The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0264-0473.htm

Received 12 April 2005
Revised 12 November 2005
Accepted December 2005

A new efficient approach for data clustering in electronic library using ant colony clustering algorithm

An-Pin Chen and Chia-Chen Chen
Institute of Information Management, National Chiao-Tung University, Hsinchu, Taiwan

Abstract

Purpose – Traditional library catalogs have become inefficient and inconvenient in assisting library users. Readers may spend much time in searching library materials via printed catalogs. Readers need an intelligent and innovative solution to overcome this problem. The purpose of this paper is to illustrate how data mining technology is a good approach to fulfill readers' requirements.

Design/methodology/approach – Data mining is considered to be the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. This paper analyzes the readers' borrowing records by using the following techniques: data analysis, building data warehouse and data mining.

Findings – The mining results show that all readers can be categorized into five clusters, and each cluster has its own characteristics. It was also found that the frequency for graduates and associate researchers to borrow multimedia data is much higher. This phenomenon shows that these readers have a higher preference for accepting digitized publication. Besides, we notice that more readers borrow multimedia data rise in years. This up trend indicates that readers are gradually shifting their preference in reading digital publications.

Originality/value – The paper proposes a technique to discover clusters by using ant colony methods.

Keywords Digital libraries, Cluster analysis, Data collection, Data analysis

Paper type Case study

1. Introduction

Database systems that collect, analyze, and transfer data are used for various mid-range and large organizations. Over time, more and more current, detailed, and accurate data are accumulated and stored in databases with various stages. These data may be related to designs, products, machines, materials, processes, inventories, sales, marketing, and performance data. They may include patterns, trends, associations, and dependencies. The collected data contain valuable information that could be integrated into the organization's strategy, and be used to improve organizational decisions. Consequently, data mining methods become important tools in today's society.

Data mining is the process of extracting valid, previously unknown, comprehensible information from large databases in order to improve and optimize organization decisions (Anand and Buchner, 1998). The term knowledge discovery in database (KDD) is used to denote the entire process of turning low-level data into

The Electronic Library
Vol. 24 No. 4, 2006
pp. 548-559

0264-0473
DOI 10.1108/02640470610689223

high-level knowledge, where data mining is considered as a single step in the process that involves finding patterns in the data (Fayyad, 1996).

Cluster analysis is a technique used to forecast and infer a great deal of data from the domain of data mining. The objective is to differentiate the data that have unknown categories. Decision managers can obtain reference information resulting from cluster analysis. Therefore, developing an efficient clustering algorithm is important for many applications.

The K-Means Algorithm is commonly used to conduct clustering as it can quickly cluster data. However, the K-Means Algorithm has many drawbacks when applied to real world clustering problems. For example, the number of clusters that the user needs to specify is not easily predicted. Clustering results may not be good if the wrong cluster numbers are assigned. Also, the K-Means Algorithm is sensitive to noise and outliers.

Accordingly, this research combines the concepts of the traditional clustering algorithm and the technique of ant colony optimization to develop a data cluster algorithm that can obtain a global optimization solution. This approach alleviates the drawbacks which lead the K-Means Algorithm to easily fall into the awkward situation where the local optimization solution is flawed. To demonstrate the benefits of our method, this research conducts experiments on several real world data sets. The result proves that the proposed cluster algorithm can obtain better cluster objective values and more accurate cluster results.

2. Literature review

This section provides a general definition of ant colony optimization, which is the main component of the proposed method as well as definitions of data mining and its task and clustering analysis. Details are as follows:

2.1 The KDD process

Before discussing data mining, we must first introduce knowledge discovery as data mining is one of the steps in the knowledge discovery process. The definition of knowledge discovery in databases (KDD), given by Fayyad *et al.* (1996), is defined as a "nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data." From their point-of-view, in a KDD process with large volumes of data processing, iterative testing and analysis, data and patterns are the starting and ending points respectively. Where the process is considered nontrivial, the data analysis goes beyond mere quantitative computing. The goal is to search for structures, models, patterns, associations or parameters. Results and patterns should be valid for new data with some degree of certainty and also be novel and potentially useful for users. The overall KDD process is outlined in Figure 1 (Han and Kamber, 2001). It is interactive and iterative involving the following steps by Fayyad *et al.* (1996):

(1) *Understanding the application domain:* includes relevant prior knowledge and goals of the application.

