ISSN:0975-9646

Review of Classification of project proposal through Ontology Text Mining.

Sunil Datir[1], Megha singh [2],
RKDF School of Engineering Indore, India, Computer Science Department affiliated to RGPV University.

***Abstract*-Research Paper Selection is important decision preparing task for the Government funding Agency, research Institutes. Ontology is Knowledge Repository in which concepts and terms defined as well as relationship between these concepts. In this paper Ontology is old research papers repository of keywords and frequencies of that keywords of the research papers of funding agencies. Ontology makes the tasks of searching similar patterns of text that is to be more effective, efficient and interactive. The current system of grouping of papers for research paper selection based on similarities of Keywords and Frequencies of research papers of ontology.**

Text mining is the extraction of useful, often previously unknown information from large document. The Research Papers in each domain are clustered using Text mining Technique. Grouped Research papers are allocated to appropriate reviewer or domain experts for peer review systematically. The Reviewer results are collected and papers are get graded based on experts review results.

***Keywords*— Document preprocessing, Clustering analysis, decision support systems, ontology, classification, research project selection, text mining.**

I. INTRODUCTION

Selection of research projects is an important and recurring activity in many organizations such as government research funding agencies. It is a challenging multi process task that begins with a call for proposals (CFP) by a funding agency. The CFP is distributed to relevant communities such as universities or research institutions. The research proposals are submitted to the funding agency and then are assigned to experts for peer review. The review results are collected and the proposals are then ranked based on the aggregation of the experts' review results.

In the National Natural Science Foundation of China (NSFC), after proposals are submitted, the next important task is to group proposals and assign them to reviewers. The proposals in each group should have similar research characteristics. For instance, if the proposals in a group fall into the same primary research discipline (e.g., supply chain management) and the number of proposals is small, manual grouping based on keywords listed in proposals can be used. However, if the number of proposals is large, it is very difficult to group proposals manually. Although there are several text-mining approaches that can be used to cluster and classify documents. TMMs (Text Mining Method) which deal with English are not effective in processing Chinese text. To solve the aforementioned problems, an ontology-based TMM (OTMM) is proposed.

Ontology Learning There is quite a long tradition in learning concept hierarchies by clustering approaches such as the ones presented in as well as by matching lexico-syntactic patterns as described in In this section we focus on the discussion of frameworks and systems designed for supporting the ontology engineering process. In the ASIUM system nouns appearing in similar contexts are iteratively clustered in a bottom-up fashion. In particular, at each iteration, the system clusters the two most similar extents of some argument position of two verbs and asks the user for validation. Bisson et al. [3] present an interesting framework and a corresponding workbench - Mo'K - allowing users to design conceptual clustering methods to assist them in an ontology building task. The framework is general enough to integrate different clustering methods. Velardi et al. [13] present the OntoLearn system which discovers i) the domain concepts relevant for a certain domain, i.e. the relevant terminology, ii) named entities, iii) 'vertical' (is-a or taxonomic) relations as well as iv) certain relations between concepts based on specific syntactic relations. In their approach a 'vertical' relation is established between a term t1 and a term t2, i.e. is-a(t1,t2), if the head of t2 matches the head of t1 and additionally the former is additionally modified in t1. Thus, a 'vertical' relation is for example established between the term 'international credit card' and the term 'credit card', i.e. is-a(international credit card, credit card).

1.1 What is Ontology?

Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. It consists of a set of concepts, axioms, and relationships 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).

1.2 Problem Definition

Research and development (R&D) project selection is an decision making task commonly found in government funding agencies, universities, research institutes, and technology intensive companies. So, here we are introducing the method for grouping proposals for research project selection is proposed using an ontology based text mining approach to cluster research proposals based on their similarities in research area. The method also includes an optimization model that considers applicants' characteristics for balancing proposals.

