Chinese Journal of Electronics
Vol.30, No.4, July 2021

Keyword Extraction from Scientific Research Projects Based on SRP-TF-IDF

WANG Zhuohao, WANG Dong and LI Qing

(*Institute of Scientific and Technical Information of China, Beijing 100038, China*)

Abstract — Keyword extraction by Term frequency-Inverse document frequency (TF-IDF) is used for text information retrieval and mining in many domains, such as news text, social contact text, and medical text. However, keyword extraction in special domains still needs to be improved and optimized, particularly in the scientific research field. The traditional TF-IDF algorithm considers only the word frequency in documents, but not the domain characteristics. Therefore, we propose the Scientific research project TF-IDF (SRP-TF-IDF) model, which combines TF-IDF with a weight balance algorithm designed to recalculate candidate keywords. We have implemented the SRP-TF-IDF model and verified that our method has better precision, recall, and F1 score than the traditional TF-IDF and TextRank methods. In addition, we investigated the parameter of our weight balance algorithm to find an optimal value for keyword extraction from scientific research projects.

Key words — Keyword extraction, TF-IDF, Scientific research project, Word vector.

I. Introduction

A keyword is usually a word or phrase that can describe the subject of a document[1]. Technology for the automatic extraction of keywords, called keyword extraction, obtains the keywords of documents by applying certain rules or mathematical models that have been designed in the field of text mining. In research on natural language processing, keywords—the smallest units that express the meaning, or theme, of a document —have played an important role in text summarization[2], text classification[3], and clustering.

Science and technology are primary productive forces in modern society. Promoting scientific and technological innovation can improve the scientific and technological level of a country, which usually needs to support scientific research projects[4]. Therefore, as the main performers of scientific research, colleges, universities, and research institutions play a growing number of roles[5]. Scientific research project information is an important resource for all types of research organizations. Scientific research project management is also an important part of research organizations’ daily management work[6]. Scientific research project management systems, which are widely used, have improved the administrative efficiency of scientific research business management and promoted the implementation of scientific research projects, to a certain extent.

A general keyword extraction model, if applied in a scientific research project management system, can extract keywords in the usual manner. However, there are some problems, such as the lack of a suitable model and algorithm, which lead to unsatisfactory results. This paper proposes a keyword extraction model, named Scientific research project Term frequency-Inverse document frequency (SRP-TF-IDF), for the field of scientific research. The model is based on the relevant data on scientific and technological resources accumulated by the Institute of Scientific and Technical Information of China (ISTIC), combined with TF-IDF and a weight balance algorithm, which is described in the following sections.

II. Related Work

1. Keyword extraction

A document is a collection of words, and keywords are words or phrases that best describe the subject of the document. Therefore, the task of keyword extraction is to filter these words from the document. There are several common keyword extraction methods, such as TF-IDF and graph-based algorithms.

The main idea of TF-IDF is that, if a word appears frequently in a document, but less frequently in other

Manuscript Received Mar. 3, 2021; Accepted Apr. 12, 2021.
DOI:10.1049/cje.2021.05.007

documents, the word has a greater effect for distinguishing the document and expressing the core content of the document, and therefore has a higher weight. TF-IDF, which was proposed by Sparck Jones, is a commonly used algorithm for text feature extraction[7,8]. It has been successfully applied in many fields. Lu *et al.* combined TF-IDF with the TextRank algorithm to extract keywords from news text by constructing a word graph model, counting the word frequency and inverse document frequency, and considering the weight of the positioning of headlines[9]. Peng *et al.* proposed an improved TF-IDF algorithm, based on Word factors TF-IDF (WF-TF-IDF), to extract keywords from Chinese medical web pages[10]. Imamah *et al.* used TF-IDF to extract keywords from social networking sites for sentiment analysis[11]. Rahmah *et al.* adapted TF-IDF to the field of technology-enhanced learning[12].

The main idea of graph-based algorithms is to treat the candidate words in the document as nodes, establish edges between the nodes according to certain rules, and finally calculate the weight of each node to obtain the keywords of the document. Biswas proposed a graph-based method to extract keywords from social networking sites[13]. Cao *et al.* proposed a method to compute importance of cooccurrence word in document and apply it to graph approach[14].

Having considered the application performance of the two methods mentioned above, we decided to use TF-IDF as the core method for extracting keywords in the field of scientific research.