(2) *Extracting the target data set:* includes selecting a data set or focusing on a subset of variables.

(3) *Data cleaning and preprocessing:* includes basic operations, such as noise removal and handling of missing data. It is an often ignored but extremely important step in the data mining process.

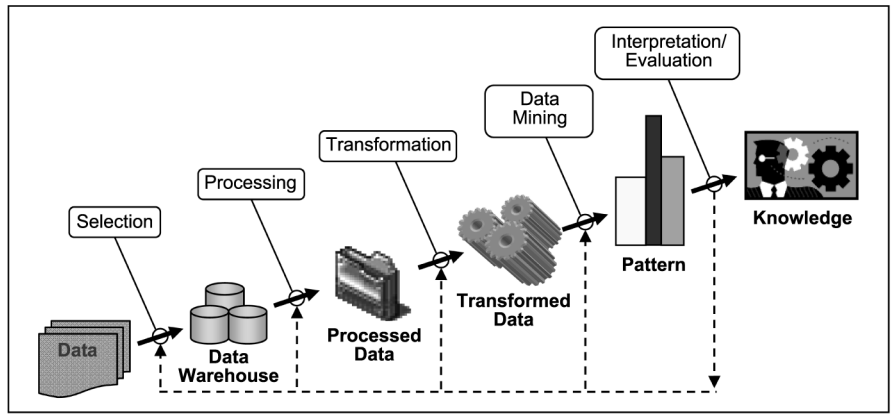

Figure 1.
An overview of KDD process

(4) *Data integration:* includes integrating multiple, heterogeneous data sources.

(5) *Data reduction and projection:* includes finding useful features to represent the data and using dimensionality reduction or transformation methods.

(6) *Choosing the function of data mining:* includes deciding the purpose of the model derived by the data mining algorithm.

(7) *Choosing the data mining algorithm(s)*: includes selecting method(s) to be used for searching patterns in data, such as deciding on which model and parameters may be appropriate.

(8) *Data mining:* includes searching for patterns of interest in a particular representational form or a set of such representations.

(9) *Interpretation:* includes interpreting the discovered patterns, as well as the possible visualization of the extracted patterns. One can analyze the patterns automatically or semi-automatically to identify the truly interesting/useful patterns for the user.

(10) *Knowledge discovered:* includes incorporating this knowledge into the performance system, taking actions based on knowledge.

2.2 Data mining

An important step in the KDD process is data mining. The literal definition of data mining is "to uncover useful information from a large amount of data." The purpose of data mining is to extract interesting knowledge from a database, data warehouse, or some other large information storage unit (Han and Kamber, 2001). From a technical viewpoint, it combines the method of gathering and cataloging information then proceeds to generate rule-like knowledge from a large amount of data.

A particular data mining algorithm is usually an instantiation of the model preference search components. The more common model functions in the current data mining process include the following (Mitra *et al.*, 2002).

- *Classification:* classifies a data item into one of several predefined categories.
- *Regression:* maps a data item to a real-valued prediction variable.

- *Clustering:* maps a data item into a cluster, where clusters are natural groupings of data items based on similarity metrics or probability density models.
- *Association rules:* describes association relationship among different attributes.
- *Summarization:* provides a compact description for a subset of data.
- *Dependency modeling:* describes significant dependencies among variables.
- *Sequence analysis:* models sequential patterns, like time-series analysis. The goal is to model the states of the process generating the sequence or to extract and report deviations and trends over time.

Resulting from the popularity of electronic commerce and personalized trends, the technique of data mining is also used extensively to analyze product items based on user purchasing. This is to determine user preference and to provide related product information to the users in order to increase sales and/or purchasing rates. For example, Amazon.com, the famous internet store, also uses data mining techniques to recommend products to users according to purchasing records as well as items they previously clicked.

