Lejiang GUO[1], Wei WANG[2], Fangxin CHEN[1], Xiao TANG[1,2], Weijiang WANG[1]

Air Force Radar Academy (1), Wuhan University (2)

A Similar Duplicate Data Detection Method Based on Fuzzy Clustering for Topology Formation

***Abstract**. The changing information technology makes data increase exponentially in all areas, the quality of the huge amounts of data is the core problems. Data cleaning is an effective technology to solve data quality problems. This paper focuses on the duplicate data cleaning techniques. It studies the quality of the data from the architectural level, the instance-level problems, the multi-source single-source problems, duplicated records cleaning application platform and the evaluation criteria. In these studies, a improved novel detection method adopts the fuzzy clustering algorithm with the Levenshtein distance combination to data cleaning .It can accurately and quickly detect and remove duplicate raw data. The improved method includes a similar duplicate records detection process, the major system framework design, system function modules of the implementation process and results analysis in the paper. The precision and recall rates are higher than several other data cleaning methods. These comparisons confirm the validity of the method. The experimental results exhibit that the proposed method is effective in data detection and cleaning process.*

*Streszczenie. Artykuł proponuje nowe metody czyszczenia danych z uwzględnieniem liczby przypadków, wielu źródeł, podwójnych rekordów i innych kryteriów oceny. Ulepszona metoda detekcji wykorzystuje algorytm rozmytego klastrowania w dystansem Levenshteina. W ten sposób szybko wykrywane są i usuwane podwójne wiersze danych. (**Metoda detekcji podwójnych danych bazująca na rozmytym klastrowaniu**)*

Keywords: Approximate Duplicate Data, Fuzzy Clustering, Data Cleaning, Levenshtein Distance
Słowa kluczowe: czyszczenie danych, rozmyte klastrowanie.

Introduction

The changing information technology makes data of all areas increase extremely fast, which makes data become the focus. The accurate analysis of data is of significance because the decision, which decision-makers made, is highly depend on it. However, the data received from external sources in the data warehouse is usually dirty, incomplete and inconsistent. There are two reasons: firstly, the lack of effective data analysis techniques; the second, the data is of not high quality, different data sources on the same data may violate the value of the data itself. The dirty data in the DW (Data Warehouse) will affect the accuracy of the extracted information, making all kinds of data analysis, such as OLAP (On-Line Analytical Processing), data mining applications have ambiguous or wrong conclusions, making decisions lose the support. Therefore, the quality of DW data is a problem can't be ignored. Data cleaning is designed to guarantee high quality data, to eliminate error data redundant data and inconsistent information in data warehouse or database, which can enhance the quality of decision. Data cleaning aim is not only on eliminating data errors, redundant and inconsistent information, its fundamental purpose is to unified a variety of data sets from different, incompatible rules, which are the necessary elements to construct DW and KDD (Knowledge Discovery in Databases). Since the data inconsistencies will be a variety of possibilities, and data is of a large amount, data anomaly detection usually relies on highly efficient data conversion algorithm [1].

Database technology and data's large amount are undergoing a revolution from quantity to quality. In the past, people complete the data query and analysis in a semi-automatic way. Those data cannot satisfy people's analysis and argument of affairs in various fields hide behind the data. KDD and data mining techniques generate and develop based on the needs of the community, it also create much new fields, such as research and application in data warehouse technology [2]. Those technologies make people use data effectively on a more convenient platform, which laid a solid foundation for decision analysis. Initializing the various data of different data sources is a very important part in the KDD; it has an important impact on the final result. Entities of the real world in different data sources often use different syntax forms. The way to remove data of different data sources is difficult since the data in different forms may lead to the situation of only-two or only-three result. The a large number of fact has proved that, similar to large-scale intelligent systems, such as data mining system, data pre-processing constitute 60% to 80% of the entire workload, the data cleaning is one of the part of data pre-processing, which has a very important role in the intelligent system. Therefore, the study on system, which could produce model from a large number of useful information, becomes increasingly important. Data cleansing is a work-related field, which work around specific areas. The data-cleaning framework for all modes is not available, which makes the research of general data cleaning system become a hot spot.

