International Journal of Database Theory and Application
Vol.9, No.8 (2016), pp.247-256
http://dx.doi.org/10.14257/ijdta.2016.9.8.23

Research and Application of Intelligent Recommendation System Based on Big Data Technology

Yongfeng Cui[1]*,Yuankun Ma[1] and Zhongyuan Zhao[2]

[1]*School of Science and Technology, Zhoukou Normal University, Zhoukou Henan 466001, China;*
[2]*College of Information Science and Engineering, Henan University of Technology, Zhengzhou, Henan 450001, China*
cuiyf@zknu.edu.cn

Abstract

With the rapid development of information technology and the Internet, people entered the era of information overload. Recommended system is an effective tool to solve the problem of information overload, it is based on the historical behavior of users and other records of interest to the user modeling, and then use the model to create user interest personalized recommendation, the interested user information, products. Online Intelligence is a new research direction, which integrates the latest achievements of artificial intelligence and information technology, greatly emphasizes on the Internet means intelligent application of data mining technology in the online intelligence research has a very important position. This paper presents a project-based collaborative filtering algorithm hierarchical similarity. Users to take advantage of some of the projects marked tags and project categories were automatically extended, to establish a hierarchy of all projects, and then use similar items tag hierarchy established between computing projects. Experimental results show that compared with traditional collaborative filtering algorithm, the ability of collaborative filtering algorithm based on similarity of item level can significantly improve the recommendation system to handle large data presented in this paper.

Keywords*: Collaborative filtering, matrix decomposition, data mining, online intelligence, intelligent recommendation systems, Bayesian network*

1. Introduction

Information age, with the extensive application of database technology continues to evolve and database management system [1]. The amount of data stored in the database increases sharply, in large amounts of data hidden behind a lot of important information, if that information can be extracted from the database and converted to a valid knowledge [2]. Companies and users will create a lot of potential value, data mining (Data Mining) concept is starting from such a business perspective born out [3]. With the growing popularity of Internet technology, the network is more widely used in people's lives, in all areas of learning, complex mass of information on the Internet for data mining technology provides a better application stage [4-5]. Since the great wealth of the interconnection line information resources, personalized information and services more and more people's attention. Gradually while people request access to information is no longer complete possession of information resources, but the resources to obtain useful information, and therefore the purpose of obtaining information from the recall into precision to meet specific user needs specific information to become information services in the new network environment targets [6].

ISSN: 2005-4270 IJDTA

Intelligent recommendation system is used on the Internet and data mining technology, to provide customers with intelligent, personalized service that can recommend products or provide information to customers, to guide customers targeted online merchandise, or other information were concerned, make the Internet from the past, "people looking for information" to "information to look for" smart stage [7]. Recommended forms include recommendations to customers of goods, providing personalized information services, issuing targeted messages and the like. Intelligent recommendation technology makes network services more personalized, policy-makers may be adjusted accordingly certain information to cater to different customers. One of the biggest researches is WEB mining data mining, and demand from all types of data for data mining internet presents new challenges, but also for the data mining technology to expand a new research platform [8].

In this paper, the algorithm for recommendation system scalability issues traitor research goal is to improve the recommendation system processing capacity data. Up to now, the collaborative filtering algorithm is the recommended system, preclude the use of the most extensive and most successful recommendation technology. Generated according to the recommended way, collaborative filtering algorithm can be further divided into collaborative filtering and heuristics-based collaborative filtering model. Among them, the collaborative filtering method due to its intuitive heuristics to achieve simplicity, robustness effect, and does not require lengthy training process to give full attention. Heuristic collaborative filtering main idea is to use the "neighbor" method, the target user for commodities score prediction relies on other users with the most similar ratings.

2. Research Status and Related Theory

2.1. Data Mining and Online Intelligence Technology

Data mining is also known as knowledge discovery in databases refers to the extraction of implicit, unknown, non-trivial and there is potential value of information or patterns from large databases or data warehouse, which is mainly used to study how massive amounts of data from extract useful knowledge for people, the integration of database, artificial intelligence, machine learning, statistical theory and technology and other fields [9-11]. Because many areas of knowledge involved, the wide range of applications, data mining has become one of the decade's most recent research and commercial organizations are areas of concern.

