Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2014, 6(7):2271-2278

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

A case retrieval algorithm based on correlation analysis

Jiang Bing, Jiankang Liu and Xiaoqiang Zeng

Business School, Sichuan Agricultural University, Dujiangyan, China

ABSTRACT

The key of case-based reasoning process (CBR) is the case retrieval. With the increase of the number of cases, the efficiency of the case retrieval decreases. In order to ensure efficiency and stability of the CBR system, related clustering algorithms are introduced to make effective classification and improve the efficiency of the case retrieval. Results of many clustering algorithms are affected by selection of initial values. For example, different results of classification may be produced with the same case library. Therefore, it is probable that the most similar cases cannot be retrieved and the optimal number of case categories cannot be determined in common clustering algorithms. A case retrieval algorithm based on correlation analysis is proposed according to the process of case-based reasoning. In the algorithm, the gray correlation analysis is adopted to classify cases stored in the case library of the CBR system, and the method oft-distribution in mathematical statistics is adopted for determining the optimal number of case categories. Then, corresponding algorithms for case classification and retrievals are designed. Finally, comparison experiments are made to verify the stability and effectiveness of the algorithm. Theoretical analysis and experimental results show that with the number of cases in the case library, the algorithm has better stability and efficiency compared with classic algorithms. The model of the algorithm has some practical value.

Keywords: Case-based reasoning; case classification; case retrieval; correlation analysis; similarity

INTRODUCTION

Case-based reasoning is a branch of artificial intelligence, it is a way that deal with the problem of reality based on past actual experience. The case is a description of the problem in the application field, which composed with two parts.It is the state and corresponding solution of the problem, actually, it is a mapping that from a state space to the solution space[1]. Current problems or conditions faced by us are called target cases, and the problem or situation of memory is calledthe source case,case-based reasoning obtain the memory of the source case by prompting of the target case, the target solving cases by the strategy for source case to guide, which overcomes the shortcomings that traditional Rule-Based Reasoning (RBR) system is difficult to obtain the knowledge and reasoning. Case-based reasoning has been widely used in industrial manufacturing, lawsuits, medicaldiagnostics, and Q&A application systems and other fields. And it hasachieved good results [2],case-based reasoning contains several typical process, namely: case retrieval, case correction, case reuse and case preservation [3]. Over the whole case retrieval is a key link in the process of case-based reasoning , the case retrieval efficiency will continue to decline with the increase of the number cases,the phenomenon is also known as swamp phenomenon. Then,how can organize the case base effectively, when increase the cases, it not reduce retrieval efficiency in case retrieval period, which is important indicator of CBR systems for evaluate the effectiveness. HuanTong Geng(2005) proposed a clustering algorithm that is applied to the case base, Feng Zheng proposeda clustering algorithm based on K-Means that is applied to maintenance for the case base, the efficiency of retrieval is improved greatly[4].Changzheng Liu,etc(2010) proposed a feature weightingC-means clustering algorithm (WF-C-means) and the index is created by clustering scheme in the process of case retrieval, since C-means clustering algorithm to adjust the weights of all attributes are included in the definition of the difference , so the retrieval and new cases have taken similar cases the differences become very objective and precise, QIAO Li (2011)put forward an improved K-Means clustering algorithm, it is a good solution to the noise caused by poly type of error, clustering algorithms have been introduced into the case retrieval,which greatly improves the efficiency of case retrieval[5][6][7]. However, The most of the clustering algorithm are globular cluster clustering that effect on chosen of random initial value,the inconsistent selectionfor clustering initial value will result in that there is a big difference on clustering, so that the accuracy of case retrieval are subject to effect greatly. Clustering algorithm used without an optimal number of categories, with the increased number of cases, classification become big, it will affect the efficiency of the algorithm; classification is too small, since each class contain too many cases, it will also affect the retrieval efficiency. In order to make more rational and correct classification, a new classification algorithms isproposed,the mathematicalcorrelation methods areintroduced for algorithm to classification and retrieval for case, finally, the application example is designed to verify the effectiveness of the algorithm. Experimental results show that the new algorithm outperforms the traditional retrieval algorithm,which has been improved greatly on retrieval efficiency and speed.

