Research Journal of Applied Sciences, Engineering and Technology 8(12): 1435-1441, 2014
DOI:10.19026/rjaset.8.1118
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: June 08, 2014 Accepted: July 19, 2014 Published: September 25, 2014

Research Article

An Efficient EM based Ontology Text-mining to Cluster Proposals for Research Project Selection

[1]D. Saravana Priya and [2]M. Karthikeyan
[1]Department of Information Technology, P.A. College of Engineering and Technology, Pollachi, India
[2]Department of ECE, Tamilnadu College of Engineering, Coimbatore, India

Abstract: Both the internet and the intranets contain more resources and they are called as text documents. Research and Development (R&D) scheme selection is a type of decision-making normally present in government support agencies, universities, research institutes and technology intensive companies. Text Mining has come out as an authoritative technique for extracting the unknown information from large text document. Ontology is defined as a knowledge storehouse in which concepts and conditions are defined in addition to relationships between these concepts. Ontology's build the task of searching alike pattern of text that to be more effectual, efficient and interactive. The present method for combine proposals for selection of research project is proposed by ontology based text mining technique to the data mining approach of cluster research proposals support on their likeness in research area. This proposed method is efficient and effective for clustering research proposals. Though the research proposal regarding particular research area is cannot always be accurate. This study proposed an ontology based text mining to group research proposals, external reviewers based on their research area. The proposed method like Efficient Expectation-Maximization algorithm (EEM) is used to cluster the research proposal and gives better results in more efficient way.

Keywords: Apriori, document clustering, ontology

INTRODUCTION

At present document clustering is a most active area of research and the development. In that one of the challenging issues is to discover the set of meaningful groups of documents. Where those within each group are more closely related to one another than documents assigned to different groups. The resultant document clusters can provide a structure for organizing large bodies of text for efficient browsing.

Research and development (R&D) project selection is an organizational decision-making task commonly found in organizations like government funding agencies, universities, research institutes and technology-intensive companies. It is a complicated and challenging task to organizations with the following reasons:

- It is very difficult to predict the future success and impacts of the candidate projects.
- It is a multi-stage multi-person decision making process involving a group of decision makers (e.g., external reviewers and panel experts).

Thus, it can be very hard to manage the decision making process, especially when the decision makers have heterogeneous decision-making strategies (Ghasemzadeh and Archer, 2000; Henriksen and Traynor, 1999; Schmidt and Freeland, 1992).

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 (Tar and Nyunt, 2011).

In the field of ontology, ontological framework is normally formed using manual or semi-automated methods requiring the expertise of developers and specialists. This is highly incompatible with the developments of World Wide Web as well as the new E-technology because it restricts the process of knowledge sharing. Search engines will use Ontology to find pages with words that are syntactically different but semantically similar (Decker *et al*., 2000; Ding and Foo, 2002; Hotho *et al*., 2001).

Traditionally, ontology has been defined as the philosophical study of what exists: the study of kinds of entities in reality and the relationships that these entities bear to one another (Berners-Lee, 1999). In Computer

Corresponding Author: D. Saravana Priya, Department of Information Technology, P.A. College of Engineering and Technology, Pollachi, India

Science Ontology is an engineering artifact describing what exists in a particular domain. Ontology belongs to a specific domain of knowledge. The scope of the ontology concentrates on definitions of a certain domain, although sometimes the domain can be very broad. The domain can be an industry domain, an enterprise, a research field, or any other restricted set of knowledge, whether abstract, concrete or even imagined. Ontology is usually constructed with a certain task in mind.

Text mining is the discovery of interesting knowledge in text documents. It is a challenging issue to find accurate knowledge (or features) in text documents to help users to find what they want. In the beginning, Information Retrieval (IR) provided many term-based methods to solve this challenge, such as Rocchio and probabilistic models (Baeza-Yates and Ribeiro-Neto, 1999), rough set models (Li *et al*., 2000), BM25 and Support Vector Machine (SVM) (Robertson and Soboroff, 2002) based filtering models. The advantages of term based methods include efficient computational performance as well as mature theories for term weighting, which have emerged over the last couple of decades from the IR and machine learning communities. However, term based methods suffer from the problems of polysemy and synonymy, where polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want.

