Project Proposals Clustering and Assignment to Reviewers Using TextMining Approach Based on Ontology

Pragya Singh[1], Archana Tandon[2], Pushpendra Kumar[3]
Assistant Professor, Assistant Professor, Assistant Professor
[1&3]Department of Computer Science & Engineering, [2]Department of Computer Application
[1&3]IIMT College of Engineering , Greater Noida,India
[2]LDC Institute of Technical Studies, Allahabad, India.

Abstract

With the continuous and quick development in the field of research work, research and development project selection is a necessary and important task for the research funding agencies, colleges and universities, research institutes, and technology intensive companies. Ontology is a repository of knowledge in which ideas and articles are defined and also the relationships between these ideas. The activities of finding similar pattern of text effectively, efficiently, and interactively is made by ontology. The task of ontology based text extraction for research project selection includes grouping of research project proposals that have been received according to their similarities in respective research area. Current methods for grouping proposals are mainly based on matchingof similar keywords and research discipline areas, but in most of the cases they cannot extract the exact research discipline areas accurately. This proposal presents an ontology based text mining approach to cluster notonly research proposals but also external reviewers based on their research area and then assigning of concernedresearch project proposals to reviewers systematically. This proposed work can provide an efficient and effective way for the clustering of research project proposals and their assignment to respective reviewer.

Keywords:

Ontology;
Text Mining;
Classification;
Research Project Proposal.

1. Introduction

For any research funding or conference arranging agencies, such as private or government agencies, the selection of research project proposals is an important and difficult task, when large numbers of project proposals are collected by the organization. The projectproposals assignment process starts with calling of proposals, then submission of those project proposals by different institutes and organizations. Now, clustering the proposals based on their similarity and assigned them to the experts for peer-review. For very large number of proposals received, need to group the proposals for peer review. The department for selection process can assign the grouped proposals to the external reviewers for evaluation and rank them based on their expertise. However, they may not have enough knowledge in all research discipline areas and the contents of many proposals may not be clear completely when the proposals were clustered. In current Text Mining Methods (TMM), keywords are not representing the complete information about the content of the proposals and they are just the partial representation of the proposals. Hence, it’s not sufficient to cluster the proposals based on keywords. In Manual based grouping, sometimes the department responsible for grouping may not have adequate knowledge regarding all the issues and areas of the project proposals. Therefore, an efficient and effective method is required to group the proposals efficiently based on its discipline areas by analyzing full text information of the proposals. An ontology-based text-mining (OTMM) approach is used for this purpose. This ontology based approach also includes a method to classify external reviewers based on their research areas and to assign grouped research proposals to reviewers systematically. Another new feature that we have proposed is a method to find similar proposals to that of proposal in which reviewers’ have interest.

The rest of this paper is organized as follows: In section 2 literature survey is represented. In section 3, implementation details of the proposed approach and its architecture is depicted. Data set and result set are presented in section 4. Finally in section 5 conclusion and future work is predicted.

2. Literature Survey

Classification of research project proposals is an important subject for research in research and development (R&D) project management. Previous works deals with specific subjects and several processes and models are developed for this purpose.Yong-Hong Sun, Jian Ma, Zhi-Ping Fan, and Jun Wang [11] proposed a group decision support approach to classify experts for R&D project selection. It is mainly concerned with criteria and their features for evaluating experts are summarized mainly on the basis of experience with the National Natural Science Foundation of China (NSFC). However, the project classification can be different in other countries. So, the proposed approach should be modified or adjusted before it can be applied to other organizations or contexts.Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon [3] developed an Automatic Topic Identification Algorithm to identify the topic for a textual document based on the chunks corresponding to each sentences in the document. By this method, they achieved 86% of matching for both total and partial matching among 200 random documents from the Wikipedia.

Cheng and Wei [10] proposed clustering-based category-hierarchy integration (CHI) technique, an extension to the clustering-based category integration (CCI) technique. This method improves the efficiency of category-hierarchy integration compared with that achieved by non-hierarchical category-integration techniques particularly homogeneous. However, common practices of organizations and individuals often place documents in intermediate categories. Therefore, the extension of the proposed CHI technique to handle such category hierarchies would be desirable.