2. Word similarity

After a long period of research on word similarity calculation at home and abroad, two methods have gradually emerged. One was developed from the vocabulary knowledge base, calculating word similarity by a semantic network representation or a tree structure of vocabulary knowledge; the other was constructed from the probability distribution of information on the context of words in a large-scale statistical corpus. Because of the lack of a vocabulary knowledge base in the field of scientific research, and because ISTIC has accumulated a sufficiently large scientific research corpus, we chose the method based on the corpus for this study.

The corpus-based similarity research theory assumes that the context information of words can determine their definition. Mikolov *et al.*[15] used word vectors trained by the skip-gram model to learn the vector representation of millions of words and phrases, and improved the effect of the calculation of word and phrase similarity. Jeffrey *et al.*[16] used the word co-occurrence matrix to determine the vector weight from the distance between the target word and the co-occuring word in the window, and proposed the Glove model. Guo *et al.*[17] proposed using a bilingual corpus to learn vector representations of word meanings, which solved the problem that a word can only be represented by a word embedding in the Neural network language model (NNLM), but cannot represent polysemous words, and made good progress in word similarity calculation.

3. Scientific research project management

At home and abroad, scientific research project management systems are widely used. Most Latin American organizations producing scientific research adopt software system to manage research projects[18]. In recent years, Japan has also established some efficient scientific research project management systems[19]. Liu *et al.* designed a scientific research project management system, which could provide full and convenient data support in the process of system operation and ensure the precise and standard operation of research projects[20]. Yan *et al.* designed a scientific research project management system based on the cloud platform[21].

Although scientific research project management systems and keyword extraction technology have been well developed and widely applied, there is still a lack of effective methods for keyword extraction in the field of scientific research, because scientific research data are not shared with the public. This paper proposes the SRP-TF-IDF model, which is based on TF-IDF and a proposed weight balance algorithm. SRP-TF-IDF can effectively extract keywords from scientific research projects, thereby providing basic data services for decision support of scientific research project management.

III. Scientific Research Project Data

1. Scientific research data structure

Scientific research data in a certain field can be represented as a four-level tree structure with the field as the root node and subproject as the leaf node, as shown in Fig.1. Scientific research data can be considered as a forest, which includes many trees: one from each research field.

data structure.png data structure.bb

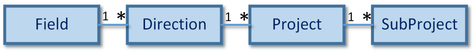

Fig. 1. Scientific research data structure in each field

Field: A scientific research field is the main characteristic of a scientific research project, such as agriculture, biology, or computer science. Each field has one or more directions.

Direction: A direction is defined by a management office to describe research plans. It comprises a direction name and an introduction. For example, data driven cloud data center intelligent management technology is a direction name of research field of computer science.

Project: According to the directions, scientific researchers create projects and prepare related information. A project contains various data, which is divided into structured data (such as project title, abstract, direction, and field) and unstructured text data in the format of .doc or .pdf[22]. We are interested in the structured data, which contribute to the keyword extraction.

Subproject: A subproject is created by the project owner, and is an extension of a project, sharing some of the properties of the project, such as the field, direction, and management department. Each project contains one or more subprojects.

2. Project data structure

According to previous research[22] and the aim of our paper, we propose a project data structure as follows: $P = \{Pid, Title, Abstract, Direction, Field\}$. Pid is the major key of the project. $Title$ is a key attribute of the project, representing the project's core meaning. $Abstract$ is a long string that includes hundreds of words to describe the main work of the project. $Direction$ is the source of the project; it is a requirement of scientific research and published by the project management office before the project application. $Field$ is a category of research directions. Therefore, we have the following constraint: if $p_i.direction = p_j.direction$, then $p_i.field = p_j.field$.

3. Subproject data structure

The subproject is the research branch of the project. We propose a subproject data structure as follows: $SP = \{Sid, Pid, Title, Abstract, Direction, Field\}$. Sid is the major key of the subproject. Pid is the major key of the project that is the parent node of the subproject. $Title$ is a key attribute of the subproject, representing the core meaning of the subproject. $Abstract$ is a long string that includes hundreds of words to describe the main work of the subproject. $Direction$ and $Field$ are inherited from the parent node. We obtain a similar constraint to the above definition: if $sp_i.pid = sp_j.pid$, then $sp_i.direction = sp_j.direction$ and $sp_i.field = sp_j.field$.

