OCMM Based clustering in Research as well as Reviewers & Optimal Allocation of Proposals to Experts by Ontology Matching

[1] T.Sahaya Arthi Jeno, [2]D.Jim Solomon Raja
[1,2]Assistant Professor
[1,2]Department of Computer Science and Engineering
[1,2]Christ the King Engineering College
[1,2]Karamadai, Coimbatore

Abstract- **Research project proposals selection is a remarkable bustle in organization, when large numbers of research proposals are collected. The common task is to group research proposal based on their similarity in disciplined area and research area. Current techniques, Ontology-based Text Mining Method (OTMM) has been used for grouping proposals based on manual matching of similar research discipline areas. In this paper, Ontology-based Concept Mining Model(OCMM) is presented to cluster Research Project Proposals as well as External Research Reviewers based on their research discipline areas and overcome the Ontology Mapping Problem to allocate grouped proposals to experts group for peer review systematically . This methodology is used to improve the efficiency and effectiveness of research project selection processes with escalating figure of proposals and external reviewers.**

Keywords-Classification, Clustering, Concept-based analysis, Ontology, Peer view, Research Project Selection, Ontology Matching, Genetic Algorithm

I. INTRODUCTION

The Research project proposals selection is an important and challenging task by the government and private funding agencies, when large numbers of research proposals are collected. Optimal allocation of research project proposals is a challenging multi -process starts with invoking research proposals by a funding agency. The proposal invoking is distributed to relevant communities such as universities or research institutions. Submission of the research proposals by many institutes and organizations are assigned to experts for peer view based on their similarity. The review results are examined, and the proposals are then ranked based on the aggregation of the experts' review results.

Figure.1 shows the processes of research project selection. Invoking proposals, Proposals submission, grouping of proposals, Assigned to experts, Peer Review, Aggregation of review, Evaluation and Ranking of proposals, Funding Decision are very similar activities involved in all the funding agencies [1].In Research Funding Agencies, the number of research proposals received has more than doubled in the past few years. To assure accurate and reliable opinions on proposals four to five reviewers are assigned to review each proposal. For very large number of proposals received by the agencies need to group the proposal for peer review [2].

Governments as well as private research funding agencies are made up of several scientific departments, bureaus, general office, and associated units. Decision-making units like scientific departments are responsible for funding recommendations and management of funded projects. Scientific departments are classified according to scientific research areas, like mathematical and physical sciences, material sciences, engineering and earth sciences, information sciences, chemical sciences, life sciences, and management sciences. Departments are further divided into several divisions with a focus on more specific research areas. Example, the Department of Management Science is further divided into the following divisions: Management Science and Engineering, Policy and Macro Management, and Business Administration. Furthermore, discipline areas of divisions are called as programs.

The department for selection process (i.e.) Division managers or program directors can assign the grouped proposals to the external reviewers for evaluation and rank them based on their aggregation. In manual based grouping, the department of selection process is responsible for grouping and they may not have adequate knowledge regarding all the issues and areas of the research proposals and the contents of many proposals were not fully understood. There are several text-mining methods used to classify and cluster the documents. These approaches are developed only with a focus on English text. Text-mining methods which deals with English text not effective in processing Chinese text. Chinese text consists of string of Chinese characters, whereas English text uses words. And also Chinese text has no delimiters to mark word boundaries but English text uses a space as a word delimiter. There are several methods were proposed to deal with Chinese text but they are not sufficiently robust to process research proposals. Therefore, there was an urgent need for an effective and feasible approach to group the submitted research proposals efficiently based on their disciplined areas by analyzing full text information of the proposals with computer supports. An ontology-based concept mining technique is proposed to solve the problem. In particular, it possesses the following advantages, the proposed strategy is fundamentally different from the existing Ontology based Text-mining methods, and it outperforms the available Text-mining methods (TMMs). Because of this essential change, the proposed strategy overcomes the drawbacks of Text-mining methods, such as manual based research proposals grouping and assignment. This proposed approach can provide us a way to easily classify and group the research proposals and reviewers and also used in funding agencies that face information overload problems. The proposed work has presented an framework on ontology based concept mining for grouping research proposals and assigning the grouped proposal to reviewers systematically as well as semantically. The Ontology based concept mining model is very user friendly and time consuming.

The residue of this paper is organized as follows. Section II reviews the research background and objective. The existing method is described in Section III and research methodologies are described in Section IV. At last, Section V provides the conclusion and its future work.