The Study and Improvement of Fuzzy Clustering Algorithm

The Status of Fuzzy Clustering Algorithm

Pattern recognition and machine recognition is regular in noisy or complex environments, which is the search in data structure. In pattern recognition, the set of data is called a cluster. In reality, data is not evenly distributed, so the rule or structure may not be accurately defined. In other words, pattern recognition is an inexact science. It is to deal with some fuzzy issue. For example, the boundary between clusters may be vague, and not certain, that is, a data may belong to two or more different levels of cluster members. So the key issue is to find a set of data points, which is a cluster. The current study on fuzzy clustering algorithm located in the following areas: the convergence local minimum of the fuzzy c-means clustering algorithm; creation of different types of objective functions; dividing the fuzzy space in different ways; leading the optimal value of real world into fuzzy clustering algorithm; minimizing the distance of fuzzy c-means clustering algorithm; clustering a large number types of data. This paper studies how to optimize the most classic fuzzy c-means clustering algorithm in fuzzy clustering algorithm, and apply the improved algorithm to the detection of approximately duplicate data.

Fuzzy c-means Clustering Algorithm

Fuzzy C-means (FCM) clustering algorithm allows a group of data belonging to two or more clusters. This method was raised by Dunn in 1973, and developed by Bezdek in 1981, and is commonly used in pattern recognition [3]. It is a clustering algorithm based on division, and it aims at led the similarity between objects which are divided into the same class (members of group) to be maximum, whereas the similarity between objects of different classes minimum. It is based on the objective function's minimizing, which list as follows:

(1) $$Jm = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\|x_i - c_j\|_2$$

where m is any real number greater than 1, which was set to 2.00 by Bezdek. u_{ij} is a j level of x_i member cluster; x_i is the i dimensional measurements; C_j is a dimensional clusters, $\|*\|$ center is to express the similarity between any criteria, any measurement center. The objective function which fuzzy partition optimized through iterative function list as equation (1), updating members u_{ij} and the cluster center C_j:

(2) $$u_{ij} = \frac{1}{\sum_{k=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}}$$

(3) $$C = \frac{\sum_{i=1}^{N} u_{ij}^{m} \bullet x_i}{\sum_{i=1}^{N} u_{ij}^{m}}$$

The iteration stops when meet the equation (4).

(4) $$max_{ij} = \left\{ \quad \left|u_{ij}^{(k-1)} - u_{ij}^{k}\right| \quad \right\} \varepsilon$$

where ε is between 0 and 1, k is the iteration step. The process converges to a local minimum or saddle point of Jm.The algorithm has the following steps:

1. transform U into matrix $\lfloor u_{ij} \rfloor, U(0)$

2. the k-step, according to equation(3) ,and calculate the matrix center vector $c(k)$

3. calculate $U(k)$, $U(k+1)$ according to equation (1) stop calculate If

$\|U(k+1) - U(k)\| < \varepsilon$, return to step 2.

This algorithm proposed in this paper will be leaded as the fundamental algorithm to detect the similar and duplicate data into the database cleaning.

The Problems of the Existing Algorithm

Research from the above algorithm, it knows that data is bound to each cluster's membership function, which represents the fuzzy way of the algorithm. To achieve this goal, it only needs to create a matrix named U , a number between 0 and 1 representing connection degree between the data and cluster canters. Given a particular data set, assuming that is an axis distribution. One-dimensional data clustering membership is shown in Fig.1, the connection degree represent the location of similar data between similar data member to a certain extent. In the FCM approach, the same origin does not belong to only one well-defined cluster, but it can be placed in middle. In this case, as the membership function shown in Fig.2, represented with a smooth curve, each base may belong to different groups of members.

