IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 17, Issue 1, Ver. I (Jan – Feb. 2015), PP 65-71
www.iosrjournals.org

Research Paper Selection Based On an Ontology and Text Mining Technique Using Clustering

[1] Snehal Shivaji Patil [2] S.A.Uddin
Alhabeeb college of Engg. & Tech. Hyderabad, India.
Alhabeeb college of Engg. & Tech. Hyderabad, India.

Abstract: *Research Paper Selection is important decision making task for the Government funding Agency, Universities, 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 method of grouping of papers for research paper selection based on similarities of Keywords and Frequencies of research papers of ontology. Text mining is method for extracting unknown information from the large documents. The Research Papers in each domain are clustered using Text mining Technique. Grouped Research papers are assigned to appropriate reviewer or domain experts for peer review systematically. The Reviewer results are collected and papers are get ranked based on experts review results.*

Keywords: *Ontology, Text Mining, Classification, Document Preprocessing, Clustering*.

I. Introduction

Research Paper Selection is important decision making task for many Organizations such as Government funding Agency, Universities, research Institutes.

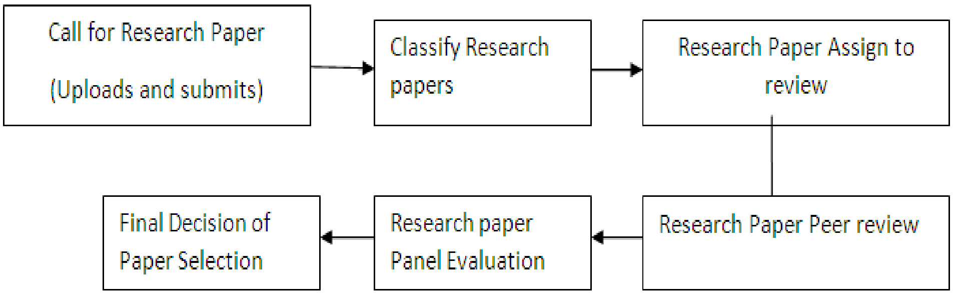
Call for Research Paper
(Uploads and submits)
Classify Research papers
Research Paper Assign to review
Research Paper Peer review
Research paper Panel Evaluation
Final Decision of Paper Selection

Fig.1. Research Paper Selection Process.

Fig. 1 shows the processes of research project selection i.e. Call for papers (CFP), paper submission, paper grouping, paper 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 papers that need to be grouped for peer review. Four to five reviewers are assigned to review each paper so as to assure accurate and reliable opinions on papers. To deal with the large volume, it is necessary to group papers according to their similarities in research disciplines and then to assign the paper groups to relevant reviewers.

In the first section we Call for Research Paper means uploading Research paper and submitting the details of that paper. Classification of research papers is based on keywords of papers similar with ontology keywords and frequencies of those keywords. Department members are classified into six groups according to their decision making in research paper selection. Decision making cooperate with each other to accomplish overall goal of selecting research paper as shown above in figure the output of one block is input to the next block.

Department is responsible for selection of research papers for classification. Department members classify research papers and assign them to external reviewer for evaluation and commentary. If department member may not have knowledge about research paper in all research domain and contents of many papers were not fully understood when papers were grouped and while assigning grouped papers to external reviewers. Therefore, there was an effective approach to group the submitted research Papers and assign the papers to external reviewers with computer supports. So we use ontology based text- mining approach is proposed to solve the problem.

Section II reviews the literature on research paper selection, grouping of papers and assigning the grouped papers to external reviewer systematically. The proposed method on an ontology based text mining framework for Research Paper Selection is described in Section III. Section IV validates and evaluates the method, and then discusses the potential application.

II. Literature Review

Selection of research projects is an important research topic in research and development (R&D) project management. Previous re- search deals with specific topics, and several formal methods and models are available for this purpose.

For example, Chen and Gorla [2] proposed a fuzzy-logic-based model as a decision tool for project selection. Henriksen and Traynor [3] presented a scoring tool for project evaluation and selection. Ghasemzadeh and Archer [4] offered a decision support approach to project portfolio selection. Machacha and Bhattacharya [5] proposed a fuzzy logic approach to project selection. Butler et al. [6] used a multiple attribute utility theory for project ranking and selection. Loch and Kavadias [7] established a dynamic programming model for project selection, while Meade and Presley [8] developed an analytic network process model. Greiner et al. [9] proposed a hybrid AHP and integer programming approach to support project selection, and Tian et al. [10] suggested an organizational decision support approach for selecting R&D projects. Cook et al. [11] presented a method of optimal allocation of papers to reviewers in order to facilitate the selection process. Arya and Mittendorf [12] proposed a rotation program method for project assignment. For example Choi and Park [13] used text-mining approach for R&D paper screening. Girotra et al. [14] offered an empirical study to value projects in a portfolio. Sun et al. [15] developed a decision support system to evaluate reviewers for research project selection. Finally, Sun et al. [16] proposed a hybrid knowledge-based and modeling approach to assign reviewers to papers for research project selection.