IV. Model and Algorithm

In this paper, we propose the SRP-TF-IDF model, which is composed of the traditional TF-IDF method and a weight balance algorithm designed for calculating the weight of keywords in scientific research projects. In addition to the text content of the scientific research project, it also makes use of extra information about the project, including the project's field and the subprojects' titles, which can help to improve the precision of keyword extraction from scientific research projects. First, the scientific research project is loaded, segmented, and cleaned by a data preprocessing workflow, which is based on the related databases. The TF-IDF model then receives the preprocessed data and generates a list of candidate keywords and their TF-IDF values. Finally, the weight balance algorithm measures the semantic similarity between these candidate keywords and the above extra information by calculating the cosine similarity between vectors, and combines the TF-IDF value with the semantic similarity to compute the score of every candidate word, to obtain the final keywords of each project. The SRP-TF-IDF model is shown in Fig.2.

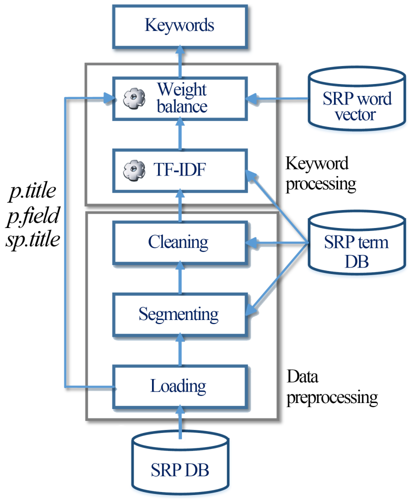

Fig. 2. SRP-TF-IDF model

1. Data preparation and preprocessing

The SRP-TF-IDF model is based on scientific data, including the SRP DB, SRP Term DB, and SRP Word Vector. These three types of data are introduced below.

1) SRP DB (SRP Database). According to the definition of our data structure of scientific research projects, we mainly use the information on fields, directions, projects, and subprojects to conduct the research for this study, and the relevant information is stored in structured databases.

2) SRP term DB. This database is composed of a scientific dictionary and list of stop words. Because there are many proper nouns in scientific research projects, the result of segmentation using common word segmentation methods is not ideal, and so we introduce a special dictionary of scientific fields. This dictionary, as well as the list of stop words, was accumulated in the process of managing scientific research projects. It contains a number of proper nouns from the field of scientific research, in addition to many common words. We also use the list of stop words to help filter the results of keyword extraction.

3) SRP word vector (word vectors SRP). In this study, the word vector model is used to calculate the semantic similarity between candidate keywords and extra project information. However, because of the existence of proper nouns in the scientific fields, the

"out of vocabulary" problem would appear if we used an existing word vector model. Therefore, we train the word vectors on tens of millions of scientific research resources using the word2vec method[15], in which the dimension of every word vector is 200.

Based on the above data, the data preprocessing module processes the project data using submodules for loading, segmenting, and cleaning. According to the four-level tree structure and data models, the data preprocessing module loads *p.title*, *p.abstract*, $p.field$, *p.direction*, *sp.title*, and *sp.abstract* from SRP DB. Following this, word segmentation is required. After word segmentation, we remove stop words, special symbols, and sentences that are too long or too short, and finally obtain the corpus that can be used in the experiment.

2. TF-IDF method

TF-IDF is a common keyword extraction method based on statistical analysis, which can be used to evaluate the importance of a word to a document. The main idea is that the importance of a word to a document is positively correlated with the number of times it appears in the document, and negatively correlated with the number of times it appears in the corpus.

Term frequency represents the frequency of a word appearing in a document. To prevent the text length from influencing the TF value, it is often normalized to the frequency of the word appearing in the document divided by the length of the document, as shown in Eq.(1).

$$Tf_{xy} = n_{xy} \Big/ \sum_{m} n_{my} \tag{1}$$

where n_{xy} denotes the frequency of the word t_x in document d_y and the denominator indicates the sum of the frequencies of all words in document d_y, which equals the length of document d_y.

Inverse document frequency represents the number of times a word appears in the whole corpus. The IDF value is negatively correlated with the frequency of the word in the corpus. The definition of IDF is as follows.

$$Idf_x = \log \frac{|D|}{|\{y : t_x \in d_y\}|} \tag{2}$$

In the above formula, $|D|$ denotes the total number of documents in the corpus and $|\{y : t_x \in d_y\}|$ denotes the number of documents containing the word t_x. To handle words not appearing in the corpus, which leads to a zero denominator, we usually add 1 to the denominator.

