Customer Behavior Pattern Discovery by Adaptive Clustering Based on Swarm Intelligence

Weihui Dai*

Abstract

Customer behavior pattern discovery is the fundament for conducting customer oriented services and the services management. But, the composition, need, interest and experience of customers may be continuously changing, thereof lead to the difficulty in refining a stable description of their consistent behavior pattern.

This paper presented a new method for the behavior pattern discovery from a changing collection of customers. It was originally inspired from the swarm intelligence of ant colony. By the adaptive clustering, some typical behavior patterns which reflect the characteristics of related customer clusters can extracted dynamically and adaptively.

Keywords : Service Management, Customer Analysis, Behavior Pattern, Adaptive Clustering, Swarm Intelligence

Received : 2009. 10. 30.   Final Acceptance : 2010. 01. 28.

※ This research was supported by the National High-tech R and D Program (863 Program) of China (No.2008AA04Z127) and Shanghai Leading Academic Discipline Project (No.B210).

* School of Management, Fudan University, e-mail : whdai@fudan.edu.cn

1. Introduction

Customer behavior pattern discovery is the fundament for conducting customer oriented services and the services management. Clustering technology is widely applied in data mining, knowledge discovery, as well as the customer behavior pattern discovery, which groups data sets into classes by similar features for each class [Wu and Wang, 2004; Ng and Huang 1999]. Traditional cluster analysis is based on some typical clustering algorithms, such as LF algorithm, K-Means algorithm and Information Entropy algorithm. It's difficult in avoiding the disadvantages of long time convergence, easy stagnation and local optimization [Liu, 2008].

Usually, customer behavior patterns and their interest are supposed to be classed into some predictive types by an empirical survey or existed experience. However, the composition, need, experience and interest of customers may be continuously changing, thereof lead to the difficulty in refining a stable description of their consistent behavior pattern [Mu, 2009]. So, the dynamically adaptive technology had to be explored for the improvement of current cluster analysis in customer oriented services.

Fortunately, swarm intelligence provides a very helpful means to solve those problems. Swarm intelligence comes of the scientists' research and their observation on the sociality of insects, ants, and fishes. The so-called swarm intelligence is that a great many of simple, unintelligent units unite into a group and express intellectual behaviors through mutual cooperating with each other [Bonabeau, et al., 1999]. Swarm intelligence exhibits a number of interesting properties such as flexibility, robustness, decentralization and self-organization [Kennedy, 2001]. It is widely used in nonlinear optimization, knowledge discovery, communication networks, data-mining, etc [Jain, et al., 1999;-Tian, et al., 2005].

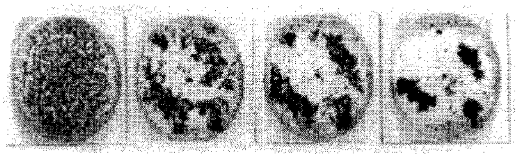
In the recent years, some scholars studied clustering problem according to the idea of swarm intelligence [Wu and Wang, 2004; Ng and Huang 1999; Handl, et al., 2005]. Inspired from the swarm intelligence of clustering behavior by ant colony, this paper presented a new method for customer behavior pattern discovery. It can handle large category dataset more rapidly, accurately and effectively, and keep the good scalability at the same time. Besides, it doesn't require to set the number of destined clusters in advance. By the cluster analysis of online shoppers, this method exhibited excellent ability in adaptive and dynamic behavior pattern discovery.

2. Clustering Algorithm based on Swarm Intelligence

2.1 Clustering Behavior by Ant Colony

<Figure 1> is a famous observation on the clustering behavior by ant colony [Bonabeau, et al., 1999].

The first picture in the left was 1500 dead ant bodies dispersed randomly on a round area of 25cm in diameter. Those bodies had been carried and accumulated into some heaps by the ant colony alive. The following pictures



〈Figure 1〉 Observation on the clustering behavior by ant colony (1500 dead ant bodies and their clusters in 4 hours, 6 hours, 26 hours)

from left to right were their clusters in 4 hours, 6 hours, and 26 hours respectively.