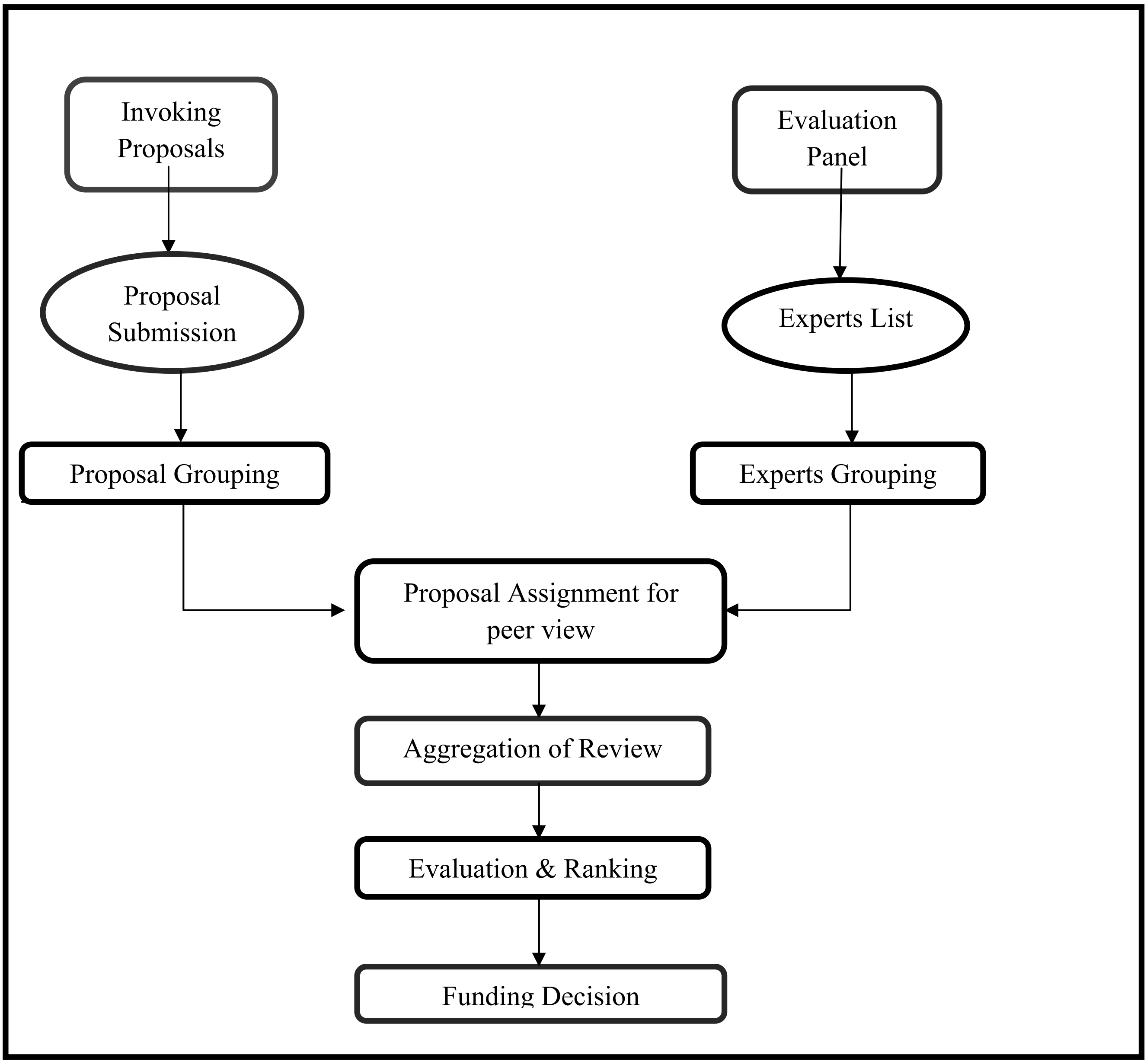


Figure.1 Research project selection processes

II. Research Background And Objectives

Research and Development (R&D) project selection is complicated and knowledge intensive decision-making process where decision models and knowledge rules play an important role. Before using OCMM several techniques are used for R&D project selection. The following are some research work that leads for proposed technique:

A R&D Project Selection Approaches

Q.Tian et al. [3] proposed Hybrid knowledge and model system for R&D project selection, which integrates mathematical models with knowledge rules. This system is designed to support the whole decision process of R&D project selection and has been used in the selection of R&D projects in the NSFC. K.Chen et al. [4] proposed Fuzzy-logic-based model as a decision tool for project selection, which smoothly aids decision makers dealing with uncertain or incomplete information without losing existing quantitative information. A.D.Henriksen et al. [5] presented An improved scrolling tool for R&D project evaluation and selection is presented that ranks project alternatives based on the criteria of relevance, reasonableness, risk and return. Algorithm scoring explicitly incorporates tradeoffs among the evaluation criteria and calculates a relative measure of project value by taking into account the fact that the value is the function of both merit and cost. W.D.Cook et al. [6] presented Peer review of research proposals and articles is an essential element in R&D processes world-wide. An integer-programming set-covering model and a heuristic procedure solves the assignment problem and maximizes the number of proposal-pairs that will be evaluated by one or more reviewers and also this approach should facilitate meaningful aggregation of partial rankings of subsets of proposals by multiple reviewers into a consensus ranking. S.Hettich et al. [7] established Prototype application deployed at the U.S.National Science Foundation for ancillary program directors in identifying reviewers for proposals. Prototype application helps program directors sort proposals into panels and find reviewers for proposals. It extracts information from the full text of proposals both to learn about the topic of proposals and the expertise of reviewers. The solution that was implemented and experience in using the solution within the workflow of NSFC. Y.H.Sun et al. [10] developed a decision support system to evaluate reviewers for research project selection.Girotra et al. [11] offered an empirical study to value projects in a portfolio.

B Ontology-based Framework

Jian Ma et al. [1] proposed an OTMM for grouping of research proposals. Research proposals ontology is constructed to categorize the concept terms in different discipline areas and to form relationships among them. It facilitates techniques like text-mining and optimization is used to cluster research proposals based on their similarities and then to balance them according to the applicants' characteristics. The OTMM improved the similarity in proposal groups as well as took into consideration the applicants' characteristics (e.g., distributing proposals equally according to the applicants' affiliations) and also, promotes the efficiency in the proposal grouping process. Preet Kaur et al. [2] developed Ontology based classification and clustering approach is used for grouping the Research Proposals and the research Reviewers. Combination of Data Mining techniques is used with the help of Ontology. It can provide a way to easily classify and group the research proposals and the reviewers. The proposed work efficiently classifies the research areas. S.Bechhofer et al [13] developed an OWL Web Ontology Language for storing the keywords.Yildiz et al [14] designed an ontoX—A method for ontology-driven information extraction. V.M.Navaneethakumar, Dr.C.Chandrasekar[2012] —A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model [15] referred this paper for clustering the proposals based on concept.N.Arunachalam et al[16] developed A knowledge based agent is appended to the proposed system for a retrieval of data from the system in an efficient way.

C Text Mining Approaches

Huan-Chao Keh et al. [12] developed Filtering measure is used for feature selection in Chinese text categorization system. Term Frequency-Inverse Document Frequency (TF-IDF) to strengthen important keywords' weights and weaken unimportant keywords' weights. We use category priority to represent the knowledge of the categories (i.e.) relationship between two different categories.

This paper has been enlarged the work of Jian Ma et al.[1] and M.Lavanya and Rajkumar[17] and also grouped proposals were assigned to grouped reviewers systematically. These methods are used to improve the efficiency and effectiveness of research project selection processes with escalating quantity of proposals and external reviewers.

III. EXISTING SYSTEM

The existing system is an Ontology-Based Text-Mining Method (OTMMs) to cluster research proposals based on their similarities in research areas. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. It consists of a axioms, relationships and set of concepts that describe a domain of interests and represents an agreed-upon hypothesis of the domains of real-world setting. Using Ontology Implicit knowledge for humans is made explicit for computers . Thus, ontology can automate information processing and can facilitate text mining in a specific domain (such as research project selection). An ontology based text mining skeleton has been built for clustering the research proposals according to their discipline areas. Text mining notice generally to the process of extracting interesting information and knowledge from unstructured text. Inequality between regular data mining and text mining is that text mining patterns are extracted from natural language text rather than from structured databases of facts.

Constructing Research Ontology, a research project proposal ontology containing the projects funded in latest five years is assembled according to keywords, and it is updated annually. While considering domain ontology research ontology is a public concept set of the research project management domain. Research ontology is used to express the research topics of different disciplines.

