GLOBAL JOURNAL OF ADVANCED RESEARCH
(Scholarly Peer Review Publishing System)

PROJECT PAPER SELECTION USING ONTOLOGY BASED TEXT-MINING

Miss. Thorat Madhuri, Miss. Naikwadi Chaitali, Miss. Gaikwad Swapnali,
Miss. Mhaske Seema &
Guided by: Mr. N. B. Kadu
Department of Computer Science,
Pravara Rural Engineering College, Loni,
India

ABSTRACT
Research projects selection is an important activity in many governmental and non-governmental organizations and also very important task in various educational institutes. The project proposals are submitted to the institutes and then are assigned to guides for review.

In the educational institutes after proposals are submitted, the next important activity is to categorize proposals and allocate reviewers for the same. Each domain should contain identical characteristics which help to group the proposals. For example, if the proposals in a domain comes into the same primary group (e.g., cloud computing) and the proposals are grouped based on keywords listed in proposals. Current methods for grouping proposals are based on manually searching the domain areas which is less accurate. There are several text-mining methods (TMM) that are used to cluster and grouping documents. TMMs which deal with only English text. To solve this problem, an ontology-based text mining method is proposed**.**

KEYWORDS: Ontology, clustering, text mining, proposal selection.

1. INTRODUCTION

Projects selection is an important activity in many governmental and non-governmental organizations and also very important task in various educational institutes. In the educational institutes after proposals are submitted, the next important activity is to categorize proposals and allocate reviewers for the same. Each domain should contain identical characteristics which help to group the proposals. There are several text-mining methods (TMM) that are used to cluster and grouping documents. TMMs which deal with only English text. To solve this problem, an ontology-based text mining method is proposed**.**

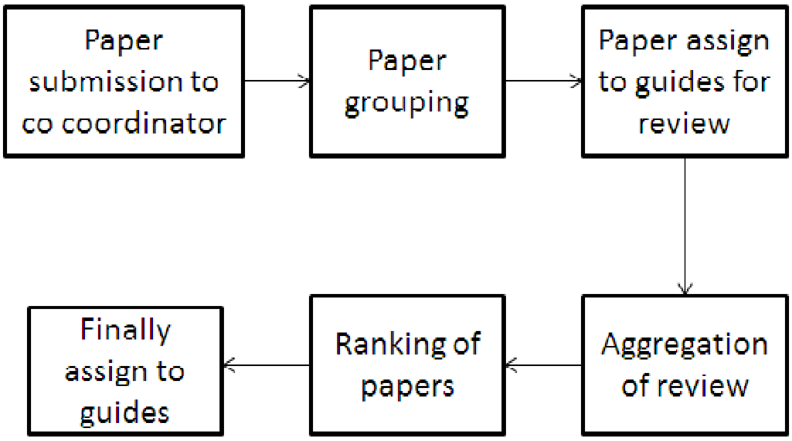

Fig 1: Existing Project paper selection process

Fig 1 shows existing process of selection of project proposals in the institutes. i.e. paper submission, grouping, papers assign to guides for peer review, governmental and non-governmental organizations such as funding agencies.

In educational institutes, numbers of project proposals are received. To deal with large data, it is necessary to cluster the proposals and assign to guides for reviews. In the educational institutes there are number of departments. Computer department is classified according to the various clouds such as networking, text mining, android, etc. Each cloud is further divided into sub types which focus on more specific area.

2. `LITERATURE REVIEW

The selection of project paper in existing i.e. in colleges is done manually means the project papers are submitted to their project coordinators and these papers according to their domains or the keywords in that paper are classified into groups (domains), this is done manually i.e.by humans. Following diagram shows the procedure of manually dividing the project papers submitted by students.

This paper presents a hybrid method for grouping educational project proposals for project selection. For this different text-mining, multilingual ontology, optimization, and statistical analysis techniques to cluster research proposals based on their similarities are used. The experimental results indicated that the method can also be used to improve the efficiency and effectiveness of the project selection process for educational institutes.

