# Personalized Recommendation Method of Power Information Operation and Maintenance Knowledge Based on Spark

Zhaoyang Qu[1], Pengfei Xu[1], and Youxue Ren[2]
[1]School of Information Engineering of Northeast Dianli University, Jilin 132012, China
[2]Jilin Power Supply Company of Jilin Power Company, Jilin 132000, China
Email: qzywww@mail.nedu.edu.cn; {893193963, 752953593}@qq.com

***Abstract*** —Power information operation and maintenance knowledge overload has become a pressing issue with the development of smart grid construction. The traditional personalized recommendation method cannot meet the demand of personalized recommendation of power information maintenance knowledge in big data environment. This paper proposes a method based on Spark which gives a personalized recommendation method of power information operation and maintenance knowledge. Firstly, an implicit rating mechanism is introduced, which can transform the learning behavior of users into implicit rating of power information operation and maintenance knowledge. Secondly, a personalized recommendation method combing knowledge features and user interests is designed. Finally, the personalized recommendation method, based on Spark, is applied to recommend power information operation and maintenance knowledge. The experimental results show that the method can effectively improve the accuracy and real-time of recommendation.

***Index Terms***—Spark, power information operation and maintenance knowledge, personalized recommendation, implicit rating, collaborative filtering

## I. Introduction

Power information operation and maintenance knowledge overload has become a pressing issue with the development of smart grid construction, which has made it difficult for users to find the knowledge that they really need from a great deal of power information operation and maintenance knowledge [1]. An effective way to enhance the power information operation and maintenance knowledge level of users is through personalized recommendation of power information operation and maintenance knowledge. This has great significance in ensuring the safe and stable operation of power enterprise information communication system.

Collaborative filtering is widely used in e-commerce, social networking, video/music on demand and other fields, and is currently the most successful personalized recommendation method [2]. It recommends the knowledge that neighbor users are interested to target users by looking for neighbor users whose habits and preferences are similar to target user. Many researchers have done intensive researches on collaborative filtering and have made many contributions in recent years. Reference [3] proposed a memory based collaborative filtering algorithm via propagation. Reference [4] proposed a recommendation of algorithm-combing item features and trust relationship of mobile users. Reference [5] proposed a collaborative filtering recommendation algorithm based on time weight and user feature. The traditional personalized recommendation method has the bottleneck of insufficient computing power, low processing efficiency and so on, which cannot meet the demand of personalized recommendation of power information maintenance knowledge in the big data environment.

The distributed or parallel algorithm can improve the efficiency of traditional collaborative filtering algorithm. Cloud computing has the advantages of high reliability, massive data processing, extendibility, and high equipment utilization, which has become the foundation of big data processing technology [6], [7]. Presently, cloud computing has been adopted to explore collaborative filtering algorithm by many industries. Reference [8] proposed an improved collaborative filtering recommendation algorithm based on hadoop. Reference [9] designed a recommendation system for E-commence based on hadoop. Collaborative filtering has a large number of iterative computation and I/O operations on the hadoop platform, which seriously reduces the performance of parallel computing. However, distributed processing framework Spark can make full use of cluster memory, which enhances the ability of rapid processing and analysis [10]. Therefore, Spark provides a new technical idea for big data processing.

In order to address issues of power information operation and maintenance overload in big data environment, a personalized recommendation method of power information operation and maintenance knowledge based on Spark is proposed in this paper. Importantly, an implicit rating mechanism is introduced. Then, a personalized recommendation method that combines knowledge features and user interests is designed. Lastly, the personalized recommendation method based on Spark

Manuscript received May 3, 2016; revised August 22, 2016.
This work was supported by the National Natural Science Foundation of China under Grant No.51277023, and the Science and Technology Development Plan of Jilin Province under Grant No.20140307008GX.
Corresponding author email: 893193963@qq.com.
doi:10.12720/jcm.11.8.785-791

is applied to recommend power information operation and maintenance knowledge.

