Design of Multi-level Teaching System Based on Association Rule Mining

https://doi.org/10.3991/ijet.v11i11.6252

Hong-yan He, Hui-ping Zhang, Hong-fang Luo
Hubei University of Technology, Wuhan, China

***Abstract*—In order to improve the overall performance for the multi-level teaching system, a system with Multi-strata teaching is designed. It divides the whole class into smaller parts based on their knowledge level and learning ability and teach students in accordance with their aptitude. The system used the model of C/S, and applies ASP in the interactive user interface. The data mining algorithm is also presented in the study. The system was tested with practical data. The results show that with the teaching system teachers can separate students into different parts and get a very good idea about how much students can learn.**

***Index Terms*—association rule mining, multi-level teaching system, multi-strata teaching**

I. INTRODUCTION

Association rule is an important method and technology in data mining; in association rule, the operating frequency of I/O will affect the efficiency of mining task. So the main method of reducing this frequency is to reduce the frequency of scanning D Dataset; from another aspect, the number of candidate item sets that need to calculate support should be reduced in order that the number can approximate to the number of frequent item sets. The reason is that the smaller number of candidate item sets can help saving the computing time and storage space. With development of the networks and economy, distributed systems become more and more popular[1]. However, Deficiency of distributed association rule Mining, with some respects to Consultation and competition between each node, Utilization of information, efficiency of network communication, become more obvious and seriously affects the application of association rule mining.

Association rule mining is a form of data mining to discover previously unknown, interesting relationships among attributes from large databases. Due to its simple form and being easy to understand, association rule mining has attracted great attention in database, artificial intelligent and statistics communities, and a lot achievements have been made in its study[2]. Compared with artificial method, such as neural network, genetic algorithm and statistics, it can processes larger dataset, on the other hand, artificial method usually processes a small set of data, and it aims at finding a model between inputs and outputs. Association rule mining can find large number of patterns among attributes. Furthermore, although large datasets can be processed in statistics, these work aims at finding data distributions or statistical model.

Big data is a broad term for data sets so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, creation, search, sharing, storage, transfer, visualization, and information privacy. The term often refers simply to the use of predictive analytics or other certain advanced methods to extract value from data, and seldom to a particular size of data set.

Analysis of data sets can find new correlations, to "spot business trends, prevent diseases, combat crime and so on." Scientists, practitioners of media and advertising and governments alike regularly meet difficulties with large data sets in areas including Internet search, finance and business informatics. Scientists encounter limitations in e-Science work, including meteorology, genomics, complex physics simulations, and biological and environmental research. Data sets grow in size in part because they are increasingly being gathered by cheap and numerous information-sensing mobile devices, aerial (remote sensing), software logs, cameras, microphones, radio-frequency identification (RFID) readers, and wireless sensor networks. The world's technological per-capita capacity to store information has roughly doubled every 40 months since the 1980s; as of 2012, every day 2.5 Exabyte (2.5×1018) of data were created; The challenge for large enterprises is determining who should own big data initiatives that straddle the entire organization.

Relational database management systems and desktop statistics and visualization packages often have difficulty handling big data. The work instead requires "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the users and their tools, and expanding capabilities make Big Data a moving target. Thus, what is considered to be "Big" in one year will become ordinary in later years. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration."

There are many ways to record the user mobility which can be Wi-Fi, Bluetooth, Infrared, GPS and GSM depending on the situation and type of intended application. Most important work regarding the location extraction based on algorithms is done in their work where they formally defined the term mobility mining to extract patterns through profiling. While the current mobility trends are studied in detail by their work, the mobility is defined as key prediction indicator of human life.

Nowadays, connotation construction becomes the central part of the development of higher vocational education which makes scientific curriculum design more important than before. The main task of higher vocational education is to improve students' specific practical skills

and the cores courses play a significant role in achieve this goal.