Applying data mining techniques in a electronic library service is also considered a trend as it can automatically filter out useful knowledge using user profiles and the function of statistical analysis. For example, filtering out popular topics from every individual's borrowing history can help promote book circulation in the library. The electronic library can also use functions of statistical analysis along with data mining to provide information on books, articles, topics and other long-term personal services for promoting circulation.

2.3 Clustering

Clustering analysis finds groups, each very different from the other. Within the group, however, all members are very similar. Unlike classification, the class label of each group is not known. Clustering is a way to naturally segment data into groups, whereas classification is a way to segment data by assigning it into groups. Briefly, a good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. However, how good a cluster is ultimately depends on the opinion of the user.

The goal of clustering analysis is to group similar objects. Cluster similarity is measured according to the mean value in a cluster, which is viewed as the cluster's center of gravity. Several clustering methods are illustrated in the following categories:

(1) *Partitioning methods.* Partitioning methods construct a partition of a database of N objects into a set of k clusters. Usually, they start with an initial partition and then use an iterative control strategy to optimize an objective function.

The K-Means Algorithm (Han and Kamber, 2001) is the well-known and commonly used clustering algorithm. It takes input parameter k and partitions data into k clusters. First, we select k objects to represent the cluster centers. The remaining objects are assigned to the cluster whose center is closest to the object. Then it computes the mean value for each cluster as new cluster centers. This process iterates until the criterion function converges.

(2) *Hierarchical methods.* Hierarchical methods create a hierarchical decomposition of the database. The method can be classified as an agglomeration or a division depending on how the hierarchical decomposition is formed. The agglomerative

approach, also called the bottom-up approach, merges the objects until all objects have been merged into one. The division approach, typically known as the top-down approach, splits a cluster into sub-clusters until each cluster has only one element. The hierarchical method does not need any input parameters, and thus overcomes the drawback of partitioning algorithms. The disadvantage of this method is that the termination condition must be specified (Sheikholeslami *et al.*, 1998).

(3) *Density based methods.* Density-based methods founded upon connectivity and density functions can filter out noise and find clusters of arbitrary shape. It creates clusters by continuously growing a cluster so long as the density of the data objects in the neighborhood exceeds a certain threshold (Han and Kamber, 2001).

2.4 Ant colony optimization

Dorigo (1997a, b) proposed an ant colony optimization algorithm (ACO), which has been successfully applied to several NP-hard problems. Amongst these successes are the traveling salesman problem (TSP) (Dorigo and Gambardella, 1997a, b) and the quality of service problem (QoS) (Caro and Dorigo, 1998; Leguizamon and Michalewicz, 1999). Just as its name implies, the ACO algorithm originates from the study of the behavior of a natural ant colony. There are three ideas from the natural ant colony that has been transferred to the artificial ant colony:

(1) the preference for paths with a high pheromone level;

(2) the higher rate of growth in the amount of pheromones on shorter paths; and

(3) the information exchanged among ants (Dorigo and Gambardella, 1997a, b).

In the ACO algorithm, the artificial pheromone trails are presented by a number, and the numeric information of pheromone trials is modified by the artificial ants. The ACO algorithm summarizes as follows:

- *Step 0:* set parameters and initialize pheromone trails;
- *Step 1:* each ant constructs its solution;
- *Step 2:* calculate the scores of all solutions;
- *Step 3:* update the pheromone trails; and
- *Step 4:* if the best solution has not been changed after some predefined iterations, terminate the algorithm; otherwise, go to step 2.

The ACO algorithm simulates the behavior of real ants. As an ant in nature moves, it leaves behind a pheromone trail. The movement of any ants following that first ant depends on the detection of the pheromones on that trail. Not only will the ant detect and follow that trail, it will also seek out newer and better paths based on the amounts of pheromones detected. This pheromone trail can be presented as a numeric value. Therefore, using these numbers, we should be able to calculate and set parameters for pheromone values.