Classifying New Research Proposals, new research proposals are classified into number of classes according to the keyword stored in ontology.

Clustering: Research Proposals Based on Similarities Using Text Mining, proposals in each discipline are clustered using the text- mining technique. The clustering process consists of five steps, text document collection, preprocessing, encoding, vector dimension reduction, and text document vector clustering. The newly submitted proposals in each discipline are clustered using a self-organized mapping algorithm (SOM).

Balancing Research Proposals and Regrouping Them by Considering Applicants' if the number of proposals in each cluster is still very large, they will be further break up into subgroups where the applicants' characteristics like affiliated universities are taken into consideration. Reviewers may feel confused and uncomfortable when evaluating proposal that may have poor decomposition so it is advisable that the applicants' characteristics in each proposal group should be as diverse as much as possible.

IV. RESEARCH METHODOLOGY

Proposed system presents a framework on ontology based concept mining approach to cluster research project proposals, and external reviewers based on their research area and to assign concerned research proposals to reviewers systematically. In the R&D, after proposals are submitted, the next challenging task is to group proposals and assign them to reviewers. The research proposals in each group should have similar trait. In case, if the proposals in a group fall into the same primary research discipline (e.g., data mining) and the number of proposals is small, the manual grouping based on keywords listed in proposals can be used and assign them to reviewer manually. However, if the number of proposals is large, that is very difficult to group proposals and assign them to reviewer manually. So the proposals and external reviewers are classified using ontology and clustered using concept based mining techniques and last it is submitted to reviewer systematically. Figure 2 shows the proposed architecture of entire system.

A. Research Ontology as well as Reviewers Ontology construction

i) Creating Research Topics:

The keywords of the supported research projects each year are collected, their frequencies are counted. And the keyword frequency is the sum of the same keywords that appeared in the discipline during the most recent five years.

