# A Novel Ontology based R&D Project Proposal Classification using Text Mining Approach

S. N. Gunjal
Department Computer Engineering,
SRES College of Engg. ( S.P Pune University),Kopargaon,
Maharashtra, India

B. J. Dange
Department Computer Engineering,
SRES College of Engg.( S.P Pune University),Kopargaon,
Maharashtra, India

A.V Brahamane
Department Computer Engineering,
SRES College of Engg.( S.P Pune University),Kopargaon,
Maharashtra, India

## ABSTRACT

Research and Development (R&D) project proposals selection is one of the decision-making task commonly found in government funding agencies, universities, research institutes, and technology intensive companies. With the rapid development of research work in projects, research project selection & classification into different domain is a necessary task for the research funding agencies. It is common to group the large number of research proposals, received by the research funding agencies based on their similarities into research discipline areas. Text Mining has emerged as a definitive technique for extracting the unknown information from large text document for the proposal classification. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. Thus, ontology can automate information processing and can facilitate text mining in a specific domain (such as research project selection). This paper presents approach towards ontology-based text-mining to cluster research proposals based on their similarities in research areas. The method also includes an optimization model for balancing proposals by geographical regions. The grouped proposals are then assign to the appropriate research experts for peer-review through system itself. The proposed method is milestone over the manual approach for classifying proposals.

## Keywords

Ontology, text mining, clustering, knowledge repository.

## 1. INTRODUCTION

Research project selection is important task for many organizations such as government funding agencies. For any research funding agencies, such as government or private agencies, the selection of research project proposals is an important and challenging task, when large numbers of research proposals are collected by the organization. The research project proposals selection process starts with the call for proposals (CFPs), then submission of the research proposals by many institutes and organizations. Now, group the Proposals based on their similarity and assigned them to the experts for peer-review.

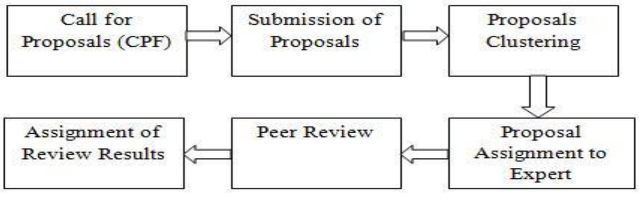

**Fig.1 Research Project Selection & Classification Process**

The decision makers classify them into groups according to their decision-making tasks in the R&D project selection process [1]. In manual, these decision-making groups coordinate with each other & select the best project proposals. Division managers or program directors then group the proposals and assign them to external reviewers for evaluation and commentary. However, they may not have adequate knowledge in all research disciplines, and contents of many proposals were not fully understood. When the proposals were grouped and assigned the grouped proposal to external reviewers. Therefore, there was an effective approach to group the submitted research proposals and assign the proposals to external reviewers experts in the specific domain with computer supports. In current methods, keywords are not representing the complete information about the content of the proposals and they are just the partial representation of the proposals. Hence, it's not sufficient to group the proposals on the basis of keywords.

Therefore, an efficient and effective method is required to group the proposals efficiently based on their discipline areas by analyzing full text information of the proposals.

## 2. LITERATURE REVIEW

Selection of research projects is an important research topic in research and development (R&D) project management. Previously research deals with specific topics, and several formal methods and models are available for this purpose. For example, Yong-Hong Sun, Jian Ma, Zhi-Ping Fan, and Jun Wang [1] [2008] proposed a group decision support approach to evaluate experts for R&D project selection. Jian Ma, Wei Xu [1] used Text-To-Onto ontology environment using supervised learning [6]. N.Arunachalam, E.Sathya [2] have been established an group decision support approach to evaluating journals. Kapil Sharma and Richa Dhiman [6] used for R&D proposal screening system based on text mining approach [8]. L. M. Meade and A. Presley [8] proposed clustering-based category-hierarchy integration (CHI) technique, which is an extension of the clustering-based category integration (CCI) technique. Yang and Lee [1] used text mining approach for automatic construction of hyper texts. Juha Vesanto and Esa Alhoniemi [7] developed an Information Retrieval application using ontologies. L. M. Meade and A. Presley [8] have been established for extract relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents using e-learning perspective. Razmerita proposed an ontology-based framework for modeling user behavior. L. L. Machacha and P. Bhattacharya [5] proposed a fuzzy-logic-based model as a decision tool for project selection. Henriksen and Traynor presented a scoring tool for project evaluation and selection. Ghasemzadeh and