3. Methodology

3.1 Definitions and notations

The following terms and notations are used throughout the algorithm:

NC:	the number of clusters;
m:	the total number of ants;
M_k:	the set M is performed by ant k;
$p_k(r, s)$:	the probability with which ant k chooses to move from node r to node;
$s\tau(r, u)$:	the amount of pheromone trail on edge (r, u);
$\overline{p}_k$:	the average of $p_k(r, s)$ for the set $\notin M_k$;
$\eta(r, u)$:	the inverse of the distance between nodes r and u;
β:	a parameter which weighs the relative importance of pheromone trail and of closeness;
q:	a value chosen randomly with uniform probability in [0,1];
q_0:	a parameter which determines the relative importance of exploitation versus exploration ($0 \leq q_0 \leq 1$);
S:	a random variable selected according to $p_k(r, s)$;
α:	the pheromone decay parameter of global updating ($0 < \alpha < 1$);
ρ:	the pheromone decay parameter of local updating ($0 < \rho < 1$);
τ_0:	the initial level of pheromone;
$O_{center}(M)$:	the center of all nodes in M;
γ:	a parameter which is the percentage of farthest nodes chosen to regroup;
CV_{intra}:	the intra-cluster variance;
CV_{intra}':	the intra-cluster variance after regrouping.

3.2 The concept of ACO algorithm

The proposed method is ACO algorithm, and its main process is shown in Figure 2.

The first step is to initialize the parameters. A set of artificial ants is positioned on the first job according to an initialization rule (e.g. randomly). Each ant constructs its own cluster. Once the ants have completed their clusters, each cluster's variance (CV_{intra}) is calculated. The γ percent of the farthest nodes are chosen to be regrouped into the cluster with the shortest distance to $O_{center}(M)$. If the new variance (CV_{intra}') is smaller than CV_{intra}, that means the nodes in the updated cluster are more similar than the nodes in the previous cluster. While applying new clusters, an ant will simultaneously update the amount of pheromone on its visited paths (by applying the local updating rule). After all of the ants have built solutions, the pheromone trails on the paths of the global best cluster are modified again (by applying the global updating rule) up to the current iterations. The process is terminated after predefined iterations.

The complete algorithm of ACO algorithm is summarized as follows:

- *Input:* n nodes.
- *Output:* the number of predefined clusters.

