Ants for Document Clustering

Priya Vaijayanthi[1], Natarajan A M[2] and Raja Murugadoss[3]

**[1] Department of CSE, Bannari Amman Institute of Technology,
Sathyamangalam 638401, Tamilnadu INDIA**

**[2] Department of CSE, Bannari Amman Institute of Technology,
Sathyamangalam 638401, Tamilnadu INDIA**

**[3] Department of Civil Engineering, Bannari Amman Institute of Technology,
Sathyamangalam 638401, Tamilnadu INDIA**

Abstract

The usage of computers for mass storage has become mandatory nowadays due to World Wide Web (WWW). This has placed many challenges to the Information Retrieval (IR) system. Clustering of documents available improves the efficiency of IR system. The problem of clustering has become a combinatorial optimization problem in IR system due to the exponential growth in information over WWW. In this paper, a hybrid algorithm that combines the basic Ant Colony Optimization with Tabu search has been proposed. The feasibility of the proposed algorithm is tested over a few standard benchmark datasets. The experimental results reveal that the proposed algorithm yields promising quality clusters compared to other ones produced by K-means algorithm.

Keywords: *ant colony, document clustering, meta Heuristic, optimization, tabu search*

1. Introduction

The amount of information that is available over Internet has increased exponentially in recent years. This is mainly because of the substantial decline in data storage cost, advancement in network technology and growth in the generation of electronic documents. Moreover the large amount of data stored contains hidden knowledge which can be used in decision making system. Data mining is mining or extraction of knowledge from large amount of data. Several data mining techniques find their application in various fields. The retrieval of relevant information from a huge collection of data is a challenging task in IR system. Clustering is the process of grouping of data that possess high similarity as a group or cluster. Clustering the search result of an IR system helps to present the user with more relevant data of search. This helps to improve the efficiency of search engines.

Clustering is known to be an unsupervised classification that groups data, patterns or feature into clusters without knowing the class label. The problem of clustering in large datasets is a combinatorial optimization problem that is difficult to solve with conventional techniques. Hierarchical and partition based clustering are the common clustering categories. ‘K’ means algorithm is the very popular partition based clustering algorithm. However, there are few limitations observed with this ‘K’ means algorithm from literature is that: (a) It fails to scale with large datasets. (b) The quality of the results of the algorithm highly depends on the initial values used for its parameters. In the past decade it has been proved by many researchers that the biologically or nature inspired algorithms (Ant Colony Optimization Algorithm, Artificial Bee Colony Algorithm, Particle Swarm Optimization, Bird Flocking Algorithm, Frog Leaping Algorithm, Genetic Algorithm) are viable tools to solve complex optimization problems like Travelling Salesman Problem, Vehicle Routing problem, Quadratic Assignment Problem and graph Coloring Problem [1] – [13].

In this paper, a new version of hybrid ant algorithm is proposed by combining the basic features of original ant colony algorithm and tabu search and is explained in section IV. The proposed algorithm is applied to document clustering which is formulated as an unconstrained optimization problem. The viability of the proposed algorithm i.e. hybrid ant algorithm is verified with the results obtained using K-means algorithm for a standard dataset. In the proposed algorithm, a powerful local search algorithm Tabu search is applied to intensify the search process in the area where better solutions are preset.

2. Related Works

Clustering is a typical unsupervised learning technique for grouping similar data points. A clustering algorithm assigns a large number of data points to a smaller number of groups such that data points in the same group share the same properties while, in different groups, they are dissimilar.