For word t_x and document d_y, after obtaining Tf_{xy} and Idf_x, $TF-IDF_{xy}$ can be calculated as

$$TF-IDF_{xy} = \ Tf_{xy} \times Idf_x \tag{3}$$

After the TF-IDF value of every word is obtained, several words $\{w_1, w_2, \cdots, w_N\}$ that have the largest values $\{v'_1, v'_2, \cdots, v'_N\}$and do not appear in the list of stop words can be screened out as candidate keywords, and then the weight balance algorithm is used to determine the final keywords.

3. Weight balance algorithm

To improve the accuracy of keyword extraction, we introduce a weight balance algorithm to further extract candidate keywords. This method uses cosine similarity between vectors to calculate the semantic similarity between the keywords produced by the TF-IDF method and the attached information, including the field and subprojects, and then combines the TF-IDF value with the semantic similarity to calculate the final weight, to obtain the final keyword according to the weight.

Algorithm 1 Weight balance algorithm

```
Input: a list of candidate keywords $\{w_1, w_2, \cdots, w_N\}$ and corresponding TF-IDF values $\{v'_1, v'_2, \cdots, v'_N\}$, project title *p.title*, project field $p.field$, and subprojects' titles $\sum sp.title$.
Output: a list of final keywords with weights
$str = concatenate\,(p.title,\ p.field,\ \sum sp.title)$
$S = \{s_1, s_2, \cdots, s_M\} = segment(str,\ stop\ words\ list)$
for $i = 1$ to $M$ do
    $Vector_{s_i} =\ get\_vector\,(s_i)$
end for
$Vector_S = \frac{1}{M}\sum_{i=1}^{M} Vector_{s_i}$
for $i = 1$ to N do
    $Vector_{w_i} = get\_vector\,(w_i)$
    $v''_i = cosine\_similarity\,(Vector_{w_i}, Vector_S\ ) = \frac{\text{Vector}_{w_i}\cdot\text{Vector}_S}{\|\text{Vector}_{w_i}\|\cdot\|\text{Vector}_S\|} = \frac{\sum_{j=1}^{d}\text{Vector}_{w_i}^{j}\cdot\text{Vector}_S^{j}}{\sqrt{\sum_{j=1}^{d}\left(\text{Vector}_{w_i}^{j}\right)^2}\cdot\sqrt{\sum_{j=1}^{d}\left(\text{Vector}_S^{j}\right)^2}} (d = 200)$
    $score_i =\ \alpha \cdot v'_i + \beta \cdot v''_i \quad (\alpha + \beta = 1)$
end for
$sort(candidate\ keywords,\ key = scores)$
Return a specified number of candidates as the final keywords
```

As shown in the above algorithm, we first concatenate *p.title*, $p.field$, $\sum$ *sp.title* to *str*, and then segment *str* and remove stop words. As a result, we obtain a set of segmented words $S = \{s_1, s_2, \cdots, s_M\}$. Each word in S is then converted to a vector by using the proposed SRP Word Vector model, and all these vectors are summed and averaged to obtain $Vector_S$, which represents S. The similarity v'' between each candidate keyword and S is then calculated using the cosine similarity method. Finally, the final score of every candidate keyword is obtained by allocating a TF-IDF value v' and semantic similarity v'' with different weights α, β and adding them together. According to the results of sorting the keywords by their scores, we filter out the specified number of final keywords.

V. Experiment

We evaluated our method on more than 5,000 scientific research projects and compared the experimental results with those of two traditional keyword extraction methods. To ensure the reliability of the experimental results, we repeated each method five times, and took the average value as the final result. All the tests were conducted on a notebook computer with Intel(R) Core(TM) i5-8265U CPU and 8GB of RAM, running 64-bit Windows 10.

1. Dataset

The dataset used in this study was constructed from more than 5,000 scientific research projects during the past five years. The projects covered most scientific fields, including medicine, physics, chemistry, and computer science.

2. Evaluation metrics

Because scientific research projects do not have their own keywords as ground truth, in this study, we constructed reference keywords. Each scientific research project has a direct connection with its direction, so we segmented the title of the direction, removed the stop words, and took the processed results as the reference keywords. We then calculated the precision P, recall R, and F1 score, which are defined as follows.

$$P=\frac{TP}{TP+FP}, R=\frac{TP}{TP+FN}, F1=\frac{2P\times R}{P+R} \quad (4)$$

3. Experimental results and analysis

We compared SRP-TF-IDF with the TF-IDF and TextRank methods through experiments, and measured the impact of parameter α on our weight balance algorithm.