Educational reform has long been started before while only few satisfied results received. According to related analysis, there are two main reasons for this phenomenon: firstly, students come from different schools with different education background; secondly, students9 learning capacity is not on the same level. Due to these above two reasons, multi-strata teaching in main curriculum becomes very necessary. This system is more easily to be applied to general courses than to the specific ones for it may be faced with difficulties. Based on research that has been done, hypothesis can be got that association rule mining is very supportive in this multi-strata curriculum designing.

II. OVERVIEW

Multi-strata teaching is to divide the whole class into smaller parts based on their knowledge level and learning ability and teach students in accordance with their aptitude. In this process, teachers can set up appropriate teaching goals and give instructions on gaining solid general basic knowledge and improving practical skills for individual students. This paper mainly explores how to design different curriculum for different students by making predication about to which extent students grasp the new knowledge. In this system, teaching materials, teachers and teaching hours should be selected as the need of students of different strata. Multi-strata Teaching system oriented for CAT majors established based on the association rules in concept lattice will play a central part in teachers 5 making predication and receiving feedbacks. To be more detailed, the system provides teaching instructions by setting up Web pages in accordance with the CAT courses and accomplishing data collecting, logical analysis and the performance predication through C/S[3-4].

Data mining (DM) is to extract knowledge from huge datasets, the purpose of which is to find the useful patterns hidden behind the data. However, most available datasets are becoming enormous in size. Also, their high dimensionality motivates the need for efficient and scalable parallel algorithms. The design of such algorithms meeting above requirements is challenging. There are two classes of association rules mining algorithms[5-6]. One is based on the A priori algorithm, which generates large candidates when counting the frequent item sets.

Hadoop and HDD algorithms are efficient and scalable parallel methods applied in the discovery of association rules in the field of data mining[7-8]. However, they become less effective due to the imbalance caused by distributing the candidates among the processors. Therefore, Hadoop and HDD are improved by means of introducing the approximate algorithms to solve the problem of load balance effectively. There are two approximate algorithms, one is called the online algorithm, and the other is named the offline algorithm. After that, we give the proof of their performance ratio; the complexity analysis of the improved Hadoop algorithm is also given[9-10]. The other class of algorithms finds the associations without candidacy. And based on the efficient FP-growth algorithm, its implementation method of constructing the frequent pattern tree and mining frequent item sets is given for the shared memory parallel formulation. However, it results the imbalance due to the distributing the work among the processors[11].

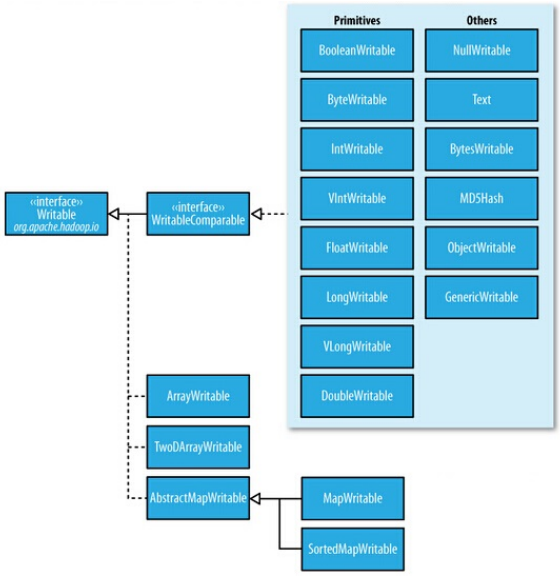

Figure 1. The basic framework of Hadoop model

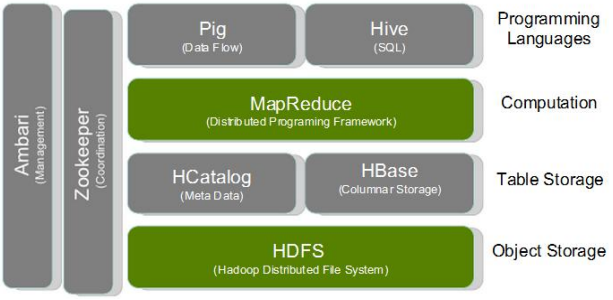

Figure 2. The Hadoop Architecture in network model

With the development of the information society, the amount of data produced daily is in exponential growth. Companies face a big problem that how to find out useful information from mass data. Data mining algorithm processing data and mining is hidden useful information, which is beneficial to the development of company to make decision. But it will take a long time to deal with mass data, or inability to deal with mass data. The figure 1 shows the basic framework of Hadoop model with the figure 2 shows the Hadoop Architecture in the network model.