[14] proposed a Novel Ant Colony Optimization algorithm for Clustering. This algorithm uses the basic ACO and constructs a connected graph, where the documents are the vertices and are connected through edges. Later the graph is disconnected which results in clusters. [15] proposed a Document clustering method based on Ant algorithm that was tested over text clustering. The author also suggested that parallelization of the algorithm would bring better results. [16] explained the phenomena of corpse clustering and larval sorting in ants. [17] modified the basic model proposed in [16] using a dissimilarity based evaluation of the local density in order to make it suitable for data clustering. They have also introduced the idea of short term memory within each artificial agent. [18] - [19] combined the stochastic principles of clustering by ants with popular K-means algorithm in order to improve the convergence of ant based algorithms. The proposed algorithm was called as AntClass. [20] - [21] proposed Ant System and ACO which is a meta-heuristic approach based on foraging behavior (a positive feedback) of real world ant species. It is based on pheromone model. [22] developed a new algorithm called “a cluster” to solve unsupervised clustering and the data retrieval problem. The algorithm was tested with text document clustering, [23] presented hybridization of ant system with Fuzzy C-means algorithm (FCM) to determine the number of clusters automatically. In this, ant based algorithm is refined using FCM algorithm. [24] presented a novel clustering algorithm called AntTree for unsupervised learning. [25] proposed an ant based clustering algorithm that was proved to be better than traditional partitioning algorithm when tested over real datasets. [26] proposed a hybrid algorithm that combines ant system with SOM and K-means for cluster analysis. This improves the robustness of traditional algorithm. [27] developed multiple ant colonies approach for data clustering. It involves parallel engagement of several individual ant colonies.

The above mentioned literature is not exhaustive but it concentrates only on the works that directly used basic ACO for solving clustering problem. However several other works have been done to solve the same problem with other techniques like Particle Swarm Optimization and Genetic Algorithm.

3. Problem Formulation

There are several methods in which a document can be represented. Among them, vector space representation is the widely used method. We have represented the document as a vector in ‘n’ dimensional space. The documents are preprocessed before they are represented. Preprocessing includes stop word removal, stemming and unique word identification. After this, each document is a list of words. A unique set of words from all the documents of the document corpus or data set is obtained. This list is used to represent each document in the dataset. The commonly used representation is through the weighted representation of the words in the documents. Thus, each document is represented as a vector of weighted values of the words in the document. The weight of each word or feature is represents the importance of the word in the document. It is calculated using the following equation.

$$W_{ij} = df\ *\ idf, \text{where} \qquad (1)$$

w_{ij} is the weight of $word_j$ in $document_i$, df is the document frequency (frequency of occurrences of $word_j$ in $document_i$). idf is the inverse document frequency (importance of $word_j$ in other documents in the document corpus).

The problem of clustering documents in a document corpus is formulated as optimization problem. The solution is a vector where each component corresponds to the cluster centre.

$$DC = (c_1, c_2, c_{3\ldots}c_k) \text{ where} \qquad (2)$$

“DC” is the document clusters and c_i is the centroids of each cluster. The resultant clusters are evaluated for their quality using DB Index.

4. Hybrid Ant Algorithm

This section provides an elaborate note on the basic ACO first and then the proposed hybrid Ant algorithm to solve the problem of document clustering.

4.1 Overview of Standard ACO Algorithm

Ants are altruistic, cooperative, and work collectively toward a common goal. The most amazing about this optimization is that ants tend to take the shortest path (i.e. route) between their nest and some external food source. This natural optimization is a part of stigmergy. After a passage of time, more ants use a particular trail to a food source and hence the trail becomes higher in concentration. The closer the food source is to the colony, the higher the number of trips made by an ant. If the food source is farther away from the nest, a less number of trips are made, and a less concentration of pheromones is applied. Therefore if the concentration of pheromones is higher, then more ants will choose the path over other ants that might be available. This iterative process achieves sub-optimal to optimal trails between the two nodes (i.e. colony and food source). Based on the above natural metaphor, ant algorithms are modeled in such a way that it shares some of the fundamental qualities of real ants. Ant algorithms share these traits in that the simulated ants (or) virtual ants within the environment work in parallel to solve a problem, and through stigmergy, help others to further optimize the solution.

The basic idea of ACO algorithm is to use a positive feedback mechanism, based on an analogy of trail-laying and trail-following behavior of real world ant species. This process of trail-laying reinforces the portions of good solutions that contribute to the quality of these solutions. A virtual pheromone, used as reinforcement, allows good solutions to be kept in memory, from where they can be used to make up better solution.

