iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** IV **Month of publication:** April 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

Volume 3 Issue IV, April 2015
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Clustering For Research Based Projects Using Ontological Text Mining

Bhagyashree .D. Shendkar[1], Mitesh Dave[2], Sagar Dave[3], Santosh Kshirsagar[4], Shreyas Puranik[5], Viresh Parkhe[6]
[1,2,3,4,5,6]*Department of Computer Engineering,*
[1,2,3,4,5,6]*Sinhgad Institute of Technology and Science, Pune.*

Abstract-This paper describes the software infrastructure (or product) which provides an overall description of research project being selection for government and private research funding agencies is an important task. When a large number of research proposals are received, they are grouped according to their similarities and requirements. These grouped projects are then taken under review. The current method is to group these projects manually with keyword referencing or field referencing. The major defect with this method is the interpretation of the concept due to human error. Text-mining methods have been proposed to solve the problem by automatically classifying text documents, mainly in English. However, these methods have limitations when dealing with non-English language texts.

I. INTRODUCTION

Selection of research project is a very important process. It starts with the call for project (CFP).The projects which are submitted are then taken for peer review. This is a manual procedure in which the researchers go through the content of the papers manually and choose a project .Usually when a call for a investment in a project is made there are a large number of entries. It is a highly laborious task and also very time consuming, therefore it would be much more convenient for a machine to do this job. Thus, with the help of this software it would be much more convenient for the investment firms to use a project with fewer amounts of efforts.

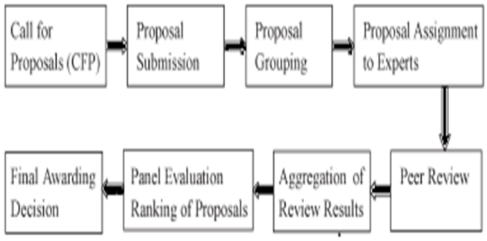


Fig. 1. Research project selection processes in the NSFC.

For instances when the National Natural Science Foundation of China (NSFC) had declared for funding of projects there were almost 110000 papers submitted for the same. It turned out to be a very tedious task for going through all the papers and choosing the most appropriate paper. For explanation of how tough it can be the most easiest example would be categorizing of vehicles. The basic distribution begins with 2-wheelers, 4-wheelers, 6-wheelers and so on. They can further be divided into air, water and land. Further on if we take an example of a land, it can be split into roadways and railways. Further if we take an example of railways this can be categorized into steam or electric locomotives. Thus if this is an example of basic vehicles, it would really get complicated when we further classify based on machine design and so on. Thus it would be much more convenient if there was a methodology by which only a mention of vehicle and a specific characteristics of that vehicle would provide all the papers which are relevant to it.

The remainder of this paper is organized as follows. Section II reviews the literature on research project selection and grouping of proposals. The proposed method is described in Section III. Section IV validates and evaluates the method, and then discusses the potential application in the NSFC. Finally, Section V provides the conclusion, and it points to future work.

II. LITERATURE REVIEW

Selection of research projects is an important research topic in research and development (R&D) project management. Previous research deals with specific topics, and several formal methods and models are available for this purpose. For example, Chen and Gorla[2] proposed a fuzzy-logic-based model as a decision tool for project selection. Henriksen and Traynor [3] presented a scoring tool forproject evaluation and selection. Ghasemzadeh and Archer [4] offered a decision support approach to project portfolio selection. Machacha and Bhattacharya [5] proposed a fuzzy logic approach to project selection. Butler *et al.* [6] used a multiple attribute utility theory for project ranking and selection. Loch and Kavadias [7] established a dynamic programming model for project selection, while Meade and Presley [8] developed an analytic network process model. Greiner *et al.* [9] proposed a hybrid AHP and integer programming approach to support project selection, and Tian *et al.* [10], identify reviewers, and assign reviewers to proposals. Current methods group proposals according to keywords. Unfortunately, proposals with similar research areas might be placed in wrong groups due to the following reasons: first, keywords areincomplete information about the full content of the proposals. Second, keywords are provided by applicants who may have subjective views and misconceptions, and keywords are only a partial representation of the research proposals. Third, manual grouping is usually conducted by division managers or program directors in funding agencies. They may have different understanding about the research disciplines andmay not have adequate knowledge to assign proposals into the right groups. Text-mining methods (TMMs) [1], [9] have been designedto group proposals based on understating the English text, but they have limitations when dealing with other language texts, e.g., in Chinese. Also, when the number of proposals and reviewers increases (e.g., 110 000 proposals and 70 000 reviewers at the NSFC), it becomesa real challenge to find an effective and feasible method to group research proposals written in Chinese. This paper presents a hybrid method for grouping Chinese research proposals for project selection. It uses text-mining, multilingual ontology, optimization, and statistical analysis techniques to cluster research proposals based on their similarities. The proposed approach has been successfully tested at the NSFC. The experimental results indicated that the method can also be used to improve the efficiency and effectiveness of the research project selection process.