Fig.1 One-dimensional data clustering membership

From the clustering membership of FCM method's data in Fig.2, there are some traditional clustering algorithms, which have been applied to the detection of similar or duplicate data by scholar, but there are some problems [4]. In the mapping process of character field, if there exists spaces in the mid-point, they will lose the corresponding effective field information. If DBSCAN algorithm is used alone in dealing with character data, the general is to add ASCII value of all character in turn, it can easily lead to misjudgement. Such as big and gibe, they can be considered to be duplicate data, those two different records may be classified into the same class, and the clustering result obtained in the cluster set is not entirely similar record.

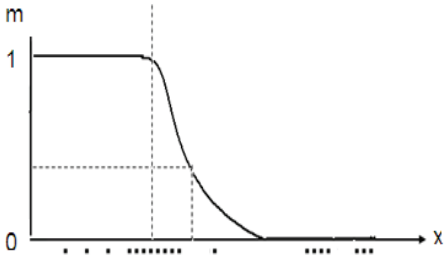

Fig.2 Data clustering method FCM membership

The Similar Duplicate Data Detection Based on Fuzzy Clustering

Through research and analysis of the previous section, these will have more or less the existence of problem by the use of a single algorithm, and it will affect the final result of data cleaning. Broadly speaking, data cleaning is the process of data sources, and making the processed data efficient and useful; in the narrow sense, data cleaning specifics the process to build DW and data mining to make the data source accurate complete consistence after that the data source is reliable [5]. Based on the comparison the previous method, this paper presents a new method of detecting similar or duplicate data. The method uses a combination of the fuzzy c-means clustering algorithm and the Levenshtein distance to process dirty data, and gets the needed data accurately effectively and quickly. So that similar or duplicate data will all be throw into trash. There are two steps in the method proposed in this paper, firstly using fuzzy c-means clustering algorithm to gather the members of greatest similarity into a class as far as possible together, and then compare the members of each class by Levenshtein distance, finally delete the same members of each class which are namely similar data [6].Levenshtein distance is the simplest edit distance which is a commonly used measure of the string distance. This method is the interpretation for minimizing the number of operations the from the source string to the destination string by insert (insert characters in a position of a string), remove (delete one character form string), replace (replace a character with another character in the string). This method has a wide range of applications in determining the similarity of two strings.

Edit distance is defined as follows: assuming all strings is created from a limited set of characters Ω. Two strings such as S, T the length of the string, are respectively $|S|, |T|$.

$D(S,T,i.j)$ represents the edit distance between the first i characters of the string S and the first j .

Characters of the string T ; Obviously, $D(S,T,0.0)=0$.

$D(S,T,i.j)$ recursively defined as follows:

$$D(S,T,i.j)=\min\begin{cases} D(S,T,i-1,j-1)+1 & T_j to S_i \\ D(S,T,i,j-1)+1 & insert T_j \\ D(S,T,i-1,j)+1 & Del S_i \end{cases}$$

the edit distance of String S , String T :

(5) $$D(S,T)=D(S,T,|S|,|T|)$$

Through the above definition, you can write a program to calculate the edit distance between two strings, the algorithm's time complexity is $O(|S|\cdot|T|)$.

For example: transform String s = "Acadammy" into a string t = "Academy", s is the source string and t is the target string, the operations are shown in table 1:

Table 1.Example edit distance

S	t	Operation
|Acadammy	|Academy	
A|cadammy	A|cademy	
Ac|adammy	Ac|ademy	
Aca|dammy	Aca|demy	
Acad|ammy	Acad|emy	
Acada|mmy	Acade|my	Replace ‘a’to ‘i’
Acadam|my	Academ|y	
Acadamm|y	Academ|y	Del ‘m’
Acadammy|	Academy|	

If define the cost of each edit operation is 1, In Table 1 it shows, edit distance between "Acadammy" and "Academy" is 2, calculate result of similarity after the standardization (to take a longer string length) is 0.75.If define the cost of each edit operation is 1, that is the Levenshtein distance. Based on the above analysis, the edit distance has a certain effect on matching of the misspelled strings. For example, edit distance between "Computer Network" and "Computer Ntwork" is only 1 and similarity value is 0.9375; it also has some effect on insertion and deletion of the short words, for example, edit distance transform between "Lenovo" and "Lenovo Co" is 3 and the similarity value is 0.6667.