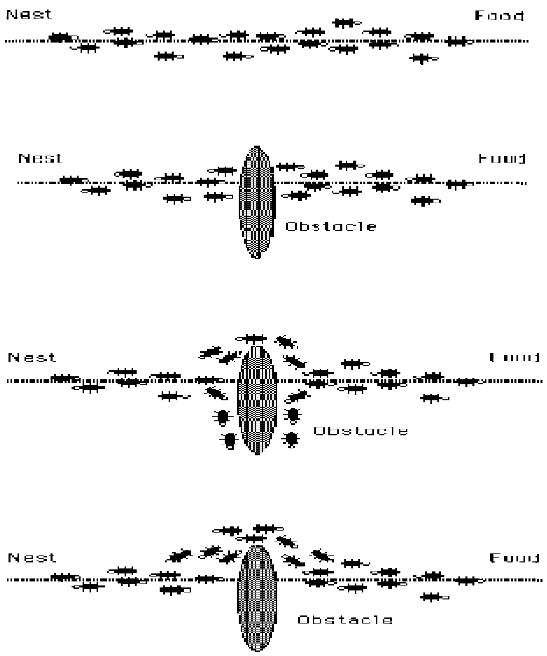


Fig.1 Physical model of ant colony optimization algorithm

4.2 Proposed Hybrid Ant Algorithm

The problem of clustering documents is formulated in the same way as Travelling Salesman Problem. Each document in the document corpus is treated as node. The amount of pheromone deposition between two documents represents the edge and is proportional to the similarity between the documents. Each virtual ant in the colony constructs a graph connecting all the documents in the corpus. The graph is disconnected using graph algorithms like minimum spanning tree which gives the resultant clusters of documents.

In the present paper, a novel hybrid ant algorithm is developed by blending the basic version of ant colony optimization algorithm with conventional K-means algorithm by maintaining a tabu list. The result obtained using K-means algorithm is taken to be initial position for the virtual ants. Each ant aims at constructing a graph that connects all the documents in the corpus. In the process of building graph, an ant at document d_i moves to document d_j if the similarity between d_i and d_j is more and if d_j is not yet visited by it. In order to avoid revisiting of documents each virtual ant maintains a list called Tabu list that contains the visited list of documents. Let 'D' represent the document corpus with 'N' documents. 'D' = $\{d_1, d_2, d_3,\ldots d_n\}$ where each d_i the document. Le t 'M' be the unique set of words in 'D'. Then, each document is represented as a vector $d_i = (w_{i1}, w_{i2}\ldots w_{im})$ where w_{i1} is the weight of term1 in document d_i. Let 'k' be the number of virtual ants engaged. The pseudo code for the Hybrid Ant algorithm is given below.

```
1.  initialization
    set iteration counter to 0
    for every edge (i, j) between documents 'i' and 'j', initialize the trail intensity
    apply K means algorithm to place m ants randomly in 'm' documents
2.  add starting document to the tabu list of the corresponding ant
3.  graph construction
    Repeat until the tabu list of all ants is full
     for each ant k =1 to m do
        Select document 'j' to move from document 'i' with probability $P_{ij}$ (t)
        Place document 'j' in tabu list of ant 'k'
        Move ant 'k' to document 'j'
     End for
   End repeat
4.  pheremone updation
    for every edge (i, j) do
              $\Delta\delta_{ij} = \sum k=1$ to m $\Delta\delta k_{ij}$
      compute $\delta_{ij}$ (t+ 1) = $(1 - \rho)\ \delta_{ij}(t) + \Delta\delta_{ij}$
      set $\Delta\delta_{ij} = 0$
    End for
5.  if stopping criterion met
      disconnect graph  to get clusters
```

Fig.2 pseudo code for hybrid ant algorithm

In the above algorithm, δ_{ij} (t) is the pheromone or trail intensity on the edge between the document 'i' and 'j' at time t. Initially when t=0, the value is set to 1/N where 'N' is the total number of documents in the document corpus as discussed earlier.