Cheng and Wei [2008] proposed clustering-based category-hierarchy integration (CHI) technique, which is an extension of the clustering-based category integration (CCI) technique. This method was improving the effectiveness of category-hierarchy integration compared with that attained by nonhierarchical category-integration techniques particularly homogeneous [16].

Methods have been developed to group papers for peer review tasks. For example, Hettich and Pazzani [2006] proposed a text-mining approach to group papers, identify reviewers, and assign reviewers to papers. Current methods group papers according to keywords. Unfortunately, papers 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 papers. Second, keywords are provided by applicants who may have subjective views and misconceptions, and keywords are only a partial representation of the research papers. 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 papers into the right groups. [17]

III. Existing System

The existing system is an Ontology-Based Text-Mining Method to cluster research papers based on their similarities in research areas. It consists of three phases. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts.

It consists of 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 papers according to their discipline areas.

Text mining refers generally to the process of extracting interesting information and knowledge from unstructured text. The main difference 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.

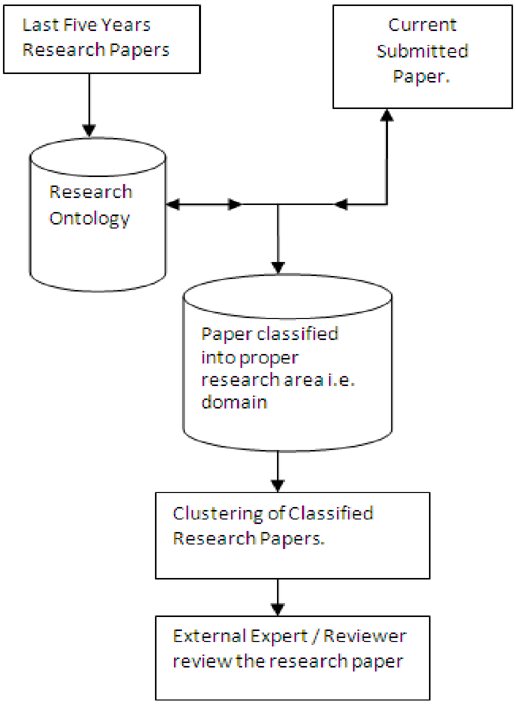
Last Five Years Research Papers
Current Submitted Paper.
Research Ontology
Paper classified into proper research area i.e. domain
Clustering of Classified Research Papers.
External Expert / Reviewer review the research paper

Fig.2 Main Framework of Proposed system.

1. **Constructing Research Ontology:** A research ontology containing the projects funded in latest five years is constructed according to keywords, and it is updated annually. As domain ontology research ontology is a public concept set of the research project management domain. The research topics of different disciplines can be clearly expressed by research ontology.

2. **Classifying New Research Papers**: New research papers are classified according to the keyword stored in ontology with the topic identified using Topic Identification Algorithm.

3. **Clustering: Research Papers Based on Similarities Using Text Mining:** After the research papers are classified by the discipline areas, the papers 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.

A. Graphs:

In the existing system we display the graph accoprding to domain, keywords and institute wise.

