International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

ONTOLOGICAL RESEARCH PAPER SELECTION USING TEXT MINING

Kunj Patel[1], Dhananjay Rajput[2], Vasudeo madane[3],Mayur Shendge[4], A.E Patil[5]

[1]*Student, IT, Rajiv Gandhi institute Of Technology, Maharashtra, India*
[2]*Student, IT, Rajiv Gandhi institute Of Technology, Maharashtra, India*
[3]*Student, IT, Rajiv Gandhi institute Of Technology, Maharashtra, India*
[4]*Student, IT, Rajiv Gandhi institute Of Technology, Maharashtra, India*
[5]*Professor, IT, Rajiv Gandhi institute Of Technology, Maharashtra, India*

Abstract - Research and development (R&D) project selection is an decision-making task commonly found in government funding agencies, universities, research institutes, For large number of proposals ,it is common to group according to their disciplines. Text Mining has emerged as a definitive technique for extracting the unknown information from large text Document. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. Ontology's make the task of searching similar pattern of text that to be more effective, efficient and interactive. This proposals are then sent to appropriate expert for peer review. Current methods for grouping proposals are based on manual matching of similar research discipline areas or keywords. This paper represents ontology based text –mining approach for clustering proposals based on similarities in research area. This method can be used to improve the efficiency and effectiveness of research proposal selection processes in government and private research agencies. A knowledge based agent is appended to the proposed system for a retrieval of data from the system in an efficient way. This method concerned with optimization model by geographical region.

Key Words: **Ontology, text mining, clustering, *Clustering analysis, R&D, and knowledge based agent*.**

1.INTRODUCTION

In computer & information science, Ontology as set of concepts i.e knowledge within domain ,& relation between the pairs of concepts. The ontology term has its origin in philosophy. It has been applied in many different ways. The core meaning of ontology within computer and information science is a model for describing the world that consists of a set of properties ,types & relationships types. The submitted research proposals are assigned to experts for review. Four to five reviewers are assigned to review each proposal so as assure accurate and reliable opinions on proposals. To deal with the large volume, it is necessary to group proposals according to their similarities in research disciplines and then assigns the proposal groups to relevant reviewers.

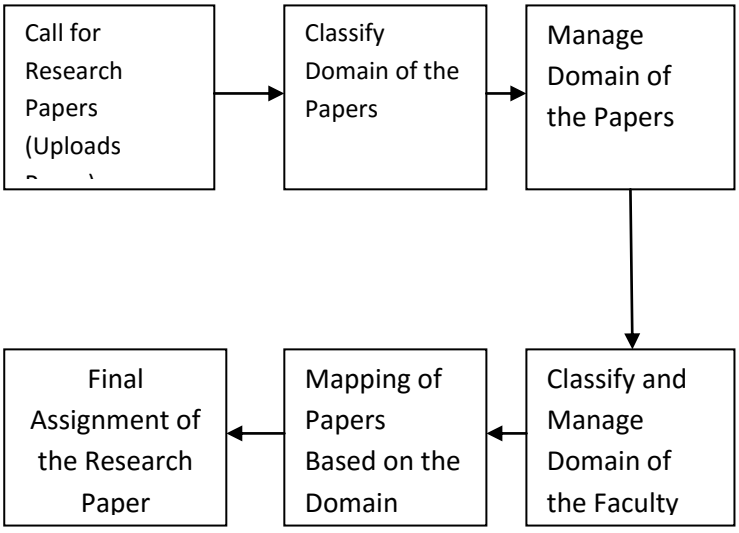

Fig. 1 shows the processes of research project selection i.e. Call for papers (CFP), paper submission, paper grouping, paper assignment to experts, peer review, aggregation of review results, panel evaluation, and final awarding decision. These processes are very similar in other funding agencies, except that there are a very large number of papers that need to be grouped for peer review. Four to five reviewers are assigned to review each paper so as to assure accurate and reliable opinions on papers. To deal with the large volume, it is necessary to group papers according to their similarities in research disciplines and then to assign the paper groups to relevant reviewers. In the first section we Call for Research Paper means uploading Research paper and submitting the details of that paper. Classification of research papers is based on

keywords of papers similar with ontology keywords and frequencies of those keywords. Department members are classified into six groups according to their decision making in research paper selection. Decision making cooperate with each other to accomplish overall goal of selecting research paper. Department members classify research papers and assign them to external reviewer for evaluation and commentary. If department member may not have knowledge about research paper in all research domain and contents of many papers were not fully understood when papers were grouped and while assigning grouped papers to external reviewers. Therefore, there was an effective approach to group the submitted research Papers and assign the papers to external reviewers with computer supports, So we use ontology based text- mining approach is proposed to solve the problem.

2. LITERATURE SURVEY

Chen and Gorla [2] proposed a fuzzy-logic-based model as a decision tool for project selection.

