research area The Ant Colony Clustering approach has a number of features that make it an interesting field of study for clustering. They have the benefit of discovering clusters without any initial partitioning, and without knowing ahead of time how many clusters will be necessary. One of the first studies of ant based clustering is found in [6]. The Ant Colony clustering was the data clustering by simulating the ant's natural behavior to cluster the data in the 2-D grid board. The general idea for Ant Colony clustering is that, every moment that the ant moved to the surrounding cells, it would either grab or drop the data based on the possibility and the similarity of the data required to be clustered. This process resulted as an effective data clustering [16].

The first proposal of a clustering algorithm inspired by ant colonies was made by Deneubourg[12], the canonical ant clustering algorithm follows the formula described by Lumer Fayette[7]. In its basic idea, the ant cluster algorithm is described to operate on cellular automata- like field, where artificial ants can, with a given probability, pick & drop objects which represent the data to be clustered. First, the ants pick isolated objects, and drop them near similar objects. As similar objects are dropped together, these "lumps" exert a stigmergic influence over ants, which in turn tend to drop more and more objects near them -thus performing clustering. Literature survey shows that Several ant clustering methods results in good cluster formation by enhancing itself with several up- dating like Eager Ants[12], Ant sleeping model[10] etc .

The algorithm of ant colony cluster [12]

```
Algorithm Ant Colony Clustering ()
{// main algorithm starts
Step1: place the object and ants randomly on the grid
Step 2: for iteration=0 to max_ iter// max_iter refers to maximum iteration for the algorithm
{
Step 3: for ant= 0 to ant_num //ant_num refers the number of ant involved in clustering process
{//Ant Picking and Droping starts
Step 4:if ant does not hold any object and the grid cell of ant hold  data object
    { //picking process starts
    if the neighbouhood data object of R*R cells of the grid satisfy the pick up threshold then Pick the object and move to the grid cel without data object.else move ant randomly to another cell with data object.
    }
  Else if ant does not hold any object and the grid cell of ant does not hold any data object
  {
  Move the ant randomly to another grid cell with data object
  }
Else if    ant  hold data object and grid cell of ant does not hold of data object

    { if the neighbouhood data object of r*r cells of the grid satisfy the drop down
  threshold then drop the object and move to the grid cell with data object. else move ant
  randomly to another cell without data object.
          }
  Else if ant hold data object and grid cell of ant hold of data object
    {
     Move ant randomly to another grid cell that does not hold any object
     }
  }//end for ant
  }//end for iteration
  Step 5: Identify the clusters based on the similarity of position of the objects on the
  grid.
  }//Main
```

6. EXPERIMENTAL IMPLEMENTATION

The sample data base of three domains that has been taken for implementation is as follows: (courtesy:UCI data repository)[19]

i)The Insurance Company Benchmark, a multivariate data base contains information on customers of an insurance company with 86 attributes.

ii) Census data set, a Multivariate data set consisting of Categorical Attribute with 68 attributes.

iii) Iris data set, consisting of 4 real types of attributes

The following assumptions were made for the threshold values of the variables, during experimental implementation.

Ideal work space for the grid	$\surd(N * 10 * 2 \ldots \ldots \surd(N * 10 * 3)$ Where N=number of data objects.

Max ants	$\frac{work\ space}{2}$
Max Iterations	(*N*/10)*ants.
Threshold value for pick and drop	[80% …. 60%]. decrement the values when iterations increase

Table-1 assumed values in the experiment

i) Iris data set, consisting of 4 real types of attributes.

ii)Census data set, a Multivariate data set consisting of Categorical Attribute with 68 attributes.

iii)The Insurance Company Benchmark, a multivariate data base contains information on customers of an insurance company with 86 attributes

The resultant clusters formed under ACO has been represented graphically as follows:

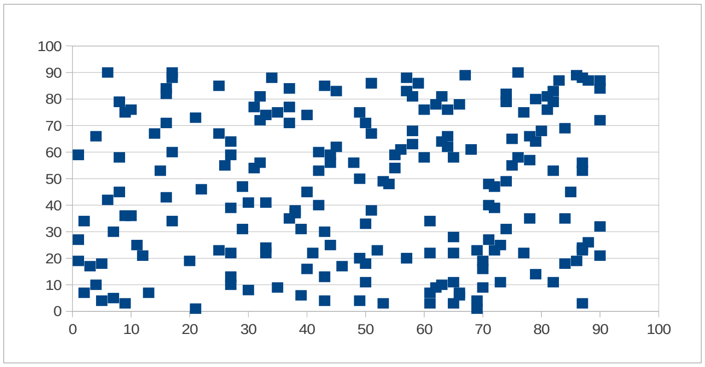

Figure-1 Initial random allocation of objects on the grid

Clusters formed under ACC for sample database is as follows:

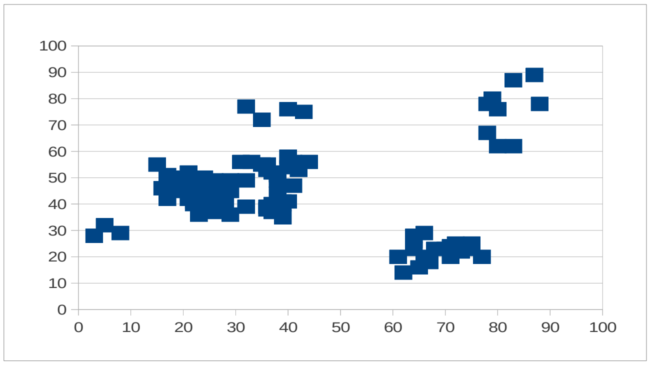

Figure-2 Ant colony clusters for insurance database

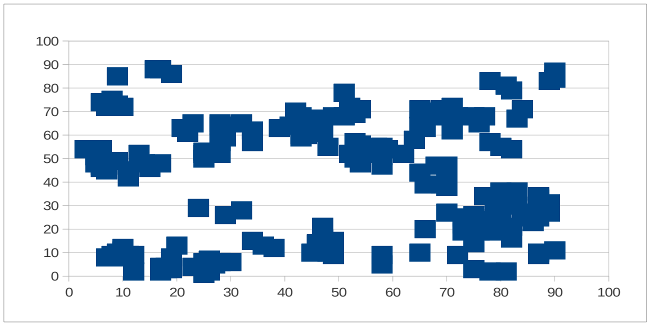

Figure-3 Ant colony clusters for census database

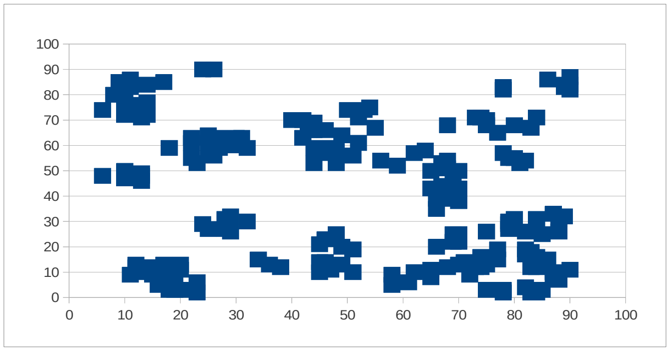

Figure-4 Ant colony clusters for insurance database

The implementation has been carried out in C language.

7. ANALYSIS OF ANT COLONY CLUSTERING(ACC):

Based on the Experimental results obtained the algorithm has been analyzed, and the following findings were made:

7.1. It has the ability to discover automatically the number of clusters. The number of clusters that can be formed under ACC need not be predetermined in advance. Based on the database and the required threshold for pick and drop given by the users, clusters are generated. Without any restriction on cluster number the clusters are generated purely on the prescribed level of similarity.

7.2. Ant Colony Clustering can be considered well as a Meta heuristic algorithm, as it provides a general algorithmic framework which is used for different optimization problem with relatively few modifications to make them adapted to a specific problem. It has been realized that the application of ACC for same data domain and for varying size of inputs needs adaptation in terms of size of the grid and the number of ants needed. The grid size and the number of ants involved in the process of pick and drop increase with the size of data base.

7.3. The distance methods applied for ACC algorithm in the literature so far, makes it necessary to convert or normalize the database into numerical data type. This may make the algorithm less suitable for real time applications with stream data type as it needs preprocessing of data items.