1. Domain wise graph:

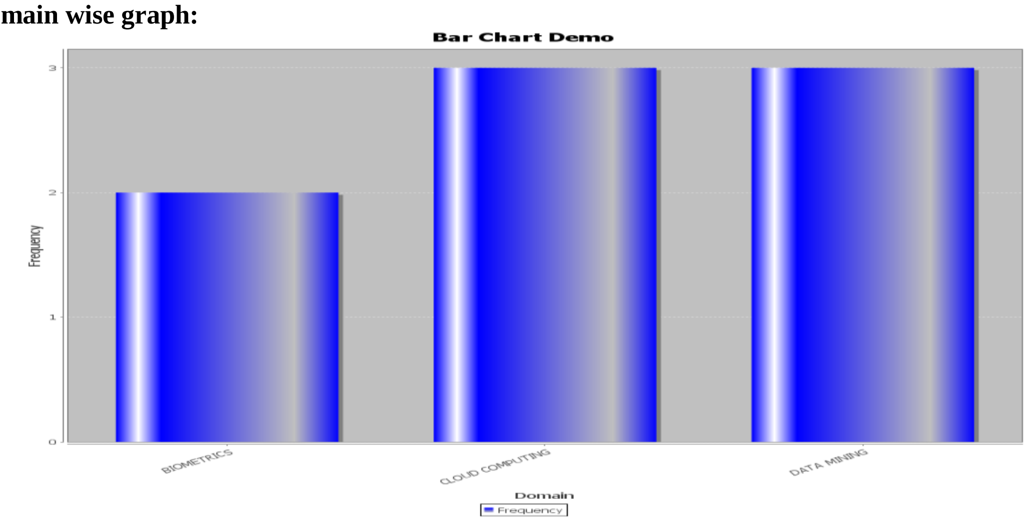

Fig.3 Domain wise graph

This bar chart shows the domain wise graph. This graph contains the domain and the frequency. AS per we submit the domain wise paper in the existing systsem after classifying & reviewing paper the graph will display the selected paper domain and domains frequency. This graph shows the cloud computing and data mining domains and freqency is higher than biometrics.

2. Keyword wise graph:

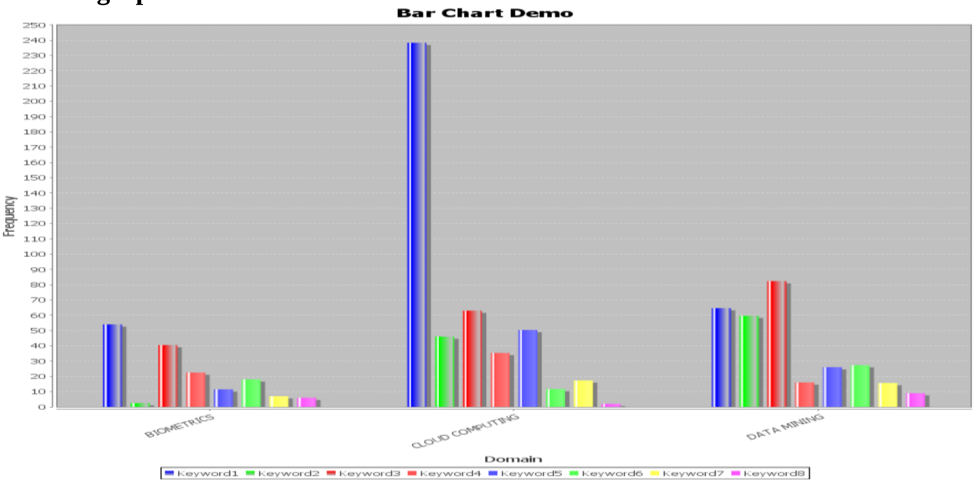

Fig.4 Keyword wise graph

This bar chart shows the keyword wise graph of selected papers.This graph contains the domain wise paper and papers keyword and the frequency. AS per we submit the domain wise paper in the existing systsem after classifying & reviewing paper the graph will display the selected paper domain and keywords of that papers and domain frequency. This graph shows the keywords of cloud computing and data mining and freqency is higher than biometrics.

3. Institute wise graph:

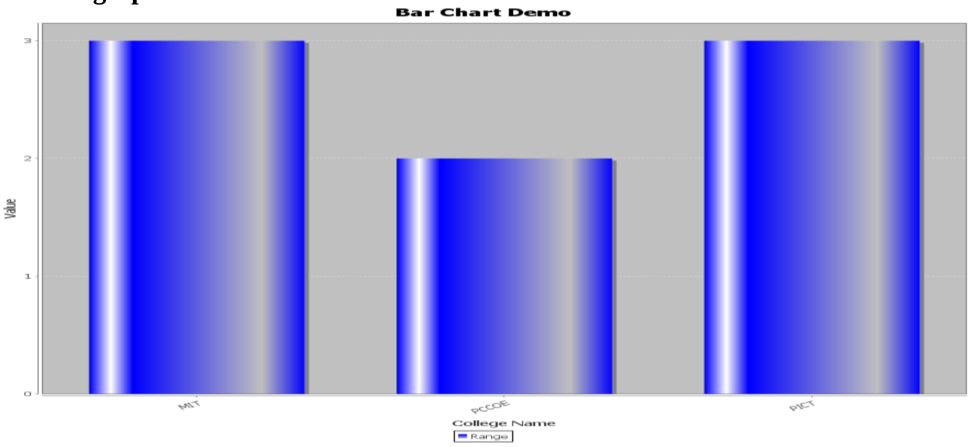

Fig.5 Institute wise graph

This bar chart shows the institute name wise graph. The graph shows the college name and value means how many papers get submitted. In this graph MIT and PICT College's student submitted lot of papers so value increased as compared other colleges.

IV. Ontology Based Text Mining Framewok

In the R&D, after papers are submitted, the next important task is to group papers and assign them to reviewers. The papers in each group should have similar research characteristics. For instance, if the papers in a group fall into the same primary research discipline (e.g., supply chain management) and the number of papers is small, manual grouping based on keywords listed in papers can be used and assign them to reviewer manually. However, if the number of papers is large, it is very difficult to group papers and assign them to reviewer manually. So the papers are classified using ontology and topic identification algorithm and then papers are clustered using text mining and last it is submitted to reviewer systematically.

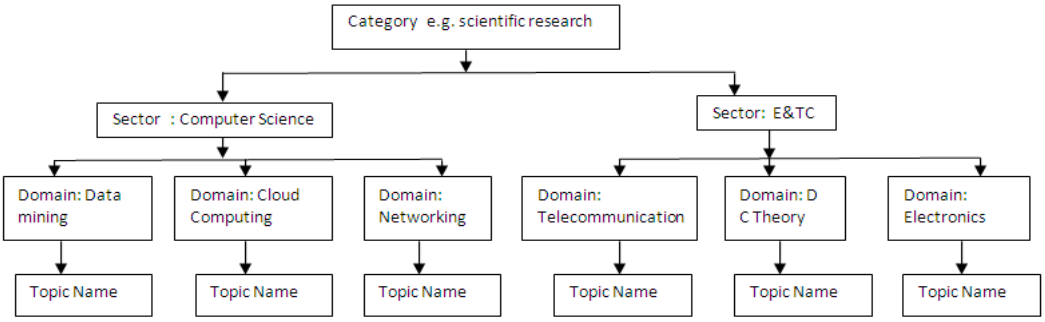

Fig.6 Structure of Research Ontology

As shown in Above Fig. It is Tree like Structure Containing Category, Sector, Domain and Topic Name of Research Papers. Considering the example Scientific Research is the Category, under that there are different sectors like Computer Science, E&TC. Sectors containing the Different Domains of the research area .Each domain contain the various topic names of the Research papers. Each topic name contains keywords of more than two domains then that topic name gets classified under domain which comes first. By using this method we solve the problem of grouping of research papers under the proper domain area.

Module1: Research Ontology Building:

Step1) 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. Creating the research topics of the discipline Ak, (k =1 ,2,...,K). The keywords and their frequencies are denoted by the feature set (Nok,IDk,year,{ (keyword1,frequency1), (keyword2, frequency2),..., (keywordk,frequencyk)}, 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 (shown in Fig. 4), and then, the feature set of Ak is denoted by (Nok,IDk,{(keyword1,frequency1)(keyword2, frequency2) (keywordk, frequencyk)}).

Step2) 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.

Sep3) 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.

Module 2: Classifying New Research Papers:

Research papers are classified by the domain area based on ontology keywords. 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 papers i(i =1 ,2,...,I), and Sk represents the set of papers which belongs to area k.

Module 3: Clustering Research Papers Based on Similarities Using Text Mining:

After the research papers are classified by the domain areas, the papers in each domain are clustered using the text-mining technique. The main clustering process consists of four steps, as shown in Fig. text document collection, text document preprocessing, text document encoding, and text vector clustering. The details of each step are as follows.

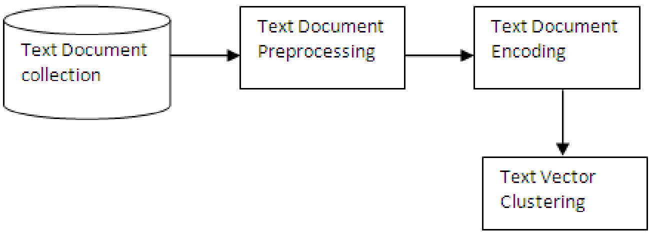

Fig.7 Process of text mining

Step 1) Text document collection: After the research papers are classified according to the discipline areas, the paper documents in each discipline Ak(k =1 ,2,...,K) are collected for text document preprocessing.