Archer offered a decision support approach to project portfolio selection. Cook presented a method of optimal allocation of proposals to reviewers in order to facilitate the selection process. Juha Vesanto and Esa Alhoniemi [7] proposed a rotation program method for project assignment. Choi and Park used text-mining approach for R&D proposal screening. Girotra offered an empirical study to value projects in a portfolio. Sun developed a decision support system to evaluate reviewers for research project selection. Finally, Sun proposed a hybrid knowledge-based and modeling approach to assign reviewers to proposals for research project selection. For example, Hettich and Pazzani 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. EXISTING SYSTEM

The existing system is an Ontology-Based Text-Mining Method to cluster research proposals based on their similarities in research areas [2]-[5]. An 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 conceptualization of the domain’s “real-world” setting. Implicit knowledge for humans is made explicit for computers by ontology. 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 framework has been built for clustering the research proposals according to their discipline areas.

The existing system of OTMM for proposals classification is desktop based. The operations at server side have to be performed manually. In the NSFC, the number of research proposals received has more than doubled in the past four years, with over 110000 proposals submitted in one deadline in March 2010. Four to five reviewers are assigned to review each propos al so as to 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 to assign the proposal groups to relevant reviewers. This task is being performed manually in the existing system.

## 4. PROPOSED SYSTEM

The basic idea for this proposed architecture is to make easier the Research Proposals Selection Process. The Ontology-based Text Mining approach is used for the selection of research project proposals for either government or private research funding agencies. The proposed system includes:

### 4.1 Client Server Model

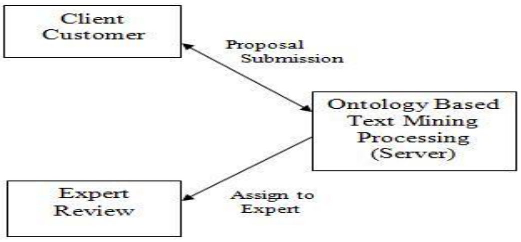

**Fig.2 Client Server Model**

Here, the client submits the proposal to the server. All the processing activities will take place at server only and then proposal will be get assigned to particular expert. Expert will get notification about assignment. The client will remotely submit the proposal to server; the received proposals are validated by system itself. Now, the valid proposals will get be stored in some directory at server. For the validation, some standard template of research proposal format is used.

There is a two way communication between client and server. It is the task of server administrator to input the proposals to system. Client will has to register first to the system and can log in for further proposal submission. After processing, proposals are assigned for expert review. This decision making task will be done by server itself. Server may notify to particular expert about their assignment for proposal in the form of mail or message. This is the atomization done in assignment of proposals for expert review.

### 4.2 Approaches towards Ontology based Proposals Classification and Clustering

#### *4.2.1 Proposals Classification and Clustering using Keyword Identification and Segmentation*

Ontology has become prominent in the research work from recent years, in the field of computer science. Ontology is a knowledge Repository which defines the terms and concepts and also represents the relationship between the various concepts. It is a tree like structure which defines the concepts.

