International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

Research Paper Forum

Ishwari Gore[1], Tushar Gosavi[2], Komal Dighole[3], Diksha Bhave[4]

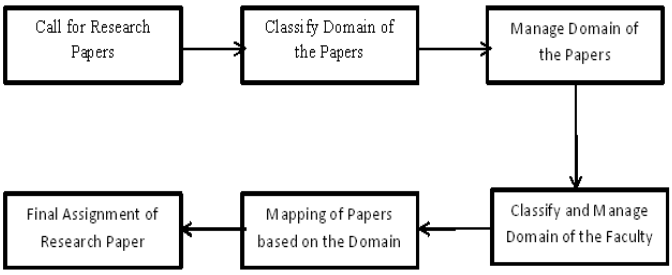
[1,2,3] *Student, COMP, Shivajirao S. Jondhale College of Engineering, Maharashtra, India*
[4]*Asst. Professor, COMP, Shivajirao S. Jondhale College of Engineering, Maharashtra, India*

Abstract - *Research and development (R&D) venture choice is a decision-making venture usually found in authorities investment agencies, universities and studies institutes. For huge range of proposals it's common to group in line with their disciplines. Contemporary methods for grouping proposals are totally based on manual matching of similar research discipline regions or key phrases. This paper represents text –mining approach for clustering proposals primarily based on similarities in research area. This approach may be used to improve the performance and effectiveness of research proposal selection tactics in government and private research agencies.*

Key Words: **Ontology, Text mining, Clustering, R&D, Domain, Keyword.**

1. INTRODUCTION

Challenge choice is a crucial activity in many governmental and non-governmental agencies and additionally very essential venture in numerous educational institutes. Inside the academic institutes after proposals are submitted, the next essential hobby is to categorize proposals and allocate reviewers for the identical. Each domain has to include same traits which assist to group the proposals. There are several text-mining strategies which might be used to cluster and grouping files.



The approach of paper selection i.e. call for papers, paper submission, paper grouping, paper venture to professionals, peer overview, aggregation of evaluate effects, panel evaluation and final awarding selection. These approaches are very similar in different investment corporations, besides that there are a completely huge quantity of papers that need to be grouped for peer evaluation. Within the first segment we call for research papers means uploading of research paper and filing the information of that paper. Classification of research papers is based totally on keywords of papers similar with ontology key phrases and frequencies of those key phrases. Department participants are labeled into six corporations in line with their decision making in research paper selection. Decision making cooperate with each other to accomplish normal intention of selecting research paper proposal. Department participants classify research papers and assign them to outside reviewer for evaluation and remark. If department member may not have whole knowledge about research paper in all research domain and contents of many papers were not fully understood by them then there may be possibilities of occurring errors while grouping and assigning grouped papers to outside reviewers.

Therefore, there was a need of developing an effective approach which will group and assign submitted research papers to external reviewers with computer support. So we have proposed ontology based text-mining approach to clear up the problem.

2. LITERATURE SURVERY

Selection of research projects is an important research topic in research and development (R&D) project management. Previous research deals with specific topics, and several formal methods and models are available for this purpose.

As an instance, Chen and Gorla [1] proposed a fuzzy logic based version as a choice tool for project selection.

Henriksen and Traynor [4] supplied a scoring device for task assessment and selection.

Ghasemzadeh and Archer [2] offered a decision support approach to project portfolio selection.

R. Feldman and J. Sanger [6] used a multiple attribute utility theory for project ranking and selection.

E. Turban et al. [8] established a dynamic programming model for project selection, while Meade and Presley [15] developed an analytic network process model.

M. Konchady [7] proposed a hybrid AHP and integer programming approach to support project selection and Tian et al. [16] suggested an organizational decision support approach for selecting R&D projects.

Kim and Lee [13] presented a method of optimal allocation of papers to reviewers in order to facilitate the selection process.

W. Fan et al. [14] proposed a rotation program method for project assignment.

Choi and Park [9] used text-mining approach for R&D paper screening.

D. Roussinov and H. Chen et al. [10] offered an empirical study to value projects in a portfolio.

Sun et al. [3] developed a decision support system to assess reviewers for research project selection. Eventually, Sun et al. [17] proposed a hybrid knowledge-based and modeling technique to assign reviewers to papers for research project selection.