Step 2) Text document preprocessing: The contents of papers are usually nano structured. Because the texts of the papers consist of characters which are difficult to segment, the research ontology is used to analyze, extract, and identify the keywords in the full text of the papers.. Finally, a further reduction in the vocabulary size can be achieved through the removal of all stop words that appeared only a few times in all paper documents.

Step 3) Text document encoding: After text documents are segmented, they are converted into a feature vector representation: V =(v1,v2,...,vM), where M is the number of features selected and vi(i =1 ,2,...,M) is the TF-IDF encoding [18] of the keyword wi. TF-IDF encoding describes a weighted method based on inverse document frequency (IDF) combined with the term frequency (TF) to produce the feature v, such that vi = tfi ∗log(N/dfi), where N is the total number of papers in the discipline, tfi is the term frequency of the feature word wi, anddf i is the number of papers containing the word wi. Thus, research papers can be represented by corresponding feature vectors.

Step 4) Text vector clustering: This step uses K-MEANS clustering algorithm to cluster the feature vectors based on similarities of re- search areas. The K-MEANS clustering algorithm is a typical unsupervised learning neural network model that clusters input data with similarities. Details of the K-MEANS algorithm.

V. Clustering Using K-Means Algorithm

After the construction of the document vector, the process of clustering is carried out. The K-MEANS clustering algorithm is used to meet the purpose of this project. The basic algorithm of K-MEANS used for the project is as following:

K-Means Algorithm Use:

For partitioning where each cluster's center is represented by the mean value of the objects in the cluster.
Input:
k: the number of clusters,
Output:
A set of k clusters.
Method:
Step 1: Choose k numbers of clusters to be determined.
Step 2: Choose Ck centroids randomly as the initial centers of the clusters.
Step 3: Repeat

3.1: Assign each object to their closest cluster center using Euclidean distance.
3.2: Compute new cluster center by calculating mean points.
Step 4: Until

4.1: No change in cluster center OR
4.2: No object changes its clusters.

VI. Conclusion And Future Work

This paper has presented a framework on ontology based text mining for grouping research papers and assigning the grouped paper to reviewers systematically. Research ontology is constructed to categorize the concept terms in different discipline areas and to form relationships among them. It facilitates text-mining and optimization techniques to cluster research papers based on their similarities and then to assign them to reviewer according to their concerned research area. The papers are assigned to reviewer with the help of knowledge based agent. Future work is needed to replace the work of reviewer by system. Also, there is a need to empirically compare the results of manual classification to text-mining classification. Also there is need to sending message on user's mobile number and also further profile schedule.

References

[1]. 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.

[2]. 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.

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

[4]. 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 andM. L. Gavrilova, Eds., 2007, pp. 660–673, Berlin,Germany: Springer-Verlag.

[5]. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang”An Ontology-Based Text- Mining Method to Cluster Papers for Research Project Selection”,IEEE Trans an systems and humans vol.42,no.3 May2012

[6]. A. Maedche and S. Staab, “The Text-To-Onto ontology learning environment,”in Proc. 8th Int.Conf. Conceptual Struct., Darmstadt, Germany,2000, pp. 14–18.

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

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

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

[10]. C. Wei, C. S. Yang, H. W. Hsiao, and T. H. Cheng, “Combining preference- and content based approaches for improving document clustering effectiveness,” Inf. process. Manage.,vol. 42, no. 2, pp. 350–372,Mar. 2006.

[11]. T. A. Runkler and J. C. Bezdek, “Web mining with relational clustering,”Int. J. Approx. Reason., vol. 32, no. 2/3, pp. 217–236, Feb. 2003

[12]. H. Li, K. Zhang, and T. Jiang, “Minimum entropy clustering and applications to gene expression analysis,” in Proc. 3rd IEEE Comput. Syst. Bioinform. Conf., Stanford, CA, 2004, pp. 142–151.

[13]. S. Gauch, J. Chaffee, and A. Pretschner, “Ontology-based personalized search and browsing,” Web Intell. Agent Syst., vol. 1, no. 3/4,pp. 219– 234, Dec. 2003.

[14]. 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.

[15]. Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon ,“An Automatic Topic Identification Algorithm,” Journal of Computer Science 7 (9): 1363-1367, 2011 ISSN 1549-3636

[16]. 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.

[17]. S. Hettich and M. Pazzani, “Mining for paper reviewers: Lessons learned at the National Science Foundation,” in Proc. 12th Int. Conf.Knowl. Discov. Data Mining, 2006, pp. 862–871.