LITERATURE REVIEW

Research and Development (R&D) project selection is a 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 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 (Arunachala *et al*., 2013).

Text or document clustering, a subfield of text data mining, is the process of automatically organizing text documents into meaningful groups in such a manner where all the documents in the same cluster have high similarity and have dissimilarity between clusters (Shawkat Ali, 2008). Text clustering techniques have wide usage in search engines (to present organized and understandable results to the user), digital libraries (clustering documents in a collection), automated (or semi-automated) creation of document taxonomies and in general, all information retrieval systems involving text. Perhaps the most popular application of document clustering is the Google News2 service, which uses document clustering techniques to group news articles from multiple news sources to provide a combined overview of news around the Web.

In the field of ontology, ontological framework is normally formed using manual or semi-automated methods requiring the expertise of developers and specialists. This is highly incompatible with the developments of World Wide Web as well as the new E-technology because it restricts the process of knowledge sharing. Search engines will use ontology to find pages with words that are syntactically different but semantically similar (Decker *et al*., 2000; Ding and Foo, 2002; Hotho *et al*., 2001). Traditionally, ontology has been defined as the philosophical study of what exists: the study of kinds of entities in reality and the relationships that these entities bear to one another (Steinbach *et al*., 2000). In Computer Science, ontology is an engineering artifact describing what exists in a particular domain. Ontology belongs to a specific domain of knowledge. The scope of the ontology concentrates on definitions of a certain domain, although sometimes the domain can be very broad. The domain can be an industry domain, an enterprise, a research field, or any other restricted set of knowledge, whether abstract, concrete or even imagined. Ontology is usually constructed with a certain task in mind. In recent years use of term ontology has become prominent in the area of computer science research and the application of computer science methods in management of scientific and other kinds of information. In this sense the term ontology has the meaning of a standardized terminological framework in terms of which the information is organized.

For many firms, especially those that depend on innovation to stay in business, the key to continued competitiveness lies in their ability to develop and implement new products and processes. For these organizations, Research and Development (R&D) is an integral function within the strategic management framework. Even firms with excellent technical skills must work within the limits of available funding and resources. R&D project selection and funding decisions, then, are critical if the organization is to stay in business. While there are many mathematical decision-making approaches proposed for this decision, literature suggests that few are actually being used. Major criticisms of these techniques include their inability to consider strategic factors and their mathematical complexity (Albala, 1975; Fahrni and Spatig, 1990; Lockett *et al*., 1986).

A number of R&D selection models and methods have been proposed in practitioner and academic

literature. Reviews of many of these can be found in Baker and Freeland (1975), Martino (1995) and Henriksen and Traynor (1999). Included in the articles reviewed in their papers are those that utilize criteria and methods such as NPV, scoring models, mathematical programming models and multi attribute approaches. Even with the number of proposed models, the R&D selection problem remains problematic and few models have gained wide acceptance. Liberatore and Titus (1983) conducted an empirical study on the use of quantitative techniques for R&D project management. They found that most R&D organizations use one or more traditional financial methods for determining project returns, often in conjunction with other methods. Mathematical programming techniques such as linear and integer programming are not commonly used in industry, primarily because of the diversity of project types, resources and criteria used. They also found that many managers do not believe that the available methods for project selection improve the quality of their decisions.

First, researchers and practitioners working in the areas of information retrieval and text mining seek to find categories of textual resources by various fully automatic methods. The approaches either:

- Predefine a metric on a document space in order to cluster 'nearby' documents into meaningful groups of documents (called 'unsupervised categorization' or 'text clustering'; (Salton, 1989).
- They adapt a metric on a document space to a manually predefined sample of documents assigned to a list of target categories such that new documents may be assigned to labels from the target list of categories, too ('supervised categorization' or 'text classification'; (Fabrizio, 2002).

Second, researchers and practitioners working mainly in the areas of thesauri (Foskett, 1997) and ontologies (Staab and Studer, 2004) predefine conceptual structures and assign metadata to the documents that confirm to these conceptual structures.