CASE-BASED REASONING PRINCIPLES AND TRADITIONS CASE RETRIEVAL ALGORITHM

Case-based reasoning principle is shown in Fig.1:

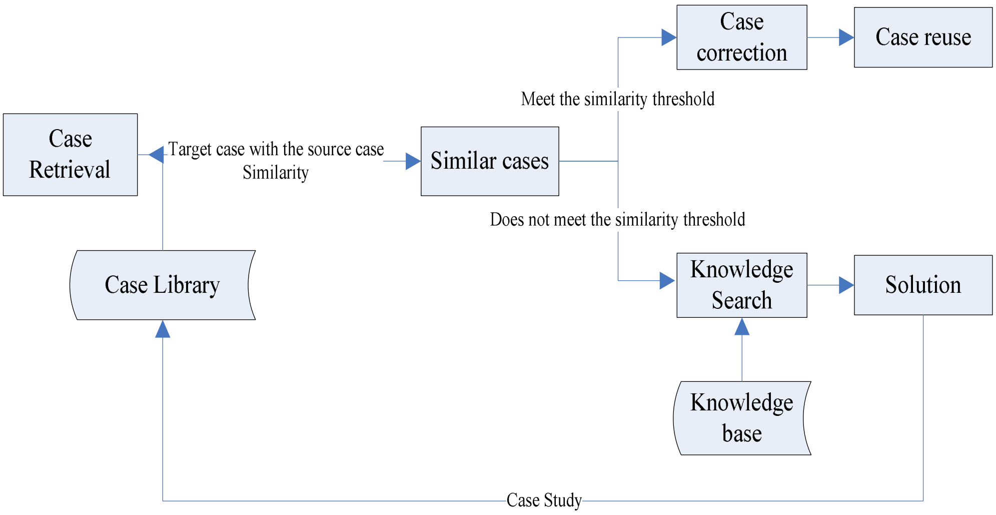

Fig.1. The working principle of case-based reasoning

Fig.1 shows over the whole working process of case-based reasoning, case retrieval is a critical step in this process, when there are fewer cases in case base, according to the similarity threshold,the success rate is lower relatively for retrieval of similar cases with the passage of running time for the system, through case studies, the case retrieval success rate is increased gradually.Traditional comparative case similarity algorithm is as follows:

1) Tversky contrast matching function

$$T_{nk} = \frac{(A^n \cap A^k)}{(A^n \cup A^k) - (A^n \cap A^k)} \tag{1}$$

Equation (1) is a probability model based on metrics.A^n and A^k is an example of n, k attributes Works, T_{nk} represents the similarity of sample between n and k. This similarity law is appliedto the application that property is binary.Improved matching method Tversky follow as:

$$S_{nk} = \frac{\sum_{i=1}^{m} w(n,i)w(k,i)V_{nk}^{i}}{\sum_{i=1}^{m} (w(n,i))^2 \sum_{i=1}^{m} (w(k,i))^2} \tag{2}$$

Equation (2) is an improved definition of the similarity matching Tversky. Wherein , w (n, i), w (k, i) denote the weights of i-th attribute come to sample n, k ; V^i_{nk} represent similarity of the i-th attributebetween n and k sample; m represent the number of all attributes about example n and k; S_{nk} representsimilarity of example. Improved Tversky algorithm considers the different weights for each attribute of two example,so that can use this method to get the similarity for same set of example attributes. For two examples with the different set of attributes , an example that contain two sample sets is created ,the k of example has i-attributes, but n of example has not, then, the w (n, i) is set to 0, the n of example has i-attributes, but k of example has not, then, the w (k, i) is set to 0. So the similarity can be obtained between sample n and k.