There are a variety of data mining methods, which are typical correlation analysis, sequential pattern analysis, classification analysis, cluster analysis. Data mining tools can analyze historical information innovative, and future trends and predict behavior, which is well supported by the people's decision. Data mining technology since the date of birth has been being constantly applied to all kinds of new areas, in terms of commerce, banking, insurance, securities, telecommunications, biotechnology and others have a wide range of applications, particularly in market forecasts and sales process among data mining can help businesses save costs, reduce waste, increase sales while also improving the quality of service to customers also benefit the greatest degree [12].

2.2. The Main Technical Methods of Data Mining

Data mining is an extract useful information "data generation" process, it dig out from a large number of incomplete, noisy, fuzzy or random data implicit, previously unknown, potentially useful for decision-making nontrivial knowledge and rules, and according to the available information on the behavior did not occur to make the results forecast for the business decision-making, provide the basis for market planning. Data mining can help companies discover business trends, explain known facts, predict future results, and help companies identify the key factors needed to complete tasks in order to increase revenue, reduce costs and make the company in a more favorable competitive advantage.

Data Mining in just ten years in different areas of the development process has been successfully used in banking, insurance, retail, customer relationship management, measurement, a variety of practical and research systems appeared in large numbers. Web data mining and data mining technology, deep application of data mining technology, the latest hotspot. Data mining has gradually entered the practical stage, but also more and more into everyday life for people to bring a variety of services and convenience. Due to the various methods has its own features and application areas (see Table 1), select the data mining technology will affect the quality and effectiveness of the final result is usually a variety of techniques will be used in combination to create complementary advantages.

Table 1. Comparison of Main Data Mining Technologies

Technical method	Main functions and features	Major application areas
Correlation analysis	Clustering classification	Retail, insurance, manufacturing *etc.*
Decision tree	Inductive classification, easy understand	Retail, manufacturing, medicine etc
Genetic algorithm	Clustering, optimization, efficiency	Finance, insurance, agriculture *etc.*
Bayesian network	Classification, clustering, prediction, easy to understand	Medicine, manufacturing, telecommunications *etc.*
Rough set method	Uncertainty classification	Retail, finance, manufacturing *etc.*
Neural network	Forecasting classification and clustering, poor interpretation	Finance, insurance, manufacturing *etc.*
Statistical analysis	Clustering, the result is accurate, easy to understand	Finance, manufacturing, medicine *etc.*

2.3. Recommended System Concept

Recommended system role is to help users more easily and quickly get the information they are interested, therefore, recommended systems and search engines, it is an automated push system does not require users to take the initiative to describe information of interest to it, it users only need to use when accessing the website recorded behavior, the use of these behaviors are recorded to build user interest model, and the model's looking for projects to build the user may be interested in, and then make recommendations.

Recommended system has three main modules: user modeling module, recommended object modeling module, a recommendation algorithm module. General recommendation system model flow shown in Figure 1. Recommended system first user based on the user behavior data modeling interests, generating user interest model, while the recommendation of the object data modeling recommendation object, and then matching user needs information of interest in the model and recommendation object model feature information, using the appropriate algorithm to calculate the recommended hoof alternative recommendations objects, and finally generate a list of recommended target users are most likely interested.