After analysis of the clustering process by ant colony, we may conclude the following three decision-making issues for each ant :

1) Should this dead body be picked up? This decision will be made by a probability function to analyze the similitude degree with existed clusters.
2) If picked up, where to put down in a near-by area with an observing radius? This problem will be solved by an optimization function to maximize the average similitude degree while giving certain clusters.
3) How to decide the observing radius? This issue will be explored by introducing a dynamic adjustment process.

Considering those problems, we established a mathematical model and developed a new clustering algorithm based on swarm intelligence [Dai, et al., 2009].

2.1 Mathematical Model

(1) Definitions :

a) Attribute Probability : Suppose data collect D contains n objects, A_1, A_2, $\cdots$, A_m are the m attributes in D. If attribute A_j of the objects $X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$is x_{ij}, x_{ij} appears q_j times in D, thus$q_j = sum,$ $\{x_{kj} = x_{ij}|X_k \in D\}$ then the attribute probability p_{ij} of object X_i on attribute A_j is :

$$p_{ij} = \frac{q_j}{n} \tag{1}$$

b) Similitude Degree : Suppose D contains n objects, the similitude degree of X_i indicates the algorithm average value of all attribute probability of each attribute, we define $f(X_i)$as the similitude degree of X_i :

$$f(X_i) = \frac{1}{m}\sum_{j=1}^{m} p_{ij} \tag{2}$$

c) Probability Conversion Function: It is the function that switches the similitude degree to the cluster object probability of the unit. The independent variable is the similitude degree, and the function field is [0, 1]. The larger the similitude degree is, the smaller the pick-up conversion probability is, while the put-down conversion probability follows the contrary rule. In this algorithm, we define probability conversion function as follows:

$$p_p = 1 - \frac{1 - e^{-cf(X_i)}}{1 + e^{-cf(X_i)}} \tag{3}$$

$$p_d = 1 - \frac{1 - e^{-cf(X_i)}}{1 + e^{-cf(X_i)}} \tag{4}$$

Here, p_p is the probability pick-up function, p_d is the probability put-down function. p_d is an up-convex function, and the convergence speed varies from different c. The larger c is, the nearer the function value is to 1, that is, the corresponding put-down function value is large. So we can expedite the convergence speed by augmenting c. In this paper, let c = 5.

(2) The characteristic of the optimized solution

The algorithm uses probability function as the basis of pick-up or put-down. Thus, the algorithm aims to increase the similitude degree of the whole system, maximizing the system's average similitude degree. We define the objective function in this algorithm as follows :

$$F(D)=\frac{1}{n}\sum_{l=1}^{k}\sum_{i=1}^{n_l}f(X_i) \tag{5}$$

$$0 \le f(X_i) \le 1$$

Here, $\sum_{i=1}^{k} n_l = n$

$f(X_i)$ is the similitude degree of X_i in cluster k.

This is a complicated optimization problem. It aims to find out clustering fashion to maximize the average similitude degree when giving certain clusters.

According to (1) and (2), (5) can be rewrite as follows:

$$F(D)=\frac{1}{n}\sum_{l=1}^{k}\sum_{j=1}^{n_l}\frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{A_i}\frac{q_{it}}{n_l}$$
$$=\frac{1}{nm}\sum_{l=1}^{k}\frac{1}{n_l}\sum_{i=1}^{m}\sum_{t=1}^{A_i}q_{it}^2 \tag{6}$$

In (6), q_{it} is the times of A_t which is the i[th] attribute in cluster l. Thus, $\sum_{t=1}^{A_i} q_{it} = n_l$. That is, the sum of the times of the attribute's value is total objects in the cluster. So we get :

$$F(D_l)=\frac{1}{mn_l}\sum_{i=1}^{m}\sum_{t=1}^{A_i}q_{it}^2 \tag{7}$$

If the attributes are independent, then we can calculate them independently :

$$F(D_{l-A_i})=\frac{1}{n_l}\sum_{t=1}^{A_i}q_t^2 \tag{8}$$

As discussed previously $\sum_{t=1}^{A_i} q = n_l$, we can get :

$$F(D_{l-A_i})=\frac{1}{n_l}\sum_{t=1}^{A_i}q_t^2=n_l\sum_{t=1}^{A_i}\frac{q_t^2}{n_l^2}\le n_l\left(\sum_{t=1}^{A_i}\frac{q_t}{n_l}\right)^2\le n_l \tag{9}$$