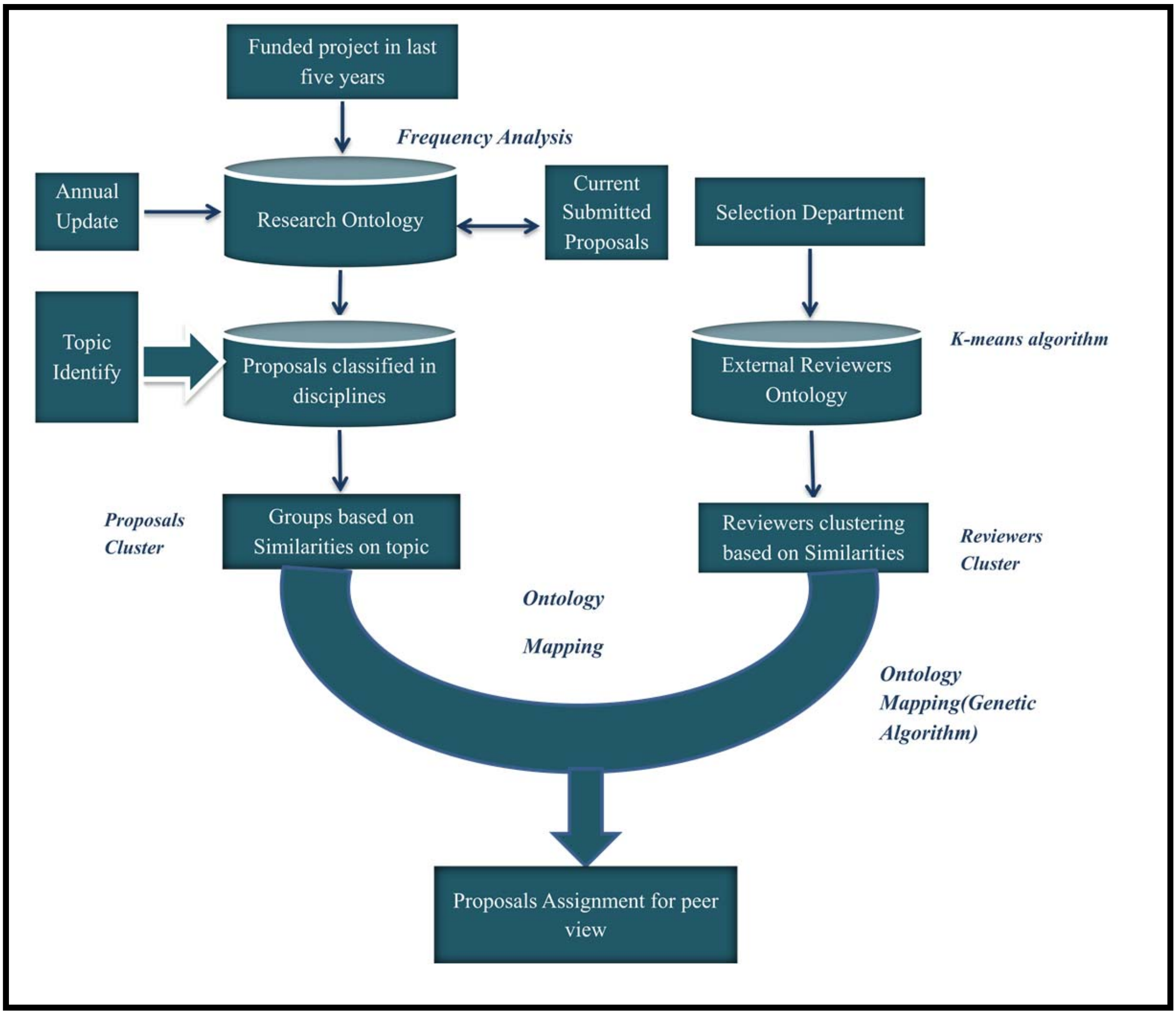

Funded project in last five years
Frequency Analysis
Annual Update
Research Ontology
Current Submitted Proposals
Selection Department
Topic Identify
Proposals classified in disciplines
External Reviewers Ontology
K-means algorithm
Proposals Cluster
Groups based on Similarities on topic
Reviewers clustering based on Similarities
Reviewers Cluster
Ontology Mapping
Ontology Mapping(Genetic Algorithm)
Proposals Assignment for peer view

Figure.2 Overall Architecture

ii) Constructing Research Ontology:

Research ontology is constructed according to scientific research areas and department of data selection process. It is then developed on the basis of several specific research areas. Next, it is further divided into some narrower discipline areas. Lastly, it leads to research topics in terms of the feature set of disciplines. In first part, there are some cross- discipline research areas (eg. data mining can be placed under Information Management in Management Sciences or under Artificial Intelligence in Information Sciences).Second part there are some synonyms used by different projects applicants, they have different names in different proposals but represent the same concepts. In Existing System, Jain Ma et al. had created ontology manually. Whereas, Here PROTEGE tool is used to create ontology
[1].

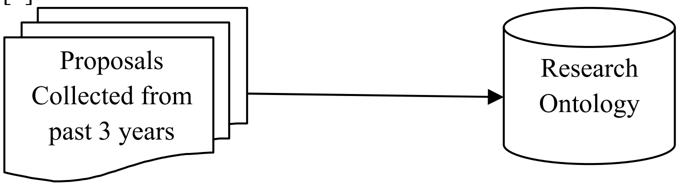


Figure.3 Research Ontology Creation

iii) Constructing Reviewers Ontology:

Reviewer's ontology is designed on the basis of all the domain areas of the reviewers either manually or using tools. In Existing System, Jain Ma et al. had created ontology manually. Whereas, Here PROTEGE tool is used to create ontology [1]

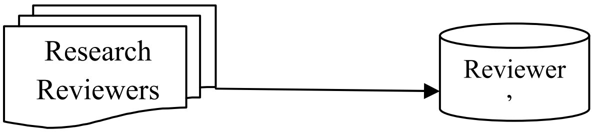


Figure.4 Reviewers Ontology Creation

iv) Update Ontologies:

Once the project funding is completed each year, revising should be done per annum via topics collected from proposals.. Reviewer's ontology is designed on the basis of all the domain areas of the reviewers.

B. Proposal classification

Research Project Proposals are classified by the discipline areas according to the keyword stored in ontology. Incoming new proposals are given as input to the research ontology. In the end, proposals are classified as accurate research area.

C. Clustering

After the research proposals are classified by the discipline areas, proposals in each discipline are clustered using the concept based text-mining technique. [15].

i) Concept based Text mining model:

Concept based mining model comprised of concept based analysis and similarity measure. In this model, the verb and the arguments are contemplated as terms. Each sentence in the document might have more than one verb argument formation. In such a cases term plays an important role. In this model, the labeled term (i.e.) repeatedly marked sentence is considered as concept. The purpose behind the concept-based analysis task is to accomplish an exact analysis of concepts on the Sentence level, Document level, and Corpus levels rather than a single-term analysis on the document only. Concept based text mining process is shown in Figure 5.