An effective method to solve problem is to transfer the traditional algorithm to cloud platform for parallel improvement. Apache Hadoop is a distributed system framework. The HDFS provides high fault tolerance and high throughput rate of file storage, reading and writing. Map Reduce provides a parallel programming framework. The user writes Map and Reduce class for distributed program without knowing distributed parallel programming details. Because of mass data storage platform and simple parallel calculation platform, Hadoop provides the basis for the traditional data mining algorithm processing mass data.

Discovery of sequential patterns is becoming increasingly useful and essential in DM fields. Many important knowledge discovery tasks in genomic require the analysis of DNA and protein sequences. The most time-consuming operation in discovering such patterns is the computation of the occurrence frequency of all sub-sequences (called

sequential patterns) in the sequence database. In particular, there are three main classes of sequential pattern discovery algorithms. In general, projection-based frequent pattern discovery algorithms have been shown to substantially outperform the others. They still require a substantial amount of time.

III. METHOD AND ALGORITHM

This system is established in the model of C/S, and applies ASP in the interactive user interface, during which process the data source will be organized by SQL Server 2000 and transmitted through ADO (ActiveX Data Objects). Three parts comprise this system: data layer, core mining module and result analysis. Data layer is the foundation of this system for it collects information for the mining module and data analysis will not be able to be processed without data layer. If this system is adequate in reading data from DBMS with the help of SQL or connecting with DBMS through ODBC and JDBC, this kind of data mining tools is very efficient in data accessing. The central part of this system lies in the mining module, for it concerns the Mining Algorithm and users can gain knowledge they need in this system.

The result analysis part makes it convenient for users to understand and make judgment on the data that they get. This tool also has advantages in making the important information manifest through its comparatively simple way in collecting information. Last but not least, the handle ability of result analysis part is also of significant importance. The system data stream will be show in the figure 3.

It is necessary to collect original data and unify the formats for they may originally functions in different kinds of documents and media. Then the data that is processed will be uploaded to SQL Server data base, in which process some noisy data will be deleted.

The general data base design should allow for the original data format and the interface with other system, while it does not consider whether the surface structure is suitable for mining algorithm in the system. Consequently, the loaded data need to be transformed in order to proceed mining, and converted data will be stored in mining base.

The mining algorithm will be programmed as the store procedure of SQL Server and work on server. Rules mined in this process will be stored on the list of rule base which will be used for user enquiry. Users can set and rearrange the mining parameter through HMI.

More attention should be paid that the general data base, mining base as well as the rule base are logically three different ones while they are physically the same one.

Based on different characteristics, the module part can be divided into the following ones: data mining module and interaction module. Important procession such as data preprocessing, mining and evaluation will be accomplished in the data mining module. This part is not only suitable for the two strata C/S, but also can be applied in the three-strata B/W/S.

Interactive module is responsible for the multi-strata teaching instructions received. Users submit data through C/S and present information to students by Web, on which data will be further processed and analysis result will clearly show to users. Meanwhile, students9 information will be stored in data mining base.