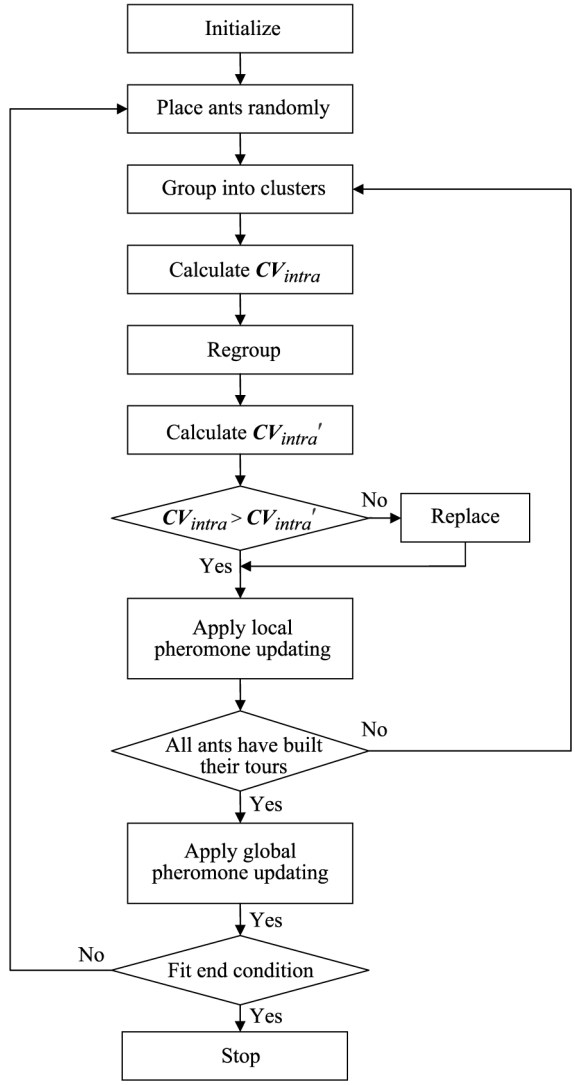

Figure 2.
The flow chart of ACO algorithm

- *Step 0.* Initialize the parameters, which include the number of ants m, and the number of clusters NC, parameters q_0, β, the pheromone decay parameter α, ρ, the percentage of farthest nodes chosen to regroup γ.
- *Step 1.* Place NC ants on the nodes randomly.
- *Step 2.* Group the collected nodes into clusters. An ant k at node r chooses the node s to move along the nodes which do not belong to its working memory M_k.

The state transition rule is applied by the following probabilistic formula, which provides a direct way to balance between exploration of new edges and exploitation of a priori and accumulated knowledge about the problem:

$$s = \begin{cases} \arg\max\limits_{u \notin M_k} \left\{ \left[\tau(r,u)\right] \cdot \left[\eta(r,u)\right]^{\beta} \right\} & \text{if } q \le q_0 \text{ (Exploitation)} \\ S & \text{otherwise (Exploitation)} \end{cases} \quad (1)$$

S is a random variable selected according to the probability distribution given in Equation (2), which favors edges that are shorter and have a higher level of pheromone trail.

$$p_k(r,s) = \begin{cases} \dfrac{\left[\tau(r,s)\right] \cdot \left[\eta(r,s)\right]^{\beta}}{\sum\limits_{u \notin M_k} \left[\tau(r,u)\right] \cdot \left[\eta(r,u)\right]^{\beta}} & \text{if } u \notin M_k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

If $p_k(r,s) \ge \bar{p}_k$, ant k collects the node s.

- *Step 3*. Calculate $O_{center}(M)$ and CV_{intra} of each cluster. The nodes that are in $\boldsymbol{\gamma}$ are chosen to be regrouped to the closest group.
- *Step 4*. Calculate $CV_{\boldsymbol{intra}}'$. If $CV_{\boldsymbol{intra}} > CV_{\boldsymbol{intra}}'$, the replaced result is adopted and local pheromone is updated on all edges according to

$$\tau(r,s) = (1-\rho) \cdot \tau(r,s) + CV_{inter}^{-1} \quad (3)$$

- *Step 5*. Global pheromone updating is intended to allocate a greater amount of pheromone. While all ants have built their tours, global pheromone is updated according to:

$$\tau(r,s) = (1-\alpha) \cdot \tau(r,s) + CV^{-1} \quad (4)$$

Where CV is the sum of the smallest CV_{intra}.

Step 6. The process is iterated until the end condition is met.

4. Example and discussion

4.1 Example

Let's take a campus electronic library as an example. Students enter the campus electronic library system and want to borrow some books or magazines. Through data clustering, it is found that students in the same cluster have the similar reading habits, and then, the library system can recommend books that are similar to those they have previously borrowed. Sample data is shown in Table I from which we can see that there are 17 tuples and two attributes in a given database. One attribute, "ID", which is the identification number of staff or faculty in a university, is classified according to

college departments. Another attribute, "CNo", is a book's category number, which denotes which category staff or faculty members are likely borrow from. In practice, it is feasible to cluster all transactions in the databases into several groups. That is, since the transactions in each group are similar with respect to the clustering variables (i.e. borrow various books), we can employ the proposed method to find out similar behaviors from the representative records. In this example, we assume that all transactions are clustered into three groups.

The proposed algorithm applied to electronic library is presented as follows.

- *Input.* Sample data from Table I.
- *Output.* The clusters of readers.
- *Step 0.* Set the number of ants $m = 12$, the number of clusters $NC = 3$, $q_0 = 0.7$, the distance weigh $\beta = 2$, local pheromone decay parameter $\alpha = 0.5$, global pheromone decay parameter $\rho = 0.5$, the percentage of farthest nodes chosen to regroup $\gamma = 0.6$.
- *Step 1.* The ants are placed randomly to be starting nodes (50, 4), (16, 4.6), and (30, 1.8).
- *Step 2.* The collected nodes in each group are shown in Figure 3.
- *Step 3.* The calculated results of $O_{center}(M)$ and CV_{intra} in each group are shown in Table II, and the 50 percent farthest nodes are chosen to be regrouped.
- *Step 4.* The calculated results of CV_{intra}' and New groups are shown in Table III
- *Step 5.* While all ants have built their tours, the smallest CV_{intra} is chosen to execute global pheromone updating.
- *Step 6.* The process is iterated until the same results are continuously produced in this example.

4.2 Results and discussions

In this paper, we compare ACCA and the well-known K-Means Algorithm by experimenting with data sets that contain different numbers of clusters. The experimental results of each phase in terms of the number of cluster and data are illustrated in Figure 4. The results are evaluated by the sum of intra-cluster variances. The parameters setting in ACCA are as listed in the following: $m = 1,000$, $q_0 = 0.7$, $\beta = 2$, $\alpha = 0.5$, $\rho = 0.5$, and $\gamma = 0.7$. It is evident that the proposed algorithm's variance is smaller than that of the K-Means Algorithm. That is, ACCA can collect

Table I.
Sample record of electronic library system

ID	CNo	ID	CNo
08	5	10	1
14	3.4	16	4.6
20	1	20	2.5
30	1.8	30	4
37	3	37	4.2
45	2	50	4
52	0.8	55	4.8
60	3	70	2

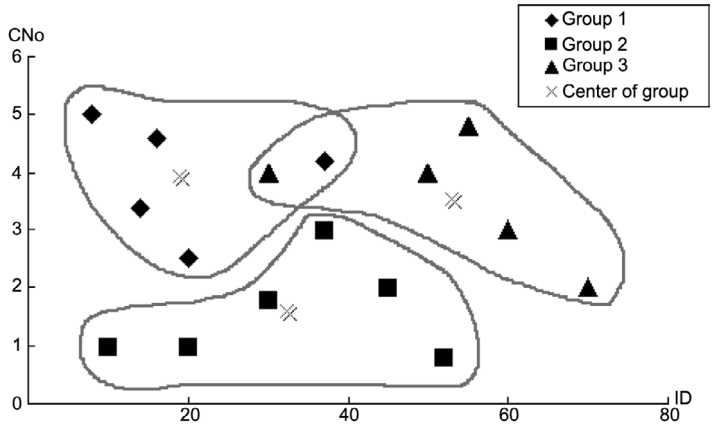

Figure 3.
The visited nodes of ants are grouped

Group name	Group 1	Group 2	Group 3
Nodes	(20, 2.5) (14, 3.4) (37, 4.2) (16, 4.6) (08, 5)	(52, 0.8) (10, 1) (20, 1) (45, 2) (37, 3) (30, 1.8)	(70, 2) (60, 3) (50, 4) (30, 4) (55, 4.8)
$O_{center}(M)$	(19, 3.9)	(32, 1.6)	(53, 3.6)
CV_{intra}	0.24	0.5	0.68

Table II.
The original groups of nodes

Group name	New group 1	New group 2	New group 3
Nodes	(20, 2.5) (14, 3.4) (30, 4) (16, 4.6) (08, 5)	(52, 0.8) (10, 1) (20, 1) (45, 2) (37, 3) (30, 1.8)	(70, 2) (60, 3) (50, 4) (37, 4.2) (55, 4.8)
$O_{center}(M)$’	(18, 3.9)	(32, 1.6)	(54, 3.6)
CV_{intra}’	0.16	0.5	0.47

Table III.
The regrouped groups of nodes

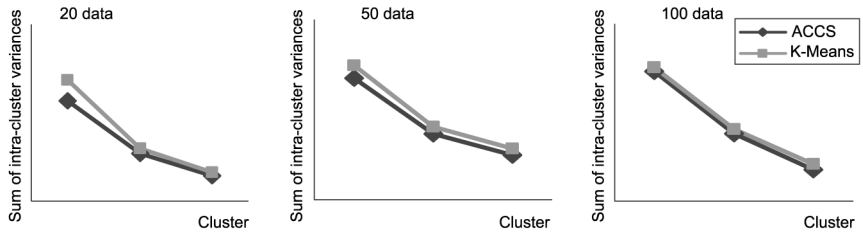

Figure 4.
The diagram of different number of data

more similar data than the K-Means Algorithm can. Furthermore, using ACCA, a campus library system can recommend books for readers to borrow.