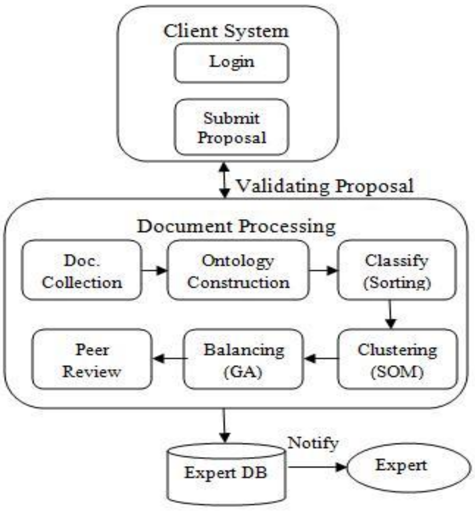

**Fig.3 Ontology Based Text Mining Process**

### *Step1) Construction of Research Ontology and Annual Update*

Construction of research ontology is one of the vital tasks of system development. It mainly depends on last five years funded projects. Ontology is basically tree like structure in which the each node denotes the particular discipline area which can be extended further. So that tree becomes narrower as it grows. Keywords for supported research area are collected each year and their frequencies are getting calculated. Discipline code for each discipline is generated. Finally, the feature set is formed using discipline code, year, keyword and frequency. Annual updating of ontology is one of the administration task that has to be carried out over time because as technology changes, new keywords has to be included.

### *Step 2) Classifying new Research Proposal*

Classification of new research proposal is totally based on sorting algorithm using keyword matching. New research proposals can also be classifying according to the keyword stored in ontology.

### *Step 3) Clustering of Research Proposals based on Similarity using Text Mining*

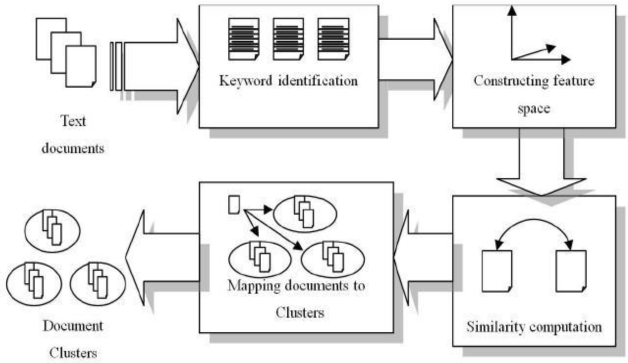

**Fig.4 Framework for Text Clustering**

After classification, next task is to clustering of research proposals in specific domain based on their similarity. It mainly consists of five stages.

### *Step 3.1) Text Documents Collection*

Research proposals are collected from client and stored in some directory at server side where main processing in carried out.

### *Step 3.2) Preprocessing of Text Documents*

It mainly involves the segmentation and stop words removal process.

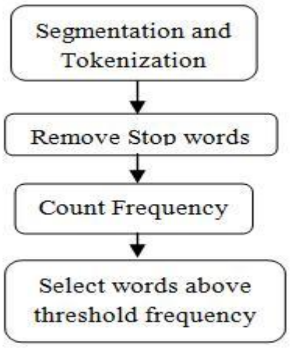
Segmentation and Tokenization
Remove Stop words
Count Frequency
Select words above threshold frequency

**Fig.5 Text Document Preprocessing Further**

Reduction in vocabulary size can be done through discarding the keywords below threshold frequency.

### *Step 3.3) Text documents Encoding*

After text documents are segmented, they are converted into a feature vector. TF-IDF encoding method 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 3.4) Vector Dimension Reduction*

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.

### *Step 3.5) Vector Clustering*

This makes use of SOM algorithm which is basically unsupervised learning neural network.[5]-[7]

**SOM Algorithm**

```
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 (wj) 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: $0 < \eta(t) \le \eta(t-1) \le 1$
```

### *Step 4) Balancing Research Proposals and Regroup*

After clustering proposals, if the size of cluster is greater than then that cluster that means dividing cluster into sub cluster by considering applicant's working area. This optimization can be achieved by implementing Genetic Algorithm (GA).