T. H. Cheng and C. P. Wei [11] presented category-hierarchy integration (chi) method, which is an extension of class integration (cci) method based on clustering. This approach became improving the effectiveness of category-hierarchy integration as compared with that attained by way of non-hierarchical category integration techniques especially homogeneous. Methods have been developed to group papers for peer review tasks.

As an instance, Hettich and Pazzani [2006] presented a text mining approach for grouping of project proposals, identify reviewers, and allocate reviewers to papers [5]. Current methods categories papers in keeping with keywords. Alas, papers with similar studies areas might be positioned in incorrect groups due to the following reasons: first, keywords are incomplete data about the full content of the papers [12].

Second, keywords are given by means of candidates who may additionally have subjective perspectives and misconceptions, and key phrases are simply a partial representation of the research papers. Third, manual grouping is generally performed by division managers or application administrators in funding agencies. They will have specific knowledge about the research disciplines and might not have adequate expertise to assign papers into the right groups.

3. PROPOSED SYSTEM

Yearly, n number of research papers are uploaded on the web by the students on various topics. The procedure behind approving every research paper seems to be unorganized and becomes a hectic job. The proposed system allows students to upload their research paper on any topic, which will be organized by the admin and assigns the research paper to the respective reviewers. The proposed system contains the illustration of domain and ontology makes the information specific for computer that is implicit for human.

Phase1: Constructing a project ontology

Step 1. Constructing domains: - It is a set of research project paper domain which is also public concept as domain ontology. Domain ontology expressed the topics of different disciplines.

Step 2. Creating keywords: - The research papers which are submitted in last five years are used to construct keyword ontology.

Step 3. Update Project ontology: - The project ontology is updated annually.

Phase 2: Classifying New Project Papers into domains using Keywords

Using Keywords research paper proposals are classified by the domains to which they belong. A simple sorting algorithm is used for classification of proposals. This is done using the project ontology as follows:

Step 1: Read PDF file
Step 2: Loop_start
Step 3: Select domain
Step 4: Select Domain specific Keywords
Step 5: Match keywords with PDF text
Step 6: If match found Increment counter
Step 7: Else do nothing
Step 8: End_loop
Step 9: Sort Domain keyword count in descending order
Step 10: Select top first domain
Step 11: Insert into project ontology

Algorithm 1: Sorting Algorithm

Phase 3: Clustering of research paper proposals Based on Similarities Using Text Mining

Once the classification based on keywords is finished. To cluster the papers in each domain text mining technique is used. Following five steps are performed for clustering of project papers:

Step 1. Text document collection- After classification of research papers according to the domain, the documents in each domain Dk (k = 1, 2. . . K) are collected for text document preprocessing. Step 2. Text document preprocessing- The contents of papers are generally unorganized. Since the writings of the papers comprise of non-English characters which are hard to section, the project ontology is utilized to investigate, extricate, and recognize the keywords in the full content of the papers. Finally, a further reduction in the vocabulary can be accomplished through evacuation of words just for few times in all archives.

Step 3. Text document encoding- After dividation of text documents, they are changed into a feature vector representation: V = (v1, v2. . . vM), where M is the number of features selected and vi(i = 1, 2, . . . , M) is the TFIDF encoding [15] of the keyword wi. TF-IDF encoding presents a weighted strategy based on inverse document frequency (IDF) combined with the term frequency (TF) to produce the feature v, such that vi = tfi ∗ log (N/dfi), where N is the total number of proposals in the discipline, tfi is the term frequency of the feature word wi, and dfi is the number of proposals containing the word wi. Thus, project proposals can be represented by corresponding feature vectors.

Step 4. Vector dimension reduction- The dimension of feature vectors is very large; thus, it is important to decrease the size of vector via consequently choosing a subset containing the most critical keywords in terms of frequency. Latent semantic indexing (LSI) is used to take care of the issue. It effectively reduces the dimensions of the feature vectors and also creates the semantic relations among the keywords. LSI is a process for substituting the original data vectors with shorter vectors in which the semantic data is stored. To diminish the dimensions of the document vectors without losing valuable data in a proposal, a term-by-document matrix is formed, where there is one column that corresponds to the term frequency of a document. Moreover, the term-by document matrix is diminished into a set of eigenvectors using singular-value decomposition. The eigenvectors that have the minimal effects on the matrix are then removed. Thus, the document vector formed from the term of the remaining eigenvectors has a too small dimension and retains almost all of the relevant original features.

