AN ONTOLOGY BASED TEXT MINING FRAMEWORK FOR R&D PROJECT SELECTION

N.Arunachalam [1], E.Sathya [2], S.Hismath Begum [2] and M.Uma Makeswari [2]

[1] Asst Professor, IT Department, Sri Manakula Vinayagar Engineering College
narunachalam85@gmail.com
[2] Students, IT Department, Sri Manakula Vinayagar Engineering College
esathyaeas@gmail.com
hismathjasmine@gmail.com
umamanickam26@gmail.com

ABSTRACT

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. Text Mining has emerged as a definitive technique for extracting the unknown information from large text document. Ontology is a knowledge repository in which concepts and terms are defined as well as relationships between these concepts. Ontology's make the task of searching similar pattern of text that to be more effective, efficient and interactive. The current 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. This method is efficient and effective for clustering research proposals. However proposal assignment regarding research areas to experts cannot be often accurate. This paper presents a framework on ontology based text mining to cluster research proposals, external reviewers based on their research area and to assign concerned research proposals to reviewers systematically. A knowledge based agent is appended to the proposed system for a retrieval of data from the system in an efficient way.

KEYWORDS

Clustering analysis, ontology, R&D, text mining, and knowledge based agent.

I. INTRODUCTION

Research project selection is important task for many organizations such as government funding agencies. The submitted research proposals are assigned to experts for review. Four to five reviewers are assigned to review each proposal 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 assigns the proposal groups to relevant reviewers.

Project selection is usually done once a year, by listing the projects, evaluating and comparing all these projects according to quantitative and qualitative criteria, and prioritizing the projects.

The funds requested by all the projects are compared with the laboratory budget and the project list is reduced according to the budgeted amount.

Some experienced R & D managers do not allocate all the budgeted funds, but keep a small percentage on reserve to take care of new projects that may be proposed during the year, after the laboratory official budget has been approved.

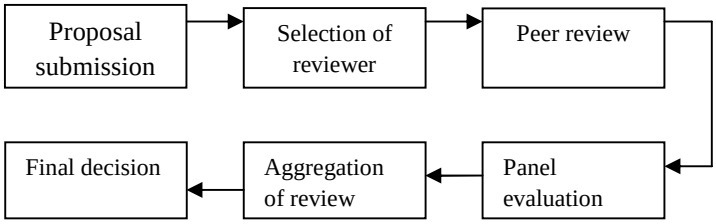

Figure 1. Major decision tasks in R&D project selection process

Fig. 1 shows the processes of research project selection at the Research and Development project selection

The decision makers are classified into six groups according to their decision-making tasks in the R&D project selection process. These decision-making groups cooperate with each other to accomplish the overall goal of selecting the best project proposals. They perform decision tasks in a certain sequence in which the outputs of one group can be the inputs to another group.
The department is responsible for the selection tasks, and it dedicates the tasks to divisions or programs. 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 while assigning 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 with computer supports. Web based ontology in text- mining approach is proposed to solve the problem.

The remaining section of this paper is described as follows. Section II reviews the literature on research project selection, grouping of proposals and assigning the grouped proposal to external reviewer systematically. The proposed method on an ontology based text mining framework for R&D Project Selection is described in Section III. Section IV validates and evaluates the method, and then discusses the potential application in the R&D .Finally, Section V provides the conclusion, and it points to future work.

II. LITERATURE REVIEW

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, Yong-Hong Sun, Jian Ma, Zhi-Ping Fan, and Jun Wang[2008] proposed a group decision support approach to evaluate experts for R&D project selection .It is mainly concerned with criteria and their attributes for evaluating experts

are summarized mainly based on the experience with the National Natural Science Foundation of China (NSFC**)**[1]**.**Henriksen, Traynor[1999] presented a scoring tool for project selection and evaluation[2].

