International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 9 5619 - 5623

OTMM based Proposal Classification and Clustering

Sunil T. Datir
Computer Science & Engg. Dept. of RKDF SOE, Indore
RKDF SOE, Indore, RGPV University
Indore, India
e-mail: sunil.datir@gmail.com

Arpit Solanki
Computer Science & Engg. Dept. of RKDF SOE, Indore
RKDF SOE, Indore, RGPV University
Indore, India
e-mail: er.arpitsolanki@gmail.com

Abstract— In the current environment, important task in any agencies (government, private) are to be selection proper search proposal. The proposal groups into the respective discipline when the large number of proposals are received .There is need to classify research proposals into proper categories automatically this will speed up the research proposal classification work. The technique used for proposals classification is OTTM Based on this respective discipline proposal assign to their expert for verification and review purpose.

Keywords*-component; **c**lassification, Clustering analysis, Document preprocessing decision support systems, ontology, classification, research project selection, text mining.*

*****

I. INTRODUCTION

The government, funding agencies, research institutes is the important activity to take a decision to the research paper selection. Functionality in NSFC first operation is call for the proposal, second proposal are submitted, third proposal grouping, fourth proposal assignment to expert for review. NSFC is the biggest government funding agency in china, with the first purpose to fund and handle the basic research. .The agency is arranged in seven scientific department, four bureaus, one general office, and three associated units. The scientific department is the process of reaching decision units responsible for funding recommendation and control the funded project. Department are categorized into respective to scientific research discipline including engineering and material science, earth science, life science, chemical science, mathematical and Physical science, etc. all the department additionally divided into forty section with attention on more respective research area. For Example the Department of Management Science is further divided into three divisions: Management Science and Engineering, Macro Management and Policy, and Business Administration. There was requirement for a feasible and effective approach to group the submitted research proposal with computer support. To solve this problem is using ontology text mining method approach to propose.

1. A research ontology storing the projects funded in latest five years is constructed according to keywords, and it is updated annually.

2. Using sorting algorithm, new research proposal classified order to discipline areas.

3. Next, with addresses to the ontology, the new proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm.

4. If the number of proposals in each cluster is still very large, they will be further decomposed into subgroups where the applicants' characteristics are taken into consideration (e.g., applicants' affiliations in each proposal group should be diverse). Here we may use of GA (Genetic Algorithm).

5. Finally, the Research project will be allocated to expert review.

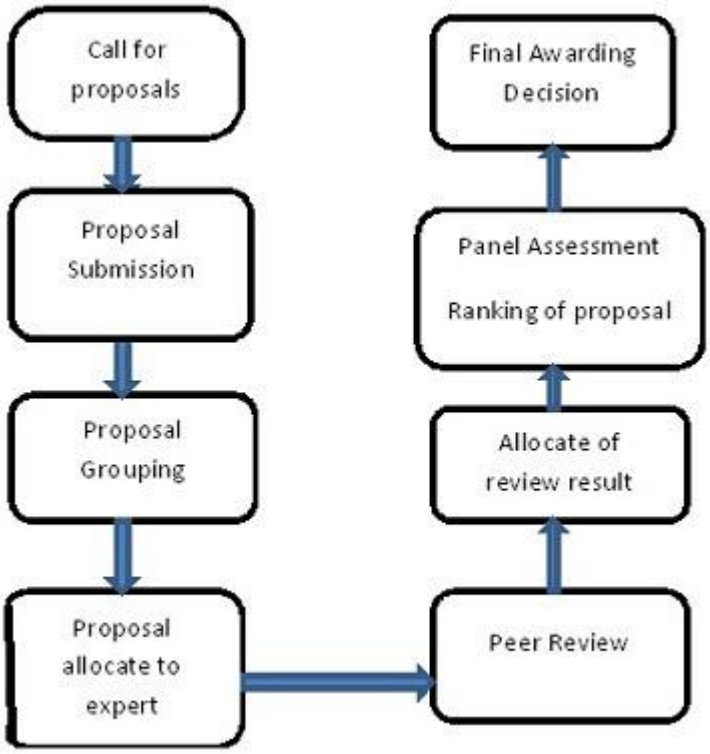

Fig 1.Research Project Selection Process in NSFC

II. LITERATURE SURVEY

There was the paper published by K. Chen and N. Gorla et al[2]. "Information System Project Selection Using Fuzzy Logic ".This paper establishes a model by incorporating