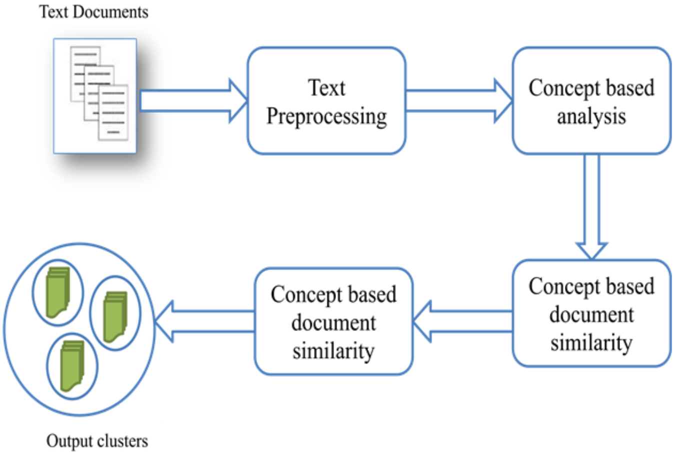


Figure.5 Concept-Based Mining Model

a. Text Document Collection

Research project proposal documents in each disciplined areas are collected for text preprocessing, after the proposals are classified by their disciplined areas

b. Text Document Preprocessing

The contents of proposals are usually unstructured. Research project Proposals comprise of Chinese characters which are problematic to segment. Research ontology is used to analyze and identify the keywords in the full text of the proposals. Finally, additional reduction in the vocabulary size can be achieved through the removal of all words that appeared only a few times (say less than five times) in all proposals.

c. Concept based Analysis

(1) Sentence-Based Concept Analysis

To survey every concept at the sentence level, the concept-based frequency assess, called the conceptual term frequency (ctf) is presented.

(2) Document-Based Concept Analysis

The term frequency is a local measure on the document level. To Analyze every concept at the document level, the concept based term frequency (tf) , the number of Occurrences of a concept (word or phrase) c in the document, is calculated.

(3)Based Concept Analysis

The df is a global measure on the corpus level. To survey concepts that can distinguish between documents, the concept-based document frequency (df) , the number of documents containing concept c, is calculated.

d. Concept Based Similarity

A concept-based similarity measure hang on on matching concept at sentence, document, and corpus instead of individual terms. First is to capture semantic structure of each sentence. Second is concept frequency that is used to measure contribution of concept in sentence as well as document level. Finally, the concepts measured from number of documents.

```
$d_{doci}$ is a new Document
L is an empty List
$s_{doci}$ is a new sentence in $d_{doci}$
Bulid concepts list $C_{doci}$ from $s_{doci}$
for each concept $c_i$ € $C_i$ do
    compute $ctf_i$ of $c_i$ in $d_{doci}$
    compute $tf_i$ of $c_i$ in $d_{doci}$
    compute $df_i$ of $c_i$ in $d_{doci}$
    $d_k$ is seen document, where k={0,1,…..,doc-1}
    $s_k$ is a sentence in $d_k$
    Build concepts list $C_k$ from $s_k$
    For each concept $c_j$ € $C_k$ do
        if($c_i$==$c_j$) then
            update $df_i$ of $c_i$
            compute ctf weight=avg($ctf_i$,$ctf_j$)
            add new concept matches to L
        end if
    end for
end for
output the matched concepts list L
```

Figure.6 Concept-Based Analysis Algorithm

e. Clustering Techniques

With the help of existing text clustering techniques we can get that which cluster is having highest priority.

f. Output Cluster

After applying the clustering techniques we can get the clustered document. That will help to find out main concepts from the text document

ii) Reviewers Clustering

With the help of reviewers ontology research reviewers are clustered based on their similarities in each discipline area or domain. A simple K-Means text mining clustering algorithm is used for this purpose.