3. PROPOSED SYSTEM

In the educational institutes, after proposals are submitted, the next important task is to group proposals and assign them to guides. The proposals in each domain should have similar characteristics. However, if the number of proposals is large, it is very difficult to group proposals manually. So the proposed system is based on the ontology in text mining .we form four phases to process the selection work. The proposed system contains the representation of domain and ontology makes the knowledge explicit for computer which is implicit for human.

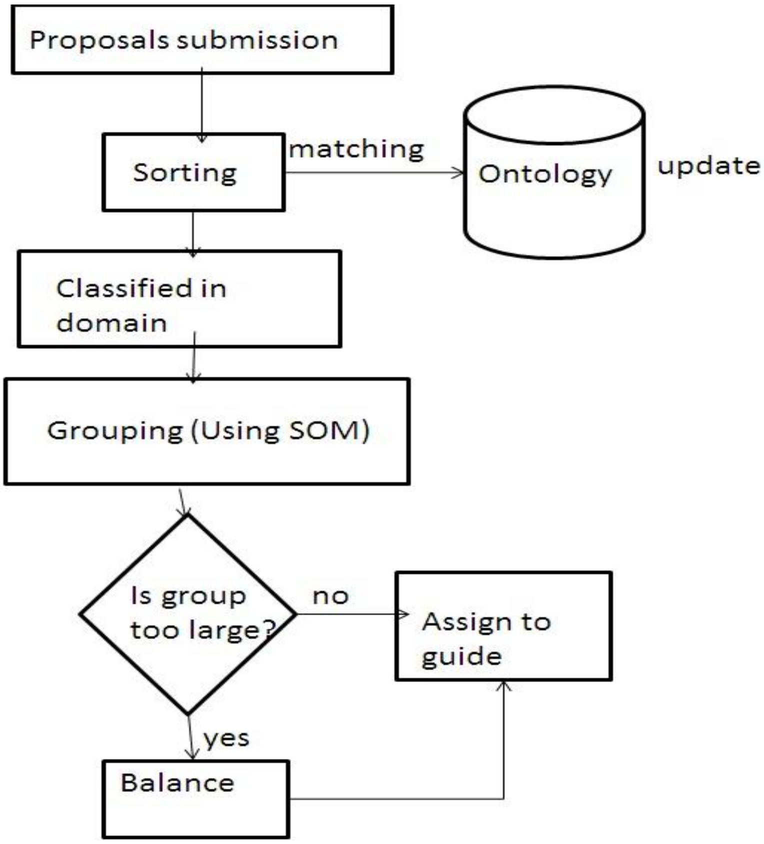

Fig 2: Flow of OTMM

Fig 2 shows phases in OTMM: (Phase 1) - an ontology containing the projects submitted in latest five years is constructed according to keywords, and it is updated annually. (Phase 2)- Then, new project proposals are classified according to domain using a sorting algorithm. (Phase 3)-Next, with reference to the ontology, the new proposals in each domain are clustered using a self-organized mapping (SOM) algorithm. (Phase 4)-Finally, if the number of proposals in each cluster is still very large, they will be further classified into subgroups. The text mining patterns are extracted from natural language text it makes the text mining different from regular data mining. This focuses on structured databases of facts. Now each phase is described in the following sections.

Phase1: Constructing a project Ontology

Step 1) Creating Project topics: The projects which are submitted in last five years are used to construct the project ontology according to keywords and it gets updated annually (As shown in fig 3).
Step 2) Creating Project ontology: It is a set of research project paper domain which is also public concept as domain ontology. Project ontology expressed the topics of different disciplines more clearly to more understand.
Step 3) Update Project ontology: Once the project is submitted, it is updated every year.