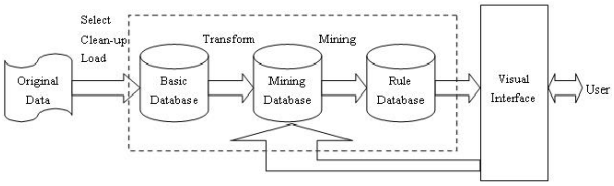

Figure 3. The system data stream

The main effect of data mining algorithm is to extract the required information from a large amount of data, including the structural data, semi-structured, and unstructured data sources, such as audio, video, data, for data algorithm. This algorithm must have model, and first search algorithm. Currently the common data mining algorithms are mainly the decision tree method, bionic global optimization of genetic algorithm and neural network, statistical analysis and row exclusive counterexample method, etc. In order to improve the effect of data mining process effectively, a detailed research should be applied to the cloud computing method. In this way, more effective implicit knowledge in the mass data information can be discovered in order to improve the effect of application of information data.

Association rules found a relationship between things and other transactions or interdependence. Assuming that $I=\{i_1,i_2,\ldots i_m\}$ is a collection and the related data task D is a collection of database transactions, in which each transaction T is a collection, and making $T \subseteq I$. Every transaction has an identifier TD. Assuming that A is a set of items, $A \subseteq T$. Association rules are the containing type of $A \Rightarrow B$, among them, $A \subset I$, $B \subset I$ and $A \cap B = \Phi$. The rules $A \Rightarrow B$ is in the transaction which sets up with support s , s is the percentage for the transaction contains $A \cup B$ in the D.

The LIPI algorithm through scanning data sets of frequency, then finding relevant data and finishing dig, its principle as follows:

Assuming that $I=\{i_1,i_2,\ldots i_n\}$ is a collection, which composed of different characteristics, the characteristics of each item as a constituted set of items. And the item set is not an empty set, but is a subset of the set of I, which can be expressed as$(x_1x_2\ldots x_m)$, every x_k is a term.

The sample variance and sample proportion of variance have established the following relationships:

$$S^{*2} = \frac{S^2}{\mu^2} \tag{1}$$

Proof: by the definition, we have:

$$\begin{aligned} S^{*2} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{\alpha_i - \overline{\mu}}{\overline{\mu}^2}\right)^2 \\ &= \frac{1}{(n-1)\overline{\mu}^2}\sum_{i=1}^{n}(\alpha_i - \overline{\mu})^2 \\ &= \frac{1}{\overline{\mu}^2}S^2 \end{aligned} \tag{2}$$

Consider delay, the L can be expressed as:

$$L^0 = \begin{Vmatrix} C^0_{ijkl} & e^0_{kij} \\ e^{0T}_{ikl} & -\eta^0_{ik} \end{Vmatrix} \quad (3)$$

These functions can be expressed in the following form:

$$C(\mathrm{x}) = C^0 + C^1(\mathrm{x}),\ e(\mathrm{x}) = e^0 + e^1(\mathrm{x}),$$
$$\eta(\mathrm{x}) = \eta^0 + \eta^1(\mathrm{x}),\ \rho(\mathrm{x}) = \rho_0 + \rho_1(\mathrm{x}) \quad (4)$$

The value with superscript of 1 represents the difference below:

$$C^1 = C - C^0, e^1 = e - e^0,$$
$$\eta^1 = \eta - \eta^0, \rho_1 = \rho - \rho_0 \quad (5)$$

The whole function can be simplified into the following integral equation set:

$$f(x,\omega) = f^0(x,\omega) + \int_V S(x-x')(\mathrm{L}^1 F(\mathrm{y}') + \rho_1\omega^2 \mathbf{g}(R)\mathrm{T}_1 f(\mathrm{y}')]S(\mathrm{y}')\mathrm{d}y' \quad (6)$$

In addition, we can introduce the abbreviated formula:

$$g(x,\omega) = \begin{Vmatrix} G_{ik}(x,\omega) & \gamma_i(x,\omega) \\ \gamma_k(x,\omega) & g(x,\omega) \end{Vmatrix},$$
$$s(x,\omega) = \begin{Vmatrix} G_{ik,l}(x,\omega) & \gamma_{i,k}(x,\omega) \\ \gamma_{k,l}(x,\omega) & g_{,k}(x,\omega) \end{Vmatrix},$$
$$L^1(x,\omega) = \begin{Vmatrix} C^1_{ijkl} & e^1_{kij} \\ e^{1T}_{kij} & -\eta^1_{ik} \end{Vmatrix},$$
$$F(x,\omega) = \begin{Vmatrix} u_{(i,j)}(x,\omega) \\ \varphi_{,i}(x,\omega) \end{Vmatrix} \quad (7)$$