III. CLUSTERING FOR RESEARCH PROPOSALS

Consider sorting of the papers based on some specific topic(eg.. Nuclear research). If the number of papers to be scanned is relatively small in number then it can be manually managed. But if the number of papers is relatively high in number then large amount of human efforts and time are required. Thus, to solve the aforementioned problems, an ontology-based TMM (OTMM) is proposed. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts [3]–[5]. It consists of a set of concepts, axioms, and relationships that describe a domain of interests and represents an agreed-upon conceptualization of the domain's "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology [2]–[4]. Thus, ontology can automate information processingand can facilitate text mining in a specific domain (such as research project selection). The proposed OTMMis used together with statistical method and optimization models and consists of four phases. First, a research ontology containing the projects funded in latest five years is constructed according to keywords, and it is updated annually (phase 1). Then, new research proposals are classified according to discipline areas using a sorting algorithm (phase2). Next, with reference to the ontology, the new proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm (phase 3). Finally, (phase 4) if the number of proposals in each cluster is still very large, they will be further decomposed into subgroups where the applicants' characteristics are taken into consideration (e.g., applicants' affiliations in each proposal group should be diverse). Each phase with its details is described in the following sections.

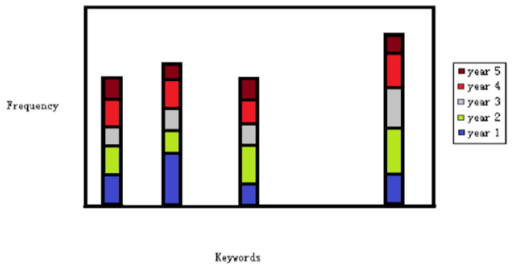


FIG 2

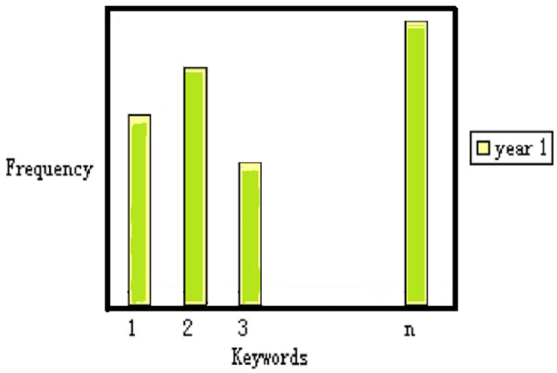


FIG 3

A. Phase 1: Constructing a Research Ontology

Different funding agencies maintain a directory of discipline areas. As a domain ontology [4],a research ontology is a public concept set of the research project management domain. The research topics of different disciplines can be clearly expressed by a research ontology. Suppose that there are *K* discipline areas, and *Bk* denotes discipline area *k*(*k* = 1, 2, . . . , *K*). A research ontology can be constructed in the following three steps to represent the topics of the disciplines.

Step 1) Creating the research topics of the discipline BN, (n= 1, 2, . . . , N).

