Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

An Ontology based Text Mining Approach to Cluster Proposals and Expert Reviewers for Selecting R & D Project

T. Suganya
Department of Information Technology
Raja college of Engineering & Technology
Madurai,Tamil Nadu

N. Shanmugakani
Department of CSE
Raja college of Engineering & Technology
Madurai,Tamil Nadu

***Abstract*— In many governments and private institutions, one of the major tasks is to select the best project proposals for allocating the fund. These funding organizations select the proposals by grouping and submitting them to the reviewers for review. Manual grouping process is too difficult when the number of projects is more. The earlier models introduced ontology based Text mining methods [1] to cluster the proposals of any language without considering the reviewer's expertise with respect to their domain. The proposed method identifies the main topic of project in an efficient manner by using ontology based topic identification algorithm. This Proposed model not only categorizes the project proposals, it also clusters the reviewers based on their domain experience by using K-means algorithm. This new approach makes the decision making process of funding organization as easier.**

***Keywords*— Ontology, textmining, Clustering. Classification, C4.5, K-means**

I. INTRODUCTION

Selection of research project proposal is an important task of most funding organizations. These organizations first collect the proposals, categorize the proposals as domain wise and then assign them to the experts for peer review. The reviewers analyze the proposals and select the best to provide funds. This categorization is done by manually in the funding agencies. When the number of proposals is very large, then manual process is too tedious because of wrong interpretation and poor subject knowledge. This problem initiates the innovation of new technology called ontology based text mining method to classify the proposals automatically and efficiently compared to manual process.

Data mining is a process of extracting knowledge from various data sources. Text mining is an application area of data mining. It identifies precious information from collection of text. The Text classification is a process which identifies the commonalities of text documents and groups them. Then, it assigns the group to predetermined label. It is a supervised learning method that can easily classify the new in coming object by predicting the predefined label based on the theme of that object. This classification method uses a trained data set for predicting the catagory label for new objects. Text clustering is a powerful method, which is used to group text documents based on their similarity. It identifies the class to which an object belongs to. It is an unsupervised learning method. It doesn't use any trained data set.The fig1 shows both text classification and clustering process.While clustering text documents, each document belongs to one catagory whereas in classification each document may belongs to different catagory.

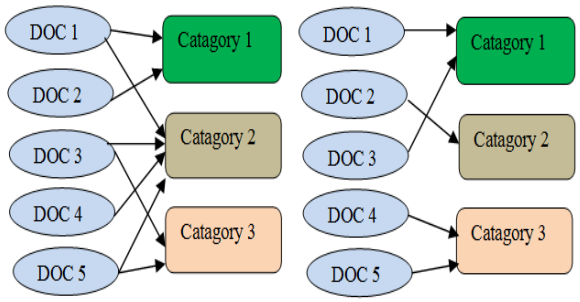

Fig. 1. Text Classification & Clustering Process

Ontology describes a concept and its properties in a domain. It is used to construct a knowledge base repository which holds some concepts of domain and its relationship with other concepts. This repository can be used with various text mining process for concept extraction and can give support to construct the automatic information learning system [9].

II. PRIOR WORK

Research project proposal selection is a very complicated and knowledge based decision making task. Researches proposed many techniques to solve this complication. The fig2 shows the traditional system in which manual grouping is done for project selection. Manual grouping is usually conducted by program directors or managers in funding agencies. They may not correctly assign the proposals into the corresponding group because of insufficient knowledge about the domain. Hence, the proposals may be placed in wrong group due to this lack of domain knowledge [3]. The text mining approach [2] uses keyword similarity to cluster proposals. In this method, same discipline projects might be placed in different group.

Text mining method (TMM) deals with English language and not effectively works with other natural languages. The keywords do not provide complete information about the project.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

In Ontology based approach [1], OTMM is combined with statistical and optimization method to categorize and group the similar discipline proposals. This model contained four phases, First, research ontology is constructed which includes trained dataset, keyword collection and classifier algorithm. Second, proposals are grouped by SOM algorithm. Third, proposals are assigned to reviewers. Finally, optimization method is used for larger clusters. Research ontology is formed based on the keyword of a research project. The keywords are not enough to identify the correct topic of research.