fuzzy logic as a decision tool, which smoothly aids decision makers dealing with uncertain or incomplete information without losing existing quantitative information. There was paper published by N.Arunuchalam, E. sathaya et al[3].”an ontology based text mining frame work for R &D project selection”. This paper presented a framework on ontology mining for grouping research proposal and assigning group proposal to reviewer systematically. A research ontology constructed to categorize to concept terms in different discipline areas and from relationship among them .it facilities TMM and make best use of situation technique to cluster research proposal based on their similarity and then to assign them to reviewer with help of knowledge based agent. There was the paper published by J. Butler, D. J. Morrice, and P. W. Mullarkey et al[5].“ A multiple attribute utility theory approach to ranking and selection”. This paper establishes a ranking and selection procedure for making comparisons of systems (e.g., project configurations) that have multiple performance measures. The procedure combines multiple attribute utility theory with Statistical ranking and selection to select the best configuration from a set of possible configurations using the indifference zone approach. There was paper published by S. Bolechdren, p. Camino et al[7]. Exploiting knowledge in textual documents is important issue in building systems for knowledge management and related task. In this paper we have presented framework. First step is natural language processing technique combined with machine learning algorithm allow to build or extend ontology in semi atomic manner this field known as ontology .in this method performed text to ONTO module to this OTTO from framework. Second, background knowledge from of ontology enhanced the perform classical text mining task text mining task such as text classification and text clustering.

There was paper published by Yiheng Chen and Bing Qin et al[6].” The Comparison of SOM and K-means for Text Clustering” this paper has presented compare two algorithm.

- Text clustering using K-Means algorithm

Partition-based clustering technique is known as K-means [6]. When k-means is applied for text clustering, total number of the documents will be place into particular position k clusters randomly, and after that the partition of clustering will be altered according to some concepts until the clustering results are firmly fixed. The basic concept of k-means for text clustering can describe as following step:

Input: N is documents clustered the cluster number k result: K clusters, and each document will be allocated to one cluster.

1) Select k documents randomly as the occurring at the beginning clustering document seeds;
2) repeat the next two steps, if the partition is firmly fixed, (then go to step 5);
3) To state as the mean vector of total number documents in each cluster, and allocate each document into most equal cluster.
4) The mean vector update of each cluster to state as the in it document vector.
5) Output the produced clusters and the partition.

- Text clustering using SOM algorithm.

When acceptable training has been done, the SOM network of output layer will be distinct into different section. And various neurons will have separate response to various input samples. as this operation is automatic. All of the input documents converted into clustered. Each of Text documents written by natural language has strong semantic feature and high-dimension. It is difficult to navigate one or more than operation documents in high-dimension scope. Taking into fact in to consideration that SOM can map total of these high-dimension documents on to 2- or 1- D space, and their association in the true space can also be kept. In extra, advantages of SOM are not very sensitive to some of the noisy documents and the clustering quality can also be reliable.

The concept of SOM for text clustering can be review as go after:

1. Initialization. Normalized and Allocate some random number for all the neurons in the output layer . The dimension number of neuron is same to the length number of all the documents;
2. Input the sample. Select randomly one document from the document collection and send it the SOM network;
3. Search the winner neuron. Determine the equality between the input document vector and the neuron vector, the neuron with the greatest equality will be the winner;
4. Modify its neighbors and the vector of the winner.

- Differentiate between SOM and K-means

To cluster documents using k-means, there is some control [6]:

1. After the occurring at beginning arrangement (document seeds and the value of k) has been obtained, the clustering results will also be calculated. But the clustering results will be separate if the initial settings are separate;
2. if the initial arrangement(the value of k and the document seeds) have been calculated, assume the clustering result of the execute again and again n+1 is similar as the iteration n, then the clustering result of iteration n+m will be equal as iteration n(m>1). Hence whether the partition has varied can be used as the stop criterion of clustering iteration. The neuron number of output layer in SOM network has close association with the class digits in the input document grouping. If the neuron digit is less than the class digit, it will be not enough to differentiate all the classes, the documents from some nearly-associated class may be combined into one class. If the neuron number is more than the class digits, the clustering results may be too good. The clustering quality and the clustering efficiency may also be adversely making a difference. K-means is very sensitive to

the occurred at begging arrangement such as k value and document seeds. Taking consideration the fact SOM can rich better text clustering excellence when the neurons in the can be fully utilized of the output layer.