## II. Implicit Rating Mechanism

Authentic and reliable rating data is the premise of personalized recommendation of power information operation and maintenance knowledge. Thus, the rating data should reflect the degree of user interests in power operation and maintenance knowledge as much as possible. Users exhibit certain learning behavior on power information operation and maintenance knowledge in the process of online learning. For example, power information operation and maintenance knowledge will be downloaded, collected, shared, and learned. Therefore, an implicit rating mechanism is introduced to track the learning behavior of users, and the learning behavior will be converted to implicit rating of power information operation and maintenance knowledge. The implicit rating can better solve the problems that users learn power information operation and maintenance knowledge, although they cannot rate, and objectively reflect the level of user interests in power information operation and maintenance knowledge with higher reliability than prediction [11].

The implicit rating can be obtained by calculating the scores of single learning behavior or combined learning behavior of users. D (Download), C (Collection), S (Share), and T (Learning Time) are used to express the learning behavior of users in this paper, and the rating value of power information operation and maintenance knowledge uses 5 points rule. The corresponding rating value of learning behavior is shown in Table I.

Table I The Corresponding Rating Value of Learning Behavior

| Learning Behavior | Rating | Learning Behavior | Rating |
|---|---|---|---|
| D | 2 | C+T | 3 |
| C | 2 | S+T | 3 |
| S | 2 | D+C+S | 4 |
| T | 2 | D+C+T | 4 |
| D+C | 3 | D+S+T | 4 |
| D+S | 3 | C+S+T | 4 |
| D+T | 3 | D+C+S+T | 5 |
| C+S | 3 | | |

## III. Recommendation Method Combing Knowledge Features and User Interests

Collaborative filtering is currently the most successful personalized recommendation method. It can be divided into two categories: collaborative filtering based on storage and collaborative filtering based on model [12]. The collaborative filtering algorithm based on storage looks for neighboring users whose habits and preferences are similar to target user by analyzing historical rating data, and then recommends the knowledge that neighboring users are interested to target user. The implementation of traditional collaborative filtering algorithm is divided into three steps: obtaining the user rating data; looking for the nearest neighbors; generating the recommendation list [13].

A personalized recommendation method combing knowledge features and user interests is designed in this paper by considering the correlation features of knowledge and the dynamic change of user interests in the personalized recommendation process of power information operation and maintenance knowledge. That is to say, knowledge correlation is introduced in the process of user similarity calculation and time function is introduced in the process of unrated knowledge prediction.

### A. User Similarity Calculation

The traditional collaborative filtering algorithm only considers the knowledge that users have commonly rated, but neglects the correlation between knowledge in the process of user similarity calculation [14]. Therefore, the knowledge that users have commonly rated may exist unrelated knowledge, resulting in inaccurate results. Knowledge correlation is introduced in the process of user similarity computation in order to reduce the interference of unrelated knowledge to user similarity calculation. There are many different methods to calculate knowledge correlation, such as cosine similarity, Pearson correlation coefficient, and conditional probability [15]. Conditional probability is used to calculate the knowledge correlation as follows.

$$sim(i, j) = \frac{P(i|j)}{Freq(i)^{\alpha}} = \frac{freq(ij)}{freq(j) + (freq(i))^{\alpha}} \quad (1)$$

where $freq(i)$ represents the number of users with rating knowledge $i$, $freq(ij)$ represents the number of users with both rating knowledge $i$ and $j$, $\alpha \in [0,1]$ represents the scaling factor.

The calculation formula of user similarity is as follows when knowledge correlation is introduced.

$$\mathrm{sim}(u,v) = \frac{\sum_{i \in I_{uv}} (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v}) \cdot sim(i, i_T)}{\sqrt{\sum_{i \in I_{uv}} (r_{u,i} - \overline{r_u})^2 \cdot sim(i, i_T)} \sqrt{\sum_{i \in I_{uv}} (r_{v,i} - \overline{r_v})^2 \cdot sim(i, i_T)}} \quad (2)$$

where $sim(u,v)$ represents the similarity between user u and v based on $i_T$, $i_T$ represents the unrated knowledge.