The keywords of the supported research projects each year are colleced, and their frequencies are counted . The keywordsand their frequencie are denoted by the feature set(*Nok*,*IDk*,*year*,{(*keyword*1,*frequency*1),(*keyword*2,*frequency*2),. . . , (*keywordk*, *frequencyk*)}), where *Nok* is the sequence number of the *k*th record and *IDk*is the corresponding discipline code. For instance, if discipline *BN*has two keywords in 2007 (i.e., "*datamining*" and "*business intelligence*") and the total number of counts for them are 30 and 50, respectively, the discipline can be denoted by (*Nok*, *IDk*, 2007, {(*data mining*, 30), (*business intelligence*, 50)}). In this way, a feature set of each discipline can be created. The keyword frequency in the feature set is the sum of the same keywords thatappeared in this discipline during the most recent five years (shown in Fig. 4), and then, the feature set of *BN*is denoted by (*Nok*, *IDk*, {(*keyword*1,*frequency*1)(*keyword*2,*frequency*2), . . . ,(*keywordk*, *frequencyk*)}).

Step 2) Constructing the research ontology.

First, the research ontology is categorized according to scientific research areas introduced in the background. It is then developed on the basis of several specific research areas. Next, it is further divided into some narrower discipline areas. Finally, it leads to research topics in terms of the feature set of disciplines created in step 1. The research ontology is constructed, and its rough structure is shown in the above figure. It is more complex than just a tree-like structure. First, there are some cross-discipline research areas (e.g., "*data mining*" can be placed under "Information Management" in "ManagementSciences" or under "Artificial Intelligence" in "Information Sciences"). Second, there are some synonyms used by different project applicants, which have different names in different proposals but represent the same concepts.Therefore, the research ontology allows more complex relationship between concepts besides the basic tree-like structure. Also, to deal with proposals with both English and Chinese text, it is designed as a multilingual ontology[5], which can process and share knowledge represented in multiple languages.

Step 3) Updating the research ontology

Once the project funding is completed each year, the research ontology is updated according to agency's policy and the change of the featureset. Using the research ontology, the submitted research proposals can be classified into disciplines correctly, and research proposal in one discipline can be clustered effectively and efficiently. The details will be given in the following two sections.

B. Phase 2: Classifying New Research Proposals Into Disciplines

Proposals are classified by the discipline areas to which they belong. A simple sorting algorithm is used next for proposals'

classification. This is done using the research ontology as follows. Suppose that there are *K* discipline areas, and *BN* denotes area *N*(*n*= 1, 2, . . . , *N*). *Pi* denotes proposals *i*(*i* = 1, 2, . . . , *I*), and *Sk*represents the set of proposals which belongs to area *k*. Then, a sorting algorithm can be implemented to classify proposals to their discipline areas.

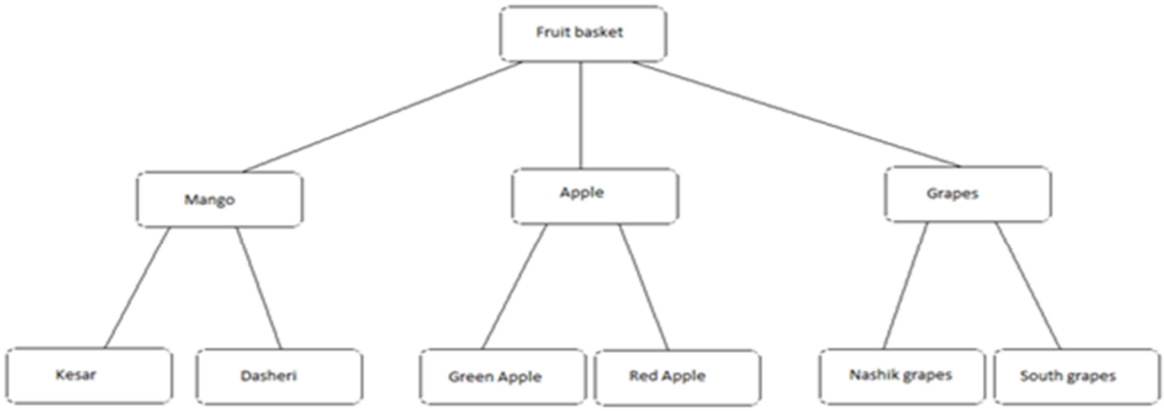


FIG 4

C. Phase 3: Clustering Research Proposals Based on SimilaritiesUsing Text Mining

After the research proposals are classified by the discipline areas, the proposals in each discipline are clustered using the text-mining technique [9], [10]. The main clustering process consists of five steps, as shown in Fig. 6: text document collection, text document preprocessing, text document encoding, vector dimension reduction,and text vector clustering. The details of each step are as follows.