Methods have been developed to cluster proposals for peer reviewing activities. For example, Hettich and Pazzani [8] proposed a text-mining approach for grouping proposals, identifying reviewers, and assignment of reviewers to proposals. Current works cluster proposals according to index terms. Unfortunately, proposals with like discipline areas might be grouped in wrong cluster. They are exploring approaches that will balance reviewer assignments across reviewers on a panel.

Matteo Gaeta, Francesco Orciuoli, Stefano Paolozzi, and Saverio Salerno [7] have presented an approach for extracting relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents using e-learning perspective. The work that they have described has several novel features. In the future improvements can be done in the approach, investigating more refined algorithms and addressing other knowledge sources.

Fabiano D. Beppler, Frederico T. Fonseca, Roberto C. S. Pacheco [2], created a ontology based framework that leads the process of engineering an IR system. They developed an instance which shows how a domain specialist without having knowledge in the IR field can also build an IR system with collaborative components. As a future work, they intend to develop a mechanism where users can define their own ontologies and configure an IR system according to their notion of reality for a specific domain.

Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang [5] proposed Text-Mining Method based on Ontology to Cluster Proposals for Project Proposal based on their similar discipline areas. This is efficient method for grouping research proposals containing English and Chinese texts. 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 experimentally compare the results of manual classification to text-mining classification.

3. Proposed Approach and Implementation Details

Basic Idea – In the proposed method of text mining based on ontology for project proposal selection it creates anontology based on previous project proposals and then applied the techniques like classification and clustering algorithms to classify the data into the disciplines using project proposal ontology and then the resultant of classification is used to make clusters of similar data. In addition to grouping of proposals the grouped proposals are also assigned to respective reviewers which are also classified similarly. We can also find the proposals similar to the proposal of our interest.

A text mining framework based on ontology has been proposed for grouping the project proposals according to its discipline areas. It consists of seven phases.

- *Construct Research Ontology:* In this module described about construction of research ontology. Initially, the ontology is categorized according to discipline areas.
- *Classifications:* In this module the input text data which are submitted project proposals, are classified into number of classes based on the discipline areas.
- *Clustering:* After classification of project proposals by the discipline areas, we need to group the proposals having similar characteristics. Clustering algorithm creates a vector of topics for each input document and measures the weight of how well the document fits into each cluster. For clustering K-means is a simple and very good method to quickly sort the data into clusters, only the need is to define the number of clusters required.
- *Re-Clustering:* In this re-clustering module we need the regrouping of very large clusters by considering the applicant's characteristics (e.g. affiliated universities) as each cluster size must be nearly same.
- *Classification of reviewers:* This module is somewhat similar to classification of project proposals in which reviewers are classified by their area of interest and their experience.
- *Assignment of proposals:* In this module the balanced cluster of project proposal is assigned to the reviewers who are having the same area of expertise (e.g. project proposals related to data mining is assigned to the reviewer having database as his area of expertise).
- *Searching of similar project proposal:* In this module similar project proposals will be searched and extracted from the cluster of project proposals to every proposal in which reviewer is interested, based on their features.

Mathematical model and Algorithm

Input set:

Set of files for ontology creation A_k = <Id_k, {kw, fre},year> which constitutes feature set of each discipline, where Id_k denotes discipline code, A_k denotes the discipline area k (k = { k_1, k_2,…,K} where K is total number of disciplines.

Set of key words kw = {kw_1, kw_2,….}

Frequency of keywords fre = {fre_1, fre_2,…..}

Reviewer set constitutes Ri < Nm, exp, aoe_1,aoe_2,…. >

where Nm = name of the reviewer, exp = years of experience and aoe_i {i=1,2,3,….} different areas of expertise of each reviewers.

Output

For New Proposal suppose that there are K areas of discipline, and A_k denotes an area k(k = 1, 2, . . . , K). P_i denotes proposals i (i = 1, 2,…..,I), and S_k represents the new proposals' set which belongs to area k.

```
For k=1 to K
    For i=1 to I
    If $P_i$ belongs to $S_k$, then $P_i$ is added to $S_k$
```

Calculate the feature vectors $V=\{v_1,v_2,\ldots v_M\}$ of each classified proposal in different domains, where M is the number of features selected and v_i (i = 1, 2, . . . , M) is the TFIDF encoding of the word set w_i of each proposal. K-means algorithm is used to cluster the feature vectors which are based on similarities of discipline area. Then each cluster *c* is assigned to each reviewer based on their *aoe* priority.