2) Nearest neighbor algorithm (k-NN)

$$\mathrm{Sim}(X, Y) = 1 - \mathrm{D\ ist}(X, Y) = 1 - \sqrt{\sum{}_i W_i D^2(X_i, Y_i)} \tag{3}$$

The main case retrieval methods contain Classification net models, templatessearch, nearest neighbor search, inductive retrieval, retrieval based on deep knowledge, neural network search,fuzzy retrieval and rough set retrieval. When carry case retrieval, the main consider the following:

(1) Effectiveness: The case that it is retrieved should has the value;
(2) Accuracy: The case of retrieval should be relate or similar possibly with the current case;
(3) Easy adjustability: Retrieved case should make easy to adjust, it is easy to produce solution of the current problem;
(4) High speed:The retrieval time of Case is shorter and faster. There is an improving method, it based on the nearest neighbor search method, the search method improve the retrieval speed greatly.

THIS CASE RETRIEVAL ALGORITHM

A. Case similarity matrix calculation

First, The cases of case base are classified, In current, there is a hypothesis that case base have n cases , each case is represented as X_i (1 ≤ i ≤ n), X_i = (a_{i1}, a_{i2}, a_{i3} ...... a_{im}), a_{im} for the first i cases the m-th attributes. N cases can be expressed as the following matrix:

$$X_{ij} = \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}, (1 \le i \le n, 1 \le j \le m) \tag{4}$$

N of these cases were pairwise correlation analysis, the i-th and the j cases related cases expressed as r_{ij}, the correlation calculation which case the following steps:

(1) Calculate the case 1 and the remaining n-1 correlation cases, completed in five steps , the first step is to sequence each case attribute a_{ij} (1 ≤ i ≤ n, 1 ≤ j ≤ m) divided by a_{i1} (1 ≤ i ≤ n), as follows:

$$X_{ij}{}' = X_{ij} / a_{i1} = \begin{pmatrix} a_{11}/a_{11} & \cdots & a_{1m}/a_{11} \\ \vdots & \ddots & \vdots \\ a_{n1}/a_{n1} & \cdots & a_{nm}/a_{n1} \end{pmatrix}, (1 \le i \le n) \tag{5}$$

The difference sequence is calculated in the second step as follows:

$$\square_i(d) = \left| x'_1(d) - x'_i(d) \right|, (1 \le d \le m, 1 \le i \le n) \tag{6}$$

The minimum and maximum values are obtained in the third step, calculated as follows:

$$M = \max_i \max_j \square_i(d), m = \min_i \min_j \square_i(d) \tag{7}$$
$$, (1 \le i \le n, 1 \le j \le m, 1 \le d \le m)$$

The correlation coefficient is calculated for the first cases and the rest of casesin the fourth step as follows:

$$r_{1i}(d) = \frac{m + \xi M}{\square_i(d) + \xi M}$$

$$,(0 \angle \xi \angle 1),(1 \le i \le n),(1 \le d \le m) \qquad (8)$$

Equation (8),set theξ= 0.5 , the correlation of the first case and the rest is calculated in the fifth step , the weight of attributes to be considered in the calculation of correlation, So the case attribute weights C_d (1≤d≤m)is setted, equation as follows:

$$r_{1i} = \frac{1}{m}\sum_{d=1}^{m} r_{1i}(d) * c_d ,(1 \le i \le n, 1 \le j \le m) \qquad (9)$$

(2) Calculation the relevance of the cases 2 to n between with the rest of n-1

Repeat the above steps, respectively, calculate the correlation operation for the second case between the rest of the n-1.Until get a n matrix by the correlation operator for the last case with the rest, it reflects the correlation scenario of case, r_{ij}represent correlation for the first i and the j cases, then, n cases resulting correlation matrix as follows:

$$r_{ij} = \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{bmatrix} ,(1 \le i \le n, 1 \le j \le n) \qquad (10)$$

In order to improve the retrieval speed and efficiency, we can classify matrix (10), the proposed classification method described in the next section.

