International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Classification and Assignment of Research Papers using Ontology based Hybrid Approach

Ratish Srivastava†* and A.B.Bagwan†

†Department of Computer Engineering, JSPM's, Rajarshi Shahu College of Engineering, Pune University, Pune, India

Accepted 06 July 2015, Available online 12 July 2015, **Vol.5, No.4 (Aug 2015)**

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 matching of 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 hybrid text mining approach to cluster not only research proposals but also external reviewers based on their research area and then assigning of concerned research 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, Hybrid Approach, Classification and 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 project proposals 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. A Hybrid Approach (which is the combination of Naive Bayes and Ontology Based Classification) is used for this purpose. This ontology based hybrid 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

*Corresponding author **Ratish Srivastava** is a PG Scholar and **Dr. A.B.Bagwan** is working as Professor & Head of Department

works deals with specific subjects and several processes and models are developed for this purpose. (Yong-Hong Sun *et al,* 2008) 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 *et al*, 2011) 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 *et al*, 2008) 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 *et al*, 2006) 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 *et al*, 2011) 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 *et al*, 2008), created an 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 *et al*, 2012) 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. Implementation Details

3.1 Basic Idea

In the proposed method of text mining using ontology based hybrid method for research paper selection it creates an ontology based on previous research 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.

We are using a hybrid approach, in which two algorithms Naive Bayes (McCallum *et al*, 1998) and Ontology based Classifier are combined for better results of classification. Using TFXIDF, Information Gain (IG) as feature selection method, results in some features that are still irrelevant. Therefore, Class Discriminating Measure (CDM), a feature evaluation metric for Naive Bayes that calculates the effectiveness of the feature using probabilities, is used. The results shown in (Chen et al. 2009), indicate that CDM is best feature selection approach than IG. Therefore, instead of using TFXIDF as feature selection method, CDM is used. The term having CDM value less than defined threshold value is ignored. It has been observed that fewer features are left for the computations, this simplifies and speedup the classification task with accuracy. And the remaining terms are used to represent the input unlabelled document; and to match the terms with domain specific ontology, to determine the class of the unlabelled document.

Step 1: For each unclassified document, remove stopwords, punctuations, special symbols, and name entities from the document and represent document as “word set”.

Step 2: For each term in the unclassified document, calculate CDM for that term using equation

$$CDM(w) = |\log P(w|Ci) - \log P(w|\overline{Ci})|$$

where P(w|Ci) = probability that word *w* occurs if class value is *i*

$P(w|\overline{Ci})$ = probability that word *w* occurs when class value is not *i*

i=1 to 7

Step 3: Term having CDM value less than threshold value is ignored. Remaining terms are represented as input document, are used to determine the class of the document.

Step 4: Calculate the frequency of document terms matched with class ontology. Assign class Data Mining to the unlabelled document, if frequency of matching terms with class Data Mining ontology is maximum.
Step 5: If no match is found or a document shows same results for two or more classes then that document is not classified into any class, and left for manual classification.

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.

K-Means Algorithm: For partitioning where each cluster's center is represented by the mean value of the objects in the cluster.
Input: k: the number of clusters,
Output: A set of k clusters.
Step 1: Choose k numbers of clusters to be determined.
Step 2: Choose C_k centroids randomly as the initial centers of the clusters.
Step 3: Repeat
3.1: Assign each object to their closest cluster center using Euclidean distance.
3.2: Compute new cluster center by calculating mean points.
Until
No change in cluster center OR
No object changes its clusters.

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 research 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. research 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.

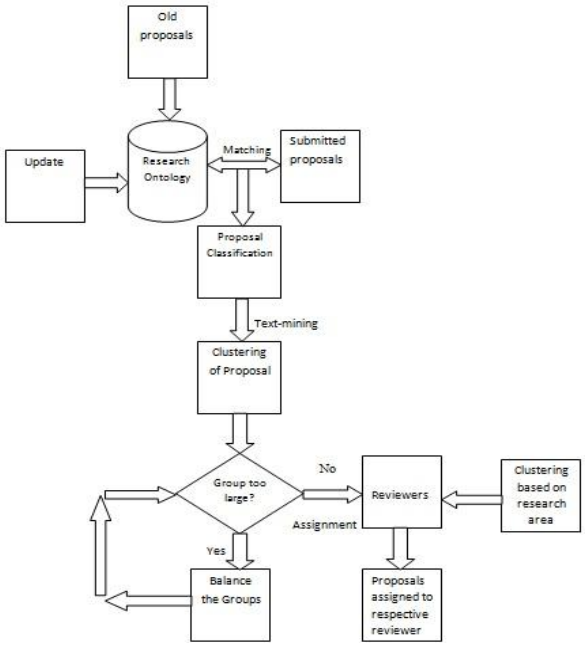

Fig.1 Proposed Framework

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
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,…).
5. After removal of stop words documents, calculate probability of each class (Ci), using equation
P(Ci) = (Total docs in Ci) / (total docs in training set)
After preprocessing and feature extraction steps, each unlabelled document are represented as list of words i.e. w1, w2 ….wn, where w_n is the nth word of the document. Then calculate
P(wj|Ci) = (1+freq. of w_j in class C_i)/(total words in C_i + total words in training set)
Using the above equation we can determine CDM (Class Discriminating Measure) which is used in our hybrid approach.
6. Then cluster the classified research project proposals 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

Precision(c,t) = $n(c,t)/n_c$(2)

Recall(c,t) = $n(c,t)/n_t$

where, *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*.

