Classification and Allocation of Project Proposals to Reviewers Using Ontology Based Hybrid Mining Technique

Ratish Srivastava[1], Prof. Dr. P.K. Deshmukh[2]
[1] *P.G. Student, Department of Computer Engineering,*
[2] *Professor, Department of Computer Engineering,*
JSPM's, Rajarshi Shahu College of Engineering, Pune University, Pune, India

***Abstract*— As the field of research work is growing quickly and continuously, 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. The activities of finding similar pattern of text effectively and interactively and classifying them are 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 work 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.**

***Index Terms*— Ontology, Hybrid Approach, Text Mining, Classification and Research Project Proposal.**

I. INTRODUCTION

In this fast growing world of electronic data the challenge to manage that data also increases. With the continuous increase in research work, the selection of research project proposals is an important and difficult task, when large numbers of project proposals are collected by any 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. Manual Text Classification is an expensive and time consuming task, as classifying millions of text documents efficiently and with accuracy is not an easy task. Therefore, automatic text classifier is constructed whose accuracy and time efficiency is much better than manual text classification. In 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. 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 (OTMM)) 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.

II. PROBLEM DESCRIPTION

Classifying millions of Research project proposals manually require lots of effort and time. Therefore, automatic Document Classifier is needed to be constructed to manage these documents efficiently in less time with minimum efforts. The problem is to classify large collection of Research Project Proposal Documents into one of the predefined classes. These classes are like: Data Mining, Computer Network,

Mobile Computing, Software Engineering, Artificial Intelligence, etc.

The operations that are needed to do in order to achieve the goals are following:

· First to remove the non essential details and extract only the relevant features from the document for better classification results.

·Then to prepare set of words for the classification task.

· Finally, create Domain (Area) specific Ontology for the Research Project Proposal documents that includes terms related to class.

III. ONTOLOGY BASED HYBRID CLASSIFIER

Pre-processing Phase

Each Uncategorized Project Proposals are represented as Word List. Before processing, stopwords, special symbols, punctuations (<,>, :,{,},[,],^,&,*,(,) etc.) are removed from the documents, as they are irrelevant to the classification task. Stopwords list is manually prepared.

Feature Extraction Phase

Input document may contain redundant or non-relevant data that increases the computations. Therefore, to reduce the feature space by extracting only the relevant features from the proposals, word lists are created.

TFXIDF weighting approach weights the frequency of a term in a document with a factor that discounts its importance if it appears in most of the documents, as in this case the term is assumed to have little discriminating power.

Processing Phase

In this phase, apply classification algorithm which is Ontology Based Hybrid Approach to relevant features extracted from feature extraction phase in order to classify the uncategorized documents into predefined classes.

Hybrid Approach

In hybrid approach, the two algorithms Naive Bayes and Ontology based Text Mining Method (OTMM) 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. 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 uncategorized document; and to match the terms with domain specific ontology, to determine the class of the uncategorized document.

For each uncategorized document, first we remove stopwords, punctuations, special symbols, and name entities from the document and represent document as word list. Then for each term in the uncategorized document, calculate CDM for that term using equation below.

$$CDM(w) = |\log P(w|Ci) - \log P(w|\overline{Ci})| \quad (4)$$

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

So the terms having CDM value less than threshold value is ignored while remaining terms are represented as input document, are used to determine the class of the document. After that we calculate the frequency of document terms matched with class ontology. Assign class Data Mining to the uncategorized document, if frequency of matching terms with class Data Mining ontology is maximum, and 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.

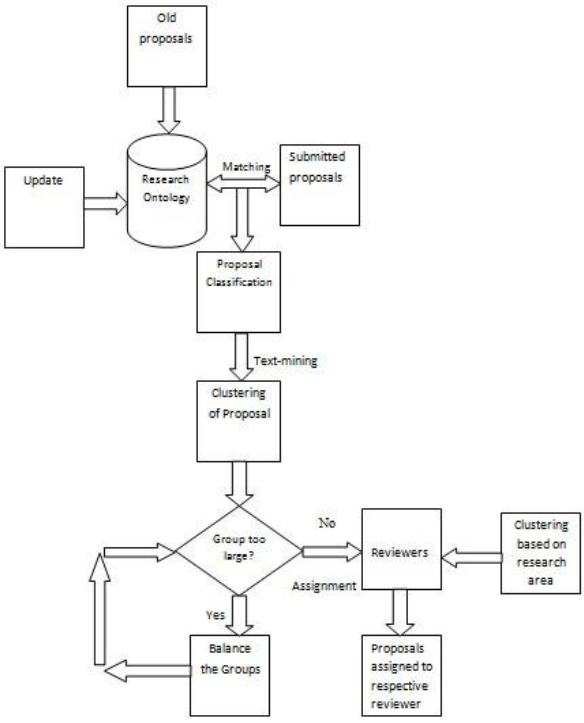

Fig 1. Proposed Framework

IV. RESULT

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.