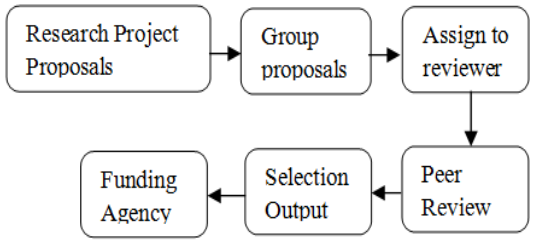

Fig. 2. Traditional System

This system has the following problems; first, it submits the proposals without considering reviewers' expertise. It is not possible for the reviewers to analyze all projects in different discipline. They may not have sufficient knowledge in all domains. Second, the classification and clustering technique depends on the results provided by research ontology. Hence, Topic identification process gets more significance in such model. The proposed model gives emphasis to Ontology Text mining process and the reviewer's experience.

III. PROPOSED WORK

The proposed system architecture is shown in fig 3. In this model, the main topic of research project proposal is identified by ontology based topic identification method. Text mining methods such as supervised and unsupervised learning algorithms are used. This approach constructs the knowledge base repository to identify the topic of proposals. Then, the proposals are categorized into the appropriate discipline areas with the help of that repository. The resultant of classification process is fed to make clusters of research papers. Simple K-Means clustering technique is used for this purpose.

First, the proposals are collected. Last year's project proposals are used to form a research ontology Knowledge base which consists of feature set of each proposal with weight values. This knowledge base is updated at every year. The proposed system includes six modules to perform the selection task.

- Topic Identification.
- Classifying the proposals based on discipline using C4.5.
- Clustering the proposals using K-means.
- Classifying the Reviewers based on domain.
- Clustering the Reviewers based on experience.
- Peer review & Proposal selection

A. Topic Identifier

In this module, the main topic of the given project proposal is identified by ontology based topic identification algorithm.

Steps:

1) The text content is divided into sentences

2) Sentences are parsed into terms, remove stop words. The terms include noun, verb and complement

3) Candidate topic is formed by the parsed terms, weight for the candidate topic is calculated based on its frequency in that paper.

4) The highest weight candidate topic is found to determine the main topic of the proposal.

5) This feature set is added in Ontology knowledge base repository for future use.

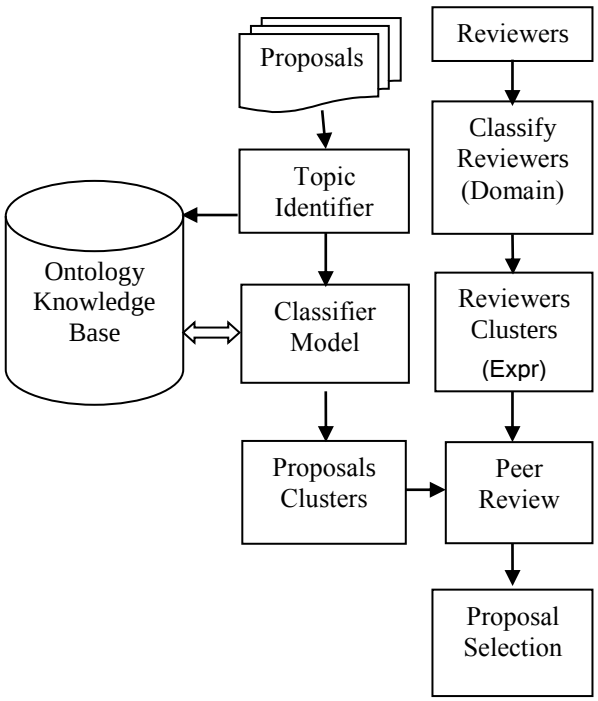

Fig. 3. Proposed System Architecture

B. C4.5 Classifier

Data mining provides several algorithms for categorizing larger dataset. C4.5 is a decision tree algorithm developed to overcome the limitations of ID3[4]. It constructs a decision tree from trained dataset like ID3.This algorithm [11] has some base cases.

- If the training dataset is empty, a tree leaf with failure label is returned.
- If the attribute list is empty, a tree leaf is returned with most frequent class label or disjunction of all classes
- If all records in a trained dataset belongs to a same class, then a tree leaf with that class label is returned

C4.5 General Algorithm:

- Check for the base cases
- Build a decision tree for the given trained dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

- For each attribute A, find the normalized information gain ratio
- Find the attribute with the highest information gain ration (A-Best)
- Split the set S into s1,s2,s3 based on the value of A-Best
- Create a decision node that splits on A-Best
- Repeat the steps for all partitions

Steps:

1) Main topic (set of terms with weight value) of research paper is fed to classifier model to identify the domain.

2) After identifying the discipline area, the proposal is added in that class.