Many types of text representations have been proposed in the past. A well known one is the bag of words that uses keywords (terms) as elements in the vector of the feature space. In Li and Liu (2003), the tf $*$ idf weighting scheme is used for text representation in Rocchio classifiers. In addition to TFIDF, the global IDF and entropy weighting scheme is proposed in Dumais (1991) and improves performance by an average of 30%. Various weighting schemes for the bag of words representation approach were given in Aas and Eikvil (1999), Joachims (1997) and Salton and Buckley (1988). The problem of the bag of words approach is how to select a limited number of features among an enormous set of words or terms in order to increase the system's efficiency and avoid over fitting (Lewis, 1992). In order to reduce the number of features, many dimensionality reduction approaches have been conducted by the use of feature selection techniques, such as Information Gain, Mutual Information, Chi-Square, Odds ratio and so on. Details of these selection functions were stated in Lewis (1992).

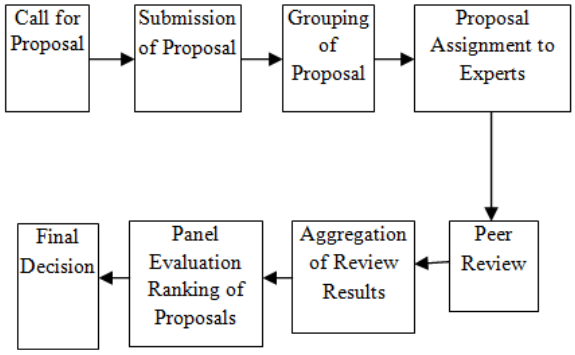

Fig. 1: Research project selection processes

Research project selection process: A number of research projects proposal received are increased in the past years according to the National Natural Science Foundation of China (NSFC). For each and every proposal nearly Four to five reviewers are allocated to review each proposal to determine their accuracy and their reliability.

Figure 1 shows the selection processes of research project at the National Natural Science Foundation of China (NSFC), that is, the CFP, proposal submission, proposal grouping, proposal assignment to experts, peer review; aggregation of review results, panel evaluation and final awarding decision is explained in Tian *et al*. (2002).

This project selection process is identical to the other funding agencies. Apart from that there are a very large number of proposals that are necessitating to be grouped for peer review in the NSFC.

METHODOLOGY

In this section that the research proposal are preprocessed (Fig. 2). Once the classification of research proposals in discipline areas, the proposals in each discipline are clustered using the text-mining technique (Choi and Park, 2006; Girotra *et al*., 2007). The main clustering process consists of five steps, as shown in Fig. 3: text document collection, text document preprocessing, text document encoding, vector dimension reduction and text vector clustering.

Text document collection: The proposal documents in each discipline $A_k(k = 1, 2, \ldots, K)$ are collected after the research proposal is classified.

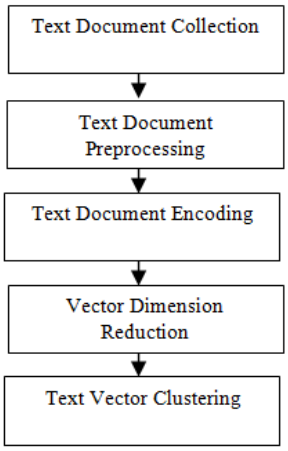

Fig. 2: Text mining process

Text document preprocessing: The contents of proposals are generally nonstructural. Because in the proposal consist of Chinese characters which are hard to segment, the research ontology is used to study, extract and identify the keywords in the full text of the proposals.

Text document encoding: After text documents are segmented, they are converted into a feature vector illustration: $V = (v_1, v_2 \ldots v_M)$, where M is the number of features selected and $v_i (i = 1, 2, \ldots, M)$ is the TFIDF encoding (Choi and Park, 2006) of the keyword w_i. 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 $v_i = tf_i * log(N/df_i)$, where N is the total number of proposals in the discipline, tf_i is the term frequency of the feature word w_i and df_i is the number of proposals containing the word w_i. Thus, research proposals can be represented by corresponding feature vectors.

Vector dimension reduction: The feature vector dimension is large so that the vector size is reduced automatically by selecting a subset which consists of more number of keywords. Latent Semantic Indexing (LSI) is used to solve the problem (Steinbach *et al*., 2000). It not only reduces the dimensions of the feature vectors also generate the semantic relations between the keywords. LSI is a technique for replacement of the original data vectors with shorter vectors in which the semantic information is conserved. Without losing the information in a proposal, a term-by-document matrix is created, where there is one column that corresponds to the term frequency of a document. Also, the matrix is decayed into a set of eigenvectors by means of singular-value decomposition. Thus, the document vector formed from the term of the enduring eigenvectors has a very small dimension and retains approximately all of the related original features.

Text vector clustering: This step uses a Efficient EM algorithm to cluster the feature vectors based on similarities of research areas.

Proposed Efficient EM algorithm (EEM): It is possible to refine the partitioning results by reallocating new cluster membership. The basic idea of the reallocation method (Rasmussen, 1992) is to start from some initial partitioning of the data set and then proceed by moving objects from one cluster to another cluster to obtain an improved partitioning. Thus, any iterative optimization-clustering algorithm can be applied to do such operation. The problem is formulated as a finite mixture model and applies a variant of the EM algorithm for learning the model.

Begin:

Initialization: Set $(Z_i)_j^{(0)}$ from a partitioning of the data and $t \leftarrow 0$.

Repeat

E-step: For each $d_i, 1 \le i \le n$ and $c_j, 1 \le j \le k$, find its new component index as:

$$(Z_i)_j^{(i+1)} = \begin{cases} 1, & \text{if } j^* = \text{argmax}_j \log\left(P^{(t)}(c_j|d_i;\theta_j)\right) \\ 0, & \text{otherwise} \end{cases}$$

M-step: Re-estimate the model parameter

$$P(c_j)^{(i+1)} = \frac{1}{n}\sum_{i=1}^{n}(Z_i)_j^{(i+1)}$$

$$m_j^{(i+1)} = \sum_{i=1}^{n} d_i (Z_i)_j^{(i+1)} / \sum_{i=1}^{n} (Z_i)_j^{(i+1)}$$

$$\sigma^{2(i+1)} = \frac{1}{n.d}\sum_{i=1}^{n}\sum_{j=1}^{k}\|d_i - m_j\|^2 (Z_i)_j^{(i+1)}$$

Until $\Delta \log L_e(\theta) < \delta$;

end

Fig. 3: Proposed Efficient EM algorithm (EEM)

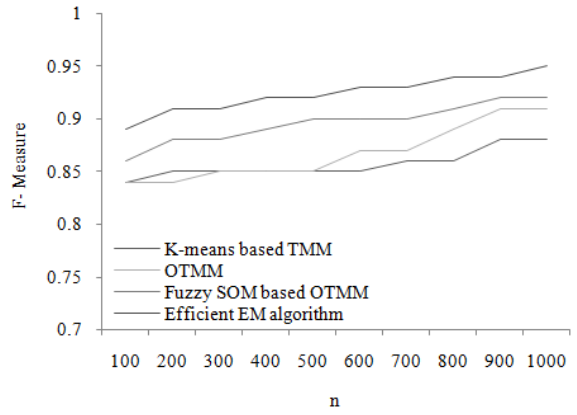

Fig. 4: Comparison of F-measurement

The most critical problem is how to estimate the model parameters. The data samples are assumed to be drawn from the multivariate normal density in Rd also assume that features are statistically independent and a component c_j generates its members from the spherical Gaussian with the same covariance matrix (Dasgupta and Schulman, 2000). Figure 4 gives an outline of a simplified version of the EM algorithm. The algorithm tries to maximize log Lc at very step and iterates until convergence. For example, the algorithm terminates when $\Delta \log Lc<\delta$, where δ is a pre defined threshold.

EXPERIMENTAL RESULTS

The demand for handling text documents in research projects has increased dramatically in recent years, owing to the rapid growth of online information. This in turn has made text clustering as one of the key techniques for handling and organizing text data. To analyze the performance of the proposed clustering algorithm, several experiments were conducted. This section explains the results obtained during performance analysis.