1) As shown in Table 1 and Fig.3, the precision, recall, and F1 score of our SRP-TF-IDF model are all higher than those of the TF-IDF and TextRank models. This demonstrates the effectiveness of the model proposed in this paper.

Table 1. Precision, recall, and F1 score of each algorithm

Algorithm	Precision	Recall	F1 score
TextRank	32.94%	33.10%	33.02%
TF-IDF	44.35%	49.94%	46.98%
SRP-TF-IDF	49.85%	56.16%	52.82%

2) As shown in Fig.4, the parameter α has a significant impact on the experimental results. As α increases from 0 to 1, meaning that the model relies more on simple TF-IDF and less on semantic similarity, the F1 score first increases and then decreases, but it is always higher than that of the simple TF-IDF method. This shows that introducing semantic similarity to help screen keywords plays a very important role in keyword extraction. When $\alpha=0$, the model only pays attention to the semantic similarity between the text and the additional information after obtaining the candidate keywords. When $\alpha=1$, the model is a simple TF-IDF model. When $\alpha=0.27$, the combined model achieves the best performance.

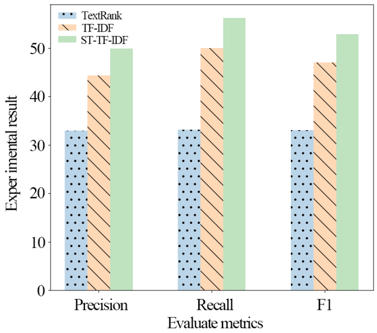

Fig. 3. Precision, recall, and F1 score of each algorithm

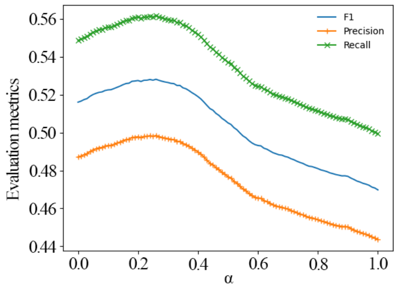

Fig. 4. Impact of parameter α on weight balance algorithm

VI. Conclusions

In this paper, we introduce a keyword extraction model SRP-TF-IDF, which integrates the traditional TF-IDF method with a weight balance algorithm that is designed to recalculate the result of the TF-IDF method. Three types of scientific data—SRP DB, SRP Term DB, and SRP Word Vector—are the basis of the SRP-TF-IDF model. The model uses a four-level tree structure to express the relations of research field, direction, project, and subproject in the scientific research domain. Furthermore, we propose a weight balance algorithm, as an extension of the traditional TF-IDF method, to extract keywords from scientific research projects. The algorithm captures candidate keywords and TF-IDF values before it calculates the final keywords of each project by using a word vector and cosine similarity method. Comparing the proposed model with the TF-IDF and TextRank methods, the experimental results show that the SRP-TF-IDF model achieves a better precision, recall, and F1 score than the traditional methods. As we measure the impact of parameter α on our weight balance algorithm, the algorithm achieves the best performance when $\alpha = 0.27$.

References

[1] Z. Jingsheng, Z. Qiaoming, Z. Guodong, *et al.*, "Review of research in automatic keyword extraction", *Journal of Software*, Vol.28, No.9, pp.2431–2449, 2017.

[2] Z. A. Merrouni, B. Frikh and B. Ouhbi, "Automatic keyphrase extraction: A survey and trends", *Journal of Intelligent Information Systems*, Vol.54, No.2, pp.391–424, 2020.

[3] A. Hassaine, S. Mecheter and A. Jaoua, "Text categorization using hyper rectangular keyword extraction: Application to news articles classification", in: *Proceedings of the Relational and Algebraic Methods in Computer Science - 15th International Conference*, Braga, pp.312–325, 2015.

[4] Y. Bai, Z. Li, K. Wu, *et al.* "Researchain: Union blockchain based scientific research project management system", *2018 Chinese Automation Congress* (*CAC*), Xi'an, China, pp.4206–4209, 2018.

[5] X. Wei and Y. Li, "Role control based workflow management for research projects", *2013 6th International Conference on Information Management, Innovation Management and Industrial Engineering*, Xi'an, China, pp.472–475, 2014.

[6] Y. Liu, Y. Yao, X. Zhang, *et al.* "Design of research management system based on workflow and rapid development platform technology", *2015 International Conference on Estimation, Detection and Information Fusion* (*ICEDIF*), Harbin, China, pp.329–334, 2015.

[7] Y. Wang, D. Zhang, Y. Yuan, *et al.* "Improvement of TF-IDF algorithm based on knowledge graph", *2018 IEEE 16th Int. Conf. on Software Engineering Research, Management and Applications*, Kunming, China, pp.19–24, 2018.

[8] P. Shanchen, Y. Jiamin, L. Ting, *et al.* "A text similarity measurement based on semantic fingerprint of characteristic phrases", *Chinese Journal of Electronics*, Vol.29, No.2, pp.233–241, 2020.

[9] L. Yao, Z. Pengzhou and Z. Chi, "Research on news keyword extraction technology based on TF-IDF and TextRank", *2019 IEEE/ACIS 18th Int. Conf. on Computer and Information Science* (*ICIS*), Beijing, China, pp.452–455, 2019.

[10] P. Sun, L. Wang and Q. Xia, "The keyword extraction of Chinese medical web page based on WF-TF-IDF algorithm", *2017 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery* (*CyberC*), Nanjing, China, pp.193–198, 2017.

[11] Imamah and F. H. Rachman, "Twitter sentiment analysis of Covid-19 using term weighting TF-IDF and logistic regresion", *2020 6th Information Technology International Seminar (ITIS)*, Surabaya, Indonesia, pp.238–242, 2020.

[12] A. Rahmah, H. B. Santoso and Z. A. Hasibuan, "Exploring technology-enhanced learning key terms using TF-IDF weighting", *2019 Fourth International Conference on Informatics and Computing (ICIC)*, Semarang, Indonesia, pp.1–4, 2019.

[13] S. K. Biswas, M. Bordoloi and J. Shreya, "A graph based keyword extraction model using collective node weight", *Expert Systems with Applications*, Vol.97, No.1, pp.51–59, 2017.

[14] J. Cao, Z. Jiang, M. Huang, *et al.* "A way to improve graph-based keyword extraction", *2015 IEEE International Conference on Computer and Communications* (*ICCC*), Chengdu, China, pp.166–170, 2015.

[15] T. Mikolov, I. Sutskever, C. Kai, *et al.* "Distributed representations of words and phrases and their compositionality", *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Nevada, United States, pp.3111–3119, 2013.

[16] J. Pennington, R. Socher and C. Manning, "Glove: Global vectors for word representation", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, Qatar, pp. 1532–1543, 2014.

[17] J. Guo, W. Che, H. Wang, *et al.* "Learning sense-specific word embeddings by exploiting bilingual resources", *Proceedings of the 25th Int. Conf. on Computational Linguistics: Technical*, Dublin, Ireland, pp.497–507, 2014.

[18] C. C. Arellano, G. J. L. Ruiz and L. M. Segundo, "Management of scientific and technological research", *2015 International Conference on Computing Systems and Telematics* (*ICCSAT*), Xalapa, Mexico, pp.1–6, 2015.

[19] Y. Komiyama and K. Yamaji, "Nationwide research data management service of Japan in the open science era", *2017 6th IIAI International Congress on Advanced Applied Informatics*, Hamamatsu, Japan, pp.129–133, 2017.

[20] K. Liu, J. Jiang, X. Ding, *et al.* "Design and development of management information system for research project process based on front-end and back-end separation", *2017 Int. Conf. on Computing Intelligence and Information System* (*CIIS*), Nanjing, China, pp.338–342, 2017.

[21] Z. Yan, G. Wei, L. Dongdong, *et al.* "University research project management system based on cloud platform", *2020 International Conference on Big Data and Informatization Education* (*ICBDIE*), Zhangjiajie, China, pp.453–456, 2020.

[22] C. C. Arellano, G. J. L. Ruiz and L. M. Segundo, "Management of scientific and technological research", *2015 International Conference on Computing Systems and Telematics* (*ICCSAT*), Xalapa, Mexico, pp.1–6, 2015.

WANG Zhuohao received the M.S. degree from the Institute of Computing Technology of the Chinese Academy of Sciences, and the Ph.D. degree in computer science from Dalian University of Technology. He is currently an associate professor in National Science and Technology Plan Management Support Center, Institute of Scientific and Technical Information of China. His research interests include cloud computing, big data, and natural language processing. (Email: wangzh@istic.ac.cn)

WANG Dong (corresponding author) received the Ph.D. degree at the School of Information and Technology, Northwest University, China, in 2016. His main research interests are in distributed systems, data warehouse, and more specifically in the area of complex event processing. (Email: wangd@istic.ac.cn)