Current system group proposals corresponding to keywords unfortunately proposals with equal research areas are located in wrong groups due to the following justification: first, the keywords are uncompleted information about the entire content of the proposals. Second, the keywords are on the condition that by applicants who may have applies for internal view sand misconceptions and keywords are only an imperfect representation of the research proposals. Third, the manual grouping is usually mange by program directors or division manager in funding agencies. They able to different understood about the research domain and may not have sufficient knowledge to allocate proposals into the correct groups. TMMs [23]-[24] have been designed.

To group proposals based on understating the English text, but they have limitations when dealing with other language texts, e.g., in Chinese. Also, when the number of proposals and reviewers increases (e.g., 110 000 proposals and 70 000 reviewers at the NSFC), it becomes a real challenge to find an effective and feasible method to group research proposals written in Chinese. This paper presents a novel approach for grouping Chinese research proposals for project selection. It uses text-mining, multilingual ontology, statistical analysis techniques and optimization to cluster research proposals depends on their equalities. The proposed approach has been successfully tested at the NSFC. The system results to show that the method can also be used to exceed the efficiency and effectiveness of the research project selection process.

III. PROPOSED SYSTEM

A .System design

Process of OTMM

The proposed OTMM is used together with statistical method and optimization models and consists of four phases, as shown in Fig.2. First, a research ontology containing the projects funded in latest five years is constructed according to keywords, and it is updated annually (phase 1). Then, new research proposals are classified according to discipline areas using a sorting algorithm (phase 2). Next, with reference to the ontology, the new proposals in each discipline are clustered using a self-organized 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 subgroups where the applicants' properties are taken into consideration (e.g., applicants' affiliations in each proposal group should be diverse).

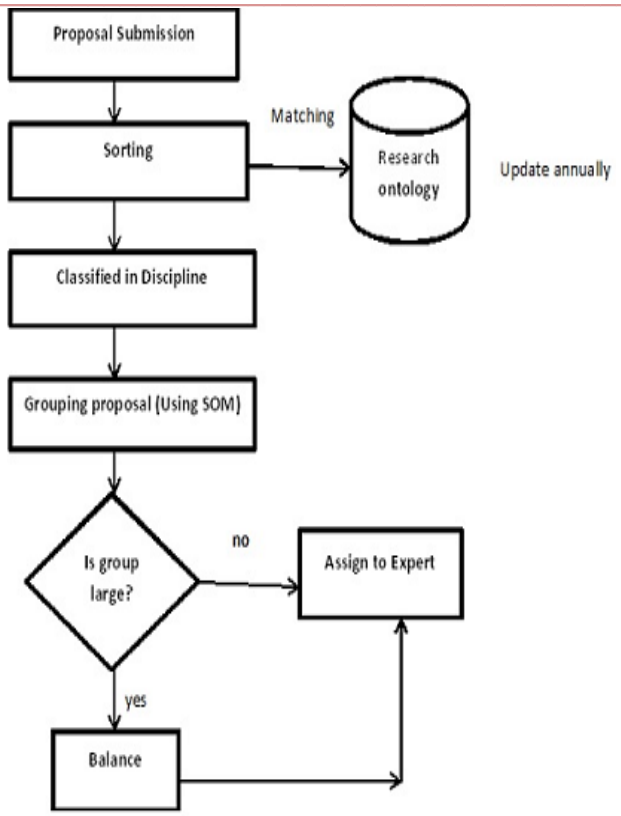

Figure 2. Process of Proposed OTMM

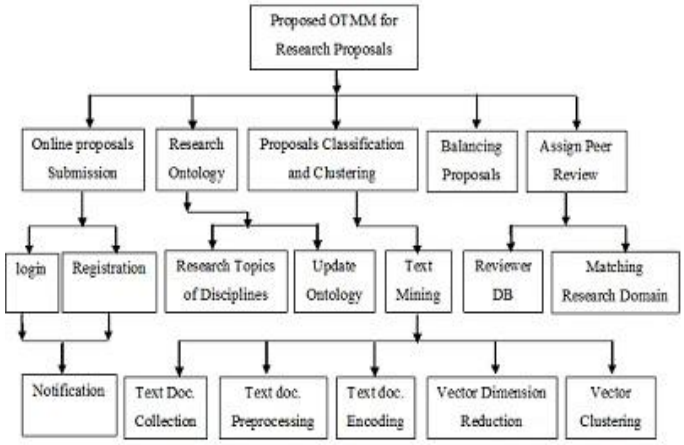

Figure 3 OTMM Project Breakdown Structure (Modules)