1.3 Relevant Theory

Below figure shows the processes of research project selection at the National Natural Science Foundation of China (NSFC) i.e. CFP, proposal submission, proposal grouping, proposal assignment to experts, peer review, aggregation of review results, panel evaluation, and final

awarding decision. These processes are very similar in other funding agencies, except that there are a very large number of proposals that need to be grouped for peer review in the NSFC. In the NSFC, the number of research proposals received has more than doubled in the past four years, with over 110,000 proposals submitted in one deadline in March 2010.

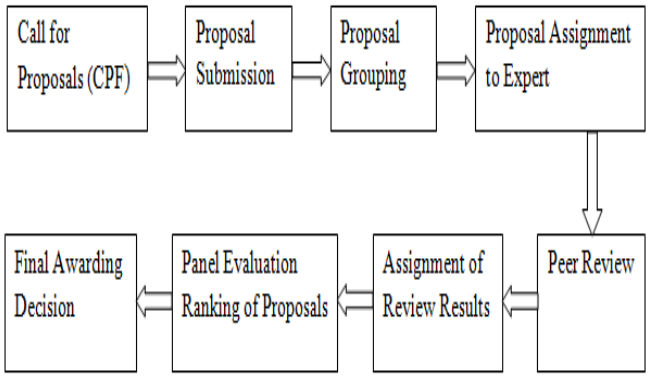


Fig 1.1 Research Project Selection Process in NSFC

Founded in 1986, the NSFC is the largest government funding agency in China, with the primary aim to fund and manage basic research. The agency is made up of seven scientific departments, four bureaus, one general office, and three associated units. The scientific departments are the decision-making units responsible for funding recommendations and management of funded projects. Departments are classified according to scientific research areas, including mathematical and physical sciences, chemical sciences, life sciences, earth sciences, engineering and material sciences, information sciences, and management sciences. These departments are further divided into 40 divisions with a focus on more specific research areas. 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 an urgent need for an effective and feasible approach to group the submitted research proposals with computer supports. An ontology-based text-mining approach is proposed to solve the problem.

1. First, a research ontology containing the projects funded in latest five years is constructed according to keywords, and it is updated annually.
2. Then, new research proposals are classified according to discipline areas using a sorting algorithm.
3. Next, with reference 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 assign to expert review.

II. LITERATURE SURVEY

Selection of research projects is an important research topic in research and development (R&D) project management. Previous research deals with specific topics, and several formal methods and models are available for this purpose. For example, Jain and Wei xu [1] proposed a fuzzy-logic-based model as a decision tool for project selection. M. Uma [2] and Archer offered a decision support approach to project portfolio selection. E. Sathya [2] and Bhattacharya proposed a fuzzy logic approach to project selection. Butler used a multiple attribute utility theory for project ranking and selection. Loch and MRS. Punitha [3] established a dynamic programming model for project selection, while Meade and J.Butter [4] developed an analytic network process model. Cook presented a method of optimal allocation of proposals to reviewers in order to facilitate the selection process. Arya and Morrice [4] proposed a rotation program method for project assignment. Jain [1] and Park used text-mining approach for R&D proposal screening. Dr.M.Punithavalli [3] 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. Methods have been developed to group proposals for peer review tasks.

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. Text-mining methods (TMMs) have been designed to group proposals based on understating the English text, but they have limitations when dealing with other language texts. The proposed approach has been successfully tested at the NSFC. The experimental results indicated that the method can also be used to improve the efficiency and effectiveness of the research project selection process.

III. PROPOSED SYSTEM

3.1 System Overview

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.