Steps of the Improved Algorithm

In this paper, the improved algorithm is divided into two phases:

The first phase: Clustering the data sets by the fuzzy c-means clustering algorithm, so that a large data set is divided into a number of small classes (groups), most of the similar or duplicate records is clustered into different small classes. And duplicate records in each class are located closely positions, which create a relatively better condition for the accurate detection of these records.

The second phase: start second cluster, it mainly improve the clustering accuracy. Most of the similar or duplicate records are clustered into different small classes in the first stage, but no accurate results of detection are acquired [7]. By using the Levenshtein distance method in the second clustering method, one can contrast the characteristics of each member one by one to find duplicate records, and make test results more accurate.

The Feasibility of the Improved Algorithm

If fuzzy c-means clustering is used individually, time complexity will directly related to the amount of input data, this will lead to that this algorithm has a big system overhead. In other words, after the first cluster, it will have a re-cluster on each class that contains similar or duplicate records, but the amount of input data has not been reduced, the time overhead is still large, and it will be very slow to reach the convergence [8]. Then compare data by using Levenshtein distance method individually, the time complexity is $O(n^2)$, it is not a good optimisation method, either. Firstly, this algorithm does cluster analysis through the fuzzy c-means, after the completion of the first cluster, another part of the error data which are caused by the change of character position are cluster into a member group (cluster). So it is needed for the Levenshtein distance method to do the re-cluster. We can know from the previous analysis, Levenshtein distance method has a certain effect on matching the misspelled strings. At the same time, the number of each class’s records has been relatively small after the first cluster, so it does not really increase the computing time. And there will be some increase in accuracy. Meanwhile, the method can not only make up for shortcomings caused by detect the similar or duplicate data individually, but also get more precise results.

The Implementation of the Similar or Duplicate Data Detecting Method Based on Fuzzy Clustering

Design of System Process

Implement data cleaning platform through the previous Improved algorithm, making data of low quality in the data warehouse can be cleaned by this data cleaning platform, making the information system data more accurate, consistent, to support the right decisions [9]. System design process chart is shown in Fig.3.

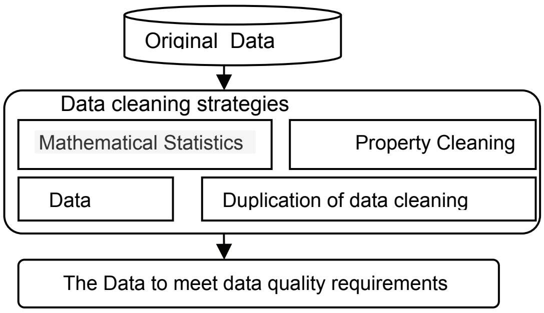

Fig.3 System Design Process

Steps of Implementation

Based on the above discussion of data cleaning, the process of cleaning data can be divided into the following steps; each step can be subdivided into a number of tasks. These steps can be described as Fig.3.

1. Data Preparation

Data analysis is the first pivotal step of data preparation. The main works of analysis are located in the source file, identification and separation of individual data elements. Data analysis makes it easy to do the transformation, standardization and data matching, because it allows comparisons between individual data components, rather than between a lot of long and complex strings. It is needed to analysis the school logo, school name and unit component into a fixed format packet in this data source. This is a key step for detection of duplicated records, and even for all of the data cleaning process. Data conversion means making data of same property have the same data type by converting, which can also be called the type transformation. If an application stores data in a data type, this format of data make sense in this system, but if these data are applied to a new system, these data may be unavailable, and even lead to errors. The main process is list as follows: identify data sources, import the data to database engine. According to data types, length, the value and discrete values, frequency, diversity, uniqueness, null values, typical string pattern of metadata and derived, providing a exact point of all kinds properties quality problems. In this paper, we regard null values, as the record is not complete. And incomplete records will be imported into the detection module to re-detect, only had this off can the data be imported into the data source and used. Do the class definition of the same type of thing, and then, each record is an object entity. After the class definition, hypostatize the data source which will be processed, each record corresponds to a single object entity.