Only when $q_s = n_l,\ q_r = 0 (r \ne s)$, the equal mark comes into existence. That is, only when all the objects in the cluster get the same value, the similitude degree gets the maximum. Furthermore, the more concentrative the value of each attribute is (the attribute of the objects is similar), the larger the average similitude degree is.

In (5), the more similar the objects in the cluster are, the larger the average similitude degree is. Contrarily, the larger the similitude degree of the sub-cluster data collect is, the larger the whole similitude degree is.

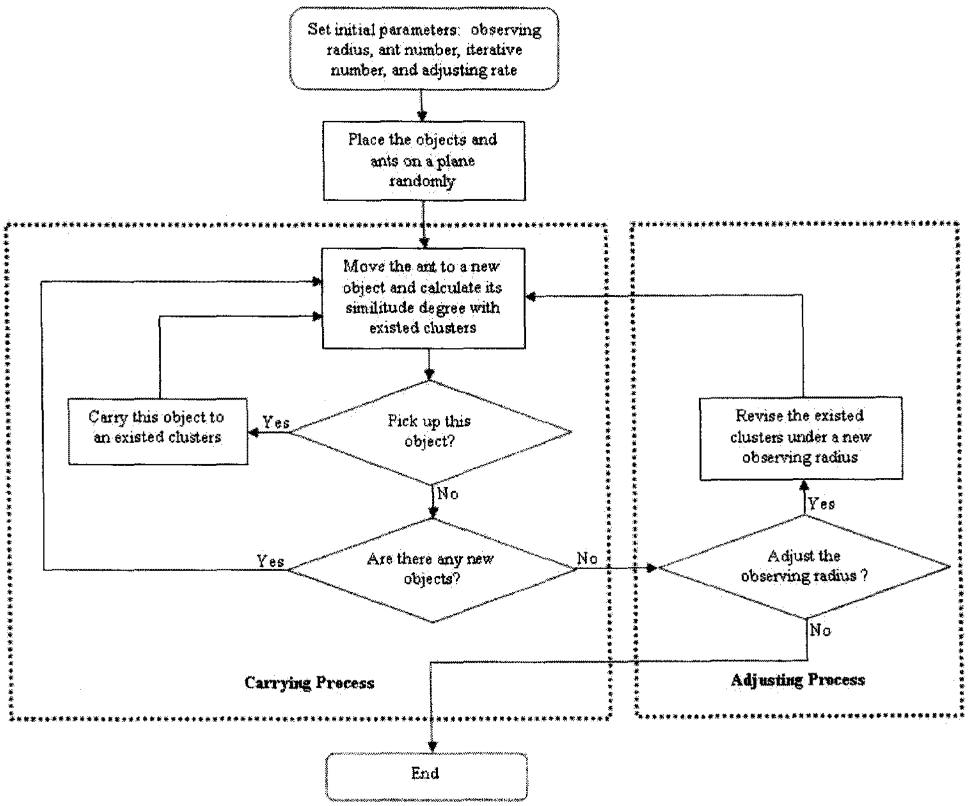
So, the characteristic of the optimized solution is that all the attributes are most concentrative. As each value of the objects' attribute is random, the optimized solution may be more than one.

(3) The Algorithm Process

<Figure 2> is the flowchart of our new algorithm. The main process of this algorithm is the ant carrying process. Ant decides whether pick-up the current object by object's probability conversion function. Similarly, when ant carries the object to the destination, it also considers the similitude degree between the current object and the surrounding objects to decide whether to put-down or not. In this process, the ant doesn't know the other ants' location distributing and load status, neither the other objects' distributing status outside its observing scope. So we can say the ant carrying process is an easy and absolute individual behavior. Yet, it is this easy individual behavior makes the objects divided into various clusters during long-time and concurrent process.