### B. Unrated Knowledge Prediction

The traditional collaborative filtering algorithm regards different time rating value of users as the same in the process of unrated knowledge prediction, but does not account for the fact that users rate knowledge in different time [16]. As user interests change over time, and it changes small in a relatively short period of time. Therefore, users are most likely interested in recent rating knowledge. Psychologist Ebbinghaus's research on the phenomenon of forgetting demonstrates that the

forgetting process of human beings is gradual and nonlinear [17]. Using the human forgetting rule, the time function is introduced in the process of unrated knowledge prediction, which can reflect the impacts of different time rating value of users on unrated knowledge. The more current the rating knowledge, the greater impact on unrated knowledge prediction. Thus, the time function should be a monotonically increasing function and function value is between 0 and 1. The exponential function is used as the time function in this paper.

$$f(t_{ui_T}) = 1/(1+\exp(-t_{ui_T})) \tag{3}$$

where $f(t_{ui_T})$ presents the time function value, $t_{ui_T}$ presents the rating time of user, $e$ presents the natural background.

The prediction formula of unrated knowledge when time function is introduced is as follows.

$$P_{u,i_T} = \overline{r}_u + \frac{\sum_{v \in N_u} \mathrm{sim}(u,v)(r_{v,i_T} - \overline{r}_v) \cdot f(t_{ui_T})}{\sum_{v \in N_u} \mathrm{sim}(u,v) \cdot f(t_{ui_T})} \tag{4}$$

where $P_{u,i_T}$ presents the rating given by target user u to unrated knowledge $i_T$.

## IV. Personalized Recommendation of Power Information Operation and Maintenance Knowledge Based on Spark

### A. Distributed Processing Framework Spark

Spark is a distributed processing framework based on memory computing, whose aim is to carry out a fast processing and analysis of big data. In order to overcome the shortcomings of hadoop in iterative calculation, Spark introduces the technology of memory computing. The advantage is that data sets are cached in memory and data are read directly from the memory, which reduces the disk I/O operations and enhances the processing speed.

Spark improves efficiency by using resilient distributed data sets [18] (RDD) in cluster computing. RDD is a type of parallel data structures, based on distributed memory, which can store data in memory and control the partition to optimize data distribution. RDD is a read-only partition sets that can be shared among multiple computing applications. RDD not only supports the application based on data sets but also has fault tolerance, local computing scheduling, and scalability [19].

The running architecture of Spark is shown in Fig. 1. Spark applications run on different nodes in the cluster with an independent executor, and the SparkContext object is used to carry out the overall scheduling in the main program. SparkContext can be connected with three cluster resource managers, such as Standalone, Mesos, or Yarn. The role of cluster resource manager is to allocate resources for different Spark applications. Spark needs to send application code to the executor of worker node to execute task in the process of executing program, whose role is to achieve the data localization calculation.

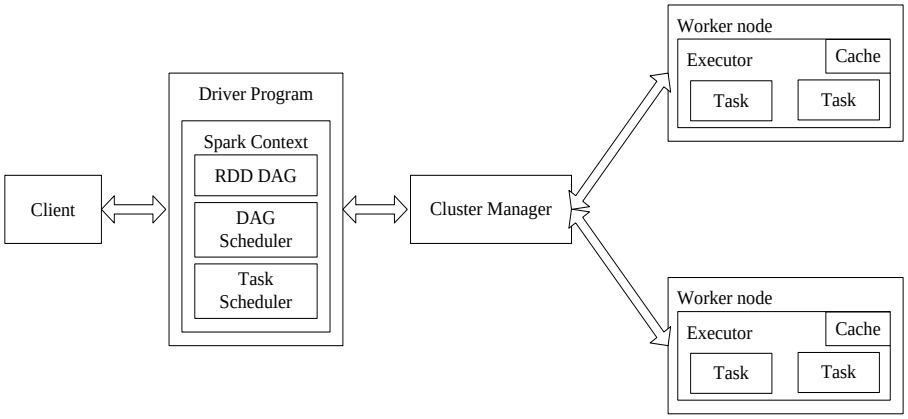


Fig. 1. The running architecture of Spark

### B. Personalized Recommendation of Power Information Operation and Maintenance Knowledge

According to the previous papers, a personalized recommendation method of power information operation and maintenance knowledge based on Spark is built in this paper, which is applied to power information operation and maintenance knowledge training platform of actually running. The process of personalized recommendation of power information operation and maintenance knowledge is shown in Fig. 2.

Recommendation engine is the central of personalized recommendation of power information and maintenance knowledge. The basic idea is that the information of users, power information operation and maintenance knowledge, and rating are transferred to the recommendation engine through the input interfaces. The personalized recommendation method combing knowledge features and user interests based on Spark proposed in this paper is applied to recommend power information operation and maintenance knowledge in recommendation module. The recommended results are presented to users through the output interfaces.

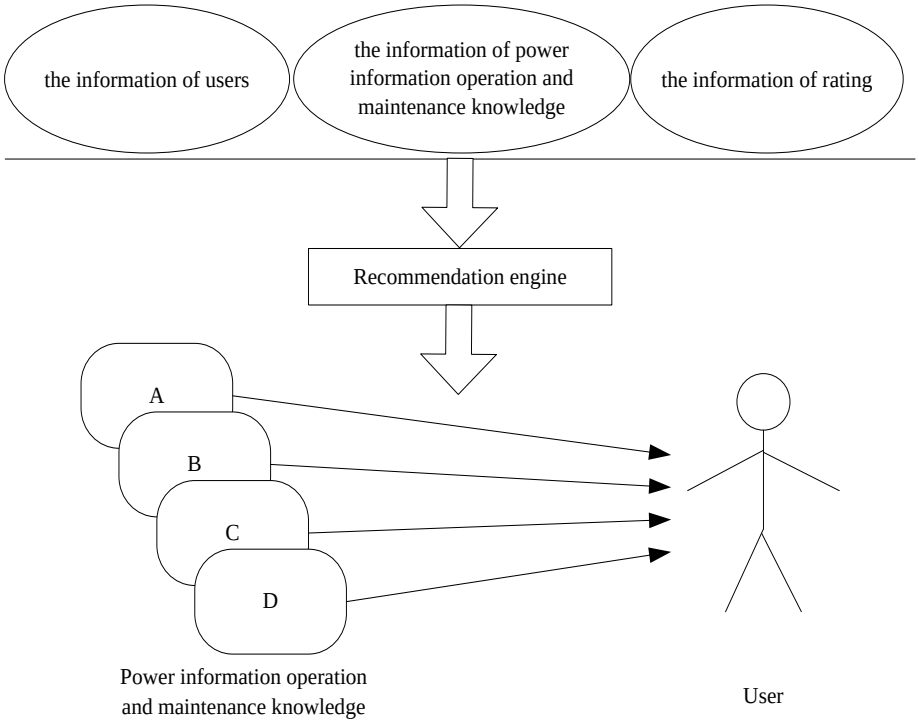


Fig. 2. Process of personalized recommendation

The parallel implementation of personalized recommendation method combing knowledge features and user interests based on Spark has also adopted the idea of "map" and "reduce" on the whole. However, the biggest difference from the MapReduce of Hadoop is that all the iterative calculation are done from the Spark memory, and the middle results do not need to interact with the disk. Its implementation process is as follows.

```
    Input: userId, knowledgeId, rating, timestamp
    Output: topNKnowledge
① sc=new SparkContext(arg[1], "Recommendation")
    lines=sc.textFile(arg[2])
    Partitions = P
② user_knowledge=lines
      .parallelize(0 until P)
      .map(parseVectorOnKnowledge)
      .groupByKey()
      .map (lambda x: sampleInteractions(x[0], x[1], 500))
      .cache()
③ pair_users=user_knowledge
      .filter(lambda  x:len(x[1])>1)
      .map(lambda x:findUserPairs(x[0], x[1]))
      .groupByKey()
    user_sims=pair_users
      .map(lambda  x:calculateSim(x[0], x[1]))
      .map(lambda  x:keyOnFirstUser(x[0], x[1]))
      .groupByKey()
      .map(lambda x:nearestNeighbors(x[0], x[1], 50))
④ user_knowledge_history=lines
      .map(parseVectorOnUser)
      .groupByKey()
      .collect()
   user_knowledge_dict={}
    for(user knowledge) in user_knowledge_history :
     user_knowledge_dict[user]=knowledge
     ukb=sc.broadcast(user_knowledge_dict)
⑤ user_knowledge_rec=user_sims
      .map(lambda x:topNRec(x[0], x[1], ukb.value, 50))
      .collect()
```

## V. Experiments and Example Analysis

### A. Experimental Environment

Experimental platform uses one Huawei server, which is configured as follows: 2XIntel (R) Xeon (R) E5-2620 V2 2.10GHz CPU, Memory 128GiB, 1000Mbit/S card, 500TB hard disk. Five CentOS virtual machines are deployed in the Huawei server, which contains a master and four slavers. The hardware configurations of each virtual machine is as follows: master has 12g memory, 2.0 core processor, 1T hard disk; each of the slaves has 8g of memory, a 2.0 core processor, and a 1T hard disk. The version of Hadoop is 2.6.0, the version of JDK is 1.7, and the program of Hadoop was written in Java language. The version of Spark is 1.0.0, the version of Scala is 2.10.4, and the program of Spark was written in Scala language.

Experimental data sets use implicit rating data, which is generated by power information operation and maintenance knowledge training platform of actually running. The data sets contain 400 users 100,000 ratings for 2,300 knowledge, and the rating score value is between 1 and 5, which shows the degree of user interests in power information operation and maintenance knowledge. In order to make the experiment more persuasive, the data sets are randomly divided into training set and test set with the ratio of 80% and 20%.

The composition of knowledge in data sets is shown in Table II.

TABLE II. THE COMPOSITION OF KNOWLEDGE IN DATA SETS

| Knowledge Type | Knowledge Quantity |
|---|---|
| Information network management | 470 |
| Information security protection | 490 |
| Host operation and maintenance | 450 |
| Database management | 460 |
| Desktop management | 430 |

## B. Example Analysis

### 1) Accuracy of recommendation

In this experiment, the personalized recommendation method combing knowledge features and user interests are compared with the traditional personalized recommendation method in order to test the accuracy of recommendation.

The study adopts mean absolute error MAE, which is easy to understand and calculate within statistical accuracy measurement methods, as the standard of accuracy of recommendation. The smaller MAE is, the more accurate recommendation achieves [20]. Setting the prediction rating sets of target user as $\{r_1, r_2, \cdots, r_n\}$, corresponding real rating sets as $\{s_1, s_2, \cdots, s_n\}$, the MAE is defined as follows.

$$MAE = \frac{\sum_{i=1}^{n} | r_i - s_i |}{n} \tag{5}$$

The number of nearest neighbors influences MAE. Therefore, we choose different size of nearest neighbors from the same data sets. The number of nearest neighbors increases from 10 to 45 with an interval of 5. The experimental result is shown in Fig. 3.

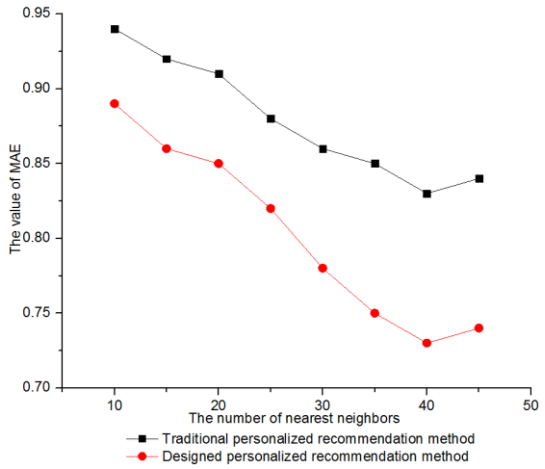


Fig. 3. The influence of nearest neighbors on the value of MAE

From Fig. 3, we can see that the value of MAE of personalized recommendation method combing knowledge features and user interests are smaller than traditional personalized recommendation method. The reason is that the knowledge correlation is introduced in the process of user similarity calculation, which reduces the interference of unrelated knowledge. The time function is introduced in the process of unrated knowledge prediction, which reflects the dynamic changes of user interests. Thus, the accuracy of recommendation is improved.

### 2) Real-time of recommendation

In this experiment, the personalized recommendation method combing knowledge features and user interests realizes on Spark and Hadoop platform in order to test the real-time of recommendation.

The study adopts running time, which is easy to understand and test, as the standard of real-time of recommendation. The smaller running time is, the higher real-time recommendation achieves. The experimental result is shown in Fig. 4.

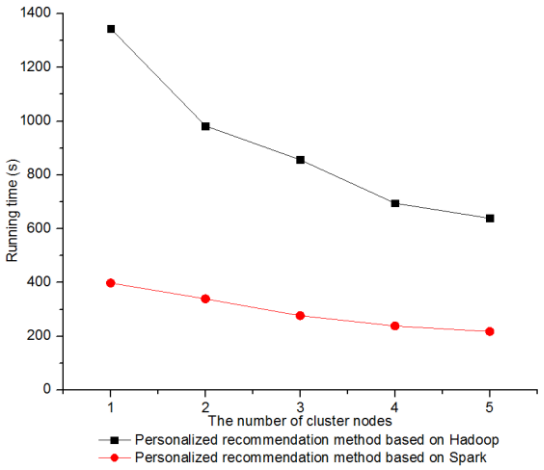


Fig. 4. The influence of cluster nodes number on running time

From Fig. 4, we can see that the running time of personalized recommendation method combing knowledge features and user interests under Spark platform is smaller than Hadoop platform. The reason is that all the iterative calculation are done in memory under Spark platform, and the middle results do not need to interact with the disk, which reduces the data transmission time. Additionally, the conversion between RDD is delayed. Thus, the real-time of recommendation is improved.

## VI. CONCLUSIONS

In order to address issues of power information operation and maintenance knowledge overload in big data environment, personalized recommendation method of power information operation and maintenance knowledge based on Spark is proposed in this paper. A personalized recommendation method combing knowledge features and user interests is designed, which considers the correlation features of knowledge and the dynamic change of user interests in the personalized recommendation process of power information operation and maintenance knowledge. The personalized recommendation method based on Spark is applied to recommend power information operation and maintenance knowledge. The experimental analysis and experiments show that the personalized recommendation

method has a better effect and enhances the accuracy and real-time of personalized recommendation of power information operation and maintenance knowledge, with a very important application value.

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China under Grant No.51277023, and Science and Technology Development Plan of Jilin Province under Grant No.20140307008GX.

## REFERENCES

[1] Z. Y. Qu, X. D. Fan, H. T. Yu, and N. Qu, "Smart Grid Text Knowledge Acquisition Model Based on Ontology," *Journal of Northeast Dianli University*, vol. 34, no. 5, pp. 60-68, Oct. 2014.

[2] Y. J. Leng, Q. Lu, and C. Y. Liang, "Survey of recommendation based on collaborative filtering," *Pattern Recognition and Artificial Intelligence*, vol. 27, no. 8, pp. 720-734, Aug. 2014.

[3] Q. Q. Zhao, K. Lu, and B. Wang, "SPCF: A memory based collaborative filtering algorithm via propagation," *Chinese Journal of Computers*, vol. 36, no. 3, pp. 672-676, Mar. 2013.

[4] X. Hu, X. W. Meng, Y. J. Zhang, and Y. C. Shi, "Recommendation algorithm combing item features and trust relationship of mobile users," *Journal of Software*, vol. 25, no. 8, pp. 1817-1830, May 2014.

[5] D. H. Liu, P. W. Peng, and H. Zhang, "Collaborative filtering recommendation algorithm based on time weight and user's feature," *Journal of Wuhan University of Technology*, vol. 34, no. 5, pp. 144-148, May 2012.

[6] Z. Y. Qu, L. Zhu, and S. C. Zhang, "Data processing of hadoop-based wide area measurement system," *Automation of Electric Power Systems*, vol. 37, no. 4, pp. 92-97, Sept. 2013.

[7] Z. Y. Qu, S. Chen, F. Yang, and L. Zhu, "An attribute reducing method for electric power big data preprocessing based on cloud computing technology," *Automation of Electric Power Systems*, vol. 38, no. 8, pp. 67-71, April 2013.

[8] B. J. Tian, C. Zhang, and Y. L. Su, "Research of improved collaborative filtering recommendation algorithm based on hadoop," *Journal of Inner Mongolia Agricultural University (Natural Science Edition)*, vol. 36, no. 1, pp. 132-138, June 2015.

[9] W. H. Li and S. R. Xu, "Design and implementation of recommendation system for e-commence on hadoop," *Computer Engineering and Design*, vol. 35, no. 1, pp. 130-136, Jan. 2014.

[10] J. Q. Che and H. W. Xie, "Hierarchical collaborative filtering algorithm based on spark," *Application of Electronic Technique*, vol. 41, no. 9, pp. 135-138, Sept. 2015.

[11] X. Sun, Y. G. Wang, and F. Y. Qiu, "Research on personalized recommendation system of online learning resources based on collaborative filtering technology," *Distance Education in China*, vol. 3, no. 66, pp. 66-71, Aug. 2011.

[12] G. H. Rong, S. X. Huo, C. H. Hu, and J. X. Mo, "User similarity-based collaborative filtering recommendation algorithm," *Journal on Communications*, vol. 35, no. 2, pp. 16-24, Feb. 2014.

[13] J. S. Lee and S. Olafsson, "Two-Way cooperative prediction for collaborative filtering recommendations," *Expert Systems with Applications*, vol. 36, no. 1, pp. 5353-5361, April 2009.

[14] D. Anand and K. K. Bharadwaj, "Utilizing various sparsity measures for enhancing accuracy of collaborative recommender systems based on local and global similarities," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5101-5109, May 2011.

[15] Z. M. Chen and Y. Jiang, "A personalized recommendation algorithm based on item rates and attributes," *Microelectronics &Computer*, vol. 28, no. 9, pp. 186-189, Sept. 2011.

[16] C. Y. Liang and Y. J. Leng, "Collaborative filtering based on information theoretic co-clustering," *International Journal of Systems Science*, vol. 45, no. 3, pp. 589-597, Sept. 2014.

[17] H. Yu and Z. Y. Li, "A collaborative filtering recommendation algorithm based on forgetting curve," *Journal of Nanjing University (Natural Sciences)*, vol. 46, no. 5, pp. 520-527, Sept. 2010.

[18] M. Zaharia, M. Chowdhury, T. Das, A. Dave, J. Ma, M. McCauley, M. J. Franklin, S. Shenker, and I. Stoica, "Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing," in *Proc. 9th USENIX Conference on Networked Systems Design and Implementation*, 2012, p. 2.

[19] J. L. Wang, *Enterprise Practice of Big Data in Spark*, Beijing: Publishing House of Electronics Industry, 2015, ch. 4.

[20] M. L. Wu, C. H. Chang, and R. Z. Liu, "Integrating content-based filtering with collaborative filtering using co-clustering with augmented matrices," *Expert Systems with Applications*, vol. 41, no. 6, pp. 2754-2761, May 2014.

**Zhaoyang Qu** was born in Jilin Province, China in 1964. He received his Doctor's degree of electrical engineering in 2010 from North China Electric Power University, Baoding. Currently he is professor in the School of Information Engineering of Northeast Dianli University. His research interests include intelligent information processing, virtual reality, and computer network.

**Pengfei Xu** was born in Shanxi Province, China in 1991. He is pursing his Master's degree of software engineering in the School of Information Engineering Northeast Dianli University. His research interests is intelligent information processing.

**Youxue Ren** was born in Jilin Province, China in 1970. He is a senior engineer of Jilin Power Supply Company of Jilin Power Company. His research interests is power informatization.