At each iteration, after graph construction, δ_{ij} is updated as per the following equation

$$\delta_{ij}(t+1)=\rho\delta_{ij}(t)+\sum_{k=1}^{m}\Delta\delta_{ij}^{k} \qquad (3)$$

$$\delta_{ij}(t+1) = \delta_{ij}(t)(1-\rho) + \Delta\delta \text{ where} \qquad (4)$$

$$\Delta\delta=\left\{\sum_{j=1}^{N_i}\left[1-\frac{\text{dist}(c_i,d_j)}{\gamma}\right]\right\}d_j\in c_i \qquad (5)$$
$$=0,\text{otherwise}$$

where $\boldsymbol{c_i}$ is the centroid vector of the i^{th} cluster, $\boldsymbol{d_j}$ is the j^{th} document vector which belongs to cluster i, dist ($\boldsymbol{c_i}$, $\boldsymbol{d_j}$) is the distance between document d_j and the cluster centroid c_i, N_i stands for the number of documents which belongs to the i^{th} cluster. The parameter 'γ' is defined as swarm similarity coefficient and it affects the number of clusters as well as the convergence of the algorithm. ψ_{ij} is a problem-dependent heuristic function for the document pair doc$_{ij}$ .It is defined as the Euclidean distance dist ($\boldsymbol{d_i}$, $\boldsymbol{d_j}$) between two documents d_i and $d_{j.}$ Ant 'k' moves (6) from document 'i' to document 'j' at t^{th} iteration by following probability $P_{ij}^{k}(t)$ defined by:

$$P_{ij}^{k}(t)=\frac{[\delta_{ij}(t)]^{g}[\psi_{ij}(t)]^{h}}{\sum_{a=1}^{|a_n|}[\delta_{ij}(t)]^{g}[\psi_{ij}(t)]^{h}}$$

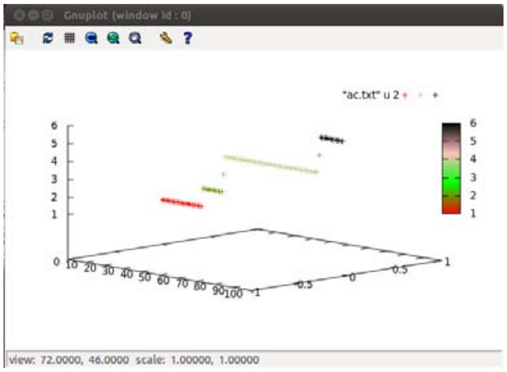
where $l\_\in$ tabu $k(t)$ means 'l' cannot be found in the tabu list of ant 'k' at time t. In other words, 'l' is a document that ant 'k' has not visited yet. The parameters 'g' and 'h' control the bias on the pheromone trail or the problem dependent heuristic function. For the proposed algorithm the termination criteria may be taken either a predefined maximum number of iterations or it the change in the average document distance to the cluster centroid between two successive iterations. In the paper the maximum number of iterations is taken as the termination criteria. The average document distance of the cluster centroid can be calculated as:

$$DB=\frac{1}{n}\sum_{i=1,i\neq j}^{n}\max\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \qquad (7)$$

where $\boldsymbol{c_i}$ is the centroid vector of the i^{th} cluster, $\boldsymbol{d_j}$ is the j^{th} document vector which belongs to cluster i, dist ($\boldsymbol{c_i}$, $\boldsymbol{d_j}$) is the distance between document d_j and the cluster centroid c_i, N_i stands for the number of documents which belongs to the i^{th} cluster. NC stands for the total number of clusters.