7.4. The method creates cluster only after allocating a data object to it and thus, none of the clusters are empty.

$$C_i \neq Ø$$

7.5 The clusters formed under the ACC are hard. That is every data object belongs to a single cluster (crisp clustering); Ci ∩ Cj= Ø, i≠j. The experimental results revealed the fact that, the inter cluster dissimilarity among the clusters may not be achieved always, as it adopts local searching method. Local searching methods take a potential solution to a problem and check its immediate neighbors (neighbourhood cells in the gridof r*r alone, during pick and drop process) in the hope of finding an improved solution. Local search methods have a tendency to become struck in suboptimal regions where many solutions are equally fit. In ACC the ants pick and drop based on the local information available in the neighborhood cells and it is not considering the whole grid. Thus the feature of object belonging to C_i may be similar to fature of objects belonging to $C_{j,}$ i≠j.

7.5. The pick and drop threshold values varies based on the number of data attributes and the distribution of data values among the attributes.

7.6. The quality of Clusters formed depends on the size of the cluster. Experimental results shows that the quality of small clusters are comparatively high than the large clusters formed. It may be due to the adaptation of local search principle.

7.7. The number of clusters formed for the same data set under different initial allocation of ants, different initial allocation data objects may vary. The data object and the ants are allocated randomly on the grid cell. This initialization is sure to have its effect on cluster formation.

7.8. The pick and drop threshold values has to be changed as iteration proceeds. Initial threshold for pick and drop is kept as high and when the iteration proceeds it can be reduced, to the accepted level of similarity. So that similar(not same) data objects will be grouped as clusters.

7.9. The pick threshold for similarity kept lower than the drop threshold similarity provides better and faster results.

8. COMPARITIVE ANALYSIS BETWEEN K-MEANS CLUSTERING AND ANT COLONY CLUSTERING METHOD:

The ACC has been compared with the traditional clustering K-Means [18] under some selected criteria's and tabulated as follows:

S.no	Criteria	K-means	Ant colony clustering
1	Initial Seed.	Specified.	Not specified.
2	Number of clusters formed.	Predetermined.	Based on pick and drop threshold value.
3	Quality of cluster	Impact of Extreme values in the cluster will be there.	Impact of Extreme values in the neighborhood cell(r*r) will be there.
4	Number of iterations for Best case.	Average.	More.
5	Number of iterations in worst case.	More.	More.
6	Space complexity.	Average, as object alone has to be represented in a grid.	Comparatively more than K-means as object and ant has to be distributed over the grid.
7	Time Complexity.	Average.	High. . The time complexity of the algorithm is O(i*a) where i represents the total number of iterations and a represents the total number of ants
8	Identification of inter cluster	May be possible based on mean average of K	Neighboring clusters formed in the grid may differ to a large

	relationships.	clusters.	extent or may be too similar, which cannot be predicted.
9	Size of data base.	Works well for small database than the larger one. (as it has to be accommodated in the grid(n*n).	Works well for small database than the larger one. (as the object and ants has to be accommodated in the grid(n*n).
11	Number of attributes.	Works well for few or average number of attributes. For more attributes, the distance computation is complex .It is more complex if the data value is of long integer type.	Works well for few or average number of attributes. For more number of attributes, the distance computation is complex, further complicated by long integer data values.
12	Mixed Types of attributes.	Has to be normalized as the distance measures suffer from handling qualitative data types.	Has to be normalized as the distance measures suffer from handling qualitative data types.

Table-2 Comparison of K-means and ACC

8. CONCLUSION

The Paper analyses the existing Bio-inspired Ant Colony Clustering algorithm through experimental results obtained by applying the model on some selected data bases. The findings have been discussed like the Ant colony clustering method can be enhanced to be applied for stream data types, can be modified to provide high quality clusters by overcoming the adverse effect of local optimization etc. It has also been compared with traditional K-means method. The Comparison results shows that in spite of more space requirement for Ant colony clustering, it works well for clustering data object for which the knowledge about the data distribution is not known in advance.