B. Case Classification

Case classification is carried out over a correlation calculated based on correlation matrix r_{ij} , n cases can be divided into a number of categories up to n categories, namely, each case as a class in the case-base ; the minimum of case is one class , that all cases is a category, set the number of cases classified as class C_i, then $1 \le C_i \le n$; case-base for case classification,in certain circumstances, the total number of cases , the optimal number of categories can be found in a comparison of the similarity threshold h, cases can be classified as a case base comparison reference values. According to statistical t-test, we get the following formula:

$$t_{0.05}(df) = \frac{h}{S_h} \qquad (11)$$

In Equation (11), the value 0.05 of $t_{0.05}$(df) means that the probability of the t-test , where df represents the attribute diversity , h represents correlation threshold , S_h is the standard error of the correlation coefficient , the formula is as follows:

$$S_h = \sqrt{(1 - h^2)\frac{1}{(m - 2)}} \qquad (12)$$

Equation (12) where m is the number of attributes in the cases, the correlation threshold h can be obtained by the formula (11) and (12) as follows:

$$h = t_{0.05}(df) * \sqrt{\frac{1}{m + t_{0.05}{}^{2}(df) - 2}} \qquad (13)$$

Equation (13), In accordance with t-test, df=m-2, The m is the number of attributes in the cases. With the correlation threshold , the correlation matrix based on r_{ij}, taking the maximum value from equation (10), R_{max}=max (r_{ij}), Judging $R_{max}\geq h$ is established , if success , then the i-th case with the j-th case is classified as a class in case-base , the class will be classified as a new case ,and use X_{ij} ($1\leq i\leq n$) represent, where n represents the total number of cases , and calculating the scenario averaging for the corresponding attribute of the i-th and j-th case .The attribute of X_{ij} is obtained to expressas follows :

$$X_{ij}=(\frac{1}{2}(a_{i1}+a_{j1}),\frac{1}{2}(a_{i2}+a_{j2}),\cdots,\frac{1}{2}(a_{im}+a_{jm})) \tag{14}$$

The merged $X_{ij}$is added to the remaining n-2 cases, repeating formula (5) to (9), the similarity of case X_{ij}between case n-2 can be obtained, sothe similarity matrix of n-1 order is obtained as follows:

$$r_{ij}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1(n-1)} \\ r_{21} & r_{22} & \cdots & r_{2(n-1)} \\ \vdots & \vdots & \vdots & \vdots \\ r_{(n-1)1} & r_{(n-1)2} & \cdots & r_{(n-1)(n-1)} \end{bmatrix},(1\leq i\leq n-1,1\leq j\leq n-1) \tag{15}$$

Also in equation (15) in the calculated R_{max}=max(r_{ij}), And determin $R_{max}\geq h$ is established, if established, then the case i and j is classified as a class, repeating the formula (14) to (15) steps until the similarity matrix obtained from the maximum likelihood value does not meet $R_{max}\geq h$, then the classification ends.

C. Case Retrieval

After classification for case of the case-base, the case is classified storage, and the corresponding indexes are established.Each of which must be calculated from a case on behalf of case A_i ($1 \leq i \leq k$), k is the case that stored in the case-base total number of categories, Assuming a case contain the number of n_i in the case-base, the case on behalf of the class isrepresented A_i, itis expressed as follows:

$$A_i=((\sum_{j=1}^{n_i} a_{j1})/n_i,(\sum_{j=1}^{n_i} a_{j2})/n_i,\cdots,(\sum_{j=1}^{n_i} a_{jm})/n_i) \tag{16}$$