Step 1) Text document collection.

After the research proposals areclassified according to the discipline areas, the proposal documents in each discipline *BN*(*N*= 1, 2, . . . , *N*) are collected for text document preprocessing.

Step 2) Text document preprocessing.

The contents of proposals are usually nonstructured. Because the texts of the proposals consist of Chinese characters which are difficult to segment, the research ontology is used to analyze, extract, and identify the keywords in the full text of the proposals. For example, “Research on behavior modeling and detectionmethods in financial fraud using ensemble learning” can be divided into word sets {“behavior modeling,” “detection method,” “financial fraud,” “ensemble learning”}. Finally,a further reduction in the vocabulary size can be achievedthrough the removal of all words that appeared only a fewtimes (say less than five times) in all proposal documents.

Step 3) Text document encoding.

After text documents are segmented,they are converted into a *feature vector* representation: *V* = (*v*1, *v*2, . . . , *vM*), where *M* is the number of features selected and *vi*(*i* = 1, 2, . . . , *M*) is the TFIDF encoding [9] of the keyword *wi*. TF-IDF encoding describes a weighted method based on inverse documentfrequency (IDF) combined with the term frequency (TF) to produce the feature *v*, such that $vi = tfi * log(N/dfi)$, where *N* is the total number of proposals in the discipline,*tfi* is the term frequency of the feature word *wi*, and *dfi* is the number of proposals containing the word *wi*. Thus,research proposals can be represented by correspondingfeature vectors.

Step 4) *Vector dimension reduction.*

The dimension of feature vectors is often too large; thus, it is necessary to reduce the vectors’ size by automatically selecting a subset containing the most important keywords in terms of frequency. Latent semantic indexing (LSI) is used to solve the problem [9]. It not only reduces the dimensions of the feature vectors effectively but also creates the semantic relations among the keywords. LSI is a technique for substituting the original data vectors with shorter vectors in which the semantic information is preserved. To

reduce the dimensions of the document vectors without losing useful information in a proposal, a term-by-document matrix is formed, where there is one column that corresponds to the term frequency of a document. Furthermore, the term-by document matrix is decomposed into a set of eigenvectors using singular-value decomposition. The eigenvectors that have the least impacts on the matrix are then discarded. Thus, the document vector formed from the term of the remaining eigenvectors has a very small dimension and retains almost all of the relevant original features.

Step 5) Text vector clustering.
This step uses an SOM algorithm to cluster the feature vectors based on similarities of research areas. The SOM algorithm is a typical unsupervised learning neural network model that clusters input data with similarities.

D. Phase 4: Balancing Research Proposals and Regrouping Them byConsidering Applicants' Characteristics
In this phase, when the number of proposals in one cluster is still very large (e.g., more than 20), the applicants' characteristics (e.g., affiliated universities) are considered. As mentioned in Sun *et al.* [8] and Fan *et al.* [5], the proposal group composition should be diverse. In the past, reviewers sometimes handled proposals improperly, having poor group composition (e.g., the same affiliation in a specific proposal group). Reviewers may feel confused and uncomfortable when evaluating proposals that may have poor group composition, so it is advisable that the applicants' characteristics in each proposal group should be as diverse as much as possible. Furthermore, the group size in each group should be similar. This may be a very complex optimization problem, and one solution method that could be used is genetic algorithm [3]. Details of the GA algorithm [6] (Fan *et al.* 2008) that are applied in our case are summarized. Conducted using the previous granted research projects. First, two experiments ($E1$ and $E2$) are constructed to evaluate the quality of clustering research projects. Second, one experiment ($E3$) is used to validate the effectiveness and efficiency of balancing research projects. In $E1$, research projects in the discipline called information management are randomly selected. In $E2$, research projects in the discipline named artificial intelligence are randomly used. In $E3$, research projects with similar topics are randomly selected. In addition, the typical criterion for text clustering F measurement is used to measure the quality of clustering research projects. As mentioned in [2], for generated cluster c and predefined research topic t, the corresponding Recall and Precision can be calculated as follows:

$$\text{Precision}(c, t) = n(c, t)/nc$$
$$\text{Recall}(c, t) = n(c, t)/nt$$

where $n(c, t)$ is the project number of the intersection between cluster c and topic t. nc is the number of projects in cluster c, and nt is thenumber of projects in topic t. F measurement between cluster c and topic t can be calculated as follows:

$$F(c, t) = (2 * \text{Recall}(c, t) * \text{Precision}(c, t)) / (\text{Recall}(c, t) + \text{Precision}(c, t)) .$$

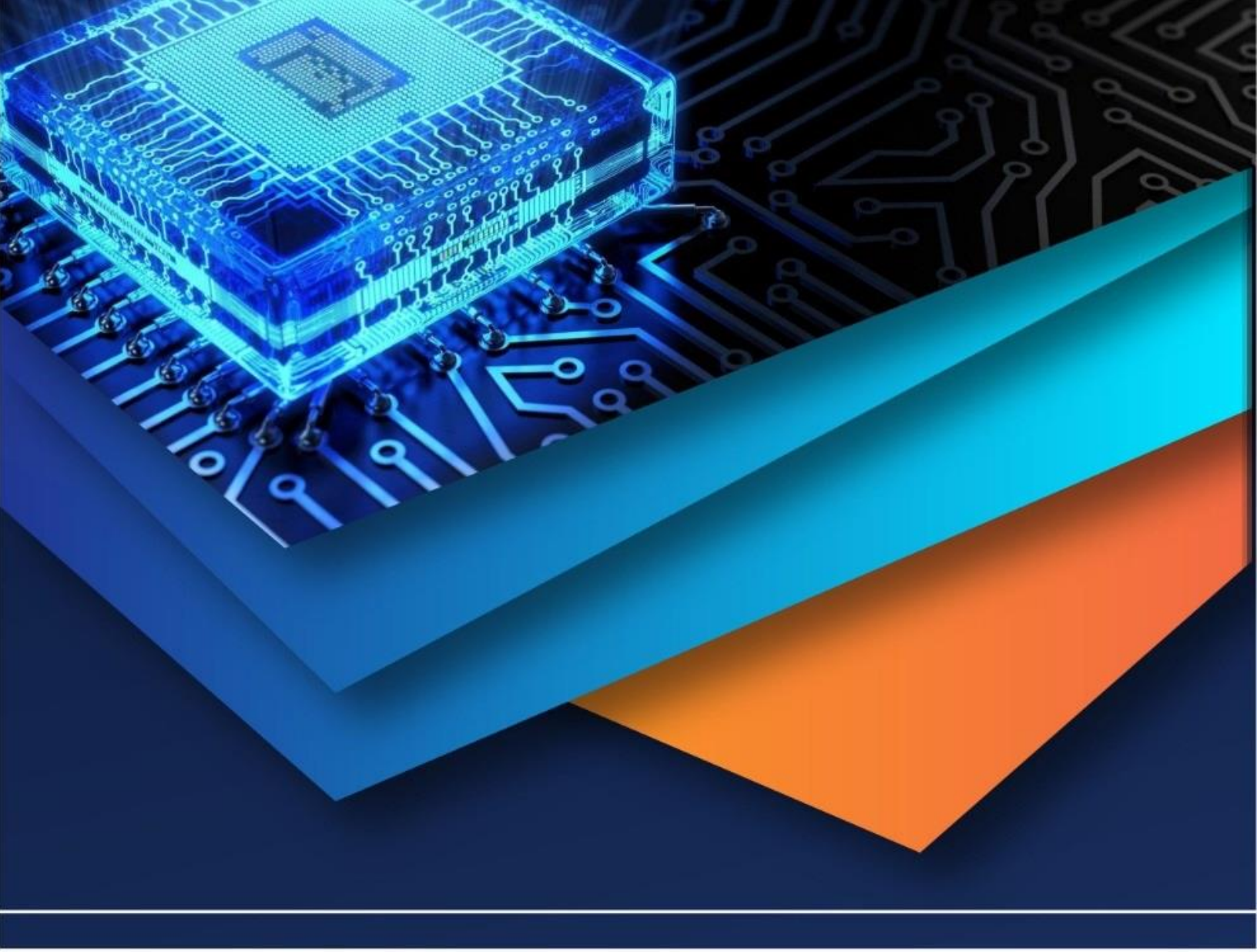
The F measurement can be given by where n is the whole number of research projects and i is each predefined research topic.
In order to compare the clustering quality of the OTMM and the general TMM, the other settings of both methods are kept the same as possible. The relations between F measurement and the number of research projects n in these two disciplines can be found , it can be seen that the performance of our proposed method is better than that of the standard TMM. Therefore, the OTMM can be an alternative for clustering research proposals. In order to validate the effectiveness and efficiency of balancing research projects, 300 research projects with similar topics are randomly selected, and the different affiliations (Northeast, Northwest, Southeast, Southwest, and Middle region) of the applicants were considered as the attribute set. The number of project groups is set to 60. The experimental result shows that the average number of different The experimental results showed that the proposed method improved the similarity in proposal groups, as well as balanced the applicants' characteristics. Therefore, the proposed method promotes the efficiency in the proposal grouping process. By manual grouping, users need to spend at least one week, while the grouping can be finished within hours using the proposed methods. Given that the method can expedite the process considerably, it can be used as the first step in a machine–human collaboration where the automatic grouping results are provided to a human that checks and then approves or modifies them.