To validate the proposed approach, several experiments are conducted using the previous granted research projects. Research projects from various disciplines are considered for the experimental evaluations. The domains and disciplines taken into consideration are information management, artificial intelligence, image processing, data mining, networking and software engineering. One of the most important issues in clusters analysis is the evaluation of the clustering results. Evaluating clustering results is the analysis of the output to understand how well it reproduces the original structure of the data. However, the evaluation of clustering results is the most difficult task within the whole clustering workflow. To evaluate the performance of the proposed model six performance metrics, as listed below, are used:

- Precision
- Recall
- F-measure

Table 1: Comparison of accuracy

Approaches	Precision	Recall
K-means based TMM	0.612	0.851
OTMM	0.741	0.911
Fuzzy SOM based TMM	0.858	0.952
Efficient EM algorithm	0.901	0.985

Precision and recall: The equation used to calculate precision (p) and recall (r) are given in Eq. (1) and (2):

$$\text{Precision}(c,t) = n(c,t)/nc \quad (1)$$

$$\text{Recall}(c,t) = n(c,t)/nt \quad (2)$$

where,

n (c, t): The project number of the intersection between cluster c and topic t

n_c : The number of projects in cluster c

n_t : The number of projects in topic t

F-measure: The F-measure is calculated using Eq. (3). F measurement between cluster c and topic t can be calculated as follows:

$$F(c,t) = (2 * \text{Recall}(c,t) * \text{Precision}(c,t)) / (\text{Recall}(c,t) + \text{Precision}(c,t))$$

The F measurement can be given by:

$$F = \sum_i \frac{n_i}{n} \max\{F(i,j)\} \quad (3)$$

where,

n : The whole number of research projects

i : Each predefined research topic

It is always desired to obtain a large F-measure, which indicates better clustering performance. In general, a larger F-measure value indicates better clustering result.

Table 1 gives the comparison of accuracy for different methods. Table 1 compared the proposed method with Ontology based Text-Mining Methods (OTMM), K-Means Based Method, Fuzzy Based TMM. From the table it is observed that the proposed method proves better accuracy compared to other. In order to compare the clustering quality of the proposed Efficient Expectation Maximation (EEM) ontology based Text Mining Method (EEM based TMM), the existing Ontology based TMM and clustering based TMM is taken for consideration. F-Measurement has to be taken.

From the Fig. 4 it is cleared that the Proposed Efficient EM (EEM) algorithm proves their clustering quality compared to other existing approaches.

CONCLUSION

In this study, Ontology based classification and clustering approach is proposed, which will be used by research funding Agencies for grouping the Research

Proposals and the research Reviewers. Here the paper presented a structure on ontology based text mining for grouping research proposals and conveying the grouped proposal to reviewers analytically. Research ontology is designed to separate the concept tasks in various regulations areas and to form a concurrent relationship with them. It assist with text-mining and optimization techniques to cluster research proposals based on their resemblance and then to assign them to reviewer according to their research area. The proposals are assigned to reviewer with the help of knowledge based agent. From the experimental result it is clearly observed that the proposed method proves Efficient EM algorithm p obtained better approaches than Existing approach. 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

Aas, K. and L. Eikvil, 1999. Text categorisation: A survey. Technical Report, Raport NR 941, Norwegian Computing Center.

Albala, A., 1975. Stage approach for the evaluation and selection of R&D projects. IEEE T. Eng. Manage., 22: 153-164.

Arunachala, E.S., S. Hismath Begum and M. Uma Makeswari, 2013. An ontology based text mining framework for R&D project selection. Int. J. Comput. Sci. Inform. Technol., 5(1).

Baeza-Yates, R. and B. Ribeiro-Neto, 1999. Modern Information Retrieval. Addison Wesley, Wokingham, UK.

Baker, N. and J. Freeland, 1975. Recent advances in R&D benefit measurement and project selection methods. Manage. Sci., 21(10): 1164-1175.

Berners-Lee, T., 1999. Weaving the Web. Harper, San Francisco.

Choi, C. and Y. Park, 2006. R&D proposal screening system based on text mining approach. Int. J. Technol. Intell. Plan., 2(1): 61-72.

Dasgupta, S. and L.J.A. Schulman, 2000. A two-round variant of EM for gaussian mixtures. Proceeding of the 16th Conference on Uncertainty in Artificial Intelligence (UAI '00). In: Craig Boutilier and Mois\&\#233;s Goldszmidt (Eds.), Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp: 152-159.

Decker, S., S. Melnik, F. Van Harmelen, D. Fensel, M. Klein, J. Broekstra, M. Erdmann and I. Horrocks, 2000. The semantic web: The roles of XML and RDF. IEEE Internet Comput., 4(5): 63-74.