The case of A_i represents a certain category of cases containing the attributes of all the representatives after averaging properties in Equation (16). When the system detects a new case X_0, according to the new property value of the case X_0, which the correlation for two cases is calculated, the formula is as follows:

$$r=\frac{\sum xy-\frac{(\sum x)(\sum y)}{m}}{\sqrt{\left[\sum x^2-\frac{(\sum x)^2}{m}\right]\left[\sum y^2-\frac{(\sum y)^2}{m}\right]}} \tag{17}$$

Σxy represent the corresponding property multiplied and summation for two cases in Equation (17), Σx^2 represents a single attribute squared and summation for one case, m represents the number of m attributes for the two cases. Corresponding case retrieval process as shown below:

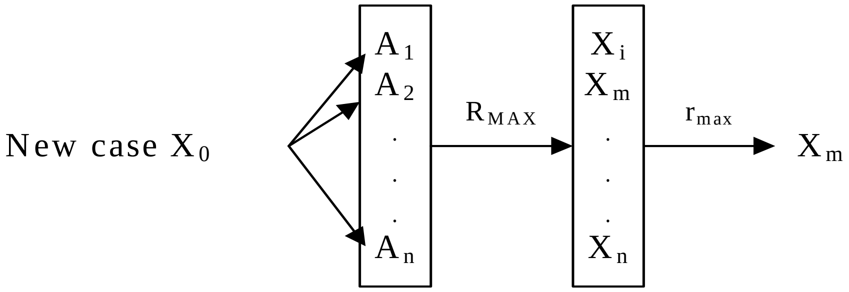

Fig.2. The process of case retrieval

In Fig.2the formula(17) is used to the similarity calculation for the new case X_0between cases of classification case-base, the greatest similarity $R_{MAX}$is obtained ,the result of retrieval contain a category from cases X_i to X_n,then, the new case computing the similarity of the most similar cases to get the final X_m.

EXPERIMENTAL RESULTS AND ANALYSIS

A. Algorithm effectiveness and stability test

In order to verify the stability and effectiveness of the algorithm, to obtain the data of land evaluation information system from the land evaluation database as the data source, C #. NET as a development tool, SQL SERVER2008 is used to design database. "Land Evaluation Information System" is used for the experimental platform .

The 100data of land index information are extracted from the database, 25 as a group, and the 100 cases are setted number from X_{001} to X_{100}, batches input system.As the original case-base. Case base on automatic classification algorithm, and then randomly select the database from the evaluation of a new case for case retrieval, each case contains the relevant information of one plots and recommendations contained species and fertilization recommendations.The experiment plots containfour properties, for example, organic matter, total nitrogen, available phosphorus and available potassium, as shown in Table 1:

TABLE 1 CASE INFORMATION SHEET

Property Number	Organic matter (g/kg)	TN (g/kg)	Phosphorus (mg/kg)	Effective potassium (mg/kg)
X_1	19.35	1.3	8.32	67.33
X_2	15.86	0.6	10.89	76.67
X_3	22.22	1.6	15.15	80.92
X_4	14.12	0.99	16.78	59.09
X_5	16.56	1.56	12.33	68.89
X_6	20.25	1.35	4.78	72.33
X_7	13.28	0.56	5.89	92.67
X_8	10.55	1.25	8.56	105.73
…	…	…	…	…

Before the experiment, the X_{012} cases calculated by hand is most similar and new cases in 100 cases. And X_{012} case is entranced into the database as the first group, case retrieval for enter batches, the k of k-means algorithm were taken 5,8,12 and 16.In addition to the proposed algorithm, the classification value of other clustering algorithm are taken 0.8, the search results ofthe algorithm and K-means algorithm as shown in Table 2:

TABLE 2ALGORITHM AND K-MEANS ALGORITHM RETRIEVESCONTRAST

Batch experiments	Number of cases	k-means Search Results	This algorithm search results
1	25	X_{012}	X_{012}
2	50	X_{012}	X_{012}
3	75	X_{036}	X_{012}
4	100	X_{062}	X_{012}

We can be drawn from Table II, with the increase of number cases in the case-base, the case that is retrieved by this algorithm are most similar with the algorithm.And the case that retrieved by k-means algorithm appeared inconsistent with the increase of number cases, through the contrast, this algorithm is stable and correct.