2. The implementation of fuzzy C-means clustering

Step through the operation on the data source data has been completed the data analysis, conversion, the process of standardization, and lack of data to do a preliminary cleaning, the same type of data encapsulated into classes, according to fuzzy c-means clustering algorithm to gather type of analysis, allows the experimental data source 4000 records are divided into two clusters x, x is much smaller than 4000, and this x number of clusters in each cluster is basically to similarity as possible between similar , as differences between different classes. In short, the implementation of fuzzy c-means clustering algorithms, large data sources can be divided into a relatively small data set, which focused on a small data set most of the duplicated records. The steps of data processing, the division has good results, but for the boundary value, can not distinguish between good, the next step will use the Levenshtein distance algorithm further identification.

3. The implementation of Levenshtein distance method

Compared using Levenshtein distance algorithm for each class of record for the second cluster, the step has been possible to repeat the record set similar to a class, so the records in each cluster similarity is relatively high. Based on the foregoing analysis of the Levenshtein distance, it determine the similarity of two strings, especially strings to have misspelled words and between the insert and remove the short side there is a wide range of applications. The data table in the same class labelled data clustering compared one by one, until the cluster from the first record to the last record so far. Then repeat for the second clustering steps until all the data in the data table comparing completed.

4. Remove duplicate records

It is well known that in our country everyone's ID number is unique on the network card's MAC address is unique, then our data in the database are unique to different entities. Follow the steps we need to think more over the members of the same class can be cleared out. In the cleaning process, you can use three different methods for a manual by a professional decision based on experience and qualifications of what information would be deleted. Progress to the second than the first semi-automatic, and it randomly selected by calculating the similarity to delete the data, the accuracy is higher. The third type is automatic, the computer do it themselves, no need to worry about you, but can not guarantee accuracy. During this step, you can based on the environment, the scope, purpose and other selection methods on the line.

Simulation and Analysis of Experimental Results

We evaluate the performance of our proposed method (EELTC) via MATLAB. For simplicity, an ideal MAC layer and error-free communication links are assumed. The simulation parameters are given in Table.2. First, we obtain maximum length of furthest level (R0) which is a predefined value and must be set primarily.

Table 2. The simulation parameter

Parameter	Value
Network size	200*200 m
BS location	100,350 m
Number of Sensors	200
Initial Energy	0.5 J
Eele	50 nJ/bit
εfs	2 10 pJ/bit/m
εmp	4 0.0013 pJ/bit/m
EDA	5 nJ/bit/singnal
Data Packet Size	4000 bits
ω 1	0.2
t0	0.008 s

In the broadcast-based mode, the nodes do not establish gradients in the interest broadcast phase. Instead, the magnetic charge is included in the data being disseminated. The receiving node can tell from the charge carried in the data where the data is from and whether to forward further down-stream. When a node receives data, it checks if it has any matching entry. If it does, it compares the magnetic charge in the entry with the magnetic charge of the data. If the former is greater than the data, it sets the magnetic charge of the data t in the entry and then broadcasts the data. This means the data is sent from the node whose magnetic charge is lower than the intermediate node, and the intermediate node repeats this process.

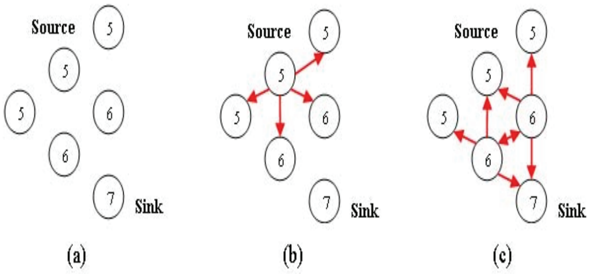

Fig.4 Data propagation with collusion.