9. REFERENCES

1. André L. Vizine Leandro N. de Castro Eduardo R. Hruschka, Ricardo R. Gudwin (2005)Towards Improving Clustering Ants: An Adaptive Ant Clustering Algorithm. Informatica29.(2005):143–154
2. Binitha S.S Siva Sathya (2012) Bio Inspired optimization algorithm. International Journal of Soft Computing and Engineering (IJSCE) Volume-2, Issue-2 :137-151. ISSN: 2231-2307
3. Bo Liu , Jiuhui Pan Bob McKay. Incremental Clustering Based on Swarm Intelligence http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=1FE62EE4E75465F9DD6E97AE2FF73FF2?doi=10.1.1.78.3449&rep=rep1&type=pdf . Accessed 4 February 2014
4. Chuck Ballard et.al.Dynamic Warehousing: Data Mining Made Easy(2007)Shroff Publishers And Distributors Pvt Ltd Mumbai.
5. Claus Aranha, Hitoshi Iba,(2006).Using Genetic Algorithms to improve Ant Colony Clustering.http://www.iba.t.u-tokyo.ac.jp/papers/2006/caranhaMPS2006.pdf. Accessed 2 February 2014
6. Deneubourg j.-L.,S.Goss,N.Franks,C.Detrain, and L.Chretien, The dynamics of Collective Sorting:Robot-Like Ant Robot, Proceedings First Conference on Simulation of Adapive Behaviour: From Animals to Animals, edited by J.A.Meyer and S.W.Wilson,356-365,Cambridg,MA:MIT Press,1991
7. E.D.Lumar and B.Faieta (1994) Diversity and adaptation in populations of Clustering ants. SAB94 Proceedings of the third international conference on Simulation of adaptive behavior: from animals to animats 3: from animals to animats 3: 501-508. ISBN:0-262-53122-4
8. Frank Neumann and Carsten Witt (2010).Bioinspired Computation in Combinatorial Optimization– Algorithms and Their Computational Complexity http://www.bioinspiredcomputation.com/self-archived-bookNeumannWitt.pdf .Accessed on 8 February 2014
9. Jiawei Han,Micheline kamber (2006) DataMining concepts and techniques. Morgan Kaufmann,San Francisco
10. Ling Chen, Xiaohua Xu, Yixin. A Novel Ant Clustering Algorithm Based on Cellular Automata. http://cse.wustl.edu/~ychen/public/IAT.pdf . Accessed 6 February 2014

11. Lior Rokach, Oded Maimon Data Mining and Knowledge Discovery HandBook Chapter 15 Clustering Methods pp 321-350
12. Magnus Erik Hvass Pedersen (2003) Ant colony Clustering & sorting. http://hvass-labs.org/people/magnus/schoolwork/swarm/project3/ant.pdf. Accessed 6 February 2014
13. Marco Dorigo, Mauro Birattari, and Thomas St¨utzle(2006) Ant Colony Optimization Artificial Ants as a Computational Intelligence. IEEE Computational Intelligence Magazine November 2006:28-39
14. Michael Steinbach, George Karypis& Vipin Kumar (2000). A comparison of Document Clustering Techniques. http://glaros.dtc.umn.edu/gkhome/node/157. Accessed 6 February 2014
15. Michael Steinbach, Levent Ertöz, and Vipin Kumar. The Challenges of Clustering High Dimensional Data http://users.cs.umn.edu/~kumar/papers/high_dim_clustering_19.pdf. Accessed 8 February 2014
16. Niphaphorn Obthong, Wiwat Sriphum(2011). Optimal Choice of Parameters for DENCLUE-based and Ant Colony Clustering. International Conference on Modeling, Simulation and Control
    IPCSIT vol.10 (2011) © (2011) IACSIT Press, Singapore
17. Parag M. Kanade and Lawrence O. Hall.Fuzzy Ant Clustering by Centroid Positioning http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.77.7146&rep=rep1&type=pdf. Accessed4 February 2014
18. Selim, Shokri Z, Ismail, M. A.(2009). K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality: 81 – 87. doi:10.1109/TPAMI.1984.4767478
19. UCI data Repository. http://archive.ics.uci.edu/ml/. Accessed 6 February 2014
20. Urszula Boryczka (2008) Ant Clustering Algorithms. Intelligent Information Systems: 377-386 ISBN 978-83-60434-44-4