Henriksen and Traynor [3] presented a scoring tool for project evaluation and selection.

Machacha and Bhattacharya [2] proposed a fuzzy logic approach to project selection.

S. Bechhofer et al[2004] developed an OWL Web Ontology Language for storing the keywords[3].

Yildiz and Miksch [2007]designed an ontoX—A method for ontology-driven information extraction[5].

Sun et al. [11] developed a decision support system to evaluate reviewers for research project selection. Finally, Sun et al. [1] proposed a hybrid knowledge-based and modeling approach to assign reviewers to papers for research project selection.

Cheng and Wei [2008] proposed clustering-based category-hierarchy integration (CHI) technique, which is an extension of the clustering-based category integration (CCI) technique. This method was improving the effectiveness of category-hierarchy integration compared with that attained by nonhierarchical category-integration techniques particularly homogeneous [6].

Maedche and Staa [2000] used Text-To-Onto ontology environment using supervised learning[7]. Turban, Zhou and Ma [2004] have been established an group decision support approach to evaluating journals[7].

Yang and Lee [2005] used text mining approach for automatic construction of hyper texts[8] .

Renata Wassermann[2005] developed an Information Retrieval application using ontologies[9].

Matteo Gaeta[2011] have been established for extract relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents using e-learning perspective[10].

Hettich and Pazzani [4] proposed 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 areas might be placed in wrong groups due to the following reasons: first, keywords are incomplete information about the full content of the proposals.

Second, keywords are provided by applicants who may have subjective views and misconceptions, and keywords are only a partial representation of the research proposals. Third, manual grouping is usually conducted by division managers or program directors in funding agencies. They may have different understanding about the research disciplines and may not have adequate knowledge to assign proposals into the right groups.

3. PROPOSED SYSTEM

The proposed system based on the ontology in Text Mining . there are several text- mining method that can used to cluster and classify the documents. But they are with a focus on English text. These methods are not effective in processing the other languages (Such as Marathi language). So many methods was proposed to deal with non-English text, but that are not efficient or sufficiently robust to process research proposals.

To achieve greater efficiency and effectiveness, an Ontology-based Text Mining Method (OTMM) is proposed.

Ontology: An ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts.

First, research ontology constructed according to keywords and it is updated annually (phase 1). Then, new proposals are classified according to discipline areas using a sorting algorithm (phase 2). Next, with reference to the ontology, the new proposals are clustered using a self organize mapping (SOM) algorithm (phase 3). Finally, (phase 4) if the number of proposals in each cluster is still very large, they will be further decomposed into consideration.

Each phase with its details is described in the following sections:

Phase 1. Construction of Research Ontology

The project are used to construct the research ontology according to keywords, A and it get updated annually . Research ontology is a set of research perfect management domain which is also public concept as a domain ontology. Research ontology expressed the topics of research of different disciplines more clearly to more understand .

The research topics of different disciplines can be clearly expressed by K discipline areas and Ak denotes discipline area k (k=1,2,……,K) . A research ontology can constructed in the following three steps :

Step 1) creating the research topics of the discipline Ak

Step 2) Constructing the research ontology-In this the research ontology is categorized according to research areas introduced in the background. Next, it is further divided into some discipline areas. Finally, it leads to research topics in terms of the feature set of disciplines created in step 1. It is more complex than just a tree-like structure. First, there are some cross-discipline research areas (e.g., "data base management system" can be placed under "Information Management" in "Management Sciences" or under "soft computing " in "Information Sciences").

Step 3) Updating the research ontology. Once the project is completed each year, the research ontology is updated according to policy and the change of the feature set. Using the research ontology, the submitted research proposals can be classified into disciplines correctly, and research proposal in one discipline can be clustered effectively and efficiently. The details will be given in the following two sections:

Phase 2: Classifying New Research Proposals Into Disciplines :-Proposals are classified by the respected areas to which they are belong. A simple sorting algorithm is used next for proposals' classification.
i (i = 1, 2, . . . , I), and Sk represents the set of proposals which belongs to area k.
This uses the algorithm and according the keyword of the paper which is match with the started keywords of specific research domain and using this the research proposals are classified .

For n= 1 to N
For j= 1 to J
If Pj belongs to Sn then
Then Pj is added to Sn
End
End

where Pi denotes proposals i (i = 1, 2, . . . , I) and Sk represents the set of proposals which belongs to area k.

Algorithm No. 1 Sorting Algorithm

Phase 3: Clustering Research Proposals Based on Similarities Using Text Mining Text mining technique is used to cluster the proposals in each discipline once the classification is done according to the discipline areas. The five steps are performed to cluster the research proposals. Which are collection of text document , Encoding of text document vector dimension reduction and vector clustering . Self-organized mapping (SOM) algorithm is used cluster the new proposals .