Ding, Y. and S. Foo, 2002. Ontology research and development: Part 1-a review of ontology generation. J. Inform. Sci., 28(2).

Dumais, S.T., 1991. Improving the retrieval of information from external sources. Behav. Res. Meth. Ins. C., 23(2): 229-236.

Fabrizio, S., 2002. Machine learning in automated text categorization. ACM Comput. Surv., 34(1): 1-47.

Fahrni, P. and M. Spatig, 1990. An application oriented guide to R&D selection and evaluation methods. R&D Manage., 20(2): 155-171.

Foskett, D.J., 1997. Thesaurus. In: Willett, P. and K. Sparck-Jones (Eds.), Reproduced in Readings in Information Retrieval. Morgan Kaufmann, San Francisco, CA, pp: 111-134.

Ghasemzadeh, F. and N.P. Archer, 2000. Project portfolio selection through decision support. Decis. Support Syst., 29(2000): 73-88.

Girotra, K., C. Terwiesch and K.T. Ulrich, 2007. Valuing R&D projects in a portfolio: Evidence from the pharmaceutical industry. Manage. Sci., 53(9): 1452-1466.

Henriksen, A.D. and A.J. Traynor, 1999. A practical R&D project-selection scoring tool. IEEE T. Eng. Manage., 46(2): 158-170.

Hotho, A., S. Staab and A. Maedche, 2001. Ontology-based text clustering. Proceeding of the UCAI-2001 Workshop on Text Learning: Beyond Supervision, Seattle.

Joachims, T., 1997. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization. Proceeding of the 14th International Conference on Machine Learning (ICML '97), pp: 143-151.

Lewis, D.D., 1992. Feature selection and feature extraction for text categorization. Proceeding of the Workshop on Speech and Natural Language, pp: 212-217.

Li, X. and B. Liu, 2003. Learning to classify texts using positive and unlabeled data. Proceeding of the International Joint Conference on Artificial Intelligence (IJCAI, 03), pp: 587-594.

Li, Y., C. Zhang and J.R. Swan, 2000. An information filtering model on the web and its application in jobagent. Knowl-based Syst., 13(5): 285-296.

Liberatore, M. and G. Titus, 1983. The practice of management science in R&D project selection. Manage. Sci., 29(8): 962-974.

Lockett, G., B. Hetherington and P. Yallup, 1986. Modeling a research portfolio using AHP: A group decision process. R&D Manage., 16(2): 151-160.

Martino, J.P., 1995. R&D Project Selection. Wiley, New York.

Rasmussen, E., 1992. Clustering Algorithms. In: Frakes, W. and R. Baeza-Yates (Eds.), Information Retrieval: Data Structures and Algorithms. Prentice Hall, Englewood Cliffs, USA.

Robertson, S. and I. Soboroff, 2002. The Trec 2002 filtering track report. Proceeding of the 11th Text Retrieval Conference (TREC, 2002). Retrieved from: trec.nist.gov/pubs/ trec11/papers/ OVER.FILTERING.ps.gz.

Salton, G., 1989. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer. Addison-Wesley, NY.

Salton, G. and C. Buckley, 1988. Term-weighting approaches in automatic text retrieval. Inform. Process. Manag. Int. J., 24(5): 513-523.

Schmidt, R.L. and J.R. Freeland, 1992. Recent progress in modeling R&D project-selection processes. IEEE T. Eng. Manage., 39(2): 189-201.

Shawkat Ali, A.B.W., 2008. K-means Clustering Adopting RBF-Kernel, Data Mining and Knowledge Discovery Technologies. In: David, T. (Ed.), IGI Pub., Hershey, pp: 118-142.

Staab, S. and R. Studer, 2004. Handbook on Ontologies. Springer, NY.

Steinbach, M., G. Karypis and V. Kumar, 2000. A comparison of document clustering techniques. Proceeding of the KDD Workshop on Text Mining'00.

Tar, H.H. and T.T.S. Nyunt, 2011. Ontology-based concept weighting for text documents. Proceeding of the International Conference on Information Communication and Management (IPCSIT, 2011). IACSIT Press, Singapore, Vol. 16.

Tian, Q., J. Ma and O. Liu, 2002. A hybrid knowledge and model system for R&D project selection. Expert Syst. Appl., 23(3): 265-271.