C. Clustering proposals

Clustering is a technique used to make group of the documents having similar features. Documents within a cluster have similar features and dissimilar to other cluster. Clustering algorithms create a vector of candidate topics for each document and measures its weight values in order to place that document into proper cluster [10]. This technology can be useful in the organization of management information systems, which may contain thousands of documents. Many clustering algorithms are used in text mining such as, K-Means, Self Organizing Maps (SOM), EM, etc. But, simple K-means is one of the best unsupervised technique for clustering larger dataset. Hence, this algorithm is used in this proposed system [5].

K-means is the best method to group larger data into clusters, Only the need is to define the number of initial clusters required. K denotes the number of clusters in which the objects are divided [6].

K-Means General Algorithm:

- Place all objects into the space.
- Define initial k number of clusters.
- Find the centroids of k clusters.
- Assign each object to the group that has the nearest centroid for that object by calculating their Euclidean distance of the object from the center point.
- When all objects have been assigned to the closest cluster, recalculate the position of the K centroids for new clusters.
- Repeat step 4 & 5 until the centroid has no longer move. Now, at this point, all objects in the dataset are separated into groups successfully [7].

Steps:

1) After the classifiying the proposals according to their domain based on main topic terms,these proposals are grouped into clusters based on similarity level by this K-Means algorithm

2) .It condiders the candidate topic terms for clustering

3) The proposals which are having similar topic terms are grouped into a cluster

4) This similar proposals are assigned to the reviwers for peer review

D. Classifying the reviewers

Reviewers' dataset includes the information about their name, id, discipline, designation and experience in various institutes. This module classifies the reviewers according to their domain or area of interest. C4.5 algorithm is applied here to categorize the reviewers.

E. Clustering the reviewers

After classifying the reviewers, they are grouped based on their experience. Due to this, the experts with similar experience reviewing the proposals efficiently. This speed up the review process.The simple K-means algorithm is used to cluster the reviewers.

F. Peer review

The proposals are assigned to the expertise group of members for peer review and the best proposal are selected according to the discipline and submitted to the funding organization.

IV. Experiment and Results

Weka is a powerful machine learning software written in java which provides many visualization tools and algorithm for analyzing the data and predicting models for data. This .tool is used to view the results of Proposals and Reviewers classification and clustering models. In proposed system, C4.5 Decision Tree Algorithms and simple K-means algorithm is applied for this purpose.. For Weka C4.5 is known as J48. Weka tool can classify the Research Proposals into classes based on the discipline areas. Sample dataset for reviewers is taken.This includes name, domain, experience information about the reviewers.The fig 4 & 5 shows the catagories of reviewers.

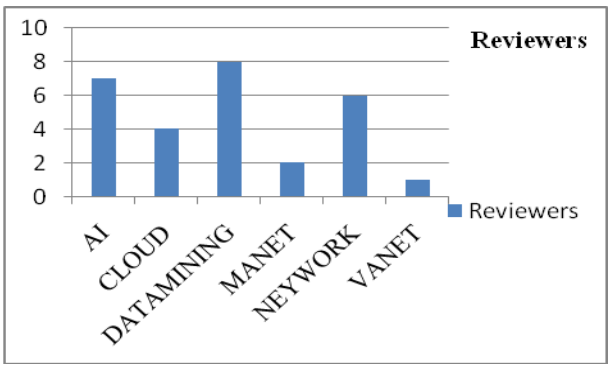

Fig. 4. Reviewers according to the domain

After classifying the reviewers, simple K-means algorithm is applied to cluster the reviewers based on experience similarity. The fig 6 & 7 shows the three clusters of experts based on years of experience. The visualization graph is obtained by weka tool.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