The pseudocode of above algorithm may be designed as followings :

For every item O_i, do
Place O_i, randomly on grid
End For
For all ants do
Place ant at randomly selected site
End For



〈Figure 2〉 Flowchart of the new algorithm

```
/* main loop */
For t=1 to t_max do
For all ants do
If ((ant unladed) and (site S_j occupied by item O_i))then
Compute the similarity f(O_i) of O_i in R×R area
Calculate the p_p and generate a random number Q
If p_p > Q then /*pick-up rule*/
    Pick up item O_i
    Remember the f(O_i) and current position
    Move the ant with the item O_i to a random site
Else
    Move the empty ant to a random site
End If
Else If ((agent carrying item O_i) and (site empty)) then
Compute the similarity f(O_i) of O_i in this place
Calculate the p_d and generate a random number Q
If p_d> Q then /* put-down rule*/
    Drop item
    End If
Move to a randomly selected neighboring site
End If
End For
If ((t > 0.5t_max) and (t meet the radius change condition)) then
Reduce the radius
Generate the clusters iteratively and calculate the cluster center
Unite the clusters with the same cluster center
Relocate the items with poor similarity
End If
End For
Print location of items /* export cluster result*/
```

Compared with traditional ant cluster algorithms, this algorithm has been improved in the followings :

Adopt an adjusting process which improves the efficiency of the algorithm, and avoids the local optimality and stagnancy as well.

Introduce the dynamic adjusting of observing radius and a new similitude degree formula.

Endow the ant with a short-term memory to reduce its repeated behaviors.

Compared with the K-Modes algorithm, Information Entropy-Based Cluster Algorithm (ECA) [Cheng, et al., 1999; Zhao, et al., 2005], and LF Algorithm, experiment results of data collect from UCI machine learning data-base [Machine Learning Repository, 2009] showed that this algorithm are excellent in both accuracy and efficiency [Liu, 2008; Dai, et al., 2009]. Besides, this algorithm doesn't require to set the number of destined clusters in advance, so it is very suitable for clustering analysis automatically and adaptively.

3. behavior pattern Discovery of online shoppers

3.1 Analysis Method

The cluster analysis of online shoppers by their behavior patterns can be divided into three steps: data preprocessing, pattern discovery, and pattern analysis. In the data preprocessing, history data about the user's IP address, ID, requested URL and access time, etc. are collected from the Web log data and thereof are purified to recognize each user and its sessions by a series of commonly used technologies [Mu, 2009]. Pattern discovery aims to discover the habit-forming behavior and its characteristics by analyzing the user's sessions, path records and other related information. But in most cases, it's difficult to refine a set of stable and common rules that suitable for a certain user cluster, because the user's requirement, interest as well as its attentive information may be dynamically changed and dispersed even in the same initial cluster. In this paper, we establish a set of "components" to describe the elements of user's behavior and its characteristics, and discover the behavior pattern by a dynamically clustering analysis of above "components." By a statistical analysis and the observation from online shoppers in China, we summarized three categories of components as from <Table 1> to <Table 3> The design of those tables had considered some special characteristics of the online shoppers in China.

Based on those tables, online shopper's behavior pattern can be obtained by a clustering analysis with three categories of components as the structured formula :

$$Patter(N): duster(m) = ((u_{ij}, \delta_{ij}\%) \cap (b_{mn}, \delta_{mn}\%) \cap (w_{st}, \delta_{st}\%) \quad (10)$$

Here,

N is the sequence number of discovered patterns ranked by the cluster size, m is the corresponding cluster number, U_{ij}, B_{μ} and C_{st} are tagged components with respective percent of $\delta_{ij}\%$, $\delta_{mn}\%$, and $\delta_{st}\%$ in this cluster. In those formula, the percent of $\delta_{ij}\%$, $\delta_{mn}\%$, and $\delta_{st}\%$ even their tagged components may be dynamically changed while the shopper's requirement, experience and interest have been shifted.

3.2 Test and Results

Test data are obtained from the basic information of 156 registered users and their browsing records in the Web log data of an online shop which contains all the components listed on <Table 3> We collect the data of 15769 browsing actions by those users during a period of continuous 30 days. To ensure the accuracy of user's information, all the data about their basic information have been updated by themselves just before the beginning of test, and the behavior components of "Attentive information about the goods" (B21~B26) and "Reason for further browsing on the goods" (B41~B47) are requested to be choose when they want to browse the details of a certain goods.

Test data are purified and scaled as a collect data set, in which each related component from <Table 1> to <Table 3> have been assigned a value of 0 or 1 to satisfy the requirement of optimized ant colony cluster algorithm. It is obviously that the user sorts are unknown before clustering analysis, and the number of user group sorts is uncertain. Thus we cannot use the algorithm that fix the cluster number like k-means algorithm and ECA, and that is also the advantage of our new algorithm.

<Figure 3> is the clustered result projected onto two-dimension plane. We see that our algorithm can divide the data collect perfectly, and all of the data are assembled to 16 clusters with the size from 4 to 23. Table 4 shows the detailed information. Adjacent field X and adjacent field Y indicate the distributing section of the object on horizontal coordinate and vertical coordinate. Clustering can differentiate the shoppers who have different browsing behavior patterns. Of the 16 clusters from clustering, we choose Cluster 5 as the example to analyze their browsing behaviors. Cluster5 contains 23 users. It has the largest size in all clusters. <Figure 4> shows the characteristic distribution of some components in this cluster.

In Cluster 5, most of shoppers have the experience of online purchase (U73) with monthly payment from 301 to 1000 RMB (U62), and their expected price of the goods is 151~250 RMB (B33). In their browsing behavior, they pay more attention to the price and introduction about interested goods (B24, B22), and usually

〈Table 1〉 Behavior Component Table

Category	Name	Tag	Description
Behavior Components	Browsing Habit	B11	browsing the information on homepage
		B12	looking for goods by the search engine
		B13	entering into a specific goods category
		B14	entering into a specific on line shop
		B15	browsing the goods in recommendation list
		B16	others
	Atytentive Information About the Goods	B21	name
		B22	introduction
		B23	advertisement
		B24	price
		B25	brand
		B26	others
	Expecter Price of the Goods	B31	<= RMB
		B32	51~150RMB
		B33	151~250RMB
		B34	251~400RMB
		B35	>400RMB
	Reason For Further Browsing on the Goods	B41	desining to qurchase
		B42	interesting advertisement
		B43	popular ranking
		B44	cheap price
		B45	comment and reputation
		B46	random choice
		B47	others

〈Table 2〉 User Component Table

Category	Name	Tag	Description
User Components	Age	U11	< + 20years
		U12	21~30years
		U13	31~40years
		U14	41~50years
		U15	51~60years
		U16	> 60years
	Gender	U21	male
		U22	famale
	Occupation	U31	manger
		U32	professional
		U33	workman
		U34	teacher
		U35	student
		U36	others
	Average Income Per Month	U41	< = 1000RMB
		U42	1001~3000RMB
		U43	3001~6000RMB
		U44	6001~10000RMB
		U45	10001~15000RMB
		U46	15001~30000RMB
		U47	>30000RMB
	Average Online Time Per Week	U51	< = 10hours
		U52	11~20hours
		U53	21~30hours
		U54	61~50hours
		U55	> 50hours
	Average Online Payment Per Month	U61	< = 300RMB
		U62	301~1000RMB
		U63	1001~2000RMB
		U64	2001~3000RMB
		U65	3001~4000RMB
		U66	4001~5000RMB
		U67	5001~60000RMB
		U68	> 6001RMB
	browsing experience on this website	U71	never browsing
		U72	browsing without purchasing
		U73	browsing and purchasing

enjoy looking for goods by the search engine (B12), while dress and home accessories (W21), book and audio-visual products (W23) and cosmetic and jewelry (W22) are poplar goods for them.