In these expression, $G_{ik}(x,\omega)$, $\gamma_i(x,\omega)$, $g(x,\omega)$ can be represented as:

$$g(x,\omega) = \frac{1}{(2\pi)^3}\int g(k,\omega)\exp(-i\mathrm{k_g x})d\mathrm{k} \quad (8)$$

$$g(k,\omega) = \begin{Vmatrix} G_{ik}(k,\omega) & \gamma_i(k,\omega) \\ \gamma^T_k(k,\omega) & g(k,\omega) \end{Vmatrix} \quad (9)$$

Where $G_{ik} = (\Lambda_{ik} + \frac{1}{\lambda}h_i h^T_k)^{-1}$,

$$g = -(\lambda + h^T_i \Lambda^{-1}_{ij} h_j)^{-1},\ \gamma_i = \frac{1}{\lambda}h^T_k G_{ki}. \quad (10)$$

The algorithm is to propose the original data processing for many times, then use the effective information contained in the original data.

The study of the data mining algorithm has great significance in improving the effect of data processing. User data information needs to be extracted in the huge amounts of data, in order to promote the development of various fields. Cloud computing is a relatively new computing mode, its application in data mining also needs to do some further researches on the existing basis, continuously improving its application efficiency and improving data information of the application efficiency.

IV. EXPERIMENT RESULT

The mining module will be designed as in figure 4:

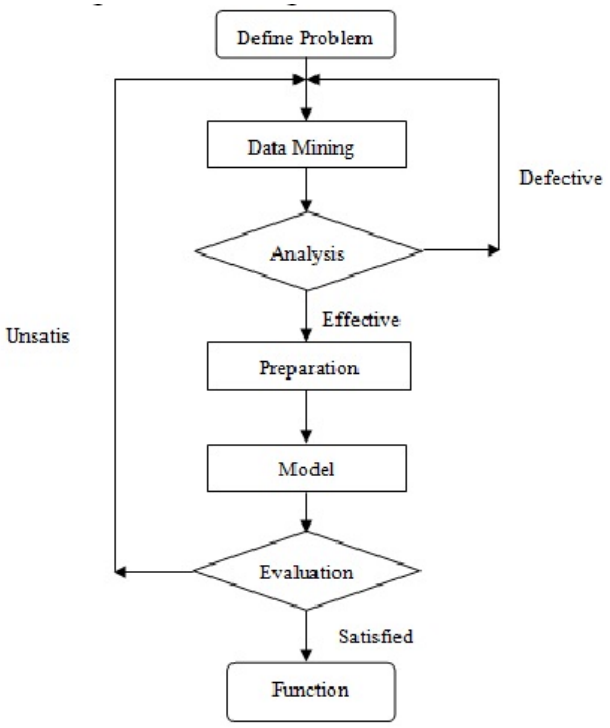

Figure 4. The mining module

Attention should be paid that the process of data mining is not a linear one, for it will need process repeatedly until a satisfactory result is got. For example, based on the problems that are observed in data analysis, data will be added or deleted and consequently a new data mining base is generated in this process.

This research mainly talks about how to achieve the multi-strata teaching goals by making teaching predication. The key point is to find the related courses which may affect the major courses study. This paper takes the CAT major course Network Database Application as example.

Mining data base is set up in the process of gradually perfection. Once the mining results functions well, changes of data will be needed to get a better result, and this process makes it necessary to repeatedly process data preparation and mining. The mined data will be stored in the same data base in this paper. Firstly, collect data for the mining data base. Data source will be "student score data base", "student personal information data base" and "enquiry data base of CAT majors".