K-means Algorithm

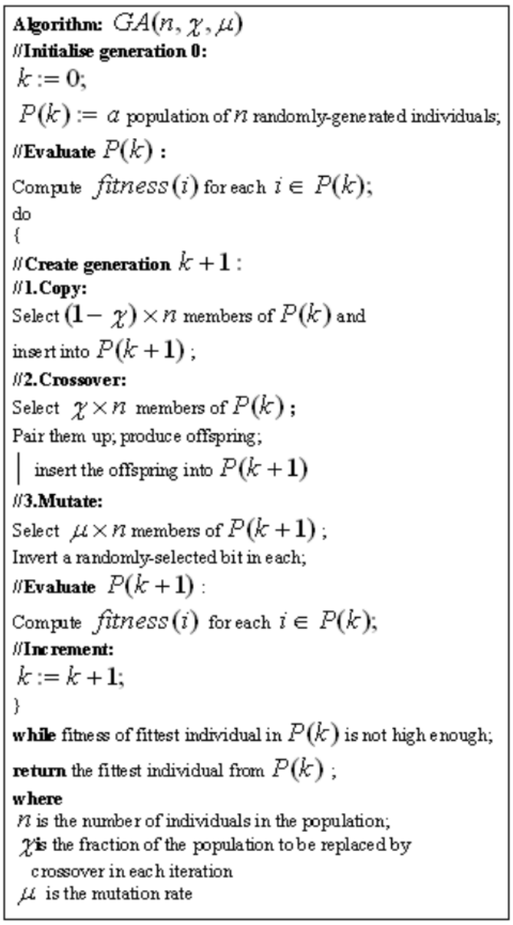
K-means is a best method to quickly sort the data into clusters, solitary the need is to define the number of clusters required. K denotes the number of clusters in which the data is divided. The algorithm works as:

1. Randomly select K-points as the initial cluster centroids.
2. Assign each object in the dataset to the closest cluster by compute their Euclidean distance of the object from the center.
3. When all objects have been assigned recalculate the position of the K centroids.
4. Repeat step 2 & 3 until the centroid no longer move. At this point clusters are separated into groups successfully.

D. Proposals Assign to Reviewers by Ontology Mapping

The Final step of this approach is to assign the Research Proposals group to the External Research Reviewers group systematically. Ontology Mapping can be done using both research and reviewers ontology. As a result ,the Proposals of the particular Discipline area is assign to the Reviewers having the same research area or domain and they can examine the proposals efficiently for the peer-review.

Usually Ontology Matching can be done by the tools like SAMBO, Falcon, DSSim, ASMOV, Anchor-Flood. Proposed technique SAMBO tool has been used for Ontology Matching.



```
Algorithm: $GA(n, \chi, \mu)$
//Initialise generation 0:
$k := 0$;
$P(k) := a$ population of $n$ randomly-generated individuals;
//Evaluate $P(k)$ :
Compute $fitness(i)$ for each $i \in P(k)$;
do
{
//Create generation $k+1$ :
//1.Copy:
Select $(1-\chi) \times n$ members of $P(k)$ and
insert into $P(k+1)$ ;
//2.Crossover:
Select $\chi \times n$ members of $P(k)$ ;
Pair them up; produce offspring;
    | insert the offspring into $P(k+1)$
//3.Mutate:
Select $\mu \times n$ members of $P(k+1)$ ;
Invert a randomly-selected bit in each;
//Evaluate $P(k+1)$ :
Compute $fitness(i)$ for each $i \in P(k)$;
//Increment:
$k := k+1$;
}
while fitness of fittest individual in $P(k)$ is not high enough;
return the fittest individual from $P(k)$ ;
where
$n$ is the number of individuals in the population;
$\chi$ is the fraction of the population to be replaced by
  crossover in each iteration
$\mu$ is the mutation rate
```

Figure.7 Genetic Algorithm

E. Performance Evaluation

The typical criterion for text clustering F measure is used to measure the quality of clustering research projects. If the value of F-measure is high then the quality of grouping is also high. Figure 8 represents the comparison between OCMM and OTMM techniques (i.e. F-measure value against number of proposals).F-measure value is calculated as harmonic mean precision and recall.

F-measure is calculated as follows,

$$F(c, t) = (2 * Recall(c, t) * Precision(c, t))/ (Recall(c, t) + Precision(c, t)) \qquad (1)$$

Where,

$$Precision(c, t) = n(c, t)/nc \qquad (2)$$

$$Recall(c, t) = n(c, t)/nt \qquad (3)$$

The Experimental result shows that the quality of grouping and assigning process has been drastically when compared to OTMM.Value of F-measure is higher for OCMM when compared to OTMM.Also,Similarity measure and Genetic algorithm improved the precision and recall values.