5. Experimental Results and Discussions

To find the viability and efficiency of the proposed hybrid ant algorithm, a standard benchmark dataset Library and Information Science Abstracts (LISA) is taken. Initially, the dataset which consists of 635 documents is preprocessed and each document is represented as a vector in 'n' dimensional space. The preprocessing involves various steps like removal of duplicate words, removal of stop words and stemming. Subsequently, a unique set of words in each document is found and the weight of term 'i', w_i is calculated using equation (1) for a particular document 'j'. Using this, a unique set of words in the entire document corpus is found. Based on the weight of each term in this set, a subset of terms is selected as features and the same is used to represent all the documents in the dataset. The reason for selecting a subset of terms as features instead of all the terms for document representation is to reduce the complexity involved in representation. As already discussed in section IV, by providing the number of clusters as input, K-means algorithm is applied to the chosen dataset and the results are plotted in Fig. 1 with 400 iterations as termination criteria. Further the quality of clusters is evaluated using DB Index as discussed in section IV. The results reveal that there is no significant improvement in the quality of clusters with increase in the number of iterations. This inherent behavior of K-means algorithm proves the inefficiency of the same by being trapped at sub-optimal points or premature convergence at early iterations. Also K-means algorithm initially chooses the cluster centers randomly. This random initialization plays a major role in finding good quality clusters. If the random selection happens to be poor, the algorithm falls at local optima at early iterations.



view: 72.0000, 46.0000 scale: 1.00000, 1.00000

Fig. 3 document clusters traced by K-means algorithm for N -100

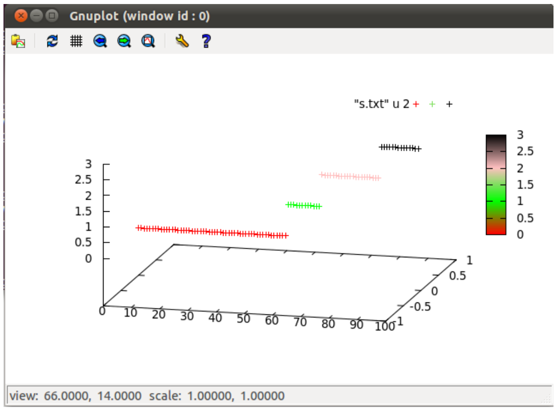


Fig. 4 document clusters traced by hybrid ant algorithm for N =100

One of the reasons for the improvement in the quality of the solutions of the proposed algorithm is the use of Tabu list. As discussed in section IV, tabu list, where each abstract ant maintains, helps to avoid the revisiting of previously visited documents. Consequently, each time, the abstract ants try to find new solutions thereby preventing the algorithm from getting trapped at poor quality solutions in addition to ‘ρ’. Based on the numerical experiments the results are shown in Figs. 3-6 by varying the number of documents from 100 to 200 in step of 50. Tables I and II, present the comparison of cluster quality for N = 100 and 200 generated by K-means and hybrid ant algorithms for different values of ‘M’ and ‘K’. It is observed from the results that the quality of clusters (the DB-Index, i.e. the average document distance of the cluster centroid) obtained using the proposed algorithm is relatively better compared to that of the other ones. This strongly substantiates the viability and efficiency of the proposed algorithm. From the experimental results the values for the parameters are suggested as M = 4 to 10, ρ = 0.1, γ = 0.4, g = 1 and h = 1.