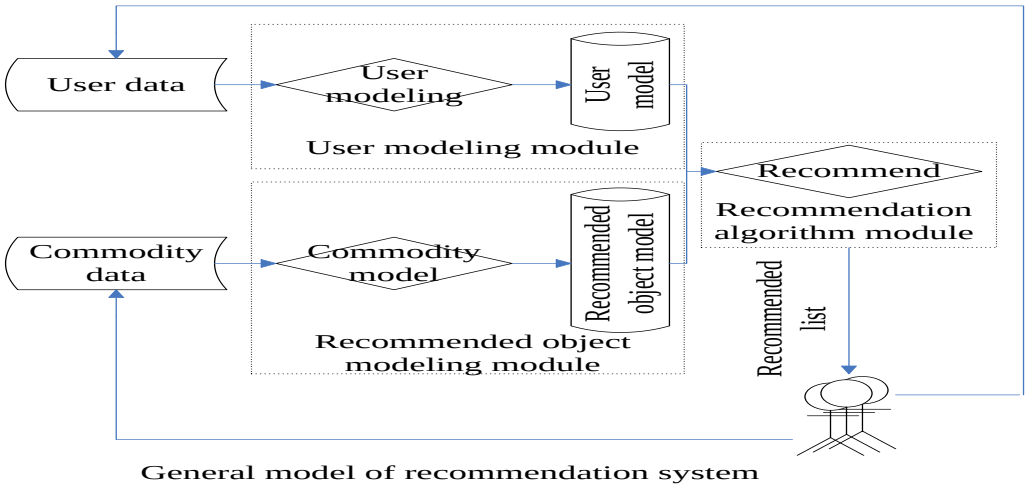

Figure 1. General Recommendation System Model Flow

3. Intelligent Recommendation System

3.1. E-Commerce Development and Demand

Fundamental goal of e-commerce is in the compression operating costs, to bring profits while giving customers greater convenience and freedom of choice. Today, e-commerce has penetrated into all walks of life in society in recent years as the Internet bubble broken, Internet companies and large-scale application of the Internet to expand their business in traditional industries company more emphasis on e-commerce among the returns, which for e-commerce maturing and development of a better external conditions.

E-commerce through the Internet to achieve business, traders and consumers of online shopping, online transactions and online electronic payment a new business model differs from traditional business operations, e-commerce with the development of INTERNET and growing up, mainly through electronic data interchange (EDI) and INTERNET. Figure 2 is the world's leading e-commerce site Amazon sales page.

Figure 2. The World's Leading E-Commerce Site Amazon Sales Page

3.2. Content-Based Recommendation

Formalized as, provided the project j kj i-th keyword, the keyword i wij is the weight in the right project i, the content feature project i can be expressed as follows vector model:

$$Content(j) = (\omega_{1j}, \omega_{2j}, \ldots\ldots) \tag{1}$$

Recommendation algorithm based on the recommended content to the user and their favorite project before a similar project on the content, therefore, needs to be of interest to the user's preferences modeled based on previous user's favorite items. Order ContentBasedProfile (M) expressed interest in user m interested vector, the vector is defined as follows:

$$ContentBased\ Profile = \frac{1}{|N(u)|}\sin_{d \in N(u)} Content(d) \tag{2}$$

User preferences defined project Content Based Profile (w) vectors and vector ContentO similarity:

$$p(u, j) = \sin(ContentBasedProfile(u), Content(j)) \tag{3}$$

Which sim (*, *) defines the similarity between two vectors. There are many ways to calculate the similarity, the method commonly used is the cosine of the angle between distance vector.

Recommended content based on commodities and commodity only consider the similarities between, regardless of the user behavior, so that it can be resolved in a collaborative filtering algorithm in new user questions, score data sparse and other issues, but also through the lists of recommended items content feature, very intuitive explanation recommended reason. In addition, based on the recommended content may also be off-line modeling, offline calculating the similarity between the goods, so that the user can quickly make a recommendation.

3.3. Collaborative Filtering Recommendation

Collaborative filtering recommendation system in the field is one of the most successful technology, which in the 1990s began research traitor and promote the prosperity of the whole recommendation system research. Up to now, the collaborative filtering recommendation system in the industrial sector is still the most widely used of personalized recommendation system.

Given user M and V, so is the collection of items like user U, V is user-friendly collection of items, so interest in M, V similarity can be simply described by the following equation:

$$s_{uv} = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|} \tag{4}$$

Formula (4) is calculated according to the formula similarity with the target user for each user, to take the highest similarity as the nearest neighbors several users. Users use neighbors to predict the target user for project evaluation method There are many, mainly include the following three methods, and briefly. Suppose now to predict the user's score on the project M, let S similarity with the user set of users is relatively high, the functional form of the forecast are:

$$r'_{u,i} = k\sum_{v \in S} sim(u, v) \cdot r_{v,i} \tag{5}$$

$$r'_{u,i} = \overline{r}_u + k\sum_{v \in S} sim(u, v) \cdot (r_{v,i} - \overline{r}_u) \tag{6}$$

Assumptions based on collaborative filtering and recommendation of different users, based on collaborative filtering assumptions of the project, a user will like his previous

projects like similar projects, based on the content of this recommendation and is somewhat similar, but they are different is that content-based when the project is calculated using the similarity between content items based on project characteristics neighbor computing project itself, and project-based collaborative filtering is based on user evaluation of the project to calculate the similarity between projects, therefore, the essential difference between the two is different data.

4. Experiment and Analysis

4.1. System Design

With the rapid development of e-commerce technology and J2EE technology, more and more J2EE-based e-commerce site architecture is proposed. These e-commerce system uses a three-layer structure shown in Figure 3, including the business logic layer, presentation layer and the data layer (enterprise information layer).