S. Bechhofer et al[2004] developed an OWL Web Ontology Language for storing the keywords[3]. Yildiz and Miksch[2007]designed an ontoX—A method for ontology-driven information extraction[4]. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang [2012] proposed An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project based on their similar discipline areas. This method is efficient and effective for clustering research proposals with both English and Chinese texts[5].

Maedche and Staa [2000] used Text-To-Onto ontology environment using supervised learning[6]. Turban, Zhou and Ma [2004] have been established an group decision support approach to evaluating journals[7]. Choi and Park [2006] used for R&D proposal screening system based on text mining approach[8]. Roussinov and Chen,[1999] proposed a Document clustering for electronic meetings an experimental comparison of two techniques[9].Wei and Yang [2006] suggested a Combining preference- and content based approaches for improving document clustering effectiveness[10]. Runkler and Bezdek, [2003] proposed Web mining with relational clustering[11]. Zhang, and Jiang [2004] offered Minimum entropy clustering and applications to gene expression analysis[12]. Gauch, Chaffee, and Pretschner[2003] established an Ontology-based personalized search and browsing[13]. Meade and Presley[2002] proposed an R&D project selection using the analytic network process[14]. Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon [2008] developed an Automatic Topic Identification Algorithm[15].

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

Methods have been developed to group proposals for peer review tasks. For example, Hettich and Pazzani [2006] 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[17].

Yang and Lee [2005] used text mining approach for automatic construction of hyper texts[21].Christian Paz-Trillo,Renata Wassermann[2005] developed an Information Retrieval application using ontologies[22].Matteo Gaeta[2011] have been established for extract relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents using e-learning perspective[24] Razmerita [2011] proposed An ontology-based framework for modeling user behavior—A case study in knowledge management[29].V.M.Navaneethakumar,Dr.C.Chandrasekar[2012]“A Consistent Web

Documents Based Text Clustering Using Concept Based Mining Model"[31]referred this paper for clustering the proposals based on concept.

III. EXISTING SYSTEM

The existing system is an Ontology-Based Text-Mining Method to cluster research proposals based on their similarities in research areas**.** It consists of four phases.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.

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.

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

2. *Classifying New Research Proposals* , new research proposals are classified according to the keyword stored in ontology with the topic identified using Topic Identification Algorithm.

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 new proposals in each discipline are clustered using a self-organized mapping (SOM) algorithm.

4. *Balancing Research Proposals and Regrouping Them by Considering Applicants' Characteristics* if the number of proposals in each cluster is still very large (e.g., more than 20), they will be further decomposed into subgroups where the applicants' characteristics(e.g.,affiliated) universities are taken into consideration. Reviewers may feel confused and uncomfortable when evaluating proposal that may have poor decomposition so it is advisable that the applicants' characteristics in each proposal group should be as diverse as much as possible.

However proposal assignment regarding research areas to experts cannot be often accurate because of manual approach Proposed system presents a framework on ontology based

text mining to cluster research proposals, external reviewers based on their research area and to assign concerned research proposals to reviewers systematically.

IV. ONTOLOGY BASED TEXT MINING FRAMEWOK

In the R&D, 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 and assign them to reviewer manually. However, if the number of proposals is large, it is very difficult to group proposals and assign them to reviewer manually. So the proposals are classified using ontology and topic identification algorithm and then proposals are clustered using text mining and last it is submitted to reviewer systematically.

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.
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.

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(eg.”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 3) Automatic topic identification approach:

3.1 Split the text into sentences: The first step in our algorithm is splitting the sentences on the given text. A sentence is a smallest text part which is capable to have a topic. Hence, we split the document into corresponding sentences. During this research we widely exploit Proxem Antelope (Proxem, 2009) which provides an open-source plenty of NLP tool. One of these tools 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 we would have a set of sentences.

3.2 Pars the sentences: In this step, the algorithm intends to pars the sentences and determines the candidate terms first to avoid any useless calculation.

We 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.

These syntactic parts are accessible through a dependency syntactic parser. In this study, we use the Stanford dependency parser (The Stanford Parser) which is an open-source tool available in Proxem Antelope package.