Step 5. Text vector clustering- This process uses Self-organized mapping to cluster the feature vectors based on similarities of research areas. It is a typical unsupervised learning neural network model that clusters input data with similarities.

Step 1: Initialization of weight vector Wj, initialize learning parameter L & parameter Nq where q is winning neuron, define neighbor function, Set n=0

Step 2: Check the condition for stopping. If it is true then stop else continue.

Step 3: For each new vector x, Continue step 4 To 7

Step 4: For given input Compute best match of the vector q(n) = max simi(n, Wj)

Step 5: For all the units belong to their specified neighbor j belongs to Nq(n), Update weight vector as T Wj (n) +L (n)[x(n)-Wj(n)] j belongs to Nq(n) Wj (n+1) = { Wj (n) j not belongs to Nq (n)

Step 6: Adjust learning parameter

Step 7: Approximately decrease topological neighbor N q (n)
Step 8: Set n=n+1, then go to Step 2

Algorithm No. 2 SOM Algorithm

Phase 4: Balancing Research Proposals and Regrouping

The qualities might be applicable to the colleges to which candidate is an associated. Less disintegration makes the perplexity and feeling cumbersome to the aides. Group size of each group ought to be same.

Step 1: Initialize population and parameters, set p=0.

Step 2: Check ceasing condition. If false, proceed; If true, stop.

Step 3: For one generation, perform Step 4 to 6.

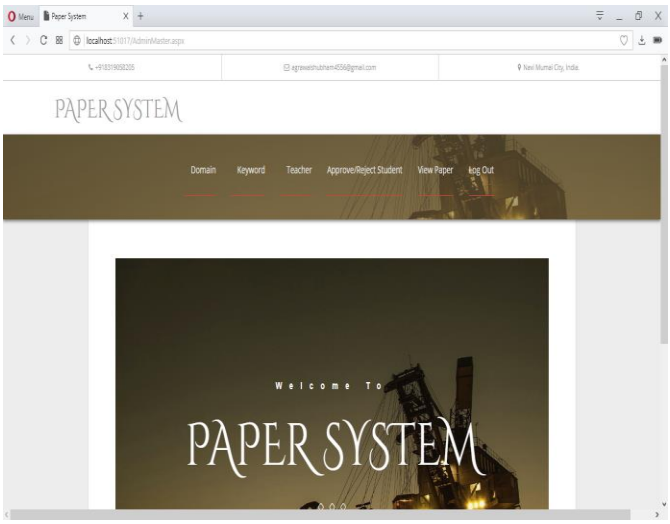
Step 4: Breed new offspring through crossover and mutation.

Step 5: Evaluate value of fitness of parents and offspring.

Step 6: for the formation of new generation, select best-ranking offspring to populate and replace worst-ranking parents.

Step 7: Set p -> p+1; then go to Step 2

Algorithm 3: GA Algorithm



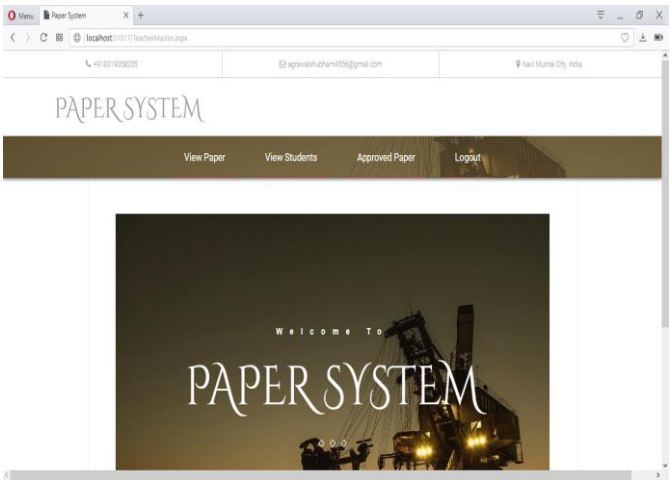
3.1 Modules

The system comprises of 3 major modules with their sub-modules as follows:

Module 1: Admin

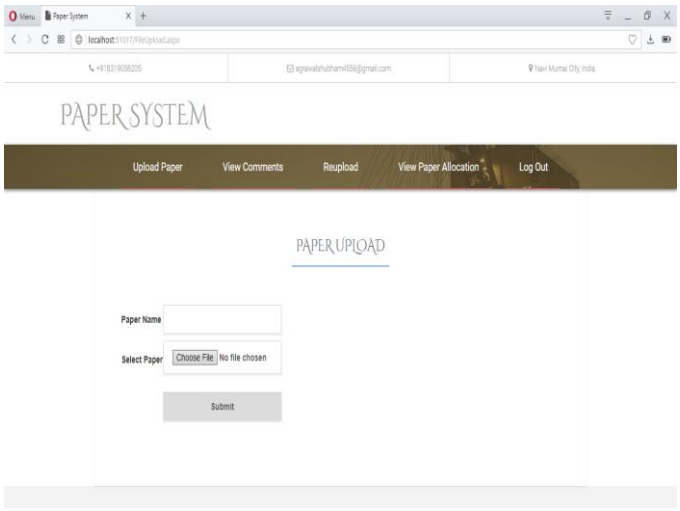
The control of whole system is in the hand of admin. Admin is responsible for creating a panel, adding Domain and

Keywords. He manages and assigns domains to faculty. He has authority to approve or reject student. Admin is able to view all the Project proposals.



Module 2: Reviewer

Reviewers will get notified through email about the allocation of new project proposals uploaded by the students. Reviewers can go through the proposal and can comment on them and can also accept or reject the proposals.



Module 3: Student

After the successful login into the system students can upload their project proposals in PDF format. They can view the status of the paper and waits till gets approved by the guide. Can get comments from reviewers and make appropriate changes to the proposal and can re-upload the proposal.

4. CONCLUSION

This paper has presented an ontology based on text mining method for organization of project papers. Project ontology is built to categorize the concept terms in distinctive area and to form relationship amongst them. It facilitates text-mining and optimization approach to cluster papers based on their similarities and then to balance them with respect to the size of the domain. The proposed method may be used to expedite and enhance the papers grouping technique. It also presents a formal procedure that enables similar papers to be grouped collectively in a professional and ethical way. The proposed technique also can be utilized in different educational institutes that face records overload problems.

5. FUTURE SCOPE

Future work is needed to cluster external reviewers predicated on their domain and to assign grouped papers to guides systematically. Subsequently, the method can be expanded to avail in finding a better match between project papers and their guides. Withal, there's a need to compare the consequences of manual classification to text-mining classification. Conclusively, there is need to supersede the work of reviewer by way of gadget.

REFERENCES

[1] 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.

[2] F. Ghasemzadeh and N. P. Archer, “Project portfolio selection through decision support,” Decis. Support Syst., vol. 29, Jul. 2000, no. 1, pp. 73–88.

[3] 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, Feb.2008, no. 1, pp. 158–170.

[4] D. Henriksen and A. J. Traynor, —A practical R&D project selection scoring tool, IEEE Trans. Eng. Manag., vol. 46, May 1999, no. 2, pp. 158–170.

[5] 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.

[6] R. Feldman and J. Sanger, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. New York: Cambridge Univ. Press, 2007.

[7] M. Konchady, Text Mining Application Programming. Boston, MA: Charles River Media, 2006.

[8] E. Turban, D. Zhou, and J. Ma, —A group decision support approach to evaluating journals, Inf. Manage., vol. 42, Dec. 2004, no. 1, pp. 31–44.

[9] C. Choi and Y. Park, —R&D proposal screening system based on text mining approach, Int. J.Technol. Intell. Plan., vol. 2, 2006, no. 1, pp. 61– 72.

[10] D. Roussinov and H. Chen, —Document clustering for electronic meetings: An experimental comparison of two techniques, Decis. Support Syst., vol. 27, Nov. 1999, no. 1/2, pp. 67–79.

[11] 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, Mar. 2008, no. 2, pp. 410–424.

[12] 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.

[13] H. J. Kim and S. G. Lee, —An effective document clustering method using user- adaptable distance metrics, in Proc. ACM Symp.Appl.Comput., Madrid, Spain, 2002, pp. 16–20.

[14] W. Fan, D. M. Gordon, and P. Pathak, "An integrated two-stage model for intelligent information routing," Decis. Support Syst., vol. 42, Oct. 2006, no. 1, pp. 362–374.

[15] L. M. Meade and A. Presley, —R&D project selection using the analytic network process,‖IEEE Trans. Eng. Manag., vol. 49, Feb. 2002, no. 1, pp. 59– 66.

[16] 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, May2005, no. 3, pp. 403–413.

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