5. Conclusion

The powerful development of information technology makes the function of a recommending mechanism more important than before. By taking users' needs as the priority, valuable and proper information should be actively served.

Since there has been a growing interest in the application of Ant Colony Optimization to difficult combinatorial problems, this paper has proposed a technique to discover clusters by using ant colony methods. The goal of this research is to develop accurate trends and prediction models to analyze a wealth of electronic library records and enable user to find a suitable match to what they are looking for more efficiently. The experimental results show the importance of using optimization method for mining. Compared with the K-Means Algorithm, ACCA Algorithm is more effective. It is an efficient approach for applying the clustering technique in a electronic library service.

References

Anand, S.S. and Buchner, A.G. (1998), *Decision Support Using Data Mining*, Prentice-Hall, Englewood Cliffs, NJ.

Caro, G.D. and Dorigo, M. (1998), "Antnet: distributed stigmergetic control for communications networks", *Journal of Artificial Intelligence Research*, Vol. 9, pp. 317-65.

Dorigo, M. and Gambardella, L.M. (1997a), "Ant colonies for the traveling salesman problem", *BioSystems*, Vol. 43, pp. 73-81.

Dorigo, M. and Gambardella, L.M. (1997b), "Ant colony system: a cooperative learning approach to the traveling salesman problem", *IEEE Transactions on Evolutionary Computation*, Vol. 1 No. 1, pp. 53-66.

Fayyad, U. (1996), "Data mining and knowledge discovery: making sense out of data", *IEEE Expert*, Vol. 11, pp. 20-5.

Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. (1996), "From data mining to knowledge discovery in database", *AI Magazine*, Vol. 2, No. 17, pp. 37-54.

Han, J. and Kamber, M. (2001), *Data Mining: Concepts and Techniques*, Morgan Kaufmann, San Mateo, CA.

Leguizamon, G. and Michalewicz, Z. (1999), "A new version of ant system for subset problems", *Proceedings of the Congress on Evolutionary Computation*, IEEE Press, Piscataaway, NJ, pp. 1459-64.

Mitra, S., Pal, S.K. and Mitra, P. (2002), "Data mining in soft computing framework: a survey", *IEEE Transactions on Neural Networks*, Vol. 13 No. 1, pp. 3-14.

Sheikholeslami, G., Chatterjee, S. and Zhang, A. (1998), "WaveCluster: a multi-resolution clustering approach for very large spatial databases", *Proceedings of the 24rd VLDB Conference, New York, NY, August 24-27*, pp. 428-439.

Further reading

Ausiello, G., Crescenzi, P., Gambosi, G., Kann, V., Marchetti-Spaccamela, A. and Protasi, M. (1999), *Complexity and Approximation: Combinatorial Optimization Problems and their Approximability Properties*, Springer, Berlin.

Dorigo, M., Maniezzo, V. and Colorni, A. (1996), "The ant system: optimization by a colony of cooperating agents", *IEEE Transactions on Systems, Man, and Cybernetics - Part B*, Vol. 26 No. 1, pp. 29-42.

Fayyad, U. and Uthurusamy, R. (1996), "Data mining and knowledge discovery in databases", *Communications of the ACM*, Vol. 39 No. 11, pp. 24-6.

About the authors

An-Pin Chen received a PhD degree in Industrial and Systems Engineering from the University of Southern California in the research areas: artificial intelligence, financial investment analysis and policy decision. He is now an Associate Professor in the Institute of Information Management at National Chiao-Tung University in Taiwan. An-Pin Chen is the corresponding author and can be contacted at: apc@iim.nctu.edu.tw

Chia-Chen Chen is a graduate student with the PhD program at the Institute of Information Management of National Chiao-Tung University in Taiwan.