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 better than 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. F-score is calculated for hybrid classification method, OTMM and TMM using equation below

F-Score = (2*Precision*Recall)/ (Precision + Recall)

Precision = (docs correctly classified in class Ci) / (total docs retrieved in class Ci)

Recall = (docs correctly classified in class Ci) / (total relevant docs in test set that belong to class Ci)

Fig. 2 represents the $F_{measure}$ against number of proposals. It also compares it with TMM and OTMM approaches.

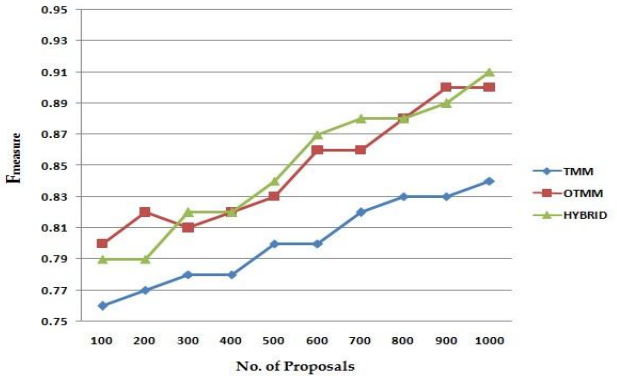

Fig.2 Relationship between $F_{measure}$ and Number of Proposals

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

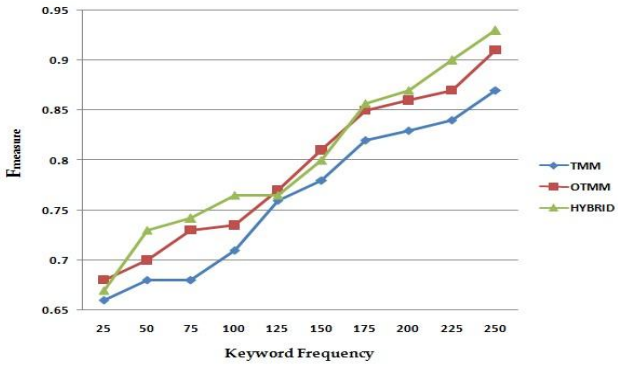

Fig.3 Relationship between $F_{measure}$ and Frequency of Keywords

In the proposed work we are focusing on clustering method for proposals and assignment method of proposals to the respective reviewers.

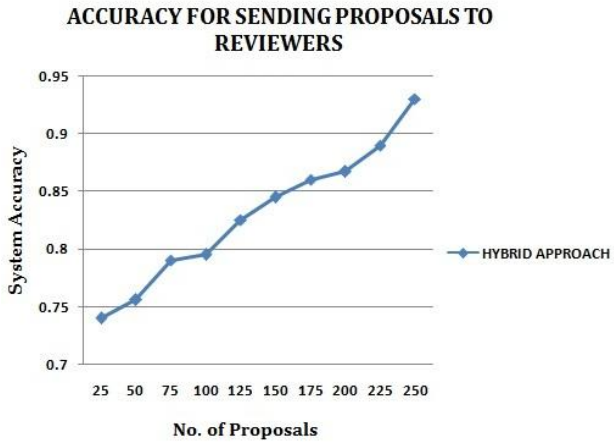

Fig.4 Accuracy of System while sending Proposals to Reviewers

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.

V. CONCLUSION

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.

ACKNOWLEDGMENT

I take this opportunity to thank Prof. Dr. P.K. Deshmukh, my Project Guide, Prof. Dr. A.B. Bagwan Head of Department of Computer Engineering, and all the teaching and non-teaching staff of Computer Engineering Department for their encouragement, support and untiring cooperation. Our sincere thanks to Principal Dr.D.S. Bormane, who is source of inspiration to everybody and always ready to extend helping hands.

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.