Properties of different bases will be shown in the following Table 1.

The enquiry data base of CAT majors is mainly built through the online teaching system. In this system, students are required to submit information with their own student number and name; therefore, data mistakes will be avoided. To make this enquiry more objective, the survey list will include the part of student self-evaluation and teaching suggestions. Then, collect student records of multi-strata teaching from data source and generate StuDM.

Figure 3 shows the establishment of data mining base and figure 4 shows the sample of database.

TABLE I.
PROPERTIES OF DIFFERENT BASES

Data source	Student score database	Student personal information database	Enquiry database of CAT majors
Owner	Teaching affairs' office	Teaching affairs' office	Department
Storage	Excel	Access	SQL server
Physical storage	Server	Server	Server
Security	High	High	Low
Accessibility	High	High	Low
Privacy	High	High	Low

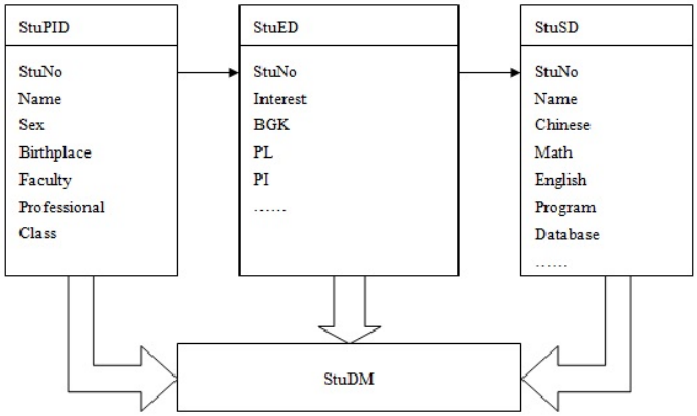

Figure 5. Establishment of data mining base

StuNO	IT	interest	BGK	sex	Pl	Chines	Math	Englis	Ph	CH	History
20010001	85	9	3	2	87	83	73	79	41	80	90
20010002	76	8	3	2	85	84	97	83	64	78	91
20010003	64	8	3	2	83	72	98	77	58	89	80
20010004	72	8	3	2	87	89	80	82	89	74	100
20010005	62	8	3	2	89	89	90	77	88	87	100
20010006	61	8	3	2	83	71	87	65	84	85	78
20010007	77	7	3	2	85	87	86	94	76	76	82
20010008	60	8	3	2	87	79	70	72	38	77	90
20010009	65	5	2	2	84	83	90	54	53	70	91
20010010	75	10	3	2	86	94	90	87	83	91	90
20010011	87	7	3	2	89	80	82	78	87	80	86
20010012	91	7	3	2	83	85	92	86	76	83	98
20010013	62	8	3	2	86	78	86	76	89	70	90
20010014	88	3	2	2	81	77	81	83	79	75	75
20010015	81	9	3	2	88	82	91	78	83	83	90
20010016	60	6	2	1	70	75	85	78	70	84	57
20010017	80	8	3	1	77	78	100	78	91	84	82

Figure 6. Sample of database

V. DISCUSSION

Students' scores cannot reflect their performance. Data needs to be processed to observe the regularity between teaching and the students1 performance. The detail operation will be substituting the scores with the ranking because the latter one is comparatively more scientific.

Data used in this research are from different recourses，such as teaching affair's department and students' affairs' office, which unavoidably include some unnecessary information. Consequently， careful selection of the data information is important. Information such as name, age, student number and some unusual data will be deleted.

Students5 scores comprise the main part of the data used in this research. Term exams can well reflect the students7 * 9 performance if questions in exam paper are scientifically organized.

In many cases, new variables are needed to generate as predictive variables form the original data and it is common that variables combined function much better than each individual. Many variables can become effective predicative ones by expanding their own scope. In the example of this paper, the students9 scores data base include terminal scores as well as the course designing scores etc. These different parts work together to get the general evaluation of this course.