The node receives duplicate data, or data whose magnetic charge is greater than that in the entry, the data will be discarded. In Fig.4 (a), the magnetic charge of every node is estab- lished; the sink's charge is 7. In Fig.4 (b), when the source wants to send data to the sink, it will broadcast the data to its neighbor nodes. In Fig.4(c), the nodes with charge strength 6 broadcast the data because the magnetic charge of the nodes is greater than that of the data. Thus, the sink receives the data. From the tools Menu, one can select the option Simulate Link Breaks. In this mode, if a link on the graph is clicked, the link is broken. If the algorithm is a parent table driven one, where the alternative parents are saved in lists, the algorithm directly changes to the alternative parent. If it is not case, the algorithm re-routes the graph without using the broken link in Fig.5. If the Simulate Link Breaks option is not selected, for the algorithms that generate parent tables, these tables can be displayed by clicking on the nodes.

In Fig.5, Simulation of link breaks. The link of node 12 towards the root node is broken. With the fuzzy Algorithm, a re-routing takes place and many nodes change their active links shown in redon the left). On the other hand, as 2-SafeLinks Algorithm is an algorithm that creates parent tables, only the node loosing the link (node 12) changes its active link, the rest of the links (in black) remain the same (on the right).

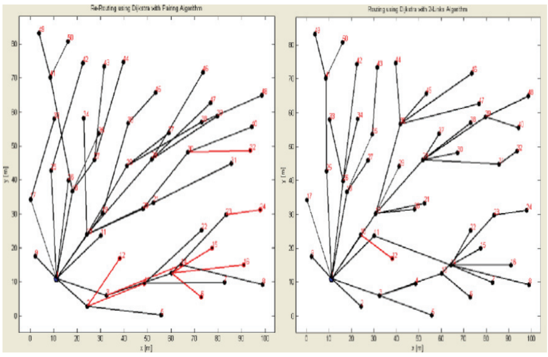
Re-Routing using Dijkstra with Pairing Algorithm
Routing using Dijkstra with 2-Links Algorithm
y [m]
x [m]
0
10
20
30
40
50
60
70
80
90
100

Fig.5 The generated routing trees with different algorithms

We compare the amount of energy consumed by CHs in three protocols for 20 rounds (Fig.6). The energy consumed by CHs per round in EELTC is much lower than that in LEACH, and is almost the same as EEUC. In LEACH because CHs send their data to the BS via single hop communication, the energy consumption is much higher. In EELTC and EEUC, CHs transmit their data to the BS via multi- hop, so a considerable amount of energy is saved.

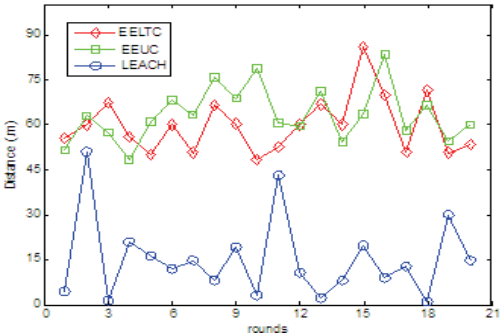

Fig.6 Minimum distance between CHs

We compare energy consumption just in setup phase without considering energy consumption in steady state of three protocols in 20 rounds in Fig.7. EELTC saved considerable amount of energy by considering of the time for each candidate to start broadcasting advertisement message. EEUC protocol consumed higher energy than LEACH and EELTC because it has more control overhead in setup phase. In EELTC clusters creation and inter-cluster multihop routing are accomplished by a single ADV message. But in EEUC this is done by too many control messages exchange. Finally, we examine the energy efficiency of three protocols by evaluating of the time until the first node dies and the time until the last node dies. It is clear that EEUC increases network life time compared with LEACH, and EEUC due to having much lower overhead and using multihop transmission data.