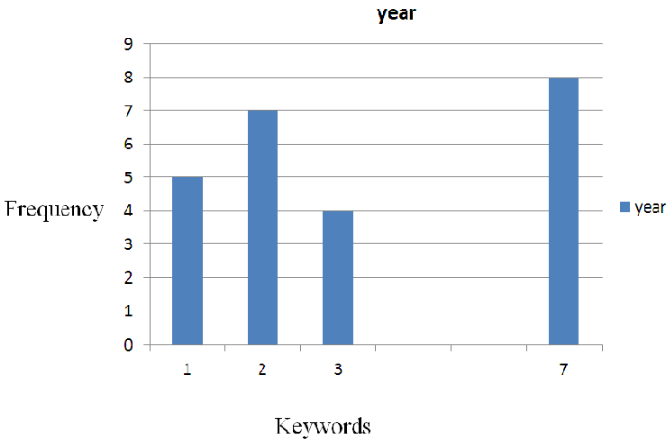

Fig 3: Keywords in year

Phase 2: Classifying New Project Papers into domains using SOM algorithm

Proposals are classified by the domains to which they belong. A simple sorting algorithm is used next for proposals' classification. This is done using the project ontology as follows.
Suppose that there are *n* discipline areas, and *Sn* denotes area *n* (*n* = 1, 2. . . *N*). *Pi* denotes proposals *j* (*j* = 1, 2. . . *I*), and *Pj* represents the set of proposals which belongs to area *n*. Then, a sorting algorithm can be implemented to classify proposals to their discipline areas. Shown below:

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

Algorithm 1: Sorting Algorithm

Phase 3: Clustering Project Proposals Based on Similarities Using Text Mining

After the classification is done according to the keywords Text mining technique is used to cluster the papers in each domain. The five steps are performed to cluster the project papers. Which are collections of 1. Text document, 2. Text document processing, 3. Encoding of text document, 4.vector dimension reduction and 5. Vector clustering. Self-organized mapping (SOM) algorithm is used cluster the new proposals. Fig 3 shows the actual process of text mining. The details of each step are as follows.

Step 1) *Text document collection*- After the project papers are classified according to the domain, the documents in each domain *Ak* (*k* = 1, 2. . . *K*) are collected for text document preprocessing.
Step 2) *Text document preprocessing*- The contents of papers are usually non structured. Because the texts of the papers consist of non-English characters which are difficult to segment, the project ontology is used to analyze, extract, and identify the keywords in the full text of the papers. Finally, a further reduction in the vocabulary can be achieved through removal of words only for few times in all documents.

Step 3) *Text document encoding*- After text documents are segmented, they are converted 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 [18] of the keyword *wi*. TF-IDF encoding describes a weighted method 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, research proposals can be represented by corresponding feature vectors.

Step 4) *Vector dimension reduction*- The dimension of feature vectors is often too large; thus, it is necessary to reduce the vectors' size by automatically selecting a subset containing the most important keywords in terms of frequency. Latent semantic indexing (LSI) is used to solve the problem. It not only reduces the dimensions of the feature vectors effectively but also creates the semantic relations among the keywords. LSI is a technique for substituting the original data vectors with shorter vectors in which the semantic information is preserved. To reduce the dimensions of the document vectors without losing useful information in a proposal, a term-by-document matrix is formed, where there is one column that corresponds to the term frequency of a document. Furthermore, the term-by document matrix is decomposed into a set of eigenvectors using singular-value decomposition. The eigenvectors that have the least impacts on the matrix are then discarded. Thus, the document vector formed from the term of the remaining eigenvectors has a very small dimension and retains almost all of the relevant original features.