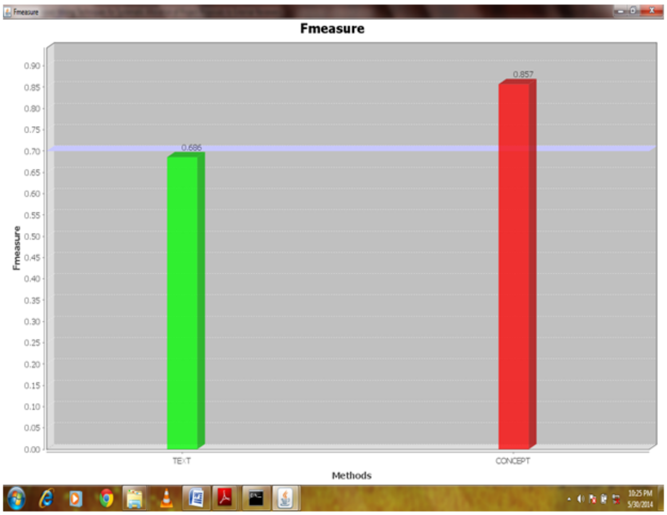


Figure.8 Comparison between OTMM Vs OCMM using F-measurement

V. Conclusion And Future Work

This paper has presented a framework on ontology based concept mining model for grouping research proposals as well as reviewers and also Genetic algorithm is used for proposal balancing and similarity measure which will be used by research funding Agencies for grouping and assigning the grouped proposal to reviewers group systematically. F-measure value illustrate that OCMM grouping was better than Existing technologies that are used for grouping. Experimental results showed that OCMM clusters proposals and reviewers in efficient manner and yield better result than previous method OTMM.

In Future, the following work can be extended to combination of Data Mining techniques along with Multilingual Ontology. Finally, using this effort improved end result has been achieved while balancing proposals.

References

[1] Jian Ma. Wet Xu, Hong Sun, Efraim Turban,ShouyangWang, 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.

[2] Preetkaur and Richasapra, ―Ontology based classification and clustering of research proposals and external research reviewers,‖ J.Inf. Sci., vol. 5, no. 1, May-June, 2013.

[3] Q. Tian, J. Ma, and O. Liu, ―A hybrid knowledge and model system for R&D project selection,‖ Expert Syst. Appl., vol. 23, no. 3, pp. 265–271, Oct. 2002

[4] 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

[5] A. 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.

[6] W. D. Cook, B. Golany, M. Kress, M. Penn, and T. Raviv, Optimal allocation of proposals to reviewers to facilitate effective ranking,‖ Manage.Sci., vol. 51, no. 4, pp. 655–661, Apr. 2005.

[7] 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.

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

[10] 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.

[11] K. Girotra, C. Terwiesch, and K. T. Ulrich, ―Valuing R&D projects in a portfolio: Evidence from the pharmaceutical industry,‖ Manage. Sci.,vol. 53, no. 9, pp. 1452–1466, Sep. 2007.

[12] D. A. Chiang, H. C. Keh, H. H. Huang, and D. Chyr, ―The Chinese text categorization system with association rule and category priority,‖ Expert Syst. Appl., vol. 35, no. 1/2, pp. 102– 110, Jul./Aug. 2008.

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

[14] 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. L. Gavril ova, Eds., 2007, pp. 660–673, Berlin, Germany: Springer-Verlag.

[15] A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model‖,V.M.Navaneethakumar, Dr.C.Chandrasekar

[16] N.Arunachalam, E.Sathya, S.Hismath Begum and M.Uma Makeswari, ―An Ontology based Framework for R&D Project Selection,J.Inf. Sci. and Technology, vol. 5, no. 1, February, 2013.

[17] M.Lavanya, N.Rajkumar-Ontology Based Clustering in research project selection & Assign Proposals to Experts by Ontology Matching,IJRET,vol.3,Special Issue:07|May-2014.

T.Sahaya Arthi Jeno received M.E degree in Computer Science and Engineering from Anna University, Regional Centre, Coimbatore, India. Her research interests include Data mining, Ontology, Artificial Intelligence and Expert System

D.Jim Solomon Raja received the M.Tech degree in Computer Science and Engineering from Karunya University, Coimbatore, India. His research interests include Data Mining, Web services, Robotics and Networking.