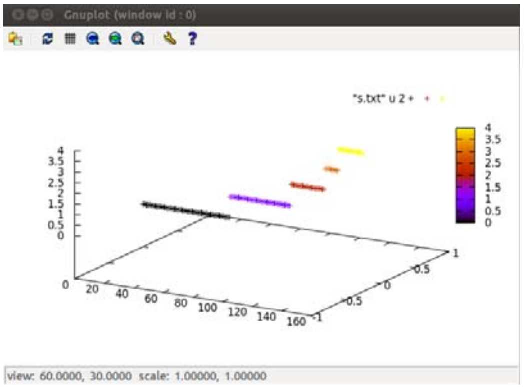


Fig. 5 document clusters traced by K-means algorithm for N -150

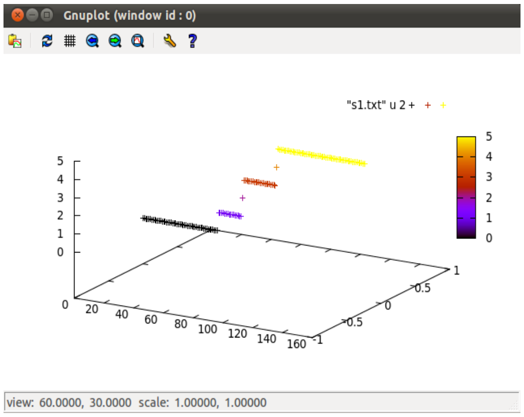


Fig. 6 document clusters traced by hybrid ant algorithm for N -150

The proposed hybrid ant algorithm for document clustering has several unique search characteristics which lead to significant improvement in the consistency and computational efficiency of its performance when compared to K-means algorithm. Another key difference between the proposed algorithm and K-means algorithm is that information stored in artificial pheromone trails represents the memory of the entire colony from all generations, whereas the information on the performance of the search is contained only in the current iteration of K-means algorithm.

TABLE I
CLUSTER QUALITY FOR N =100

M	K	DB- Index	DB- Index
		(Proposed)	(K-means)
4	6	0.5146	0.7340
6	8	0.6292	0.7190
8	12	0.7431	0.8253
10	14	0.6674	0.7112

TABLE II
CLUSTER QUALITY FOR N =200

M	K	DB- Index	DB- Index
		(Proposed)	(K-means)
4	6	0.6984	0.7324
6	5	0.7321	0.7962
8	10	0.6781	0.8472
10	14	0.6819	0.8942

In general, the proposed hybrid ant algorithm yields relatively better quality clusters compared to that of K-means algorithm. It indicates that the hybrid ant algorithm has greater potential to solve the problem of document clustering in IR systems.

6. Conclusion

A Hybrid algorithm that uses Tabu search with the basic ACO has been proposed to solve the problem of Document Clustering. ACO has been proved to be an effective optimization technique to solve combinatorial optimization problems. Tabu search, an efficient local search procedure helps to explore the solutions in different regions of solutions space. Also, it helps to avoid revisiting previously visited solutions. The K-means algorithm is the well known and widely used partition based clustering algorithm. However it suffers from local optima problem. The proposed Hybrid Algorithm is a blended technique that combines features of basic ACO and Tabu search. This novel algorithm is combined with K-means algorithm to bring out the effective solutions by combining the features of both the algorithms. The proposed algorithm is tested with standard benchmark dataset and the quality of solutions produced is compared with that of K-means algorithm. The quality of solutions obtained by Hybrid Algorithm strongly substantiates the effectiveness of the algorithm for document clustering in IR system.

References

[1] J. A. Bland, “Optimal structural design by ant colony optimization,” Engineering Optimization, vol. 33, pp. 425 – 443, 2001.

[2] E. Bonabeau, M. Dorigo and G. Theraulaz, “Swarm intelligence: From Nature to Artificial Systems,” New York: Oxford University Press, 1999

[3] M. H. Botee and E. Bonabeau, “Evolving ant colony optimization,” Adv. Complex Systems, vol. 1, pp. 149-159, 1998

[4] A. Colorni, M. Dorigo and V. Maniezzo, “An investigation of some properties of an ant algorithm,” *in 1992 Proc. Parallel Problem Solving from Nature, Amsterdam, Elsevier, pp. 509-520*

[5] D. Costa and A. Hertz, “Ants can colour graphs,” Journal of Operational Research Society, vol. 48, pp. 295-305, 1997

[6] M. Dorigo, “Ant algorithms solve difficult optimization problems,” *in 2001 Proc. Advances in Artificial Life: Artificial Life Conf., Springer Verlag, pp. 11-22*

[7] M. Dorigo and L. M. Gambardella, “Ant colony system: A cooperative learning approach to the traveling salesman problem,” IEEE Trans. Evol. Comp., vol. 1, pp. 53-66, 1997

[8] M. Dorigo and T. Stuzzle, “An experimental study of the simple ACL algorithm,” *in 2001 Proc. WSES Evolutionary Computation Conf.,* WSES-Press International, pp. 253-258

[9] M. Dorigo and T. Stuzzle, “Ant colony optimization”, England: MIT Press, 2004

[10] M. Dorigo, G. Di Caro and L. M. Gambardella, “Ant algorithms for discrete optimization,” Artificial Life, vol. 5, pp. 137 – 172, 1999

[11] M. Dorigo, V. Maniezzo and A. Colorni, “The ant system: optimization by a colony of cooperating agents,” IEEE Transactions on Systems, Man, and Cybernetics-Part-B, vol. 26, no.1, pp. 1-13, 1996

[12] J. Holland, “Concerning efficient adaptive systems,” Self – organizing systems, Washington, D. C.: Spartan Books, pp. 215-230, 1962

[13] J. Holland, “Adaptation in natural and artificial systems”, Ann Arbor: University of Michigan Press, 1975

[14] Yulan He, Siu Cheung Hui, and Yongxiang Sim, “A novel ant based clustering algorithm for document clustering,” Asia Information Retrieval Symposium, pp. 537 – 544, 2006

[15] Lukasz Machnik, “ACO based document clustering method,” Technical report , Annales UMCS Informatica AI 3, pp 315-323, 2005

[16] J. L. Deneubourg, S. Gross, N. Franks, A. Sendova, C. Detrain and L. Chretien, “The dynamics of collective sorting: robot like ants and ant like robots,” *in 1991 Proc. First International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, MIT Press: Cambridge, MA, pp. 356-363

[17] E. D. Lumer and B. Faieta, “Diversity and adaptation of populations of clustering ants,” in *1994 Proc. Simulation of Adaptive Behavior Conf.,* pp. 501-508

[18] N. Monmarche, M. Silmane and G. Venturini, “On improving clustering in numerical databases with artificial ants,” Advances in Artificial Life, pp. 626-635, 1999

[19] N. Monmarche, “On data clustering with artificial ants,” “Data mining with evolutionary algorithms: research directions”, AAAI Workshop, AAAI Press, pp. 23-26, 2005

[20] M. Dorigo, E. Bonabeau and G. Theraulaz, “Ant algorithms and stigmergy,” Future Generation Computer Systems, vol. 16, no. 8, pp. 851-871, 2000

[21] M. Dorigo, G. Di Caro and L. M. Gambarella, “Ant algorithms for discrete optimization,” Artificial Life, vol. 5, no. 3, 137-172, 1999

[22] V. Ramos and J. J. Merelo, “Self organized stigmergic document maps: environment as mechanism for context learning,” in *2002 Proc.*

Evolutionary and Bio-inspired Algorithms Conf., pp. 284-293

[23] P. Kanade and L. O. Hall, "Fuzzy ants as a clustering concepts", *in 2003 Proc. North American Fuzzy Information Processing Society Conf.,* pp. 227-232.

[24] H. Azzag, N. Monmarche, M. Slimane and G. Venturini, "Ant Tree: a new model for clustering with artificial ants," Evolutionary Computation, vol. 4, pp. 2642-2647, 2003

[25] P. S. Shelokar, V. K. Jayaraman and B. D. Kulkarni, "An ant colony algorithm for clustering,"Analytica Chemica Acta, vol.509, no. 2, pp. 187-195, 2004.

[26] S. Chi and C. C. Yang, "Integration of ant colony SOM and K-means for clustering analysis," Knowledge based Intelligent Information and Engineering Systems, LNCS, Springer, vol. 4251, pp. 1-8, 2006

[27] Yan Yang and Mohamed S. Kamel, "An aggregated clustering approach using multi-ant colonies algorithms," Pattern Recognition, vol. 39, no. 7, pp. 665-671, 2006

R. Priya Vaijayanthi is currently working as Assistant Professor in the Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalm. She had completed her Masters in computer Science and Engineering. She has a teaching experience of 6 years. Her area of interest includes Programming languages, Data mining and web Technologies.

Dr A M Natarajan is working as Professor in the Department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalm. He has a vast experience of more than 40 years in teaching. He has guided more than 15 Phds. He has a long list of publications in National and International Journals and Conferences.

Dr J Raja Murugadoss is Professor and Head of Department of Civil Engineering, Bannari Amman Institute of Technology, Sathyamangalm. He has a vast research experience in the area of optimization. He has more than 20 publications in National and International Conferences and Journals.