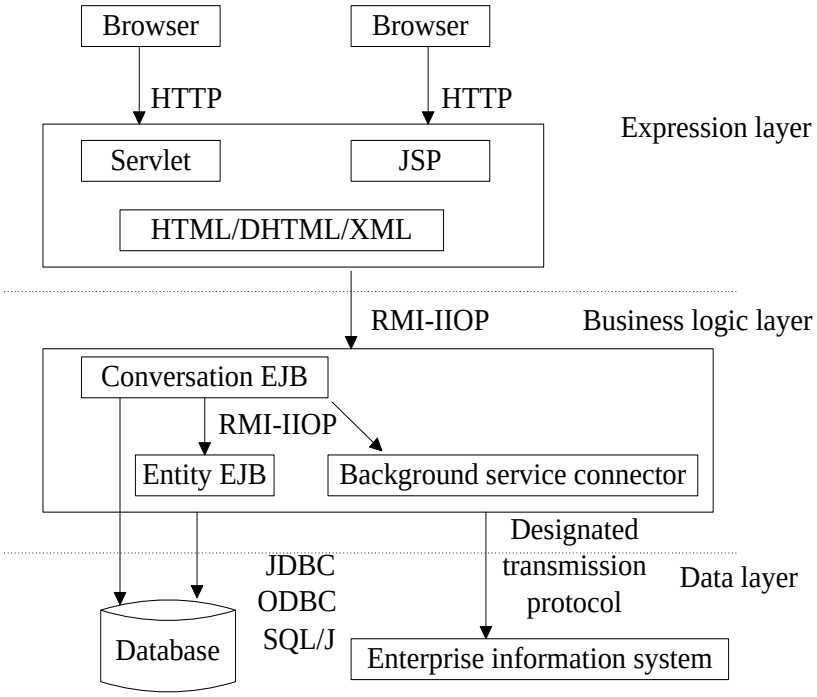

Figure 3. The E-Commerce System Uses Three-Layer Structure

Expression layer contains one or more Web servers, each of which will affect the end user. Presentation layer displays the requested information to end users through HTML, it can read and interpret the user's product selection and calls the business layer EJB components, the presentation layer via RMI- communication, presentation layer implementation would IIOP and business logic layer use Java Servlet, JSP or any other Java application implementation.

Standard structure alarm network was shown in Figure 4. Input node network were among the patient's breathing, pulse, temperature, allergic reactions, blood volume reduction and environmental conditions and other physical parameters, the output node is the rescue, treatment of patients should be taken, and so on. This can be seen by the network structure is a sparse graph, as in this instance, as sparse Bayesian networks are among the most practical lives of maps, such as graphic representation of conditional independence more information.

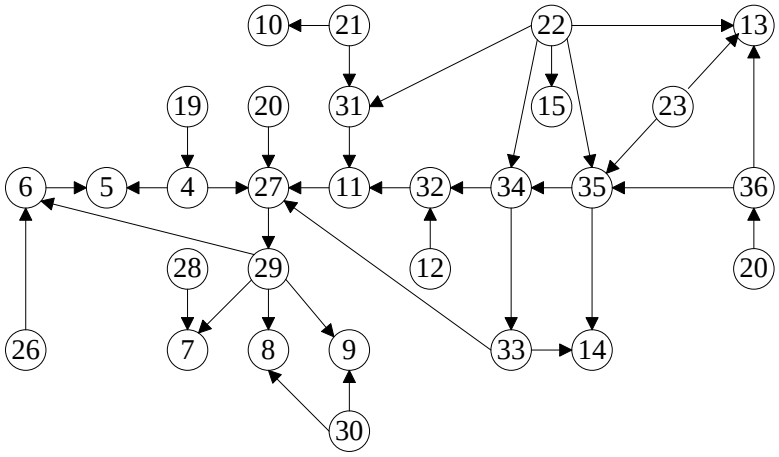

Figure 4. Standard Structure Alarm Network

4.2. Forecast Accuracy

Forecast accuracy is a quantitative indicator, the main recommendation algorithm used to measure the user's ability to predict behavior. Usually we need to get off the experiment. First, the historical behavior of the user data preprocessing into an offline data set, and then generate a separate set of offline data into training and test sets, use the recommended method to establish the recommended model in the training set, and finally recommended the use of the model created in the test predict the user's behavior on the set, and calculate the predicted fit user behavior and user's behavior between the goodness of fit is the prediction accuracy.