IV. CONCLUSION

This paper has presented an OTMM for grouping of research proposals. A research ontology is constructed to categorize the conceptterms in different discipline areas and to form relationships among them. It facilitates text-mining and optimization techniques to cluster research proposals based on their similarities and then to balance them according to the applicants' characteristics. The experimental results at the NSFC showed that the proposed method improved the similarity in proposal groups, as well as took into consideration the applicants' characteristics (e.g., distributing proposals equally according to the applicants' affiliations). Also, the proposed method promotes the efficiency in the proposal grouping process. The proposed method can be used to expedite and improve the proposal grouping process in the NSFC and elsewhere. It uses the data collected from a research social network (Scholar Mate http://scholarmate.com) and extends the functions of the Internet based Science Information System (https://isis.nsfc.gov.cn). It also provides a formal procedure that enables similar proposals to be grouped together in a professional and ethical manner. The proposedmethod can also be used in other government research funding agencies that face information overload problems. Future work is needed to cluster external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Finally, the method can be expanded to help in finding a better match between proposals and reviewers.

REFERENCES

[1] D. Fensel, Ontologies: A Silver Bullet for Knowledge Management andElectronic Commerce. Berlin, Germany: Springer-Verlag, 2004.
[2] J. Plisson, P. Ljubic, I. Mozetic, and N. Lavrac, "An ontology for virtual organization breeding environments," IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 37, no. 6, pp. 1327–1341, Nov. 2007.
[3] M. Cai, W. Y. Zhang, and K. Zhang, "ManuHub: A semantic web system for ontology-based service management in distributed manufacturing environments," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 3, pp. 574–582, May 2011.
[4] Y. Liu, Y. Jiang, and L. Huang, "Modeling complex architectures based on granular computing on ontology," IEEE Trans. Fuzzy Syst., vol. 18, no. 3,pp. 585–598, Jun. 2010.
[5] L. Razmerita, "An ontology-based framework for modeling user behavior—A case study in knowledge management," IEEE Trans.Syst., Man,Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 772–783,Jul. 2011.
[6] F. M. Ham and I. Kostanic, Principles of Neurocomputing for Science &Engineering. New York: McGraw-Hill, 2001.
[7] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," IEEE Trans. Neural Netw., vol. 11, no. 3, pp. 586–600, May 2000.
[8] D. E. Goldberg, Genetic Algorithms in Search, Optimization and Machine Learning. Redwood City: Addison-Wesley, 1989.
[9] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge Univ.Press, 2007.
[10] M. Konchady, Text Mining Application Programming. Boston, MA: Charles River Media, 2006.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY