An Ontology Based Text Mining

Kuwar Aditya, Bhalekar Arjun, Bade Ankush

Department of Computer, DYPIET, University of pune , India

***Abstract*-Research project selection is important task for government and private research agencies. When a large number Of research proposals are received, it is common to group them according to their similarities in research discipline areas. The grouped proposals are then assigned to the appropriate experts for peer review. Current methods for grouping proposals are based on manual matching of similar research discipline areas or keywords. However, the exact research discipline areas of the proposals cannot be determined accurately by the applicants due to their subjective views and possible misinterpretations. Therefore, rich information in the proposals' full text can be used effectively. Text mining methods have been proposed to solve problem by automatically classifying text documents. This paper presents an ontology based text mining approach to cluster research proposals effectively based on their similarities in research discipline areas. This method can be used to improve the efficiency and effectiveness of research proposal selection processes in government and private research agencies.**

***Index Terms*- Ontology, research project selection, text mining, clustering**

I. Introduction

Selection of research projects is an important activity in many organizations such as government research funding agencies. It is a challenging multi-process task that begins with a call for proposals (CFP) by a funding agency. The CFP is distributed to relevant communities such as universities or research institutions. The research proposals are submitted to the funding agency and then are assigned to experts for peer review. The review results are collected, and the proposals are then ranked based on the aggregation of the experts review results. Fig. 1 shows the process of research project selection at private and government research agencies, that is CFP, proposal submission, proposal grouping, proposal assignment to experts, peer review, aggregation of review results, panel evaluation and final awarding decision. Generally the department is responsible for the selection tasks and it dedicates the tasks to division or programs. Division managers or program directors then groups the proposals and assign them to external reviewers for evaluation and commentary. However, they may not have adequate knowledge in all research disciplines and contents of many proposals ware not fully understood when the proposals were grouped. Therefore, there was an urgent need for an effective and feasible approach to group the submitted research proposals with computer supports. An ontology-based text mining approach is proposed to solve the problem.

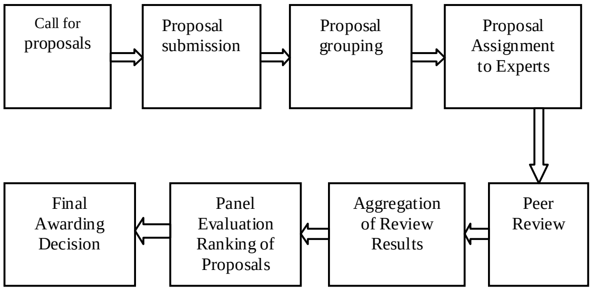
Call for proposals
Proposal submission
Proposal grouping
Proposal Assignment to Experts
Peer Review
Aggregation of Review Results
Panel Evaluation Ranking of Proposals
Final Awarding Decision

Fig.1 Research Project Selection Process.

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 For example, Chen and Gorla Machacha and Bhattacharya [1] proposed a fuzzy logic approach to project selection. Butler *et al.* [2] used a multiple attribute utility theory for project ranking and selection. Cook *et al.* [3] presented a method of optimal allocation of proposals to reviewers in order to facilitate the selection process.

Methods have been developed to group proposals for peer review tasks. For example, Hettich and Pazzani [4] proposed a text-mining approach to group proposals, 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 are incomplete 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 and may not have adequate knowledge to assign proposals into the right groups. Text-mining methods (TMMs) [5], [6] have been designed to group proposals based on understating the English text, but they have limitations when dealing with other language texts.

III. ONTOLOGY- BASED TEXT MINING TO CLUSTER RESEARCH PROPOSALS

In the Research Funding Agencies, after proposals are submitted, the next important task is to group proposals and assign them to respective reviewers. The proposals in each group should have similar research characteristics. For instance, if the proposals in a group fall into the same primary research discipline (e.g. computer science) and the number of proposals is small, then manual grouping based on keywords listed in proposals can be used. However, if the number of proposals is large, it is very difficult task to group the proposals manually.

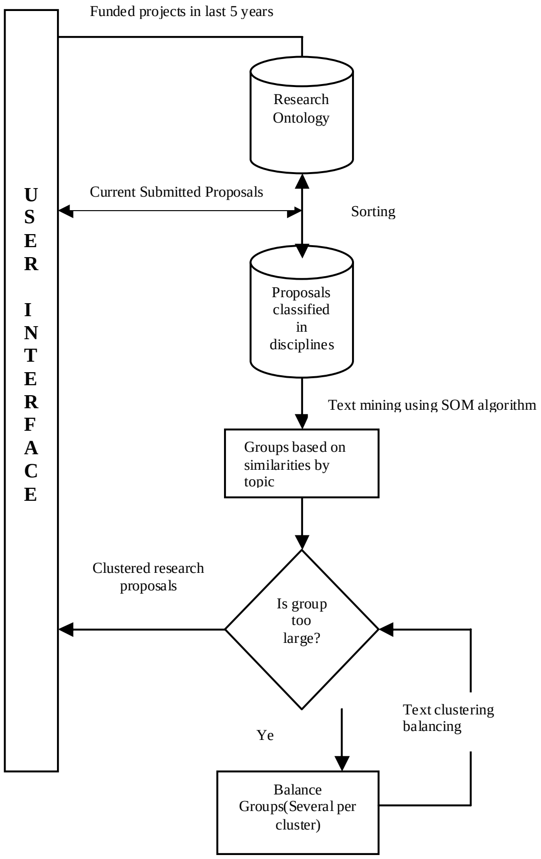
Funded projects in last 5 years
Research Ontology
Current Submitted Proposals
Sorting
U S E R I N T E R F A C E
Proposals classified in disciplines
Text mining using SOM algorithm
Groups based on similarities by topic
Clustered research proposals
Is group too large?
Text clustering balancing
Ye
Balance Groups(Several per cluster)

Fig. 2. Process of the proposed OTMM.

Although there are several text- mining approaches that can be used to cluster and classify the documents. But they are developed with a focus on English text. These methods are not effective in processing the other languages (Such as Chinese language). Several methods were proposed to deal with non-English text, but they are not efficient or sufficiently robust to process research proposals.

To achieve greater efficiency and effectiveness, an Ontology-based Text Mining Method (OTMM) is proposed.

Ontology: An ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts.

It consists of a set of concepts, axioms and relationships that describe a domain of interest and represents an agreed-upon conceptualization of the domain's "real-world" setting. Implicit knowledge for humans is made explicit for computers by ontology. Thus, ontology can automate information processing and facilitate text mining in a specific domain (such as research project selection). The proposed OTMM is used together with statistical method and optimization models and consist of four phases, as shown in fig.2. First, research ontology containing the projects funded in last 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 (phase 2). Next, with reference to the ontology, the new proposals in each discipline are clustered using a self-organize 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 consideration. Each phase with its details is described in the following sections.

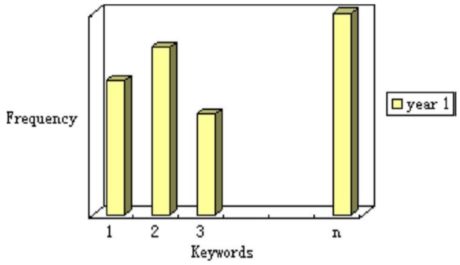

Fig. 3. Keywords of A_k in a year.

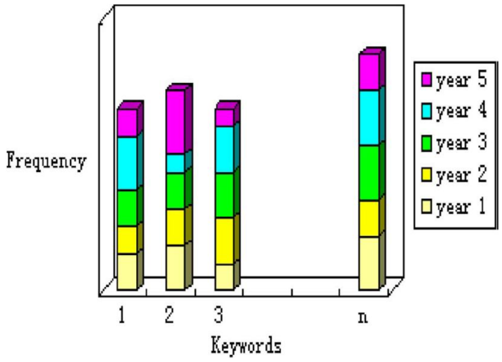

Fig. 4. Feature set of A_k.

A . Phase 1: Constructing a Research Ontology

Funding agencies maintain a directory of discipline areas that form a tree structure. As domain ontology, 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 A_k denotes discipline area *k* (*k=1,2,……,K*) . A research ontology can be constructed in the following three steps to represents the topics of the disciplines.

Step 1) *creating the research topics of the discipline* A_k , (*k=1,2,……,K*). the keywords of the supported research projects each year are collected and their frequencies are counted (fig 3). The keywords and their frequencies are denoted by the feature sets $(No_k, ID_k, year, \{(keyword_1, frequency_1), (keyword_2, frequency_2), .., (keyword_k, frequency_k)\})$, where No_k is the sequence number of the k^{th} record and ID_{ki} is the corresponding discipline code. For instance, if discipline A_k has two keywords in 2007 (i.e., “*data mining*” and “*business intelligence*”) and the total number of counts for them are 30 and 50, respectively, the discipline can be denoted by (No_k, ID_k, 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 that appeared in this discipline during the most recent five years (shown in Fig. 4), and then, the feature set of A_k is denoted by $(No_k, ID_k, \{(keyword_1, frequency_1)(keyword_2, frequency_2), ... ,(keyword_k, frequency_k)\})$

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 Fig. 5. 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 “Management Sciences” 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.

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 feature set.

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 *Ak* denotes area *k* (*k* = 1, 2, . . . , *K*). *Pi denotes* proposals *i* (*i* = 1, 2, . . . , *I*), and S_k represents the set of proposals which belongs to area *k*. Then, a sorting algorithm can be implemented to classify proposals to their discipline areas, as shown in Table I.

C. Phase 3: Clustering Research Proposals Based on Similarities Using Text Mining After the research proposals are classified by the discipline areas, the proposals in each discipline are clustered using the text-mining technique [6]. 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*. Once the documents are classified into different discipline areas, we collect all the documents for further processing. The collection of documents is important for preprocessing because preprocessing does not requires loading of documents again and again.

Step 2) *Text document preprocessing*. Text document preprocessing involves removing of unwanted and less frequent words from the collected documents to reduce the vocabulary size. The document preprocessing consist of following two steps:

(i) Reduction in the vocabulary can be achieved by removing the stop words from the documents. Stop words are the general English words which often comes in the documents such as ‘what’, ‘it’, ‘is’, ‘the’, etc.

(ii) Further reduction in vocabulary can be achieved by removing the less frequent words occurred in the documents. In this step words occurred in the document less than some frequency (say 5) can be removed to reduce the vocabulary.

Step 3) *Text document encoding*. In this step all documents are converted into feature vector representation.

$$V = (v_1, v_2, ….., v_M)$$

Where M is number of features selected and v_i is the term frequency-inverse document frequency (TF-IDF) encoding of the keyword w_i. The TF-IDF encoding of keyword w_i can be given by $V_i = tf_i * \log(N/df_i)$, where N is the total number of proposals in the discipline, tf_i is the term frequency of feature word w_i and df_i is the number of proposals containing the word.

Step 4) *Vector dimension reduction*. The number of features selected in the feature vectors for a documents is called as the dimension of that feature vector. The dimension of feature vectors is often too large; thus, it is necessary to reduce the vectors' size by selecting a subset containing the most important keywords in terms of frequency. This can be done by selecting the features with the higher tf-idf encoding values and removing the features with lower tf-idf encoding value (say 10).

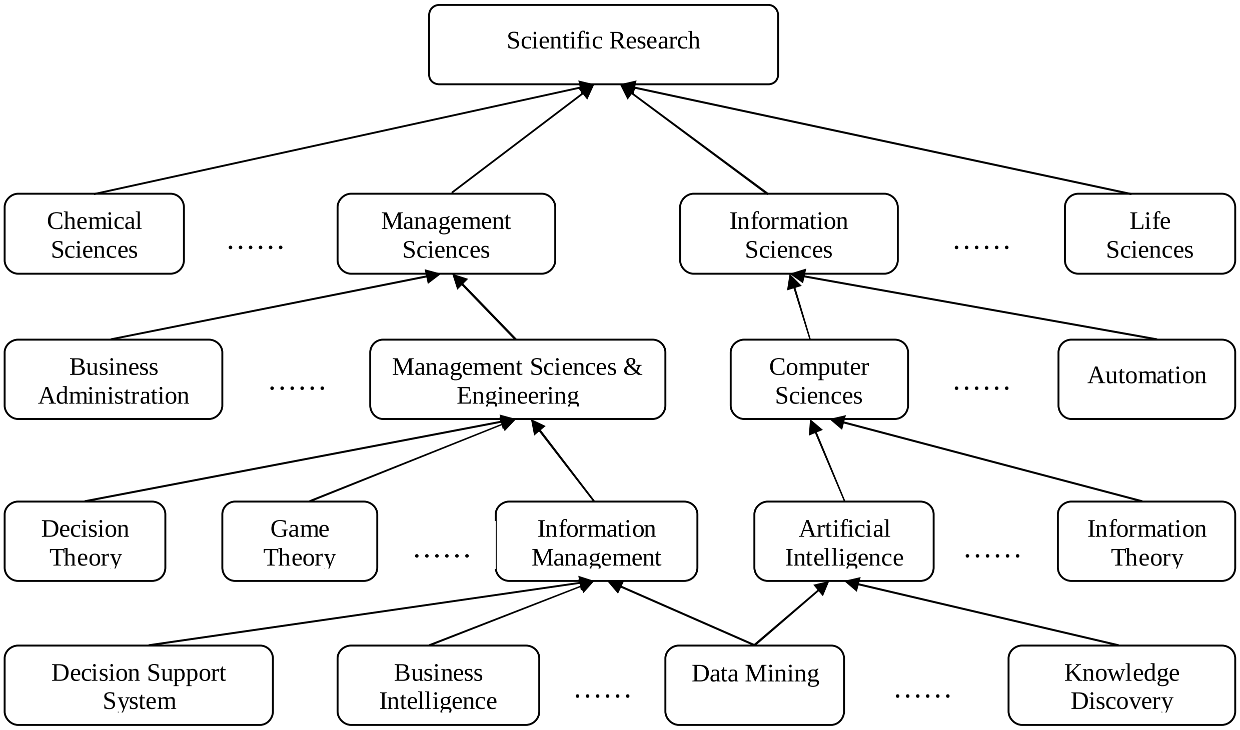

Fig. 5.Structure of the research ontology

TABLE I
SUMMARY OF THE SORTING ALGORITHM

```
For  $k = 1$ to $K$
    For  $i = 1$ to $I$

         If  $Pi$ belong to $S_k$ then
              $P_i$ is added to $S_k$ .
     End
End
```

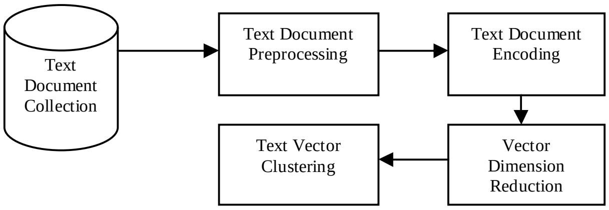

Fig. 6.Main process of text mining

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. Details of the SOM algorithm can be summarized as shown in Table II.

IV. CONCLUSION

This paper has presented an OTMM for grouping of research proposals. Research ontology is constructed to categorize the concept terms 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 proposed method can be used to expedite and improve the proposal grouping process in the funding agencies and elsewhere. Currently our approach outperform well enough but at some Extent we have kept it to a side where more research is required because it has not shown the experimental results.

TABLE II
SUMMARY OF THE SOM ALGORITHM

```
Step 1: Initialize network weight vectors $w_i$, initialize
        Learning rate parameter, define topological
        Neighborhoods functions and initialize parameter
        $N_q$, set $k = 0$.
Step 2: Check stopping condition. If false, continue: If
        true, stop.
Step 3: For each training vector $x$, perform steps 4 to 7.
Step 4: Compute the best match of a weight vector with
        input
          $q(x)=max\ sim(x,w_i)\ \forall i$
          where sim can be calculated as cosine value of the
          angle between vectors.

Step 5: For all units in the specified neighborhood where
        $q$ is the winning neuron, update the weight vectors
        according to,

          $w_i(k+1)=\begin{cases} w_i(k)+\mu(k)[x(k)-w_i(k)] & I \in N_q(k) \\ w_i(k) & i \notin N_q(k) \end{cases}$

          where $0< \mu(k)<1$ (the learning parameter)
Step 6: Adjust the learning rate parameter.
Step 7: Approximately reduce the topological
          Neighborhood
        $N_q(k)$
Step 8: set k $\rightarrow k + 1;$ then go to step 2.
```

The proposed method 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] K. Chen and N. Gorla, "Information system project selection using fuzzy logic," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 28, no. 6, pp. 849–855, Nov. 1998.

[2] L. L. Machacha and P. Bhattacharya, "A fuzzy-logic-based approach to project selection," *IEEE Trans. Eng. Manag.*, vol. 47, no. 1, pp. 65–73, Feb. 2000.

[3] J. Butler, D. J. Morrice, and P. W. Mullarkey, "A multiple attribute utility theory approach to ranking and selection," *Manage. Sci.*, vol. 47, no. 6, pp. 800–816, Jun. 2001.

[4] Q. Tian, J. Ma, J. Liang, R. Kowk, O. Liu, and Q. Zhang, "An organizational decision support system for effective R&D project selection," *Decis. Support Syst.*, vol. 39, no. 3, pp. 403–413, May 2005.

[5] S. Hettich and M. Pazzani, "Mining for proposal reviewers: Lessons learned at the National Science Foundation," in *Proc. 12th Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 862–871.

[6] 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.

[7] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban,Shouyang Wang, and Ou Liu "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection" *IEEE transactions on systems, man, and cybernetics—part a: systems and humans*, vol. 42, no. 3, may 2012