A Novel Ant-Based Clustering Approach for Document Clustering

Yulan He, Siu Cheung Hui, and Yongxiang Sim

School of Computer Engineering, Nanyang Technological University
Nanyang Avenue, Singapore 639798
{asylhe, asschui, S8137640I}@ntu.edu.sg

Abstract. Recently, much research has been proposed using nature inspired algorithms to perform complex machine learning tasks. Ant Colony Optimization (ACO) is one such algorithm based on swarm intelligence and is derived from a model inspired by the collective foraging behavior of ants. Taking advantage of the ACO in traits such as self-organization and robustness, this paper proposes a novel document clustering approach based on ACO. Unlike other ACO-based clustering approaches which are based on the same scenario that ants move around in a 2D grid and carry or drop objects to perform categorization. Our proposed ant-based clustering approach does not rely on a 2D grid structure. In addition, it can also generate optimal number of clusters without incorporating any other algorithms such as K-means or AHC. Experimental results on the subsets of 20 Newsgroup data show that the ant-based clustering approach outperforms the classical document clustering methods such as K-means and Agglomerate Hierarchical Clustering. It also achieves better results than those obtained using the Artificial Immune Network algorithm when tested in the same datasets.

1 Introduction

Nature inspired algorithms are problem solving techniques that attempt to simulate the occurrence of natural processes. Some of the natural processes that such algorithms are based on include the evolution of species [1,2], organization of insect colonies [3,4] and the working of immune systems [5,6]. Ant Colony Optimization (ACO) algorithm [3,4] belongs to the natural class of problem solving techniques which is initially inspired by the efficiency of real ants as they find their fastest path back to their nest when sourcing for food. An ant is able to find this path back due to the presence of pheromone deposited along the trail by either itself or other ants. An open loop feedback exists in this process as the chances of an ant taking a path increases with the amount of pheromone built up by other ants. This natural phenomenon has been applied to model the Traveling Salesman Problem (TSP) [3,4].

Early approaches in applying ACO to clustering [7,8,9] are to first partition the search area into grids. A population of ant-like agents then move around this 2D grid and carry or drop objects based on certain probabilities so as to

H.T. Ng et al. (Eds.): AIRS 2006, LNCS 4182, pp. 537–544, 2006.

categorize the objects. However, this may result in too many clusters as there might be objects left alone in the 2D grid and objects still carried by the ants when the algorithm stops. Therefore, Some other algorithms such as K-means are normally combined with ACO to minimize categorization errors [10,11,12]. More recently, variants of ant-based clustering have been proposed, such as using inhomogeneous population of ants which allow to skip several grid cells in one step [13], representing ants as data objects and allowing them to enter either the active state or the sleeping state on a 2D grid [14].

Existing approaches are all based on the same scenario that ants move around in a 2D grid and carry or drop objects to perform categorization. This paper proposes a novel ant-based clustering approach without relying on a 2D grid structure. In addition, it can also generate optimal number of clusters without incorporating any other algorithms such as K-means or AHC. When compared with both the classical document clustering algorithms such as K-means and AHC and the Artificial Immune Network (aiNet) based method [15], it shows improved performance when tested on the subsets of 20 Newsgroup data [16]. The rest of the paper is organized as follows. Section 2 briefly describes the Ant Colony Optimization (ACO) algorithm. The proposed ant-based clustering approach is discussed in Section 3. Experimental results are presented in Section 4. Finally, section 5 concludes the paper and outlines the possible future work.

2 Ant Colony Optimization

The first Ant Colony Optimization (ACO) algorithm has been applied to the Traveling Salesman Problem (TSP) [3,4]. Given a set of cities and the distances between them, the TSP is the problem of finding the shortest possible path which visits every city exactly once. More formally, it can be represented by a complete weighted graph $G = (N, E)$ where N is the set of nodes representing the cities and E is the set of edges. Each edge is assigned a value d_{ij} which is the distance between cities i and j. When applying the ACO algorithm to the TSP, a pheromone strength $\tau_{ij}(t)$ is associated to each edge (i, j), where $\tau_{ij}(t)$ is a numerical value which is modified during the execution of the algorithm and t is the iteration counter.

The skeleton of the ACO algorithm applied to the TSP is:

```
procedure ACO algorithm for TSPs
    set parameters, initialize pheromone trails
    while (termination condition not met) do
        Tour construction
        Pheromone update
    end
end ACO algorithm for TSPs
```

At first, each of the m ants is placed on a randomly chosen city. At each *Tour construction* step, ant k currently at city i, chooses to move to city j at the tth iteration based on the probability $P_{ij}^{k}(t)$ which is biased by the pheromone trail strength $\tau_{ij}(t)$ on the edge between city i and city j and a locally available

heuristic information η_{ij}. Each ant is associated with a *tabu list* in which the current partial tour is stored, i.e. $tabu_k(s)$ stores a set of cities visited by ant k so far at time s. After all the ants have constructed their tours, *Pheromone update* is performed by allowing each ant to add pheromone on the edges it has visited. At the end of the iteration, the tabu list is emptied and each ant can choose an alternative path for the next cycle.