GA 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;
```

## 4.3 Proposals Classification and Clustering using Open Source NLP Tools

This is another technique for classifying and clustering of proposals. There are plenty of NLP tools such as Stanford NLP, Apache OpenNLP, NLP Engine are more popular.

### *Step 1: Research Ontology building*

### *Step1.1) Creating the research topics*

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

### *Step1.2) Constructing the research ontology*

First, the research ontology is categorized according to scientific research areas. It is then developed on the basis of several specific research areas. [2] Next, it is further divided into some narrower discipline areas. Finally, it leads to research topics in terms of the feature set of disciplines. First, there are some cross- discipline research areas (e.g. ”data mining“ can be placed under ”Information Management” in “Management Sciences” or under “Artificial Intelligence” in “Information Sciences”).second there are some synonyms used by different projects applicants, which have different names in different proposals but represent the same concepts.

### *Step 1.3) Automatic topic Identification Approach*

### *A. Split the Text into Sentences.*

The first step in topic identification algorithm is splitting the sentences on the given text. [3] A sentence is a smallest text part which is capable to have a topic. Hence, it splits the document into corresponding sentences. One of the NLP tool is Text Splitter which splits a text into sentences. The Text splitter tool used to split the text files into chunks. By performing this tool have a set of sentences.

### *B. Parse the Sentences*

In this step, the algorithm intends to pars the sentences and determines the candidate terms first to avoid any useless calculation. Believe that syntactic parts like Noun Phrase (NP) and Verb Phrase (VP) are playing most important roles to present the meaning of the sentence and therefore we should consider them instead of grammatical roles like noun and verb to identify the candidate topic for each sentence.

### *C. Select the Candidate Parts*

At this step select noun phrase (NP) and the head of a Verb Phrase (VP) instead of just pairs of nouns and noun-verb. [3] Assume that the most important parts from a sentence are the NP’s that function as subject or complement and the head of the VP. The combination of three topics is considered as a candidate topic.

### *D. Calculate the weight for each candidate topic*

At this moment calculate the IDF and SNV for only required syntactic parts. By this way, there is no need to calculate these amounts for irrelevant parts and in fact, avoid any calculation overhead. The calculation formula is SNV (NP, head (VP)) = IDF (NP). IDF (head (VP)) / D (NP, head (VP))

### *E. Select the final topic*

When we determine the candidate topic and its associated weight for each sentence, select the most weighted one and consider it as the main topic for the whole document. In case there are more than one candidate topics with greatest weight, we consider all of them as the main topic.

### *Step 1.4) Updating Research Ontology*

Once the project funding is completed each year, the research ontology is updated according to agency’s policy.

### *Step 2) Proposal Classification*

Proposals are classified by the discipline areas according to the keyword stored in ontology and the topic identified using Topic Identification Algorithm.

### *Step 3) Proposals Clustering*

After the research proposals are classified by the discipline areas, the proposals in each discipline are clustered using the concept based text-mining technique. [2]

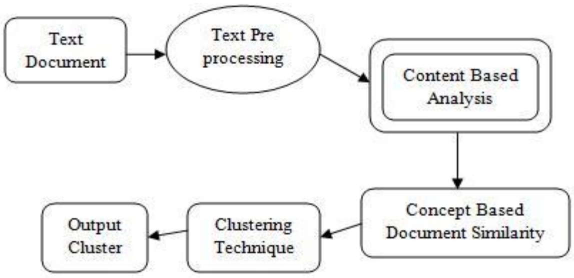

**Fig 6: Concept Based Mining Model**

### *Step 3.1) Sentence-based Concept Analysis*

To examine every concept at the sentence level, a novel concept-based frequency assess, called the conceptual term frequency (ctf) is computed. The ctf is the number of concept c happened in verb argument structures of sentence S. The concept c, which normally emerges in diverse verb argument structures of the similar sentence S, has the prime job of contributing to the significance of S.

### *Step 3.2) Document based Text Clustering using the Concept Based Mining Model*

Document based clustering is done through a concept based mining model and identify the similarity measure of the document by analyzing each concept at the document level, based on the type of markup language formats and the number of occurrences of document are also being identified and discriminated. [3] The analysis of document text clustering is done by the proposed document based text clustering algorithm.

### *Step 4) Information Retrieval*

A knowledge-based agent used for information retrieval. It includes a knowledge base and an inference system. A knowledge based agent is used in this system to retrieve the grouped proposal and assign to external reviewer systematically.

*Step 5) Assign to External Reviewer*

The information retrieved by knowledge based agent is assigned to external reviewers where reviewer's research area, experience will be collected before. According to their research area and experience the reviewer will be clustered, [8] but there may be some ambiguous because the reviewer may be specialized in more than one domain.

## 5. RESULT AND PERFORMANCE ANALYSIS

Result & performance analysis of Proposals Classification and Clustering using Keyword Identification and Segmentation approach .Let the set of proposals relevant to the research domain denoted by {Relevant} and set of proposals accepted from users denoted by {Retrieved}. Set of proposals that are both relevant and retrieved, are denoted as: {Relevant} ∩ {Retrieved}.

### 5.1 Effect of Increase in Number of Proposals

To validate the proposed approach, several experiments are conducted using the previous granted research projects. First, two experiments (E1 and E2) are constructed to evaluate the quality of clustering research projects. Second, one experiment (E3) is used to validate the effectiveness and efficiency of balancing research projects. In E1, research projects in the discipline called information science are randomly selected. In E2, research projects in the discipline named chemical science are randomly used. In E3, research projects with similar topics are randomly selected. In addition, the typical criterion for text clustering F measurement is used to measure the quality of clustering research projects. For generated cluster c and predefined research topic t, the corresponding Recall and Precision can be calculated as follows:

Precision(c, t) =n(c, t)/nc

Recall(c, t) =n(c, t)/nt

where n(c, t) is the project number of the intersection between cluster c and topic t.

nc is the number of projects in cluster c,

nt is the number of projects in topic t.

F measurement between cluster c and topic t can be calculated as follows:

F(c, t) = (2 ∗ Recall(c, t) ∗ Precision(c, t)) / (Recall(c, t) + Precision(c, t)) .

In order to compare the clustering quality of the OTMM and the general TMM, the other settings of both methods are kept the same as possible. The relations between F measurement and the number of research projects n in these two disciplines can be found in below figures. It can be seen that the performance of our proposed method is better than that of the standard TMM. Therefore, the OTMM can be an alternative for clustering research proposals.

**Table 2 Calculations of E1**

| Number of Proposals | Precision | Recall | F_measure |
|---|---|---|---|
| 5 | 0.20 | 1 | 0.33 |
| 10 | 0.10 | 1 | 0.19 |
| 15 | 0.15 | 1 | 0.23 |
| 20 | 0.25 | 1 | 0.40 |

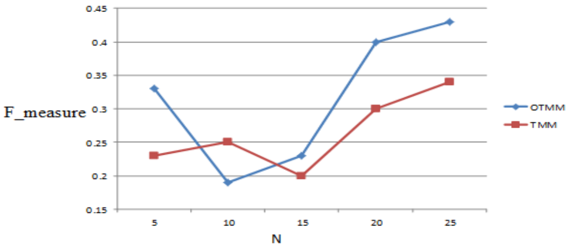

**Fig. 7 Graph of E1**

**Table 3. Calculations of E2**

| Number of Proposals | F_measure |
|---|---|
| 5 | 0.20 |
| 10 | 0.27 |
| 15 | 0.29 |
| 20 | 0.23 |

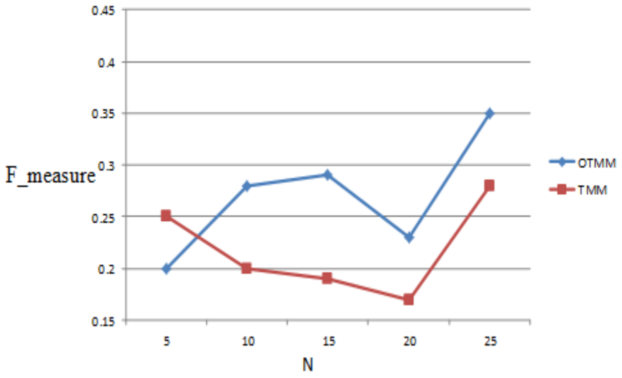

**Fig.8 Graph of E2**

Similarly, tested our system for large number of proposals. If subjected our system to number of proposals like 100, 200,…,1000, the results may get as shown in below graph.

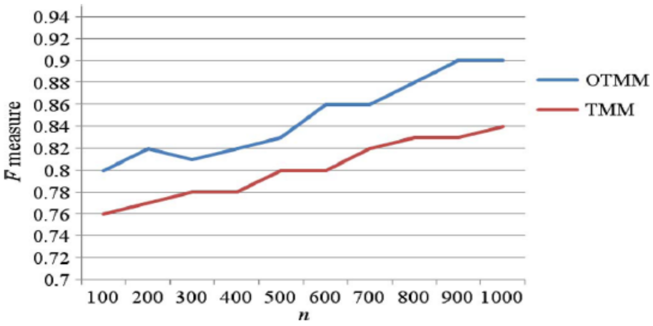

**Fig. 9 Relative Graph**

The experimental results showed that the proposed method improved the similarity in proposal groups, as well as balanced the applicants' characteristics. Therefore, the proposed method promotes the efficiency in the proposal grouping process. By manual grouping, users need to spend at least one week, while the grouping can be finished within hours using the proposed methods. Given that the method can expedite the process considerably, it can be used as the first step in a machine–human collaboration where the automatic grouping results are provided to a human that checks and then approves or modifies them.

## 6. CONCLUSION AND FUTURE WORK

In this paper, Ontology based classification and clustering approach is proposed, which is used by research funding agencies for grouping the Research Proposals and the research Reviewers. It also facilitates text-mining and optimization techniques to cluster research proposals based on their similarities and then to balance them. This Proposed approach can provide us a way to easily classify and group the research proposals and the reviewers.

Future work can be done for enhancements in the proposed system such as provide the mail indication to reviewer to which research proposals for peer review are assigned. System may also store the reviewer's history and assignment on basis of that. The situation might occur where a reviewer may have expertise in more than one domain. So, priority has to be deciding for assignment of proposals to reviewers. It may be possible that system automate the work of reviewer.

## 7. REFERENCES

[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 Systems, Man, and Cybernetics*, vol. 42, No. 3, May 2012.

[2] N.Arunachalam, E.Sathya, S.Hismath Begum and M.Uma Makeswari ,an ontology based text mining framework,,*ijcsit* vol 5, no 1, february 2013.

[3] Preet Kaur, Richa Sapra, ontology based classification and clustering of research proposals and external research reviewers, *International Journal of Computers & Technology Volume 5*, No. 1, May -June, 2013.

[4] Habiba Drias, Ilyes Khennak, A Hybride Genetic Algorithm for Large Scale Information Retrival,978-1-4244-4738-1/09/ 2009 *IEEE.*

[5] 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.

[6] Kapil Sharma and Richa Dhiman, implementation and evaluation of k means, kohonen-som, and hac data mining algorithms based on clustering, *International Journal of Computer Science Engineering & Information Technology Research (IJCSEITR)* ISSN 2249-6831 Vol. 3, Issue 1, Mar 2013, 165-174.

[7] Juha Vesanto and Esa Alhoniemi, Clustering of the Self-Organizing Map,*IEEE transactions on neural networks*, vol. 11, no. 3, may 2000.

[8] 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.

## 8. AUTHOR'S PROFILE

**S.N. Gunjal** received the bachelor's degrees in computer engineering from the SRES COE, Kopargaon from University of Pune and Master degrees in computer science & engineering from the RITS,Bhopal from Rajiv Gandhi Proudyogiki Vishwavidyalaya,Bhopa and working as Asst. professor of computer engineering department at SRES COE Kopargaon, University of Pune . His research interests include Pattern recognition, data mining, knowledge-based systems, and web information exploration. He is LMISTE.

**B.J. Dange** received the bachelor's degrees in computer engineering from the SRES COE,Kopargaon from University of Pune,Pune and working as Asst. professor of computer engineering department at SRES COE Kopargaon,University of Pune ,Pune His research interests include data mining, knowledge-based systems, and web information exploration,Computer Network.He is LMCSI and ISTE

**A.V. Brahmane** received the bachelor's degrees in computer engineering from the PREC COE, Loni from University of Pune and working as Asst. professor of computer engineering department at SRES COE Kopargaon,University of Pune . His research interests iclude distributed system, data mining, knowledge-based systems, and web information exploration,Computer Network. He is LMCSI and *ISTE*