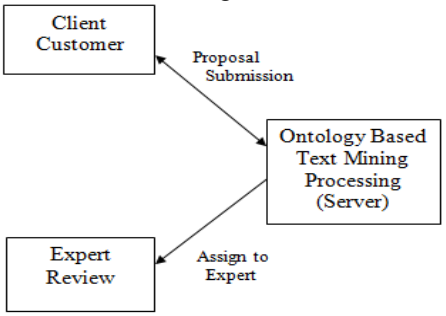


Fig 3.1.1. Client-Server Model

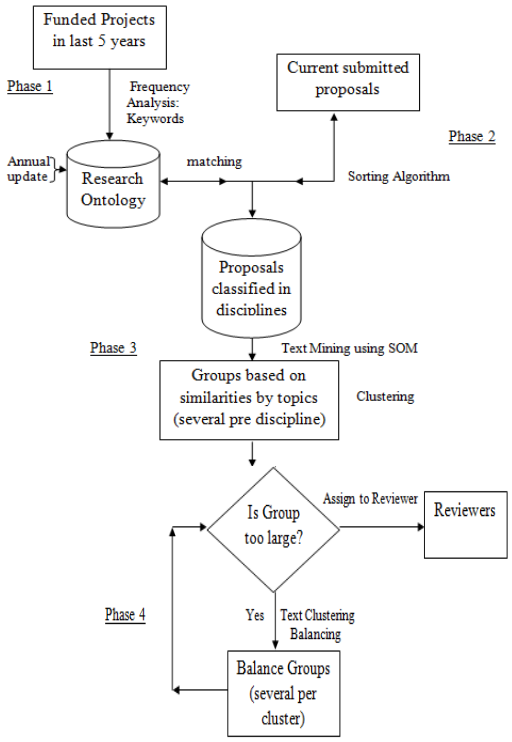


Fig 3.1. 3Process of Proposed OTMM

The proposed OTMM is used together with statistical method and optimization models and consists of four phases, as shown in Fig.1.3. 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' characteristics are taken into consideration (e.g., applicants' affiliations in each proposal group should be diverse).

Phase 1: Constructing a Research Ontology

Funding agencies such as the NSFC maintain a directory of discipline areas that form a tree structure. As a domain ontology, a research ontology is a public concept set of the research project management domain. Suppose that there are K discipline areas, and Ak denotes discipline area k (k = 1, 2, . . . , K).

Research ontology can be constructed in the following three steps to represent the topics of the disciplines. The example model of structure of research ontology:

Creating the research topics of the discipline Ak, (k = 1, 2, . . . , K)

The keywords of the supported research projects each year are collected, and their frequencies are counted. The keywords and their frequencies are denoted by the feature set (Nok, IDk, year, {(keyword1, frequency1),(keyword2,frequency2),. . . , (keyword, frequency)}), where Nok is the sequence number of the kth record and IDk is the corresponding discipline code. For instance, if discipline Ak has two keywords in 2007 (i.e., "data mining" and "business intelligence") and the total number of counts for them are 30 and 50, respectively, the discipline can be denoted by (Nok, IDk, 2007, {(data mining, 30), (business intelligence, 50)}). In this way, a feature set of each discipline can be created. The keyword frequency in the feature set is the sum of the same keywords that appeared in this discipline during the most recent five years, and then, the feature set of Ak is denoted by (Nok, IDk, {(keyword1, frequency1)(keyword2, frequency2), . . . ,(keyword, frequency)}).

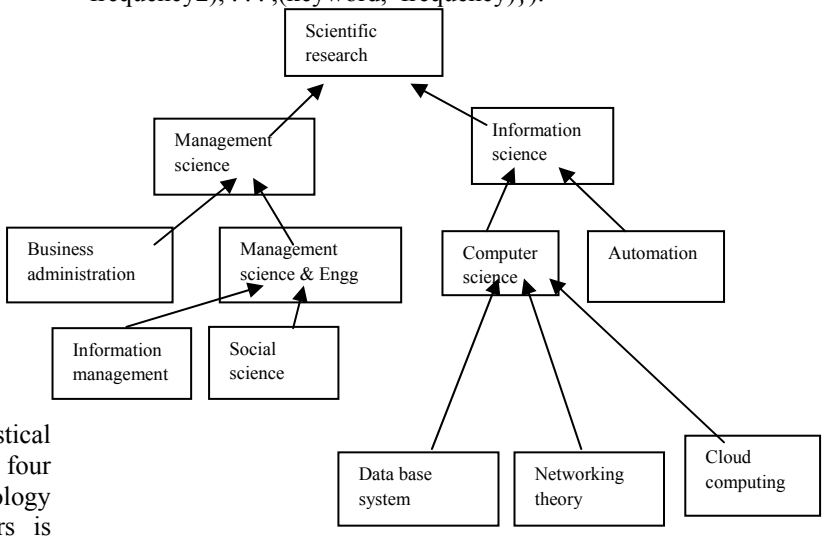


Fig. 3.1.2 Structure Of Research Ontology

- **Constructing the research ontology**

First, the research ontology is categorized according to scientific research areas introduced in the background. It is then developed on the basis of several specific research areas. Next, it is further divide into some arrowed discipline areas. Finally, it leads to research topics in terms of the feature set of disciplines created in. It is more complex than just a tree-like structure. There are some cross-discipline research areas (e.g., "data mining" can be placed under "Information engagement" in "Management Sciences" or under "Artificial Intelligence" in "Information Sciences"). Therefore, the research ontology allows more complex relationship between concepts besides the basic tree-like structure. Also, to deal with proposals in English text, there are some synonyms used by different project applicants, which have different names in different proposals but represent the same concepts. Therefore, the research ontology allows more complex relationship between concepts besides the basic tree-like structure.

- **Updating the research ontology**

Once the project funding is completed each year, the research ontology is updated according to agency's policy and the change of the feature set.

Phase 2: Classifying New Research Proposals Into Disciplines

Proposals are classified by the discipline areas to which they belong. A simple sorting algorithm is used next for proposals' classification. This is done using the research ontology as follows. Suppose that there are K discipline areas, and Ak denotes area k (k = 1, 2, . . . , K). Pi denotes proposals i (i = 1, 2, . . . , I), and Sk represents the set of proposals which belongs to area k. A sorting algorithm can be implemented to classify proposals to their discipline areas.

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

After the research proposals are classified by the discipline areas, the proposals in each discipline are clustered using the text-mining technique. The main clustering process consists of five steps: text document collection, text document preprocessing, text document encoding, vector dimension reduction, and text vector clustering. The details of each step are as follows.

- **Text document collection**

After the research proposals are classified according to the discipline areas, the proposal documents in each discipline Ak (k = 1, 2, . . . , K) are collected for text document preprocessing.

- **Text document preprocessing**

The contents of proposals are usually non-structured. Because the texts of the proposals consist of Chinese characters which are difficult to segment, the research ontology is used to analyze, extract, and identify the keywords in the full text of the proposals. For example, "Research on behavior modeling and detection methods in financial fraud using ensemble learning" can be divided into word sets {"behavior modeling," "detection method," "financial fraud," "ensemble learning"}. Finally, a further 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 proposal documents.

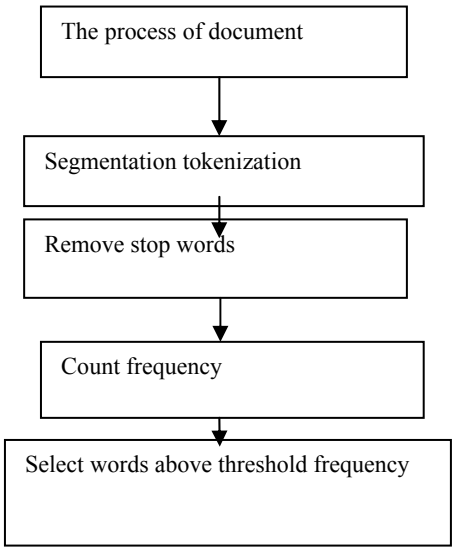


Fig. 3.1.4 Preprocessing steps of text document for text clustering.

- **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 encoding of the keyword wi. 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.

- **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. 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. Ruther more, the term-by document matrix is decomposed into a set of eigenvectors using angular-value decomposition. The eigenvectors that have the least impacts on the matrix are then discarded.

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 'n' supervised learning neural network model that clusters input data with similarities.

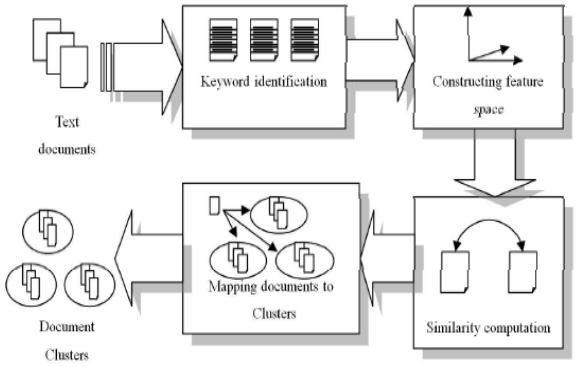


Fig. 3.1.5 Text clustering system

SOM Algorithm:

All problems are coming under NP category. SOM problem comes under NP-Complete area. SOM is the unsupervised learning neural network. Hence, we have to take decision about winning neuron and such kind of decision problems falls under NP-Complete.

In unsupervised learning, training of network is entirely data driven and no target result for input data vectors is provided. Input data vector may fall under any of the output unit which is non-deterministic. SOM provides a topology preserving mapping from high dimensional space to map units. Map units or neurons, usually forms two dimensional (2D) lattice and thus mapping is mapping from high dimensional space onto plane.

- **Input:**
  - ✓ Training data: vectors, X
    – Vectors of length *n*
    (x1,1, x1,2, ..., x1,i,…, x1,n)
    (x2,1, x2,2, ..., x2,i,…, x2,n)
    …
    (xj,1, xj,2, ..., xj,i,…, xj,n)
    …
    (xp,1, xp,2, ..., xp,i,…, xp,n)
    – Vector components are real numbers
- **Outputs**
  – A vector, Y, of length *m*: (y1, y2, ..., yi,…, ym)
  Sometimes m < n, sometimes m > n, sometimes m = n
  – Each of the *p* vectors in the training data is classified as falling in one of *m* clusters or categories
  – That is: Which category does the training vector fall into?
- **Generalization**
  – For a new vector: (xj,1, xj,2, ..., xj,i,…, xj,n)
  – Which of the *m* categories (clusters) does it fall into?

Network Architecture

Two layers of units:
– Input: *n* units (length of training vectors)
– Output: *m* units (number of categories)
Input units fully connected with weights to output units.::

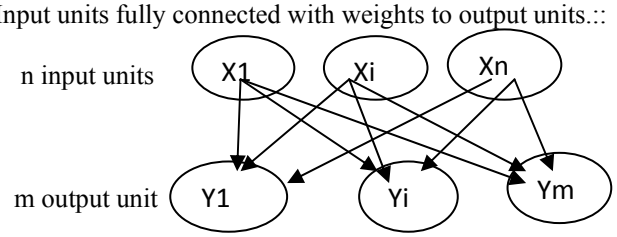


Fig 3.1.6Network Architecture

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.

$$\sum_{k=1}^{n}(i_{l,k} - w_{j,k}(t))^2$$

   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:

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

Phase4: Balancing Research Proposals and Regrouping Them by Considering Applicant's Characteristics

In this phase, when the number of proposals in one cluster is still very large (e.g., more than 20), the applicants' characteristics (e.g., affiliated universities) are considered. The proposal group composition should be diverse. Reviewers may feel confused and uncomfortable when evaluating proposals that may have poor group composition, so it is advisable that the applicant's characteristics in each proposal group should be as diverse as much as possible. Furthermore, the group size in each group should be similar. This may be very complex optimization problem and one solution method that could be use is Genetic Algorithm.

GA is used for optimization of clusters and optimization problems generally come under NP-Hard category. NP-Hard problems are more complex and more than simple polynomial.

GA is based on mechanics of biological evolution. GA provides solution for high complex search space.

GA Operators:

- Population ⇨ set of solutions
- Fitness ⇨ quality of solution
- Chromosome ⇨ encoding for solution
- Gene ⇨ part of encoding solution
- Reproduction ⇨ crossover
- Offspring ⇨ parent's child
- Mutation ⇨ change is genetic structure results in variant form

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;
```

Encoding

The process of representing a solution in the form of a string that conveys the necessary information is encoding. Each bit in the string represents a characteristic of the solution. Most common method of encoding is binary coded. Chromosomes are strings of 1 and 0 and each position in the chromosome represents a particular characteristic of the problem.

Fitness function

A fitness function value quantifies the optimality of a solution. The value is used to rank a particular solution

against all the other solutions. A fitness value is assigned to each solution depending on how close it is actually to the optimal solution of the problem. $F(d,h)=c((\pi d^2/2)+\pi dh)$ … Fitness equation

Crossover

The crossover operator is used to create new solutions from the existing solutions. This operator exchanges the gene information between the solutions in the mating pool. The most popular crossover selects any two solutions strings randomly from the mating pool and some portion of the strings is exchanged between the strings. The selection point is selected randomly.

- **Simple Crossover**

  It is similar to binary crossover.

P1 = [8 6 3 7 6] ⇨ C1 = [8 6 3 7 6]
P2 = [2 9 4 8 9] C2 = [2 9 4 8 9]

Linear Crossover

Parents: **(**x1,…,xn **)** and **(**y1,…,yn **)**

Select a single gene (*k*) at random

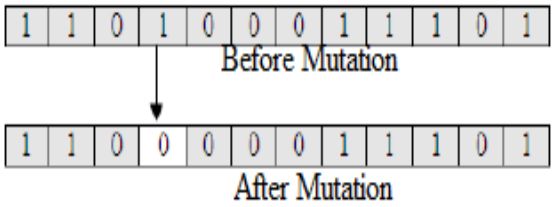
Three children are created as,

(x1, ...,xk , 0.5*yk+0.5*xk , ...,xn)
(x1, ...,xk , 1.5*yk-0.5*xk , ...,xn)
(x1, ...,xk ,- 0.5*yk+1.5*xk , ...,xn)

From the three children, best two are selected for the next generation.

Mutation

Mutation is the occasional introduction of new features in to the solution strings of the population pool to maintain diversity in the population. Though crossover has the main responsibility to search for the optimal solution, mutation is also used for this purpose. Mutation operator changes a 1 to 0 or vice versa.



Research proposal assignment

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

CONCLUSION

In this paper we uses text mining multilingual ontology, optimization and statistical analysis technique to cluster research approach based on their similarities .proposed system indicate that improvement of the efficiency and effectiveness of the research project selection. The current system of grouping of papers for research paper selection based on similarities of Keywords and Frequencies of research papers of ontology.

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 System,man, and cybernetics May,2012.

[2] 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

[3] 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.

[4] 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. [5] S. Hettich and M. Pazzani, "Mining for proposal reviewers: Lessons learned at the National Science Foundation," in *Proc. 12th Int. Conf. Knowl. Discov. Data Mining*, 2006, pp. 862–871.

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

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

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

[9] 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.

[10] 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.

[11] T. H. Cheng and C. P. Wei, ―A clustering-based approach for integrating document-category hierarchies,‖ IEEE Trans. Syst., Man, Cybern.A,Syst., Humans, vol. 38, no. 2, pp. 410–424, Mar. 2008.

[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] P. Velardi, P. Fabriani, and M. Missikoff. Using text processing techniques to automat-ically enrich a domain ontology. In Proceedings of the ACM International Conference on Formal Ontology in Information Systems, 2001