3 Ant-Based Algorithm to Document Clustering

In document clustering, the vector-space model is usually used to represent documents and documents are categorized into groups based on the similarity measure among them. For each document d_i in a collection $\mathcal{D}$, let $\mathcal{W}$ be the unique

1. **Initialization.**
   set the iteration counter $t = 0$
   For every edge (i, j), set an initial value $\tau_{ij}(t)$ for trail intensity and $\Delta\tau_{ij} = 0$.
   Place m ants randomly on the n documents.
2. Set the tabu list index $s = 1$.
   for $k = 1$ to m do
   Place starting document of the kth ant in tabu$_k(s)$
   end for
3. **Tour Construction.**
   repeat until tabu list is full
   Set $s = s + 1$
   for $k = 1$ to m do
   Choose the document j to move to with probability $P_{ij}^k(t)$
   Move the kth ant to the document j
   Insert document j into the tabu list tabu$_k(s)$
   end for
   end repeat
4. **Pheromone Update.**
   for every edge (i, j) do
   $\Delta\tau_{ij} = \sum_{k=1}^{m} \Delta\tau_{ij}^k$
   compute $\tau_{ij}(t + 1) = (1 - \rho)\tau_{ij}(t) + \Delta\tau_{ij}$
   set $\Delta\tau_{ij} = 0$
   end for
5. Set $t = t + 1$
   if a stopping criteria is met, then
   print clustering results, stop
   else
   empty tabu lists, go to 2
   end if

Fig. 1. Ant-based document clustering algorithm

word items occurring in $\mathcal{D}$ and $M = |\mathcal{W}|$, then document d_i is represented by the vector $\boldsymbol{d_i} = (w_{i1}, w_{i2}, \cdots, w_{iM})$ where w_{ij} denotes the appearance of word w_j in document d_i which is normally weighted by *term frequency X inverse document frequency* (TFIDF).

Fig. 1 shows the ant-based document clustering algorithm. The design of the ant-based algorithm involves the specification of the following:

- $\tau_{ij}(t)$ represents the amount of pheromone associated with the document pair doc_{ij} at iteration t. The initial amount of pheromone deposited at each path position is inversely proportional to the total number of documents which is defined by $\tau_{ij}(0) = \frac{1}{N}$ where N is the total number of documents in the collection $\mathcal{D}$.
  At every generation of the algorithm, τ_{ij} is updated by $\tau_{ij}(t+1) = (1-\rho)\tau_{ij}(t) + \Delta\tau$ where $\rho \in (0, 1]$ determines the evaporation rate and the update of pheromone trail; $\Delta\tau$ is defined as the integrated similarity of a document with other documents within a cluster which is measured by:

$$\Delta\tau = \begin{cases} \sum_{j=1}^{N_i} [1 - \frac{\text{dist}(\boldsymbol{c_i}, \boldsymbol{d_j})}{\gamma}] & d_j \in c_i \\ 0 & otherwise \end{cases} \tag{1}$$

  where $\boldsymbol{c_i}$ is the centroid vector of the ith cluster, $\boldsymbol{d_j}$ is the jth document vector which belongs to cluster i, $\text{dist}(\boldsymbol{c_i}, \boldsymbol{d_j})$ 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 ith 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.
- η_{ij} is a problem-dependent heuristic function for the document pair doc_{ij}. It is defined as the Euclidean distance $\text{dist}(\boldsymbol{d_i}, \boldsymbol{d_j})$ between two documents d_i and d_j.
- Ant k moves from document i to document j at tth iteration by following probability $P_{ij}^k(t)$ defined by:

$$P_{ij}^k(t) = \begin{cases} \frac{[\tau_{ij}(t)]^\alpha \cdot [\eta_{ij}]^\beta}{\sum_{l \notin \text{tabu}_{k(t)}} [\tau_{il}(t)]^\alpha \cdot [\eta_{ij}]^\beta} & if \quad j \notin \text{tabu}_k(t) \\ 0 & otherwise \end{cases} \tag{2}$$

  where $l \notin \text{tabu}_k(t)$ 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 α and β control the bias on the pheromone trail or the problem-dependent heuristic function.
- Finally, a stopping criteria needs to be carefully set. It can either be a predefined maximum number of iterations or it can be the change in the average document distance to the cluster centroid between two successive iterations. The average document distance to the cluster centroid is defined as:

$$f = \frac{\sum_{i=1}^{N_C} \{ \frac{\sum_{j=1}^{N_i} \text{dist}(\boldsymbol{c_i}, \boldsymbol{d_j})}{N_i} \}}{N_C} \tag{3}$$

where $\boldsymbol{c_i}$ is the centroid vector of the ith cluster, $\boldsymbol{d_j}$ is the jth 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 ith cluster. N_C stands for the total number of clusters.