3.3 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. We 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 topic is considered as a candidate topic.

3.4Calculate the weight for each candidate topic: At this moment we can 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, we avoid any calculation overhead. The calculation formula is

SNV (NP, head (VP)) = IDF (NP) . IDF (head (VP)) / D (NP, head (VP))

3.5Select the final topic: When we determine the candidate topic and its associated weight for each sentence, we 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 4)Updating the research ontology: Once the project funding is completed each year, the research ontology is updated according to agency's policy.

Module 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.

Module 3: 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. [31]

The concept-based mining model for document's text clustering, a raw text document is given as the input. Each document has definite sentence restrictions. Each sentence in the document is marked repeatedly and might have one or more marked verb argument formation. The amount of labeled information is totally reliant on the information present in the sentence. The sentence contained many marked verb argument formation comprises many verbs connected with their arguments. The labeled verb argument structures are examined by the concept-based mining model on sentence and document levels. The process of concept based mining model is shown in fig.

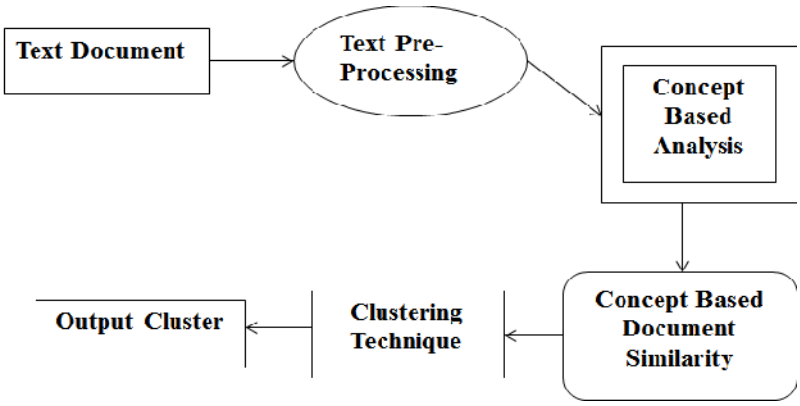

Fig 2 .Concept Based Mining Model Process

The purpose of following the concept-based analysis task is to realize a precise examination of concepts on the sentence, in the document relatively, than a single-term study on the document only.

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.

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. The analysis of document text clustering is done by the proposed document based text clustering algorithm.