Online proposal submission

Here user is going to submit the proposal as client of system. For that the http connection wants to establish with server. This module may include the registration of each before proposal submission with intent of keeping track of each user.

Notification of email

Once the proposal is submitted to the system the client will get the email notification. The email notification will include the domain of the paper as per SOM algorithm, the name of guide to which paper is forwarded. This Email notification will work through SMTP protocol which will

automatically direct the email from user to guide and vice versa. With the help of this students can keep the track of their research paper.

SMTP is one of the parts of the application layer of the TCP/IP protocol. It works like "store and forward," SMTP send and receive your email on and over the networks. It works similar something called the Mail Transfer Agent (MTA) to send your communication to the specific computer and inbox of email. SMTP explain step-by-step and directs how your email transferred between more than one computer MTA. Using that "store and forward" feature described before, the message travel in number steps from your computer to its destination computer. Simple Mail Transfer Protocol is doing its work at each step.

Construction of research ontology

A research ontology contain the projects funded in latest five years is constructed according to keyword and it is updated annually. As domain ontology research ontology is a public concept set of the research ontology is a public concept set of the research project management domain. The research topic of different disciplines can be already expressed by research ontology.

Proposals classification and clustering

New research proposals are classified according to keyword stored in ontology with topic identified using keywords and features set matching with repository disciplines using with sorting algorithm. After research proposals are classified by discipline areas, next new proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm.

Balancing research proposals and regroup them by considering applicant's characteristics

If number of proposals in each cluster is still very large, they will be further decomposed into subgroups where applicant's characteristic universities are taken into consideration. One solution method that could be used is genetic algorithm [15].

Assign peer review

Here system trying to become fully automated we are going to maintain separate reviewer's repository it maintain external reviewers based on their research area and to assign concerned proposals to reviewers

B. System Methodology:

1. Sorting algorithm:

In sorting algorithm given proposals get categorized by the domain to which the below, research ontology used it
Suppose,
k- Discipline/domain.
A_k- Area (k=1,2,…….K).
P_i-proposals (i=1,2……I).
S_k- set of proposals which belong to area of K.
Sorting algorithm:

```
Start
    for k=1 to K
        for i=1 to I
          if (P_i belongs S_k )
          then P_i -> S_k
        end of for
    end of for
stop.
```

2. SOM Algorithm:

New Proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm [16]-[17].

1. Select output layer network topology.
   1.1 Initialize current neighborhood distance, D(0), to a positive value.
2. Initialize weights from inputs to outputs to small random values
3. Let t = 1
4. While computational bounds are not exceeded do
   4.1 Select an input sample
   4.2 Compute the square of the Euclidean distance of from weight vectors (w_j) associated with each output node.
   $\sum^{n}_{k=1}(i_{l,k} - w_{j,k}(t))^2$
   4.3 Select output node *j** that has weight vector with minimum value from step 2.
   4.4 Update weights to all nodes within a topological distance given by D(t) from *j**, using the weight update rule:
   $w_j(t+1) = w_j(t) + \eta(t)(i_l - w_j(t))$
   4.5 Increment t
5. End while.

Learning rate generally decreases with time:

$$0 < \eta(t) \leq \eta(t-1) \leq 1$$

3. *Genetic Algorithm*

Balancing and regroup the proposal using by Genetic algorithm.