By the structured formula in (10), we can ex-

〈Table 3〉 Web Site Component Table

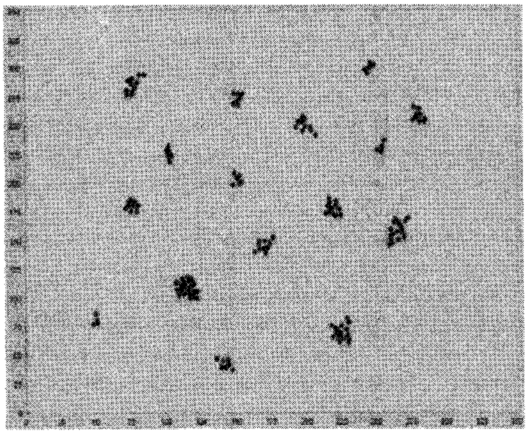
Category	Name	Tag	Description
Web Site Components	ontent item	W11	homepage
		W12	search engine
		W13	goods category
		W14	on-line shop list
		W15	recommendation goods list
		W16	others
	goods category	W21	dress and home accessories
		W22	cosmetic and jewelry
		W23	book and audio0visual products
		W24	ticket and phone card
		W25	electronic products
		W26	matemal and child products
		W27	food and health products
		W28	others

〈Table 4〉 Clustered Result Lists

Cluster	Size	Adjacent field on X-axis	Adjacent field on Y-axis
cluster 1	11	(69, 83)	(274, 296)
cluster 2	6	(98. 105)	(219, 233)
cluster 3	7	(148, 155)	(267, 279)
cluster 4	10	(193, 208)	(242, 257)
cluster 5	23	(105, 123)	(98, 121)
cluster 6	4	(48, 52)	(75, 86)
cluster 7	5	(242, 249)	(295, 306)
cluster 8	6	(147, 156)	(198, 211)
cluster 9	5	(249, 255)	(226, 240)
cluster 10	11	(213, 226)	(172, 189)
cluster 11	9	(67, 81)	(174, 189)
cluster 12	10	(163, 176)	(137, 152)
cluster 13	16	(217, 233)	(58, 72)
cluster 14	9	(275, 285)	(252, 269)
cluster 15	7	(133, 148)	(37, 49)
cluster 16	17	(257, 274)	(147, 173)

press the browsing behavior pattern as following :

$$\begin{aligned}&Patter(1): cluster(5) = \\ &((U_{54}, 65.39\%), (U_{62}, 62.75\%), (U_{73}, 78.89)) \cap \\ &((B_{12}, 42.64\%), (B_{24}, 35.97\%), (B_{22}, 24.35\%), \\ &(B_{33}, 54.50\%), (B_{41}, 64.15\%)) \cap ((W_{21}, 29.27\%), \\ &(W_{23}, 18.29\%), (W_{22}, 15.85\%))\end{aligned} \quad (11)$$



〈Figure 3〉 Clustered Result

In (11), only those tagged components whose percent are more than the average are listed. Similarly, the second pattern and the third pattern can be expressed with Cluster 16 and Cluster 13. Through dynamic analysis of those patterns and their changes, we can compartmentalize different shopper's types by their characteristics; thereof build adaptive websites and advertising strategies to match their behaviors. Further experiments have shown that our

〈Figure 4〉 Characteristic Distribution of Components in Cluster 5

method is very suitable to be applied in self-adaptive pattern discovery in the smart targeting system for online advertising [Dai, et al., 2009].

3.3 Conclusion and Expectation

In order to overcome the disadvantages of current algorithm in customer behavior pattern discovery, this paper presented a new method based on swarm intelligence. It has been successfully applied in dynamic clustering analysis of online shoppers' behaviors, and showed prominent superiority to traditional method.