Once the ant-based clustering algorithm has run successfully, a fully connected network of nodes will be formed. Each node represents a document, and every edge is associated with a certain level of pheromone intensity. The next step is essentially to break the linkages in order to generate clusters. Various methods can be applied such as minimum spanning trees. Here, the average pheromone strategy is used. The average pheromone of all the edges is first computed and then edges with pheromone intensity less than the average pheromone will be removed from the network and results in a partially connected graph. Nodes will then be separated by their connecting edges to form clusters.

4 Performance Evaluation

In this section, the performance of the proposed ant-based clustering algorithm is evaluated. The experimental setup is first explained followed by a comparative account of the results generated from different experiments conducted. This section also attempts to obtain the optimal parameters of ant-based clustering and reason intuitively its performance on clustering in comparison with other algorithms.

4.1 Experimental Setup

Experiments have been conducted on the 20 Newsgroup data set [16] which is in fact a benchmarking data set commonly used for experiments in text applications of machine learning techniques. A few combinations of subsets of documents are selected for experiments based on various degrees of difficulty. Table 1 lists the details of various subsets used. Each subset consists of 150 randomly chosen documents from various newsgroups. All newsgroup articles have their headers stripped and main body pre-processed only. Once transformed into term-document vectors, they are fed into the clustering engine.

The tunable parameters in the ant-based clustering algorithm include number of iterations i, number of ants m, rate of decay ρ, swarm similarity coefficient γ,

Table 1. Statistics on experimental data

Dataset	*Topics*	*No. of Docs Per Group*	*Total No. of Docs*
1	sci.crypt, sci.space	150, 150	300
2	sci.crypt, sci.electronics	150, 150	300
3	sci.space, rec.sport.baseball	150, 150	300
4	talk.politics.mideast, talk.religion.misc	150, 150	300

and the parameters α and β that control the bias on the pheromone trail. A set of experiments have been conducted and it was found that the optimal value or range of each parameter is $i = 100$, $14 \leq m \leq 19$, $0.1 \leq \rho \leq 0.3$, $0.4 \leq \gamma \leq 0.5$, $\alpha = 1$, and $\beta = 1$. From this section onwards, subsequent experiments will be carried out using these selected optimized values for the parameters.

4.2 Clustering Accuracy

This section evaluates the performance on clustering accuracy based on F-measure [17]. The 10-fold cross-validation technique is applied to each of the four sample data sets and the average F-measure score will be used as a comparative observation. The experimental results given in Table 2 shows that the ant-based clustering method performs much better than AHC and K-means in all four sample data sets. In AHC and K-means, information on the expected number of clusters must be supplied, but ACO is able to predict the cluster structure accurately by generating the exact number of clusters in each sample data set.

Table 2 also compares the published results from aiNet [15], a technique based on Artificial Immune System (AIS), with ant-based clustering using the identical sets of data. Belonging to the same class of nature inspired algorithms as ACO, AIS performs an evolutionary process on raw document data based on the immune network and affinity maturation principles. The proposed method uses Agglomerate Hierarchical Clustering (AHC) and K-means to construct antibodies and detect clusters. Also, Principal Component Analysis (PCA) is introduced for dimensionality reduction in a bid to improve clustering accuracy. From the published experimental results, none of the proposed AIS methods scored is close to the results of ant-based clustering in all four sample data sets. In fact, the results from the ant-based clustering algorithm have shows a relative improvement in performance over the aiNet$_{pca}$_K-means by 5% to 34% and a relative improvement over the aiNet$_{pca}$_AHC by 7% to 35%. Furthermore, it is observed that the ant-based approach is far more stable by producing consistent F-measure scores of higher than 0.8, while the aiNet scored varying F-measure results, ranging from 0.6 to 0.8.

Although both ACO and AIS belong to the same family of nature inspired algorithms, these two methods use entirely different approaches to model problem solving techniques. This results in a performance difference when applying both to the same problem domain. The evolution stage in ACO makes use of stochastic ants as decision tools for choosing a path to move based on pheromone trail intensity. The final network generated has edges of varying amounts of pheromone that represent the differences among documents in a collection. Such differences will allow easy partitioning of semantically similar documents by using the averaged pheromone level as a threshold for searching connected sub-graphs in the network. This entire evolving process models after the document clustering task almost perfectly.