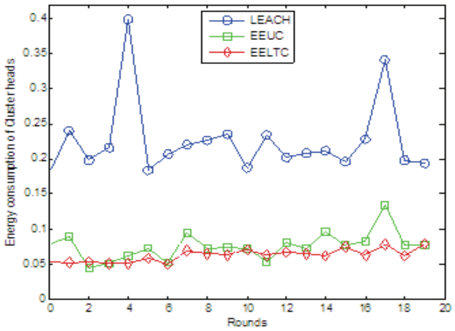

Fig.7 The amount of energy consumed by CHs

Conclusion

With the development of information technology, data cleaning process has become an important part of data integration . Similar to the detection of duplicate records in a data cleansing has been the focus of the study. Existing duplicate record detection algorithm is similar in most cases, Field matching algorithm and the record matching algorithm is the core of this approach. Existing field matching algorithm does not have the versatility, and record matching algorithms usually find field values using weighted method to calculate the similarity of accuracy is poor.In this paper, we develop an accurate and comprehensive energy model for the front-end of a wireless netowork. For a disjoint multipath configuration whose patterned fail- ure resilience is comparable to that of braided multipaths, the braided multipaths have about 50% higher resilience to isolated failures and a third of the overhead for alternate path maintenance.This more in-depth study of the existing duplicated records detection method to analyze the advantages and disadvantages of various detection methods. Existing detection methods for calculating the similarity values recorded shortcomings, the improved fuzzy clustering algorithm is introduced to improve the detection of approximately duplicate records, for specific issues, constructed for the detection of duplicated records cleaning platform to improve a similar accuracy of detection of duplicate records.

REFERENCES

[1] Stojmenovic, X. Lin. Power-aware localized routing in wireless networks, (2001) 12,No.11,1122-1133

[2] Huseyin Ozgur Tan, Ibrahim Korpeogle. Power Efficient Data Gathering and Aggregation in Wireless Sensor Networks, (2003) 32,No.4, 66-71

[3] Jian Yu,Miin-Shen Yang.A Generalized Fuzzy Clustering Regularization Model With Optimality Tests and Model Complexity Analysis, IEEE Transactions on Fuzzy Systems, (2007) 15,No.5, 904-915

[4] Chatzis, S.,Varvarigou, T..Factor Analysis Latent Subspace Modeling and Robust Fuzzy Clustering Using t-Distributions, IEEE Transactions on Fuzzy Systems , (2009) 17, 505-817

[5] M. Cardei, J. Wu, M. Lu. Improving network lifetime using sensors with adjustable sensing ranges, Sensor Networks, (2006) 10,No.2,41-49

[6] Luo H., Luo J., Liu Y.,, Das S. K..Adaptive Data Fusion for Energy Efficient Routing in Wireless Sensor Networks, IEEE Trans. on Computers, (2006) 18,No.4, 1286-1299

[7] Yao Shen, Yunze Cai, Xiaoming Xu.A shortest-path-based topology control algorithm in wireless multihop networks, Computer Communication Review, (2007) 37,No.5, 29-38

[8] M. Zuniga , B. Krishnamachari. Analyzing the transitional region in low power wireless links, IEEE Secon'04, 2004.

[9] K. Seada,M. Zuniga,A. Helmy,B.Krishnamachari,Energy efficient fowwarding strategies for geographic routing in wireless sensor networks,in ACM Sensys'04, Baltimore, MD, Nov. 2004.

***Authors**: dr Lejiang Guo, a Senior IEEE member, the department of Early Warning Surveillance Intelligence, Air Force Radar Academy, Wuhan, 430019, China. E-mail: radar_boss@163.com.* dr Wei Wang, school of power and mechanical engineering,wuhan university ,Wuhan, 430072,China. *E-mail: leezbaby@163.com. Prof Fangxin Chen, the department of Early Warning Surveillance Intelligence, Air Force Radar Academy, Wuhan,430019, China. E-mail:pxial@msn.com.*