B. The efficiency comparison of algorithm retrieval

In order to verify the efficiency of the algorithm, there is a comparison to the K-means algorithm, variable weights KNN algorithm and this algorithm. The threshold of other classification algorithms are set 0.8, there is no similar threshold to comparison, only the most similar case as the search results. The similarity comparison of two cases adopt to euclidean distance as the result of calculation,contrasting results as shown in Table3:

TABLE 3 BATCH INPUT CASE RETRIEVAL ALGORITHM EFFICIENCY COMPARISON

Retrieval algorithm	Input 25Case	Input 50Case	Input 75Case	Input 100Case
K-means algorithm Similarity values	0.9968	0.9968	0.9577	0.9876
K-means algorithm Retrieval time (S)	0.0356	0.0462	0.0533	0.0698
Variable weights KNN algorithm Similarity values	0.9865	0.9865	0.9865	0.9865
Variable weights KNN algorithm Retrieval time (S)	0.0332	0.0398	0.0469	0.0503
This algorithm is similar Value	0.9998	0.9998	0.9998	0.9998
Retrieval algorithm Time (S)	0.0358	0.0369	0.0395	0.0401

The result can be shown from Table 3, The proposed algorithm is more stable and efficient with the increase of number cases. The efficient of algorithm has not decreased significantly, As shown in Fig.3:

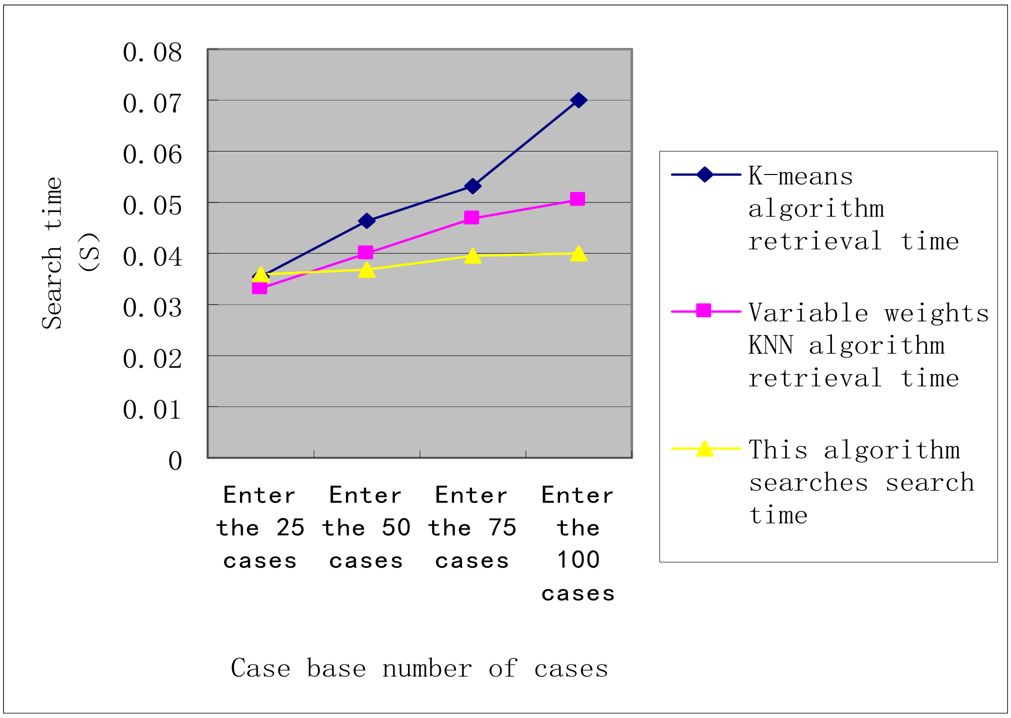

Fig.3. retrieval efficiency of algorithm comparison

CONCLUSION

This paper presents a case retrieval based on correlation analysis algorithm, using correlation analysis algorithms for classifying case of case-base and designing specific classification and retrieval procedures.The first time search is executed in typical cases of one classified to retrieve one cluster of most similar case, in which next retrieval step for final outcome will be done in last cluste. The experiments show that, this algorithm is stable and efficient compared with other traditional algorithm in the case retrieval with the increase of the number cases. The model is of practical value. The algorithm requires further research for the implementation of large-scale data cases and for non-numeric attributes applications. In order to further improvement of the algorithm.

Acknowledgments
The authors wish to thank the Youth Fund of Sichuan Province Education Department of China for contract 14ZB0005, the Youth Fund of Sichuan Province Education Department of China for contract 11ZB058, under which the present work was possible.

REFERENCES

[1]Ahn H,Kim KJ. *Expert Syst Appl*, v.36: pp.724-734, **2009**.
[2]Hu Ming, Heyan Ning. *Journal of Information*,n.2,pp.129-136,**2009**.
[3] YANG Shanlin, *Nee Machine Learning and Intelligent Decision Support System [M]. Beijing: Science Press*,pp.79 -112,**2004**.
[4]Yuan Guo,Jie Hu,Yinghong Peng. *Computer-Aided Design*,v.44,pp.496-508,**2012**.
[5] Liu Lianxi, Xing Tong, Xu Hao, Wang Wei, Gao Kai. *Hebei University of Technology*,v.33,n.2,pp.150-153 ,**2012**.
[6] Chen Ling, Cheng Zhonghua, Zeng Hui Yan. *Computer Engineering and Design*, , v.33,n.2, pp.581-585,**2012**.
[7]Zheng Yonghe,Kou Yingzhan,Zhang Weijun. *Journal of sichuan Ordnance*,v. 30,n.7,pp.122-124,**2009**.
[8] Lu Yang, He Xin, Jane Du. *Computer Engineering*, v.34,n.9,pp.28- 30,**2008**.
[9] Wang tomorrow, LIU Fu-Yun, Huang hair. *Computer Applications*,v.26,n.11, pp.4084-4086,**2009**.
[10] MA Huimin, Chunming. *Computer Engineering*,v.35 ,n.6,pp.206-209,**2009**.
[11]Liu Chunqing, Yang Xin Yuan, Zhang Ying, *Systems Engineering and Electronics*,v.29,n.6, pp.1017-1020,**2007**.
[12]Mykola Galushka , David Patterson. *Knowledge-Based Systems*, v.19,pp.625–638,**2006**.
[13] Guillermo Cortes Robles, Ste´phane Negny, Jean Marc Le Lann. *Chemical Engineering and Processing*,v.48,pp. 239–249 ,**2009**.

[14]Negny Stephanen, RiescoHector,LeLannJeanMarc. *Engineering Applications of Artificial Intelligence*.v. 23,pp.880–894,**2010**.
[15]ShenTsu Wang,WenTsann Lin. *Expert Systems with Applications*,v.37,pp.4544–4555,**2010**.
[16]I.Pereira, A.Madureira. *Applied Soft Computing*, v.13,pp.1419–1432,**2013**.
[17]Hu wei. *Computer Engineering and Applications*,v.49,n. 2,pp.157-159,**2013**.
[18]Elkan C.*Using the triangle inequality to accelerate meals[C]//Proceedings of the 2rid International Conference on Machine Learning (ICML-2003).* Menlo Park: AAAI Press,pp.147-153,**2003**.
[19]Kanungo T,MountDM. *Compumtional Geometry*, v.28,n.3,pp.89-112,**2004**.
[20]Patrice C. Roy, *Applied Artificial Intelligence*,v. 25,pp.883–926,**2011**.