Algorithm

1. An ontology of project proposal from previous years is created according to discipline area and keywords.
2. New research proposals are classified according to the keyword stored in ontology with the topic identified using Topic Identification Algorithm.
3. Collecting all the proposals of each discipline A_k (k=1,2,…K).
4. Spilt the text into word sets W $(w_1,w_2,\ldots)$.
5. After removal of stop words documents are converted into a feature vector $V = (v_1, v_2,\ldots v_M)$.

where M is the number of features selected and v_i (i=1,2,…,M) is the TF-IDF encoding of the keyword $w_{i.}$

$$v_i= tf_i * \log(N/df_{i,}) \qquad (1)$$

N is the total number of proposals, tf_i is the term frequency of each feature word w_i and df_i is the number of proposals containing the word $w_{i.}$

6. Then cluster the feature vectors which are based on the research area similarities using K-means.
7. Balance the bigger cluster (taking threshold e.g. 20) according to applicants' characteristics.
8. Calculate F-measure for measuring the quality of clustering

$$\text{Precision}(c,t) = n(c,t)/n_c \qquad (2)$$

$$\text{Recall}(c,t) = n(c,t)/n_t \qquad (3)$$

$$F(c,t) = (2* \text{Recall}(c,t) * \text{Precision}(c,t)) / (\text{Recall}(c,t) + \text{Precision}(c,t)). \qquad (4)$$

$$\text{F-measurement } (F) = \sum_i (n_i/ n) \max \{F(i,j)\} \qquad (5)$$

where n is the number of research project proposals and i is each predefined research topics. n(c,t) is the project number of the intersection between cluster, c is the cluster and n_c is the number of projects in cluster c and n_t is the number of projects in topic t.

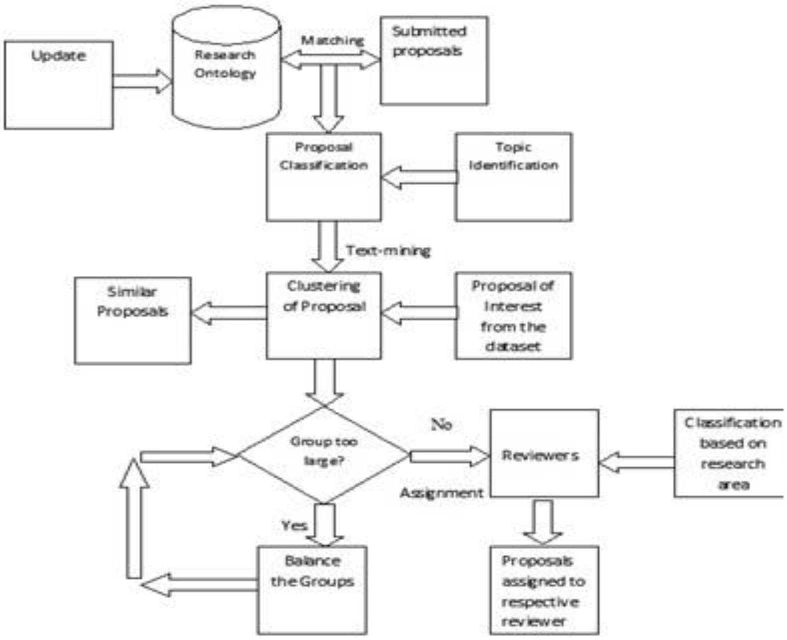


Fig 1. Proposed Framework

9. Reviewers are classified according to their area of expertise (aoe).
10. Balanced clusters are assigned to respective reviewers accordingly.
11. Based on the keywords, project proposals which are similar to the proposal of interest are selected.

Output Set:

F-Measure to evaluate quality of clustering of project proposal, Accuracy of assignment of proposals to reviewers, and Accuracy of extraction of similar papers.

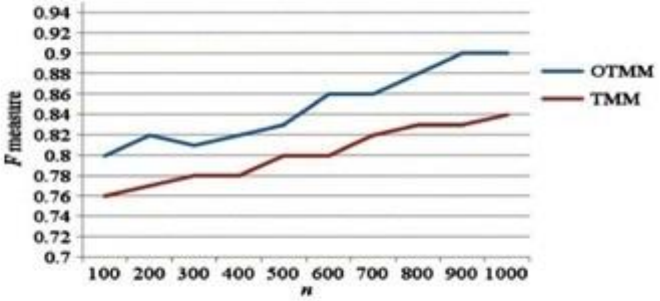
4. Result Analysis

A. Dataset

For clustering and assignment we require two data sets one containing project proposals and other containing reviewers' details. Firstly, using the dataset files of the Research project proposals and the reviewers having 1000 records, the ontology is generated. From proposal data sets first all stop words and low frequency words (say less than 5 words) are removed. The resulting feature vectors' dimension is further reduced using latent semantic indexing. After applying Clustering Techniques to the resultant data, the Research Project Proposals belong to same discipline area can be in single cluster approximately of size 20 and having different areas belongs to other clusters. For evaluation of the performance of the proposed work, we use data sets of project proposal papers from different scholarly sites. The proposed work will assign the resultant of the proposal data sets to the reviewers' data set accordingly.

B. Results Set

As shown in Fig. 2 F-measure is being used for evaluating the efficiency of clustering and comparing the F-measure for proposed work and previous work as the numbers of proposals are increased. This proposed approach can provide us a way to easily classify and group the research proposals and the reviewers. In the other experiment we are checking the accuracy of assignment of project proposals to respective reviewers.



Number of proposals (n)

Fig 2. Relationship between F-measurement and n

In the proposed work we are focusing on clustering method for proposals and assignment method of proposals to the respective reviewers. Also selection method of similar project proposal to that of paper of reviewer's interest is considered for efficiency.

4. Conclusion

This paper has presented a text mining method based on ontology for clustering of project proposals and assigning the clustered proposal to reviewers accordingly. Research proposal ontology is created to categorize the keywords in different discipline areas and to form association among them. It provides mining of text and optimization techniques to improve the proposal grouping process based on its similarities. This proposed approach can provide us a way to easily classify and group the research proposals and the reviewers. It also provides a procedure that allows finding similar proposals to every project proposal in which the reviewers are interested. The proposed work encourages the efficiency in the proposal clustering process.

In future work can be done in this assignment of the proposals such as the proposals are assigned on the basis of different features such as their experience. Also work can be done to remove the role of reviewers also from the system.

References

[1] D. E. Johnson,F. J. Oles,T. Zhang and T. Goetz," A decision-tree-based symbolic rule induction system for textcategorization", IBM Systems Journal, Vol 41, No 3, 2002.

[2] Fabiano D. Beppler,"An Architecture for an Ontology-Enabled Information Retrieval".

[3] Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon,"An Automatic Topic Identification Algorithm," Journal of Computer Science 7 (9): 1363-1367, 2011 ISSN 1549-3636.

[4] http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html

[5] Jian Ma. Wet Xu, Hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu, "An Ontology Based Text Mining Methods to Cluster Proposals for Research Project Selection", IEEE Transactions on Systems, Man, and cybernetics-Part A: System And Humans, Vol.42, No.3, May 2012.

[6] Juanying Xie, Shuai Jiang School of Computer Science Shaanxi Normal University Xi'an, Shannxi Province, P.R.China, "A simple and fast algorithm for global K-means clustering" 2010 Second International Workshop on Education Technology and Computer Science.

[7] Matteo Gaeta, "Ontology extraction for knowledge reuse the e-learning perspective", *IEEE* Trans on systems, man, and cybernetics—part a: systems and humans, vol. 41, no. 4, July 2011.

[8] 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.

[9] Tapas Kanungo, David M. Mount, Nathan S. Netanyahu, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu," An Efficient k-Means Clustering Algorithm: Analysis and Implementation", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 24, No. 7, July 2002

[10] T. H. Cheng and C. P. Wei, "A clustering-based approach for integrating document-category hierarchies," IEEE Trans. Syst., Man, Cybern.A,Syst., Humans, vol. 38, no. 2, pp. 410–424, Mar. 2008.

[11] Y. H. Sun, J. Ma, Z. P. Fan, and J. Wang, "A group decision support approach to evaluate experts for R&D project selection," IEEE Trans Eng. Manag., vol. 55, no. 1, pp. 158–170, Feb.2008.