Step 1) Text Document collection: After the classification of research proposal according to the discipline area, the proposal in each discipline Rn(n = 1, 2, . . . , N) are collected for document preprocessing.

Step 2) Text document preprocessing. Text document preprocessing content removing of unwanted and less frequent words from the collected documents to reduce the vocabulary size. The preprocessing consist of following two steps:
(i) Reduction in the vocabulary can be achieved by removing the stop words from the documents. Stop words are the general English words which are often comes in the documents such as 'what', 'it', 'is', 'the', etc.
(ii) Further reduction in vocabulary can be obtain by removing the less frequent words occurred in the documents. In this step words occurred in the document less than some frequency (say 5) can be removed to reduce the vocabulary.

Step 3) Text document encoding. In this step all documents are converted into feature vector representation. F = (f1, f2, ….., fM) ,Where M is number of features selected ,and fi is the term frequency-inverse document frequency (TF-IDF) encoding of the

keyword wi. The TF-IDF encoding of keyword wi can be given by Fi = tfi * log(N/dfi),
where N is the total number of proposals in the discipline, tfi is the term frequency of feature word wi and dfi is the number of proposals containing the word.

Step 4) Vector dimension reduction. It is must to reduce the vector's size because of The dimension of feature vectors is too large by selecting a subset containing the most important term words in terms of frequency. To solve this problem, Latent semantic indexing (LSI. This can be done by selecting the features with the higher tf-idf encoding values and removing the features with lower tf-idf encoding value.

Step 5) Text vector clustering. This step uses an SOM algorithm to cluster the feature vectors based on similarities of research areas. The SOM algorithm is a typical learning neural network model that clusters the given input data with their similarities. Details of the SOM algorithm can be summarized as follow:

Step 1: Initialize network weight vectors wi, initialize Learning rate parameter, define topological Neighborhoods functions and initialize parameter Nq, set k = 0.
Step 2: Check stopping condition. If false, continue: If true, stop.
Step 3: For each training vector x, perform steps 4 to 7.
Step 4: Compute the best match of a weight vector with Input q(x)=max sim(x,wi) i
where sim can be calculated as cosine value of the angle between vectors.
Step 5: For all units in the specified neighborhood where q is the winning neuron, update the weight vectors according to,

$$wi(k+1)=\begin{cases} wi(k)+\mu(k)[x(k)-wi(k)] & I \in Nq(k) \\ wi(k) & \notin Nq(k) \end{cases}$$

where $0< \mu(k)<1$ (the learning parameter)
Step 6: Adjust the learning rate parameter.
Step 7: Approximately reduce the topological Neighborhood Nq(k)
Step 8: set k →k + 1; then go to step 2.

Algorithm No. 2 SOM Algorithm

Phase 4: *Balancing of Research Proposals and Regrouping* if the number of proposals in each cluster is still very large, they will be further decomposed into consideration.. it rearranged according the applicants characteristics to balance the each group or cluster.

3. CONCLUSION AND FUTURE SCOPE

This paper has presented a framework on ontology based text mining for grouping research papers and assigning the grouped paper to reviewers systematically. Research ontology is constructed to categorize the concept terms in different discipline areas and to form relationships among them. It facilitates text-mining and optimization techniques to cluster research papers based on their similarities and then to assign them to reviewer according to their concerned research area. The papers are assigned to reviewer with the help of knowledge based agent.

Future work is needed to replace the work of reviewer by system. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Also there is need to sending message on user's mobile number and also further profile schedule.

REFERENCES

1) 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.

2) K. Chen and N. Gorla, "Information system project selection using fuzzy logic," *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 28, no. 6, pp. 849–855, Nov. 1998.

3) S. Bechhofer et al., OWL Web Ontology Language Reference, W3C recommendation, vol.10, p. 2006-01, 2004.

4) Q. Tian, J. Ma, J. Liang, R. Kowk, O. Liu, and Q. Zhang, "An organizational decision support system for effective R&D project selection," *Decis. Support Syst.*, vol. 39, no. 3, pp. 403–413, May2005.

5) B. Yildiz and S.Miksch, —ontoX—A method for ontology-driven information extraction,‖ in Proc.ICCSA (3), vol. 4707, Lecture Notes in Computer Science, O. Gervasi and M. L. Gavrilova, Eds., 2007, pp. 660–673, Berlin,Germany: Springer-Verlag.

6) 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.

7) A. Maedche and S. Staab, ―The Text-To-Onto ontology learning environment,‖in Proc. 8th Int.Conf. Conceptual Struct., Darmstadt, Germany,2000, pp. 14–18.

8) H. C. Yang and C. H. Lee, "A text mining approach for automatic construction f hypertexts,'Expert Syst. Appl., vol. 29, no. 4, pp. 723–734,Nov. 2005.

9) Christian Paz- Trillo, Renata Wassermann , "An Information Retrieval application using ontologies.

10) 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.

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