IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 2, Ver. II (Mar-Apr. 2016), PP 31-36
www.iosrjournals.org

Automatic Ontology Creation for Research Paper Classification

Ankita

Abstract: *As a large number of research proposals are received at different journal or research institute, it is common to group them according to their similarities in research disciplines and the grouped proposals are then assigned to the appropriate experts for peer review. Grouping in Current scenario is done by manual matching of similar research discipline areas and/or keyword. This work implement Text-mining methods to solve the problem by automatically classifying text documents have pattern based ontology text-mining approach where one put paper and year of submission, then it make pattern, then cluster research proposals based on their similarities in research areas. It can be efficient and effective for clustering research proposals with English texts as most of research paper are in English language.*

IndexTerm: *Information Extraction, Text Analysis, Ontology, feature extraction, text categorization, clustering*

I. Introduction

For many research funding agencies, international journals, national journals, such as either government or private agencies, the selection of research project proposals is an important and challenging task, when large numbers of research proposals are collected by the organization. The Research Project Proposals Selection Process starts with the call for proposals, then from different research scholars, scientist, etc. from many institutes and organizations submit there research proposals. As there is single point of contact for researchers from different area so, group the proposals based on their similarity and assigned them to the experts for peer-review. The review results are examined and proposals are ranked based on their aggregation of experts result. So the simple steps of the Research Project Selection Process, these processes are very similar in all research funding agencies.[2] For very large number of proposals received by the agencies need to be group the proposals for peer review. The department for selection process can assign the grouped proposals to the external reviewers for evaluation and rank them based on their aggregation. As they may not have adequate knowledge in all research discipline areas and the contents of many proposals were not fully understood when the proposals were grouped, there may be short of time for doing this so doing evaluation for whole in detail manually is tough. In current Methods, keywords are not representing the complete information about the content of the proposals and they are just the partial representation of the proposals. Hence, it’s not sufficient to group the proposals on the basis of keywords. In Manual based grouping, sometimes the department responsible for grouping may not have adequate knowledge regarding all the issues and areas of the research proposals. Therefore, an efficient and effective method is required to group the proposals efficiently based on their discipline areas by analyzing full text information of the proposals. So an ontology is construct for text-mining that will effectively used for this purpose.

II. Related Work

Ontology patterns were introduced by Blomqvist and Sandkuhl in 2005 [4]. Later the same year, Gangemi presented his work on ontology design patterns [5]. Such patterns, encodings of best practices, were intended to reduce the need of extensive experience when developing ontologies. Rainer Malik et. al. have used a combination of algorithms of text mining to extract keywords relevant for their study from various databases and also identified relationships between key terminologies using PreBIND and BIND system (Donaldson et al., 2003; Bader et al., 2003). Boosting classifier was used for performing supervised learning and used on the test data set. 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.

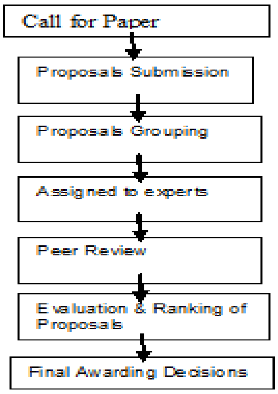

Fig. 1. Paper selection procedure

Cook *et al.* [11] presented a method of optimal allocation of proposals to reviewers in order to facilitate the selection process. Arya and Mittendorf [12] proposed a rotation program method for project assignment. Choi and Park [13] used text-mining approach for R&D proposal screening. Girotra *et al.* [14] offered an empirical study to value projects in a portfolio. Sun *et al.* [5] developed a decision support system to evaluate reviewers for research project selection. Finally, Sun *et al.* [6] 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. For example, Hettich and Pazzani proposed a text-mining approach to group proposals, identify reviewers, and assign reviewers to proposals. Current methods group proposals according to keywords. Unfortunately, proposals with similar research areas might be placed in wrong groups due to the following reasons: first, keywords are incomplete information about the full content of the proposals. Second, keywords are provided by applicants who may have subjective views and misconceptions, and keywords are only a partial representation of the research proposals. Third, manual grouping is usually conducted by division managers or program directors in funding agencies.

III. Background

This paper using the concept of ontology with Text Mining techniques such as Classification and Clustering algorithms. The proposed approach builds the research ontology and then applies Decision Tree Algorithm to classify the data into the disciplines using research ontology and then the resultant of classification is used to make clusters of similar data.

3.1 Ontology

Ontologies have several technical advantages over other types of data models or knowledge representation languages - they are flexible and easily accommodate heterogeneous data, they are platform and programming-language independent, and being based on description logics they can easily be computed on by classifier software, allowing for the inferencing of new knowledge based on that which is already known. This computability capability can also help ensure the consistency and quality of information encoded in ontology languages.Uses of ontologies in information logistics range from competence modeling [1] to requirements management [2] to general knowledge fusion architectures [3]. 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.[5] An ontology in the paper is create by supplying the Research project/paper year wise as project/paper are containing the keywords which are representation of the overall research project/paper. Then creating list of the keywords from that specific area is ontology of the area. Here creating list of the words area wise is necessary as on that behave we will train the network for number of words appear in the paper for finding the correct area.