Input: Fitness function f(), maximum number of iteration max_tier

Output: best found solution,

```
begin
    Generate at random initial population of solution;
    i:=0;
    while i<= max_tier and stop_cond.= false do
```

begin
– evaluate each solution with f();
– apply crossover on selected solution;
– mutate some of the new obtained solutions
– add new solution to population;
– remove less adopted solutions according to f() from
population;
– i:= i+1;
end;

- return best found solution;
end;

IV. CONCLUSION

Sometimes a large number of research proposals are received, agencies like government, private institution, group this proposal into their respective discipline and then assign to respective reviewer. The propose system introduced OTMM based research proposal selection. In this sorting classified the proposals after that proposal group together using clustering algorithm that is SOM and finally balancing the proposals using GA. There for in proposed system achieved high accuracy and efficiency.

REFERENCE

[1] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban,Shouyang Wang, and Ou Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection," IEEE Transaction On System,man, and cybernetics May,2012.

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

[3] N.Arunachalam, E.Sathya , S.Hismath Begum and M.Uma Makeswari, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection," International Journal of Computer Science & Information Technology (IJCSIT) Vol 5, No 1, February 2013.

[4] MS. K.Mugunthadevi, MRS. S.C. Punitha, Dr..M. Punithavalli, "Survey on Feature Selection in Document Clustering," International Journal on Computer Science and Engineering (IJCSE), Feb. 2000.

[5] J. Butler, D. J. Morrice, and P. W. Mullarkey, "A multiple attribute utility theory approach to ranking and selection," Manage. Sci., vol. 47, 6,Jun.2001.

[6] Yiheng Chen and Bing Qin et al." The Comparison of SOM and K-means for Text Clustering "School of computer Science and Technology vol 3,no 2;May 2010.

[7] S.Bloehdom and p.cimiano et al. "An Ontology –based framework for Text Mining" Institute AIFB,July 28, 2004.

[8] Tu, S.W., Tennakoon, L., Das, A.K.: Using an Integrated Ontology and Information Model for Querying and Reasoning about Phenotypes: The Case of Autism. AMIA Annual Symposium, pp. 727–731, Washington, DC (2008)

[9] Hus, V., Pickles, A., Cook, E.H., Risi, S., Lord, C.: Using the Autism Diagnostic Interview-Revised to Increase Phenotypic Homogeneity in Genetic Studies of Autism. Biol Psychiatry. 61(4), 438–448 (2007)

[10] McGuinness, D.L., van Harmelen, F.: OWL Web Ontology Language Overview. W3C Recommendation, (2004)

[11] Horrocks, I., Patel-Schneider, P.F., Boley, H., Tabet, S., Grosof, B., Dean, M.: SWRL: A Semantic Web Rule Language Combining OWL and RuleML. (2004)

[12] Open Directory Project, http://www.dmoz.org/ Google support on snippets, http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35624

[13] Cooper, W.S.: Fact Retrieval and Deductive Question-Answering Information Retrieval Systems. J. ACM. 11(2), 117–137 (1964)

[14] Miliaraki, S., Androutsopoulos, I.: Learning to Identify Single-snippet Answers to Definition Questions. 20th International Conference on Computational Linguistics, Geneva, Switzerland (2004)

[15] D.E Goldberge ,Genetic Algorithms in search , Optimization and machine Learning. Redwood city :Addision – Wesely,1989

[16] F.M Ham and I,kostanic , Principle of Neurocomputing for Science and Engineering .New york :McGraw-Hill,2001.

[17] J. vesanto and E.Alhoniemi ,"Clustering of the self - organizing map" IEEE Trans. Neural network ,vol. 11,no.3,pp.586-600,May 2000.

[18] L.Razamerita , "An ontology –based framework for modeling user behavior –A case study in knowledge management "IEEE Trans. Syst. Man Cybern, A, Humans ,vol 41,no. 4,pp. 772-783,jul,2011.

[19] Q. Liang, X. Wu. E. K. park ,T.M. Khoshgoftaar ,and C.H Cho,"Ontology –based business process customization for composite web services "IEEE Trans Syst.,Man, cybern. A, Syst. Humans , vol. 41, no.4 pp.717-729,jul,2011

[20] 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.

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

[22] L. L. Machacha and P. Bhattacharya, "A fuzzy-logic-based approach to project selection," IEEE Trans. Eng. Manag., vol. 47, no. 1, pp. 65–73, Feb. 2000.

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

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

[25] S.Bloehdom and p.cimiano et al. "An Ontology –based framework for Text Mining" Institute AIFB,July 28, 2004.