On the other hand, using the AIS to model the same clustering task may not be an ideal case. The clonal selection theory is only capable of generating an

Table 2. Comparison of clustering accuracy of AHC, K-means, aiNet, and ant-based clustering

	F-Measure			
Method	*Subset 1*	*Subset 2*	*Subset 3*	*Subset 4*
AHC	0.665	0.654	0.700	0.631
K-means	0.794	0.580	0.513	0.624
aiNet_AHC	0.810	0.640	0.718	0.641
aiNet$_{pca}$_AHC	0.815	0.735	0.715	0.640
aiNet_K-means	0.807	0.628	0.630	0.639
aiNet$_{pca}$_K-means	0.836	0.661	0.631	0.646
Ant-Based	**0.874**	**0.811**	**0.803**	**0.865**

immune network based on affinity between antibodies and antigens (documents). Detection of clusters in this network requires assistance from other methods such as AHC or K-means, which may disrupt the biological ordering of antigens. Therefore, the performance of AIS is limited by the artificial clustering applied on its network.

5 Conclusions

This paper has proposed a novel ant-based clustering algorithm and its application to the unsupervised data clustering problem. Experimental results showed that the ant-based clustering method performs better than K-means and AHC by a wide margin. Moreover, the ant-based clustering method has achieved a higher degree of clustering accuracy than the Artificial Immune Network (aiNet) algorithm which also belongs to the same family of nature inspired algorithms.

In future work, it would be interesting to investigate the behavior of the ant-based algorithm using other sources of heuristic functions and pheromone update strategies. In addition, more intelligent methods of breaking linkages among documents can be devised to replace the existing average pheromone approach.

References

1. E. Yu and K.S. Sung. A genetic algorithm for a university weekly courses timetabling problem. *International Transactions in Operational Research*, 9(6):703–717, 2002.
2. E.K. Burke, D.G. Elliman, and R.F. Weare. A genetic algorithm based university timetabling system. In *Proceedings of the 2nd East-West International Conference on Computer Technologies in Education*, pages 35–40, Crimea, Ukraine, September 1994.
3. M. Dorigo, V. Maniezzo, and A. Colorni. Positive feedback as a search strategy. Technical report 91-016, Politecnico di milano, 1991. Dip. Elettronica.
4. 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*, 26(1):29–42, 1996.

5. L.N. de Castro and F.J. Von Zuben. Learning and optimization using the clonal selection principle. *IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems*, 6(3):239–251, 2002.
6. D. Dasgupta, Z. Ji, and F. Gonzlez. Artificial immune system (ais) research in the last five years. In *Proceedings of the International Conference on Evolutionary Computation Conference (CEC)*, Canbara, Australia, December 2003.
7. J. L. Deneubourg, S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, and L. Chretien. The dynamics of collective sorting robot-like ants and ant-like robots. In *Proceedings of the first international conference on simulation of adaptive behavior on From animals to animats*, pages 356–363, Cambridge, MA, USA, 1990. MIT Press.
8. Lumer E. D. and Faieta B. Diversity and adaptation in populations of clustering ants. In Cli D., Husbands P., Meyer J., and Wilson S., editors, *Proceedings of the Third International Conference on Simulation of Adaptive Behaviour: From Animals to Animats 3*, pages 501–508, Cambridge, MA, 1994. MIT Press.
9. Kuntz P., Layzell P., and Snyers D. A colony of ant-like agents for partitioning in vlsi technology. In P. Husbands and I. Harvey, editors, *Proceedings of the Fourth European Conference on Artificial Life*, pages 417–424. MIT Press, 1997.
10. N. Monmarche. On data clustering with artificial ants. In Alex Alves Freitas, editor, *Data Mining with Evolutionary Algorithms: Research Directions*, pages 23–26, Orlando, Florida, 18 1999. AAAI Press.
11. B. Wu, Y. Zheng, S. Liu, and Z. Shi. Csim: a document clustering algorithm based on swarm intelligence. In *Proceedings of the 2002 congress on Evolutionary Computation*, Honolulu, USA, 2002.
12. Y. Peng, X. Hou, and S. Liu. The k-means clustering algorithm based on density and ant colony. In *IEEE International Conference in Neural Networks and Signal Processing*, Nanjing, China, December 2003.
13. Julia Handl and Bernd Meyer. Improved ant-based clustering and sorting. In *PPSN VII: Proceedings of the 7th International Conference on Parallel Problem Solving from Nature*, pages 913–923, London, UK, 2002. Springer-Verlag.
14. L. Chen, X. Xu, and Y. Chen. An adaptive ant colony clustering algorithm. In *Proceedings of the Third International Conference on Machine Learning and Cybernetics*, pages 1387–1392, Shanghai, China, August 2004.
15. Na Tang and V. Rao Vemuri. An artificial immune system approach to document clustering. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 918–922, New York, NY, USA, 2005. ACM Press.
16. *20 Newsgroups Data Set*, 2006. http://people.csail.mit.edu/jrennie/20Newsgroups/.
17. M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In *In KDD Workshop on Text Mining*, 2000.