F(c,t) = (2* Recall(c,t) * Precision(c,t)) / (Recall(c,t) + Precision(c,t)).

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

where *n* is the number of research project proposals and *i* is each predefined research topics .
9. Reviewers are classified according to their area of expertise (aoe).
10. Balanced clusters are assigned to respective reviewers accordingly.

Output Set:
F-Measure to evaluate quality of clustering of project proposal, Accuracy of assignment of proposals to reviewers.

4. Result

4.1 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 respective ontology is generated. From new proposals data sets first all stop words and low frequency words are removed and then classified according to project proposal ontology. After applying Clustering Technique 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.

4.2 Experimentation and Results

We have performed different experiments using our Hybrid Approach. As per study of previous work, it has been found that Hybrid Approach is way better that previously proposed methodologies. As shown in Fig. 2 and Fig. 3 $F_{measure}$, the measure to indentify the accuracy of proposal clustering, is experimented against different parameters by our Hybrid approach.
Fig. 2 represents the $F_{measure}$ against number of proposals. It also compares it with TMM and OTMM approaches.

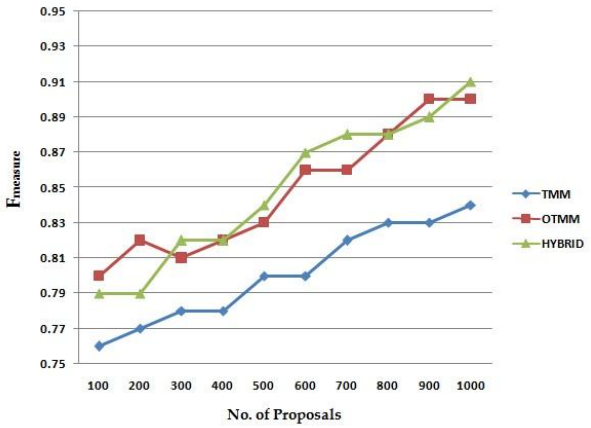

Fig.2 Relationship between $F_{measure}$ and Number of Proposals

Fig. 3 represents the $F_{measure}$ against frequency of keywords. It also compares it with TMM and OTMM approaches.

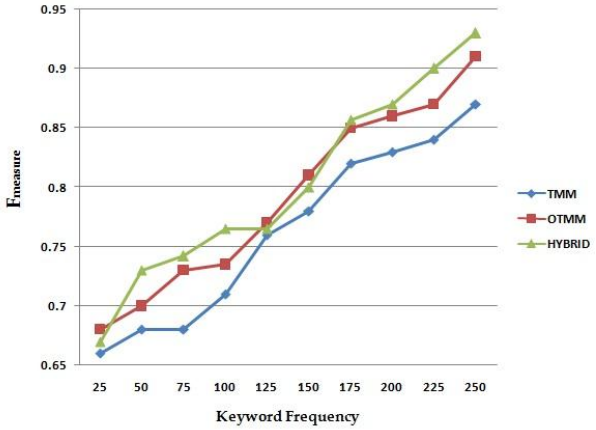

Fig.3 Relationship between $F_{measure}$ and Frequency of Keywords

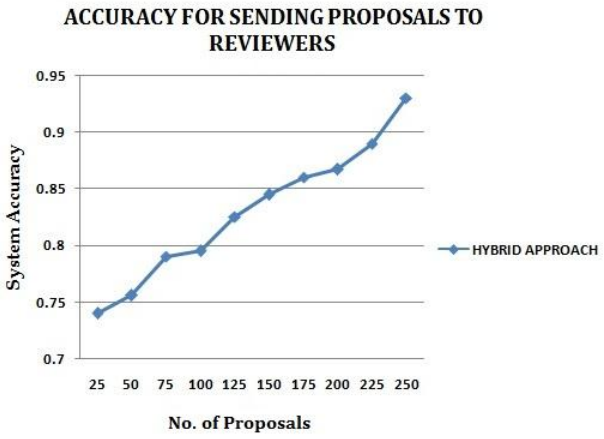

Fig.4 Accuracy of System while sending Proposals to Reviewers

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. 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.

The idea of sending the papers to different reviewers by classifying the category of research paper and the appropriate experts is not proposed in previous researches.

So Fig. 4 represents an individual study on correctness of the proposed system for sending the proposals to relevant reviewers. This study represents that depending upon the increase in number of proposals, the system will learn more and it will become more and more accurate.

Conclusion and Future Works

This paper has presented a text mining method using an ontology based hybrid approach for classification and clustering of research project proposals and assigning the clustered proposals to reviewers accordingly. Research project 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

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

Fabiano D. Beppler (2008), An Architecture for an Ontology-Enabled Information Retrieval, ACM 978-1-59593-753-7/08/0003.

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

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

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

Jian Ma. Wet Xu, Hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu (2012), 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.

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

Matteo Gaeta (2011), 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.

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

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

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

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

McCallum, A. and Nigam, K. (1998), A comparison of event models for naive Bayes text classification. AAAI-98 workshop on learning for text categorization. Technical Report WS-98-05. AAAI Press. pp 41-48.

Chen Jingnian, Huang Houkuan, Tian Shengfeng and Qu Youli (2009), Feature selection for text classification with Naive Bayes. Expert Systems with Applications: An International Journal, Volume 36 Issue 3, Elsevier.