However, our method mainly aims at classified data, but we should disperse the value data in the dealing process. So, further research on this method may be expected to be improved in the future.

References

[1] Wu, Q. D. and Wang, L., Intelligent Ant Colony Algorithm and Application. Shanghai : Technology Press, 2004.

[2] Wu, Q. D. and Wang, L., Intelligent Particle Swarm Optimization Algorithm Research and Application. NanJing : Jiangsu Education Press, 2005.

[3] Ng, M. K. and Huang, Z., "Data-mining massive time series astronomical data : challenges, problems and solutions", *Information and Software Technology*, Vol. 41, No. 9, 1999, pp. 545-556.

[4] Liu, S. J., Ant clustering algorithm based on swarm intelligence and its application. Shanghai : Fudan University, 2008.

[5] Mu, T., Research and implementation to clustering technology of browsing behavior based on swarm intelligence, Shanghai : Fudan University, 2009.

[6] Bonabeau, E., Dorigo, M., and Theraulaz, G., Swarm Intelligence : From Natural to Artificial Systems. New York : Oxford University Press, 1999.

[7] Kennedy, J., Eberhart, R. C., Swarm Intelligence, San Francisco : Morgan Kaufmann, 2001.

[8] Jain, A. K., Murty, M. N., and Flynn, P. J., "Data clustering : a review", *ACM Computing Surveys*, Vol. 31, No. 3, 1999, pp. 264-323.

[9] Ma, H., Lu, Y. Q., and Yin, H. B., "Customer segmentation based on the share of the customer", *Journal of Wuhan University of Technology Information and Management Engineering*, Vol. 25, No. 3, 2003, pp. 184-187,

[10] Tian, Y. M. Meng, L. N., and Mei, L. X., "Region-based image retrieval system using integrative feature description", *Computer Engineering*, Vol. 31, No. 19, 2005, pp. 167-169.

[11] Handl, J. L., Knowles, J., and Kell, D. B., "Computation cluster validation in post- genomic data analysis", *Bioinformatics*, Vol. 21, No. 15, 2005, pp. 3201-3212.

[12] Dai, W. H., Liu, S. J., and Liang, S. Y., "An Improved Ant Colony Optimization Cluster Algorithm Based on Swarm Intelligence", *Journal of Software*, Vol.4, No. 4, 2009, pp. 299-306.

[13] Cheng, C. H., Fu, A. W., and Zhang, Y., "Entropy-based subspace clustering for mining numerical data", *Proceedings of The Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1999. pp. 84-93.

[14] Zhao, W. L., Sun, Y. X., and Zhang, Z. G., "An improved ant colony clustering algorithm based on information entropy", *Journal of Shenyang Institute of Chemical Technology*,

Vol. 29, No. 4, 2005, pp. 296-300.
[15] Machine Learning Repository, http://archive.ics.uci.edu/ml/, Feb. 6, 2009.
[16] Dai, W. H., Dai, X. Y., and Sun, T., "A smart targeting system for online advertising", *Journal of Computers*, Vol. 4, No. 8, 2009, pp. 778-786.

Author Profile

Weihui Dai

received his B.S. degree in Automation Engineering in 1987, his Master degree in Automobile Electronics in 1992, and his Ph.D. in Biomedical Engineering in1996, all from Zhejiang University, China. He worked as a post doctor at School of Management, Fudan University from 1997 to 1999, a visiting scholar at Sloan School of Management, M.I.T from 2000 to 2001, and a visiting professor at Chonnam National University, Korea from 2001 to 2002. He is currently an Associate Professor at the Department of Information Management and Information Systems, School of Management, Fudan University, China.

Dr. Dai has published more than 100 papers in Software Engineering, Information Management and Information Systems, Complex Adaptive System and Socioeconomic Ecology, Digital Arts and Creative Industry, etc. Dr. Dai became a member of IEEE in 2003, a senior member of China Computer Society in 2004, and a senior member of China Society of Technology Economics in 2004.

◈ 이 논문은 2009년 10월 30일 접수하여 2010년 1월 28일 게재확정되었습니다.