International Journal of Computers & Technology
Volume 5, No. 1, May -June, 2013, ISSN 2277-3061

ONTOLOGY BASED CLASSIFICATION AND CLUSTERING OF RESEARCH PROPOSALS AND EXTERNAL RESEARCH REVIEWERS

Preet Kaur
Lovely Professional University, Phagwara.
preetkaur2119@gmail.com

Richa Sapra
Lovely Professional University, Phagwara.
richa.16859@lpu.co.in

ABSTRACT

With the rapid development of research work in projects, research project selection is a necessary task for the research funding agencies. It is common to group the large number of research proposals, received by the research funding agencies. Based on their similarities in Research Discipline areas. The grouped Proposals are then assign to the appropriate Research experts for peer-review. In current methods, which are manual based, proposals assigned to experts may not have adequate knowledge about all discipline areas. In this paper, Ontology-based Text Mining Method is presented to classify Research Project Proposals, as well as External Research Reviewers and then group them based on their research discipline areas and assign the particular research proposal group to the appropriate reviewer group. This approach provides an efficient and effective way for the selection of research project proposals with the increasing number of research proposals and reviewers.

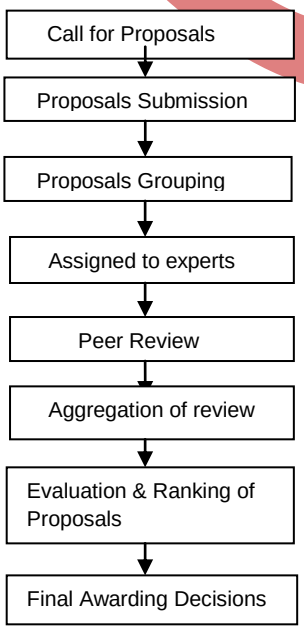
Keywords

Ontology-based Text Mining Method, Clustering, Classification, Research Project Selection.

1 INTRODUCTION

For any research funding agencies, such as either government or private agencies, the selection of research project proposals is an important and challenging task, when large numbers of research proposals are collected by the organization. The Research Project Proposals Selection Process starts with the call for proposals (CFPs), then submission of the research proposals by many institutes and organizations. Now, group the proposals based on their similarity and assigned them to the experts for peer-review. The review results are examined and proposals are ranked based on their aggregation of experts result. [1] Fig1 represent the steps of the Research Project Selection Process, these processes are very similar in all research funding agencies.[2] For very large number of proposals received by the agencies need to be 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 aggregation. However, they may not have adequate knowledge in all research discipline areas and the contents of many proposals were not fully understood when the proposals were grouped. In current Methods, 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 group the proposals on the basis of keywords. In Manual based grouping, sometimes the department responsible for grouping may not have adequate knowledge regarding all the issues and areas of the research proposals. Therefore, an efficient and effective method is required to group the proposals efficiently based on their discipline areas by analyzing full text information of the proposals. An ontology-based text-mining approach is used for this purpose.



International Journal of Computers & Technology
Volume 5, No. 1, May -June, 2013, ISSN 2277-3061

Fig1 Research Project Proposals Selection Process

This paper is organized as following. Section2 describes some related work.Section3 represents the basic idea of the proposed method. Section4 will discuss the proposed Architecture. And Section 5 shows the Experimental Results and Section6 concludes the paper.

2 RELATED WORK

Many methods have been developed for the selection of research project proposals. Few of them proposed a fuzzy-logic basic model as a decision tool for project selection [1].Hendrickson and traynor presented a scoring tool for project evaluation and selection and decision-support approach for the project portfolio selection. Cook et al. [4] 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 this proposes 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 discipline areas might be placed in wrong groups. Due to some of the following reasons: First, keywords not provide complete information regarding the full text of the proposals. Second, by manually grouping proposals, they may not have adequate knowledge about different discipline areas and particular proposal is assigned into the right group. But, if the number of proposals is large, it is difficult to group proposals manually.[2] Several Text-mining methods have been designed to group proposals based on understanding the English text, not the Non-English texts, i.e. Chinese. There was another limitation of Text Mining approach, when number of proposals and reviewers were increases, it becomes a real challenge to efficiently group the proposals in Chinese.[3]

3 BASIC IDEA

This paper using the concept of ontology with Text Mining techniques such as Classification and Clustering algorithms. The proposed approach builds the research ontology and then applies Decision Tree Algorithm to classify the data into the disciplines using research ontology and then the resultant of classification is used to make clusters of similar data. K-means Clustering technique is used for this purpose.

3.1 Ontology

Ontology has become prominent in the research work from recent years, in the field of computer science.

Ontology is a knowledge Repository which defines the terms and concepts and also represents the relationship between the various concepts. It is a tree like structure which defines the concepts.[5]

3.2 Classification

In Classification, the input text data can be classified into number of classes based on that data. Various Text-Mining techniques are used for classification of text data such as Support Vector Machine, Bayesian, Decision Tree, Neural Network, Latent Semantic Analysis, Genetic Algorithm, etc. In this paper, Decision Tree is used for the Classification of Research Proposals as well as to classify Research Reviewers.

3.2.1 C4.5 Decision Tree Algorithm

C4.5 is considered for determining the best Decision tree using both categorical and numeric feature values. It is the extension of earlier ID3 algorithm. It works at two stages –first stage generates a decision tree based on training dataset and Second stage has pruned the decision tree based on validating samples.[6][7] [12]

3.3 Clustering

Clustering is a technique used to make group of the documents having similar features. Documents within a cluster have similar objects and dissimilar objects as compared to any other cluster. Clustering algorithms creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. This technology can be useful in the organization of management information systems, which may contain thousands of documents. Several Text Mining Algorithms used for clustering are K-Means, Self-Organizing Maps (SOM), EM, etc. But simple K-means Text mining clustering Technique is used for this research work.[9]

3.3.1 K-means Algorithm

K-means is a best method to quickly sort the data into clusters, only the need is to define the number of clusters required. K denotes the number of clusters in which the data is divided.[10]The algorithm works as:

1. Randomly select K-points as the initial cluster centroids.
2. Assign each object in the dataset to the closest cluster by compute their Euclidean distance of the object from the center.
3. When all objects have been assigned recalculate the position of the K centroids.
4. Repeat step 2 & 3 until the centroid no longer move. Now. At this points clusters are separated into groups successfully.[13]

International Journal of Computers & Technology
Volume 5, No. 1, May -June, 2013, ISSN 2277-3061

4 PROPOSED ARCHITECTURE

The basic idea for this proposed architecture is to make easier the Research Proposals Selection Process. For this purpose, the combination of C4.5 Decision Tree and k-means Clustering techniques is used to assign the particular proposals to particular reviewers based on their domain. The Ontology-based Text Mining approach is used for the selection of research project proposals for either government or private research funding agencies. The specific Approach is as follows:

4.1 For the research proposals

From the dataset of the research proposals scientific research ontology is designed. With the help of this research ontology, the research proposals are classified into discipline areas using C4.5 Decision tree Algorithm of text-mining.[6][7] Now, in each discipline area, the research proposals are clustered based on their similarities using K-means clustering algorithm.

4.2 For the external research reviewers

From the dataset of the reviewers the ontology is designed on the basis of all the domain areas of the reviewers. Then, with the help of ontology, the reviewers are classified into discipline areas as expert areas based on their interest of knowledge. Using text mining clustering algorithms, grouping the research reviewers based on their similarities in each discipline area or domain.

4.3 Proposals Assign to Reviewers

The Final step of this approach is to assign the Research Proposals to the External Research Reviewers. The Proposals of the particular Discipline area is assign to the Reviewers having the same research area or domain. For example, all the Research proposals related to the computer Networks is assigned to the Reviewers having the Networks as a research domain. So, they can examine the proposals efficiently for the peer-review.

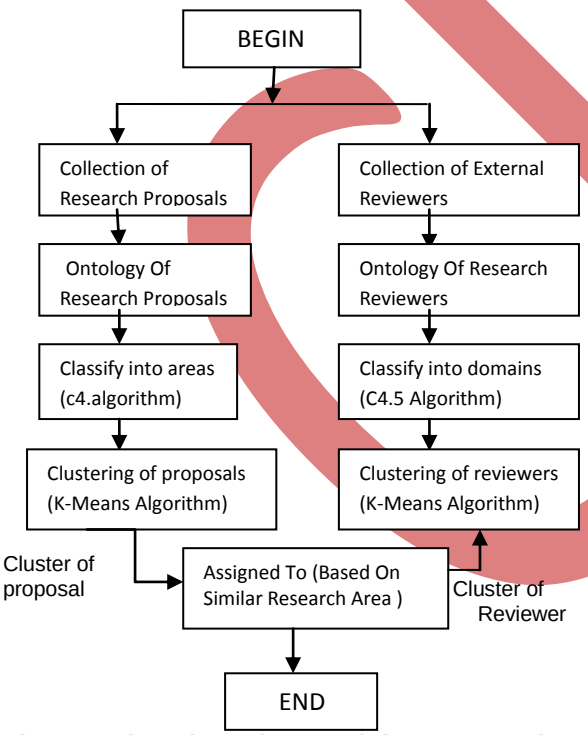


Fig2 Ontology based Text Mining Approach

5 EXPERIMENTAL RESULTS

Firstly, using the dataset files of the Research Proposals and the reviewers having 500 records, the ontology is generated. After constructing ontology the, the Weka tool is used for Classification of Proposals and Reviewers using C4.5 Decision Tree Algorithms. For Weka C4.5 is known as J48. [8] Weka tool can classify the Research Proposals into classes based on their discipline areas.

International Journal of Computers & Technology
Volume 5, No. 1, May -June, 2013, ISSN 2277-3061

Fig3 Reviewers Ontology

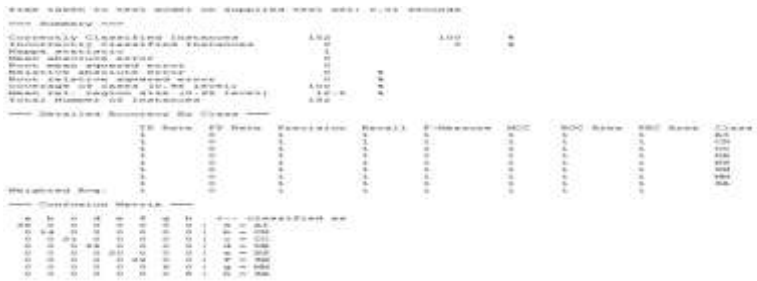
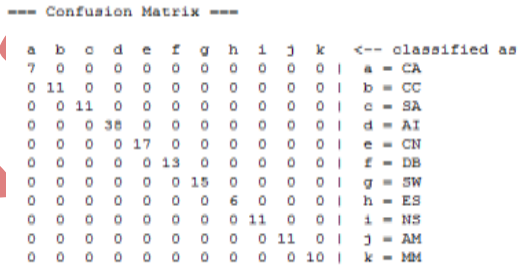


Fig4 Classification result of Proposals

Similarly, External Research Reviewers can be classified into the domain to which they belong.



```
=== Confusion Matrix ===

  a  b  c  d  e  f  g  h  i  j  k   <-- classified as
  7  0  0  0  0  0  0  0  0  0  0 |  a = CA
  0 11  0  0  0  0  0  0  0  0  0 |  b = CC
  0  0 11  0  0  0  0  0  0  0  0 |  c = SA
  0  0  0 38  0  0  0  0  0  0  0 |  d = AI
  0  0  0  0 17  0  0  0  0  0  0 |  e = CN
  0  0  0  0  0 13  0  0  0  0  0 |  f = DB
  0  0  0  0  0  0 15  0  0  0  0 |  g = SW
  0  0  0  0  0  0  0  6  0  0  0 |  h = ES
  0  0  0  0  0  0  0  0 11  0  0 |  i = NS
  0  0  0  0  0  0  0  0  0 11  0 |  j = AM
  0  0  0  0  0  0  0  0  0  0 10 |  k = MM
```

Fig5 Classification of Reviewers

After Classifying the proposals into classes of different discipline areas. Apply K-means Clustering Techniques to the resultant data. The Research Proposals belong to same discipline area can be in single cluster and having different areas belongs to other clusters. Each Cluster belongs to a particular area and it contains all the research proposals related to that particular areas. Fig6 represents 8 clusters of research proposals in which Cluster0 contains 18%,cluster1 contain 5%, cluster2 and cluster3 contains 7 % ,cluster4 contains 13% ,cluster5 contains 25%,Cluster6 contains 9 % and cluster7 contains 17%.

Fig6 Clusters of Research Proposals

By using the Classification result of the Research reviewers, the clusters of the particular domain of the external reviewers are created. For example, all the research Reviewers having Network Security As a research area they all are cum under the cluster termed as Network Security.

Fig7 clustering result of Reviewers

International Journal of Computers & Technology
Volume 5, No. 1, May -June, 2013, ISSN 2277-3061

6 CONCLUSIONS AND FUTURE WORK

In this paper, Ontology based classification and clustering approach is proposed, which will be used by research funding Agencies for grouping the Research Proposals and the research Reviewers. This approach is very user friendly and time consuming. In this proposed approach, the combination of Data Mining techniques – Classification using Decision tree and Clustering using K-means is used with the help of Ontology. This Proposed approach can provide us a way to easily classify and group the research proposals and the reviewers. The proposed work gives 100% accurate result for classification i.e. efficiently classifies the research areas.

In future work, a focus is required to find a systematic way which represents the research proposals are assigned to the appropriate research Reviewer. And also required some more work done in this assignment of the proposals such as the proposals are assigned on the basis of their experience.

7 REFERENCES

[1] 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.

[2] 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.

[3] D. A. Chiang, H. C. Keh, H. H. Huang, and D. Chyr, "The Chinese text categorization system with association rule and category priority," *ExpertSyst. Appl.*, vol. 35, no. 1/2, pp. 102–110, Jul./Aug. 2008.

[4] W. D. Cook, B. Golany, M. Kress, M. Penn, and T. Raviv, "*Optimal allocation of proposals to reviewers to facilitate effective ranking*," *Manage. Sci.*, vol. 51, no. 4, pp. 655–661, Apr. 2005.

[5] Hmway Hmway Tar, Thi Thi Soe Nyunt ***"****Ontology-Based Concept Weighting for Text Documents",* International Conference on Information Communication and Management IPCSIT vol.16 2011 IACSIT Press, Singapore

[6] Lei Zhang, Zhichao Wang. "*Ontology-based clustering algorithm with feature weights"*, 2010 Journal of Computational Information Systems 6:9 (2010) 2959-2966.

[7] A. Maedche and V. Zacharias, *Clustering "Ontology-based Metadata in the Semantic Web"*. In Proceedings of the 6th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'02), Helsinki, Finland, pp. 342-360, 2002.

[8] http://home.unpar.ac.id/~integral/Volume%208/Integral%208%20No.%202/C45%20Algorithm.PDF

[9] Jain, A. K., and Dubes, R. C., "Algorithms for clustering data", Prentice-Hall.,1988.

[10] 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

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

[12] http://cis.poly.edu/~mleung/FRE7851/f07/decisionTrees.pdf

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