Recommended by the system and the number of customers end up purchasing goods and recommendation set among the total number of goods ratio, called the accuracy of the recommended goods, namely:

$$\text{Precision} = \frac{\left|E_{buy} \cap R\right|}{\left|R\right|} \tag{7}$$

The ratio of the total number of goods recommended by the system and the number of customers end up purchasing goods and purchased by the user, called the coverage recommendation of goods, namely:

$$\text{Coverage} = \frac{\left|E_{buy} \cap R\right|}{\left|E_{buy}\right|} \tag{8}$$

Description system accuracy recommendation engine generates the correct recommended level set, and coverage metrics recommendation engine producing ability may be purchased by all users of goods. These two measures is to evaluate the effectiveness of e-commerce recommendation system essential commodities technical indicators. Because e-commerce marketing, the low accuracy rate is very susceptible to receiving an incorrect user recommended product of anger or frustration, and low coverage will result in users roaming disoriented at the site, so that businesses lose cross-selling (cross-selling) opportunities. At the same time these two parameters also mutual restraint, high accuracy will certainly lead to drop coverage, and high coverage will naturally make accurate rate decreased, so the combination of these two parameters can recommend a reasonable result comprehensive evaluation.

When sampling the sample is small, the probability value obtained may be inaccurate, especially when the parent node are more likely to attribute values and node more time (Table 2). When learning conditional probability tables from the sampler samples may be

introduced prior probability (which may be designated or experts drawn from the previous test), prior to a priori probability weights, when statistical computing with a new training set, based on the weight of the original conditional probability corrected, so that the new training set larger role prior probability is smaller.

Table 2. Get CPT According to the Sampling Data

A1	A2	A3	A4	A5	Sample quantity
True	True	True	True	True	10
False	True	False	True	False	200
False	True	False	False	False	100
False	False	False	True	False	20
True	True	False	True	False	500
True	True	False	False	False	50

Since the recommendation system scoring matrix is usually very sparse, and the degree of sparse data have a significant impact on the efficiency TopKS algorithm, this paper introduces the concept of sparsity, which is defined as the user entry H scoring matrix, users of the project's ratings the percentage of the total number of user entries H (ie the number of users and the number of product items H).

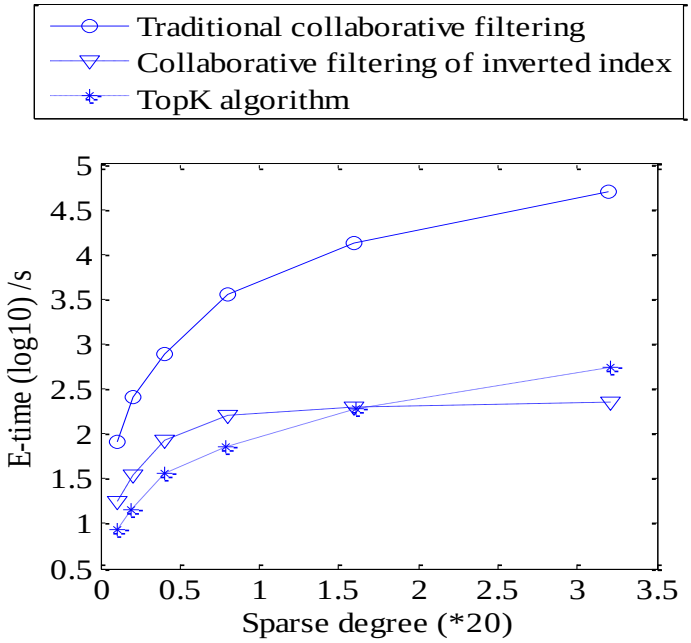

Figure 5. Topks Algorithm and Traditional Traversal Algorithm

Figure 5 shows, TopKS algorithm is better than the traditional traversal algorithm, there is always 10 times more efficiency. The more sparse data sets, efficiency is more obvious. But when the sparsity above a certain value, the use of inverted index method is slightly better than TopKS algorithm, which is due to the time consuming process of the algorithm determines a greater cause. However, due to the general user can score the number is limited, and as the number of users and projects continued rapid growth, data collection sparsity recommendation system are much lower than 0.1, so TopKS in the practical application still can play good results.