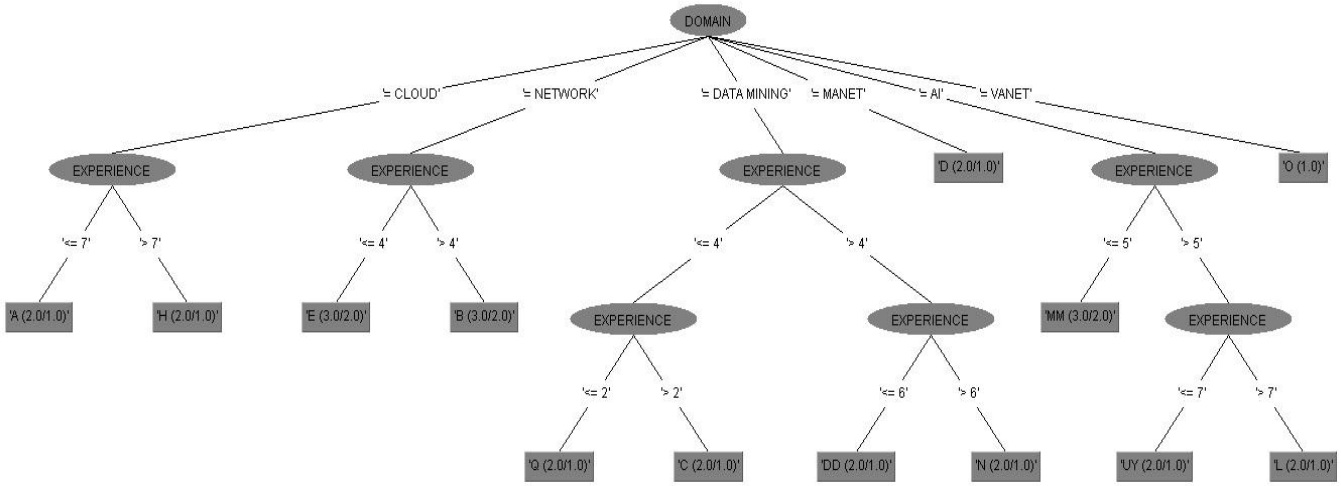

Fig. 5. Classification of Reviewers using J48 in Weka

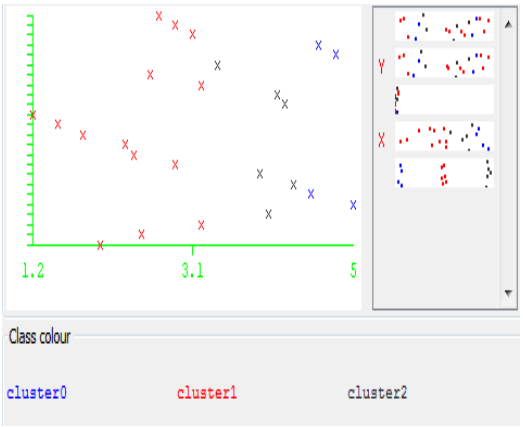

Fig. 6. Reviewers in AI domain

The fig 7 shows the clusters of reviewers in Artificial Intelligence domain with experience less than five years. Three clusters are formed using k-means procedure.

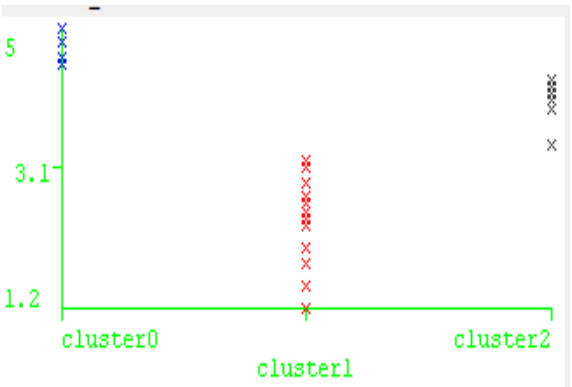

Fig. 7. Clusters of Reviewers in AI with experience <=5

When new reviewer is entered, he is first assigned to the domain class and then placed in the corresponding cluster according to his years of experience and once again the centroids are adjusted to form new clusters.

V. CONCLUSION

In this paper, Ontology based text mining approach is used to cluster both proposals and reviewers. The c4.5 algorithm is used to categorize the proposals with the help of ontology topic identification algorithm. Reviewers are classified based on the domain area and with similar experienced reviewers are grouped into clusters by applying simple K-means algorithm. The categorized proposals are grouped and then assigned to the experts based on their experience. Hence, best project proposals are selected by the reviewers which help the funding organizations to make better decision. This proposed work can also be applied to other natural language proposals. In future, optimization techniques can be applied to improve the efficiency of proposed model.

REFERENCES

[1] ian 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 JIEEE,2012

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

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

[4] Badr HSSINA, AbdelkarimMERBOUHA Hanane EZZIKOURI, Mohammed ERRITALI -A comparative study of decision tree ID3 and C4.5-*(IJACSA), Special Issue on Advances in Vehicular Ad Hoc Networking and Applications*

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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 5 Issue 06, June-2016

[6] 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,July 2002

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

[8] https://en.wikipedia.org/wiki/C4.5_algorithm

[9] S.C.Punitha,K.Muguntha devi, M.PunithaValli,India-Impact of Ontology based approach on document clustering .their advantages.

[10] Srivastava A., Sahami M., Text mining classification, clustering, and applications, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series 2009

[11] A. S. Galathiya, A. P. Ganatra, C. K. Bhensdadia- Classification with an improved Decision Tree Algorithm, International Journal of Computer Applications (0975 – 8887)Volume 46– No.23, May 2012