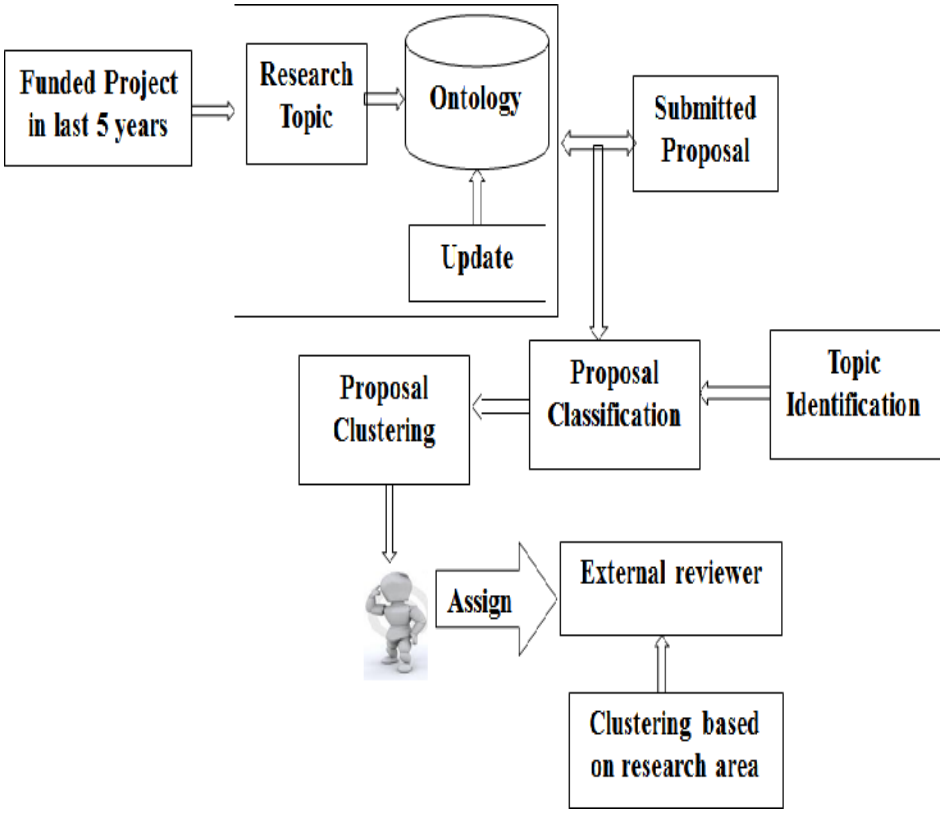

Fig 3. Proposed Framework

Module 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.

Module 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 . But there may be some ambiguous because the reviewer may be specialized in more than one domain For example, the reviewer will be specialized in data mining and network security. So while clustering the reviewer, their priority of research area are taken into consideration.

VI. CONCLUSION AND FUTURE WORK

This paper has presented an framework on ontology based text mining for grouping research proposals and assigning the grouped proposal to reviewers systematically. A 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 proposals based on their similarities and then to assign them to reviewer according to their concerned research area. The proposals 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.

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 Proposals 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 proposal 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 proposal reviewers: Lessons learned at the National Science Foundation," in Proc. 12th Int. Conf.Knowl. Discov. Data Mining, 2006, pp. 862–871.

[18] 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.

[19] R. C. Wang and S. J. Chuu, "Group decisionmaking using a fuzzy linguistic approach for evaluating the flexibility in a manufacturing system,"Eur. J. Oper. Res., vol. 154, no. 3, pp.563–572, 2004.

[20] O.Liu and J. Ma, "A multilingual ontology framework for R&D project management systems," Expert Syst. Appl., vol. 37, no. 6, pp.4626–4631,Jun. 2010.
[21] H. C. Yang and C. H. Lee, "A text mining approach for automatic construction f hypertexts,"Expert Syst. Appl., vol. 29, no. 4, pp. 723–734,Nov. 2005.
[22] ChristianPaz-Trillo,RenataWassermann,"An Information Retrieval application using ontologies
[23] M. Nagy and M. Vargas-Vera, "Multiagent ontology mapping framework for the semantic web," IEEE Trans. Syst., Man, Cybern. A, Syst., Humans,vol. 41, no. 4, pp. 693–704, Jul. 2011
[24] Matteo Gaeta, "Ontology extraction for knowledge reuse the e-learning perspective", IEEE Trans on systems, man, and cybernetics—part a: systems and humans, vol. 41, no. 4, july 2011.
[25] Fabiano D. Beppler,"An Architecture for an Ontology-Enabled Information Retrieval"
[26] "Automatic Ontology Generation:State of the Art "Ivan Bedini, Benjamin Nguye, Orange Labs
[27] "OntoSearch: An Ontology Search Engine",Yi Zhang, Wamberto Vasconcelos, Derek Sleeman
[28] Falcons Concept Search: A Practical SearchEngine for Web Ontologies IEEE transcation onsystem, man and cybernetics,vol,41,no4,july2011.7
[29] L. Razmerita, "An ontology-based framework for modeling user behavior—A case study in knowledge management," IEEE Trans.Syst., Man,Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 772–783,Jul. 2011.
[30] "Construction of Ontology-Based SoftwareRepositories by Text Mining", Yan Wu, Harvey Siy, Mansour Zand, and Victor Winte
[31] "A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model", V.M.Navaneethakumar, Dr.C.Chandrasekar