4. Conclusions

With the rapid development of information technology and the Internet, people by the lack of information age into the age of information overload. Traditional network applications such as portals, search engines and professional data index cannot meet people's individual needs, because they are essentially a means to help users to filter the information, without taking into account individual needs, therefore, these traditional network application tools We cannot solve the problem of information overload. Recommended system is an effective tool to solve the problem of information overload, it is based on the historical behavior of users and other records of interest to the user modeling, and then use the model to create user interest personalized recommendation, the interested user information, products, *etc.* recommended to the user. And search engine compared to the recommended system through the user's interest in research hobbies, personalize calculated by the system users find points of interest, in order to guide users to find their own information needs.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (Grant no. U1504602), the Key Research Projects of Universities in Henan Province for contract 15A520035 and 15A520124, the Science and Technology Plan Projects of Henan Province for contract 162102310614, under which the present work was possible.

References

[1] N. K. Roy, W. D. Potter and D. P. Landau, "Polymer property prediction and optimization using neural networks", IEEE Transactions on Neural Networks, vol. 17, no. 4, **(2006)**, pp. 1001-1014.
[2] L. Wang, J. C. Fang and Z. Y. Zhao, "Application of backward propagation network for forecasting hardness and porosity of coatings by plasma spraying", Surface & Coatings Technology, vol. 201, no. 9-11, **(2007)**, pp. 5085-5089.
[3] L. C. K. Liau, T. C. K. Yang and M. T. Tsai, "Expert system of a crude oil distillation unit for process optimization using neural networks", Expert System with Applications, vol. 26, no. 2, **(2004)**, pp. 247-255.
[4] R. Babuska, H. B. Verbruggon and H. J. L. Bancan, "Fuzzy modeling of enzymatic penicillin-G Anversion", Engineering Application of Artificial Intelligence, vol. 12, no. 1, **(1999)**, pp. 79-72.
[5] L. C. Jun, P. Binay and L. J. Min, "Stochastic nonlinear optimization for robust design of catalysts", Industrial and Engineering Chemistry Research, vol. 50, no. 7, **(2011)**, pp. 3938-394.
[6] P. T. Boggs and J. W. Tolle, "Sequential quadratic programming for large-scale nonlinear optimization", Journal of Computational and Applied Mathematics, vol. 124, no. 1, **(2000)**, pp. 123-137.
[7] P. Zikopoulos and C. Eaton, "Understanding big data: Analytics for enterprise class hadoop and streaming data", McGraw-Hill Osborne Media, **(2011)**.
[8] S. Chen, W. Li and M. Li, "Latest Progress and Infrastructure Innovations of Big Data Technology", Cloud Computing and Big Data (CCBD), 2014 International Conference on. IEEE, vol. 2014, **(2014)**, pp. 8-15.
[9] S. LaValle, E. Lesser and R. Shockley, "Big data, analytics and the path from insights to value", MIT sloan management revie., vol. 21, **(2013)**.
[10] T. J. Hui and Z. Quan, "Design and implementation od the crying voice detection circuit in the baby's supervision system", Review of Computer Engineering Studies, vol. 1, no. 1, **(2014)**, pp. 13-16.
[11] W. Yao and S. Qin, "Aircraft diagnosis by solving map exactly", Review of Computer Engineering Studies, vol. 2, no. 1, **(2015)**, pp. 1-8.
[12] N. F. Xie, Z. F. Zhang and W. Sun, "Research on Big Data Technology-Based Agricultural Information System", International Conference on Computer Information Systems and Industrial Applications. Atlantis Press, **(2015)**.

Authors

Yongfeng Cui, received the BS degree in Computer Science and Technology from Henan Normal University and the MS degree in Computer Application Technology from Huazhong University of Science and Technology, China in 2000 and 2007 respectively. He is currently researching on Computer Application Technology (CAT). (E-mail: cuiyf@zknu.edu.cn)

Yuankun Ma, received the BS degree in South-Central University for Nationalities and the MS degree in Shandong University of Science and Technology, China in 2011 and 2014 respectively. He is currently researching on Machine Learning and Big Data Analytics. (E-mail: mayk@zknu.edu.cn)

Zhongyuan Zhao, Graduate student, Henan University of Technology. Received the BS degree in computer science and technology from Zhoukou Normal University, China in 2009. He is currently researching on Computer Application Technology (CAT). (E-mail: zhaozy@zknu.edu.cn)