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 unsupervised learning neural network model that clusters input data with similarities. Details of the SOM algorithm can be summarized.

```
Step 1: Initialization of weight vector yj, initialize learning parameter u & 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, Yj)
Step 5: For all the units belong to their specified neighbor j belongs to Nq(n), Update weight vector as T
Yj (n) +u (n)[x(n)-Yj(n)] j belongs to Nq(n)
Yj (n+1) = {
Yj (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 2: SOM Algorithm

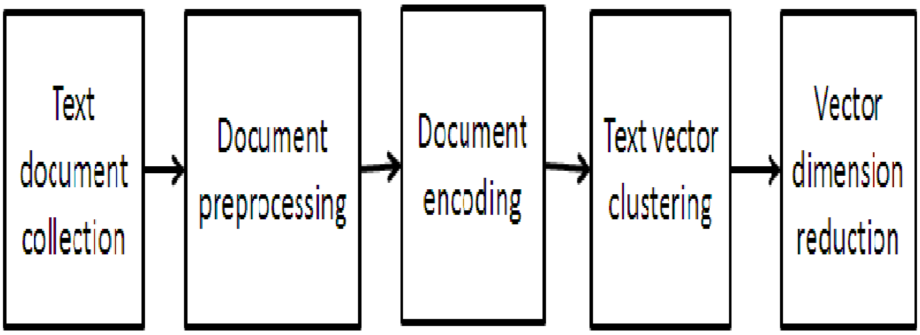

Fig 3: Text-mining process

Phase 4: Balancing Research Proposals and Regrouping

If the each group get large (i.e. max than 20) due to number of papers in it. It rearranged according the applicants characteristics to balance the each group or cluster. The characteristics may be the universities which applicant is an affiliated. Less decomposition generates the confusion and feeling uncomfortable to the guides. And group size of each group should be same.

```
Step 1: Initialize population and parameters, set k=0.
Step 2: Check stopping condition. If false, continue; If true, stop.
Step 3: For one generation, perform Step 4 to 6.
Step 4: Breed new offspring through crossover and mutation (genetic operation).
Step 5: Evaluate the fitness values of parents and offspring.
Step 6: Select best-ranking offspring to populate and replace worst-ranking parents to form a new generation.
Step 7: Set k -> k+1; then go to Step 2
```

Algorithm 3: GA Algorithm

4. VALIDATING THE PROPOSED METHOD

To validate the proposed paper, several experiments are conducted using the previous accepted projects. First, two experiments ($E1$ and $E2$) are constructed to evaluate the quality of clustering project papers. Second, one experiment ($E3$) is used to validate the effectiveness and efficiency of balancing project papers. In $E1$, papers in the domain called networking are randomly selected. In $E2$, papers in the domain named text mining are randomly used. In $E3$, papers with similar topics are randomly selected.

5. CONCLUSION

This paper has presented an Ontology based on Text Mining Method for grouping of project papers. A project ontology is constructed to categorize the concept terms in different domain and to from relationship among them. It facilitates text-mining and optimization technique to cluster papers based on their similarities and then to balance them according to the size of domain. The proposed method can be used to expedite and improve the papers grouping process. It also provides a formal procedure that enables similar papers to be grouped together in a professional and ethical manner.

The proposed method can also be used in other educational organizations that face information overload problems. Future work is needed to cluster external guides based on their domain and to assign grouped papers to guides systematically. Also, there is a need to compare the results of manual classification to text-mining classification. Finally, the method can be expanded to help in finding a better match between project papers and their guides.

6. REFERENCES

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

[2] 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.

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

[4] 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.

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

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

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

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

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

[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] 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.

[12] 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.

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

[14] G. H. Lim, I. H. Suh, and H. Suh, “Ontology-based unified robot knowledge for service robots in indoor environments,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 3, pp. 492–509, May 2011.

[15] C. Lu, X. Hu, and J. R. Park, “Exploiting the social tagging network for web clustering,” *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, vol. 41, no. 5, pp. 840–852, Sep. 2011.

[16] J. C. Trappey, C. V. Trappey, F. C. Hsu, and D. W. Hsiao, “A fuzzy ontological knowledge document clustering methodology,” *IEEE Trans. Syst., Man, Cybern. B, Cybern*, vol. 39, no. 3, pp. 806–814, Jun. 2009.

[17] M. Zhang, Z. Lu, and C. Zou, “A Chinese word segmentation based on language situation in processing ambiguous words,” *Inf. Sci.*, vol. 162, no. 3/4, pp. 275–285, Jun. 2004.

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

[19] O. Liu and J. Ma, “A multilingual ontology framework for R&D project management systems,” *Expert Syst. Appl.*, vol. 37, no. 6, pp. 4626–4631, Jun. 2010.

[20] T. Ong, H. Chen, W. Sung, and B. Zhu, “Newsmap: A knowledge map for online