Association rules are appropriate to be applied in finding out what courses are related to the Network Database Application for the rules concern about the relevance between values. Research needs to be done to decide to which extent these related courses affect the major one.

This paper makes full use of algorithm of extracting minimal rule-generating sets in research.

Being part of multi-strata teaching system, interactive module is responsible for making predication and receiving feedback. Based on the analysis result concluded from data mining module, teachers can separate students into different parts and get a very good idea about how much students can learn.

Apart from the above effect, instructions for students to make progress in major course through better performance in other related course. This module functions under the frame work of C/S, in which users submit students9 information and receive analysis results on the enquiry web page of Network Database Application.

VI. CONCLUSION

Data mining technology is commonly used in business, finance, manufacturing and marketing, while it is not made perfect use in teaching. This paper proposes a CAT oriented multi-strata teaching system based on the association rules. The system is instructive and will be a brave attempt in the application of mining technology in teaching area.

REFERENCES

[1] I. Jugo, B. Kovačić, and V. Slavuj, "Increasing the Adaptivity of an Intelligent Tutoring System with Educational Data Mining: A System Overview," *International Journal of Emerging Technologies in Learning,* vol. 11, no.3, pp. 67-70, March 2016. https://doi.org/10.3991/ijet.v11i03.5103

[2] H. Jing, "The Study on the Impact of Data Storage from Accounting Information Processing Procedure," *International Journal of Database Theory and Application*, vol. 8, no.3, pp. 323-332, June 2015. https://doi.org/10.14257/ijdta.2015.8.3.28

[3] N. Pereira,F. Ribeiro, G. Lopes, et al., "Jorge. Autonomous golf ball robot design and development," *The Industrial Robot,* pp. 396-402, 2012.

[4] Z. Zhang, D. He, L. Jing, et al., "Robot Arm Trajectory Planning Method," *Sensors & Transducers,* pp. 1621-1626, 2014.

[5] Z. Zhang, J. Tang, I. Huang, et al., "Research on Kinematics for Inhibition Fluttering of Robot Arm," *Sensors & Transducers*, pp. 161-171, 2013.

[6] L. Wang, L. Brodbeck, and F. Iida, "Mechanics and energetics in tool manufacture and use: a synthetic approach," *Journal of the Royal Society Interface*, pp. 111-114, 2014. https://doi.org/10.1098/rsif.2014.0827

[7] D. Zou, W. Zhang, W. Qiang, et al., "Design and implementation of a trusted monitoring framework for cloud platforms," *Future generations computer systems: FGCS,* vol. 29, no.8, pp. 2092-2102 , 2013. https://doi.org/10.1016/j.future.2012.12.020

[8] X. Zhang, et al., "Rotation-based privacy-preserving data aggregation in wireless sensor networks," *ICC 2014 - 2014 IEEE International Conference on Communications,* pp. 4184-4189, 2014. https://doi.org/10.1109/icc.2014.6883977

[9] J. Zheng, et al., "Auction-based adaptive sensor activation algorithm for target tracking in wireless sensor networks," *Future Generation Computer Systems,* vol. 39, no. 1, pp.88-99, 2014. https://doi.org/10.1016/j.future.2013.12.014

[10] Z. T. Li, et al., "A low latency, energy efficient MAC protocol for wireless sensor networks," *International Journal of Distributed Sensor Networks*, vol. 10, no. 6, pp.1-9, 2015. https://doi.org/10.1155/2015/946587

[11] H. Jing, "Routing optimization algorithm based on nodes density and energy consumption of wireless sensor network," *Journal of Computational Information Systems*, vol. 11, no.14, pp. 5047-5054, July 2015.

AUTHOR

Hong-yan He, Hui-ping Zhang, and **Hong-fang Luo** are with Hubei University of Technology, Engineering and Technology College, Wuhan, China (e-mail: 395127920@qq.com).

Submitted 09 September 2016. Published as resubmitted by the authors 18 October 2016.