3.2 Classification

In Classification, the input text data can be classified into number of classes based on that data. Various Text-Mining techniques are used for classification of text data such as Support Vector Machine, Bayesian, Decision Tree, Neural Network, Latent Semantic Analysis, Genetic Algorithm, etc.

3.3 Clustering

A cluster is comprised of a number of similar objects collected or grouped together. Everitt documents some of the following definitions of a cluster (Everitt, 1974):

1. A cluster is a set of entities which are alike, and entities from different clusters are not alike.
2. A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it.
3. Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points.

Making sense of data is an ongoing task of researchers and professionals in almost every practical endeavor (Pedrycz, 2005). The age of information technology, characterized by a vast array of data, has enormously amplified this quest and made it even more challenging. Data collection anytime and everywhere has become the reality of our lives. Understanding the data, revealing underlying phenomena, and visualizing major tendencies are major undertakings pursued in intelligent data analysis, data mining, and system modeling. Clustering is a technique used to make group of the documents having similar features. Documents within a cluster have similar objects and dissimilar objects as compared to any other cluster. Clustering algorithms creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. This technology can be useful in the organization of management information systems, which may contain thousands of documents. Several Text Mining Algorithms used for clustering are K-Means, Self-Organizing Maps (SOM), EM, etc.

IV. The Proposed Approach

In this paper research project/paper are clustered into specific area using ontology of the different areas. So following are the modules of approach. From raw paper collections to classified as per area.

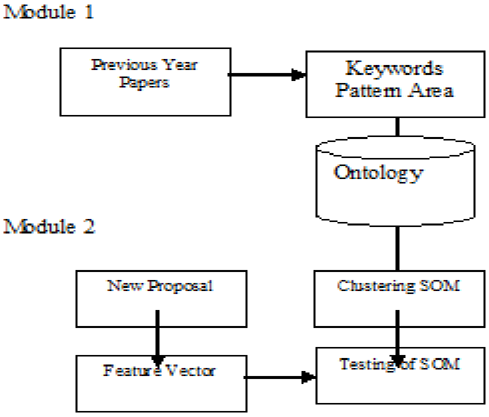

Fig.2 Different module of the proposed work

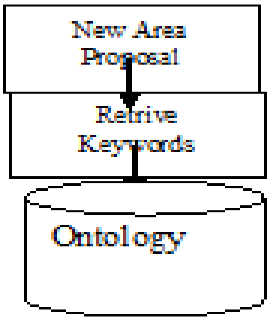

Fig.2 Different module of the ontology updation work

Module 1:

In order to create ontology previous year research papers are selected which may be of different. Here from each paper keywords are fetch which is mention in the keyword or index term portion of the paper, one more thing is store the words in form of the pattern. This can be understand by an example ‘data’, ‘mining’ are two keywords but “Data Mining” is a patternas they are known to the area which they belong, so information for clustering is create in this way. The research topics of different disciplines can be clearly expressed by a research ontology. Suppose that there are K discipline areas, and Ak denotes discipline area $k(k = 1, 2, \dots, K)$.

Here feature vector of the different paper is create which is the collection of one identification number, then keyword pattern. It look like a vector {Id, Area, Pattern....Pattern-n}. For this step only research paper is submit with small detail like year of submission, area. After that it automatically search keyword in the project/keyword and add that keyword in the corresponding area if exist or simply add new area if not exist.

Finally all the keywords pattern with there area is save in a depository for further analysis. One can easily update the ontology as new proposals if required the method of updation is same as passing new paper in the proposal then it automatically learn new keywords for the area or even it will learn new area for the mention keywords.

Module2.

In this module Clustering of new Research Proposals done Based on Similarities of the created ontology with the existing paper. So following are the generic strategy for text classification is the main steps involved are

i) document preprocessing
ii) feature extraction / selection
iii) model selection
iv) training and testing the classifier.

Pre-Processing: Data preprocessing reduces the size of the input text documents significantly. It involves activities like sentence boundary determination, natural language specific stop word elimination and stemming. Stop-words are functional words which occur frequently in the language of the text (for example, „a‟, ‟the‟, ‟an‟, ‟of‟ etc. in English language), so that they are not useful for classification. Here we read whole project and put all words in the vector. Now again read the file which contain stop words then remove similar words from the vector. Once the data is pre-process then it will be the collection of the words that may be in the ontology list. For example let one paper of the image class is taken and its text vector is Rough_text = {a1, f1, s1, a2, s2, a3, a4, f2…………..an} and let the stop words collection is stop_text = {a1,a2,a3,………….am}. Then the vector obtain after the Pre-Processing is processed_text = {f1, s1, s2, f2,……….fx}.
[processed_text] = [Rough_text] – [stop_text]
After getting the processed_text vector then proper feature selection should be done from the vector which contain large number of texts.

Feature Extraction: The vector which contain the pre-processed data is use for collecting feature of that document. This is done by comparing the vector with vector KEY (collection of keywords) of the ontology of different area. So the refined vector will act as the feature vector for that research project/proposal. To understand this let us take an example of the vector obtain after the pre-processing is processed_text = {f1, s1, s2, f2,……….fx}. Now let features are KEY = f1, f2,….fx then comparing those from those from the ontology we will find a feature vector of the inserted proposal/paper which will act as the testing_feature vetor.
[testing_feature] = [processed_text] ∩ [KEY]
In this way testing_feature vector is created from the inserted testing proposal.

Assign Proper Area as per SOM

Now the way by which that paper is categorize into the research area is the clustering of the Proposal/paper this is done by many approach, this paper use neural network approach by Self Organizing mapping (SOM). Here the created ontology is use for training the SOM neural network (Self Organized Mapping). Here the feature vector is pass into the network in form of vector of the keywords frequency. Here we pass the created ontology feature vector that will train the neurons as per the different research area.

Testing: Here the created research projects feature vectors are transfer in form of input as the testing data to the SOM network for training and then this trained network is test with different proposal/paper feature vector so one can obtain the belonging class of the proposal/paper. This can be understand as fix length vectors of the different class is transfer to the SOM network of same number of output as in the input vector class. So it will generate output to the corresponding class whose vector is more closer to the new proposal feature vector of the same size as the size of the input vector.

I. Proposed Algorithm

Ontology Creating Algorithm

Input: Dataset D[n], Area, Year where n is number of paper/project
Output: Updated Ontology Data OD. For each area and year repeat this algorithm.

```
Loop I = 1: n
    K ← Read_keyword(D[I])
    P[t] ← Pattern(K)
    Loop J = 1: m
        Loop k = 1:t
            If NotEqual( O[J] , P[t] )
                O[m+1] = P[t]
            Endif
        EndLoop
    EndLoop
EndLoop
OD ←{Area, Year, O}
```

Testing Algorithm

Input: Research Paper R, Ontology Dataset OD

Output: Research Paper Area RA

```
R ← Pre-processing(R)
F ← Frequent_keywords(R)
F ← Feature_extraction(F, OD)
NN ← SOM (OD)  // NN is neural network
RA ← NN(F)
```

V. Experiment And Result

Data Set : Inorder to implement this work research paper are collect of different field. This include 100 research paperin word format as the file need to read and find relative word from it in a pattern. Then for initial ontology creation one has to divide the dataset into training part and other one for testing part.

Evaluation Algorithm: Here ontology base text mining algorithm has been developed on the bases of [13]. Here they develop similar approach but without pattern of the keywords, they just use keywords for there ontology.

Evaluation Parameter: To test outcomes of the work following are the evaluation parameter such as accuracy of the text mining approach. Then to find Precision, Recall and F-score.

Precision = TP / (TP+ FP)

Recall = TP / (TP + TN)

F-score = 2 * Precision * Recall / (Precision + Recall)

Where TP : True Positive
TN : Treue Negative
FP: False Positive

Results: As the dataset contain four different field paper so the values obtain from all the field of different evaluation parameter are averaged. Average is taken because as the paper of some field are common and can be easily detected.

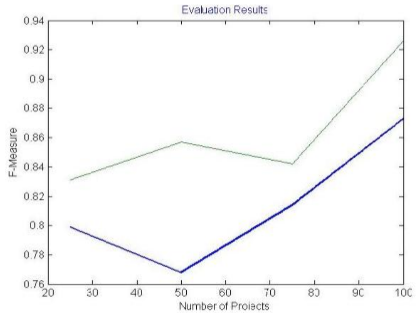

Fig. 5 graph of the F-Score for different number of papers Broad for Term base and narrow for Pattern base

Average of 4 Area	Term Base	Pattern Base
Precision	0.857	0.893
Recall	0.705	0.84
F-Measure	83.85	86.04

Table 1. Results of the different measure for average of 4 different area.

From above table 1 it has been find that proposed pattern base approach for document of research field procedure work in a refine manner and can do the separation of the paper in the respected area in accurate manner. It has been observe in the graph as well that as the training data increases then the number of f-measure score is also increases, because the pattern number increases.

VI. Conclusions

Exploiting knowledge present in textual documents is an important issue in building systems for knowledge management and related tasks. In this paper pattern base, Ontology is created for research paper classification and clustering as per the type of matter is in the paper. This approach is very user friendly and less time consuming as time at which one submit the paper can be categorize and result displayed. This proposed method work well for different research paper categorization which has seen by f-measure value of 0.86. With the combination of both text mining and neural network approach new bridge of learning is develop for paper classification. This same approach can be use for story, article, topic, classification without any manual interference.

References

[1]. E. M. Voorhees. Implementing Agglomerative Hierarchical Clustering for use in Information Retrieval,Technical Report TR86–765, Cornell University, Ithaca, NY, July 1986.

[2]. Young, L., Tu, S.W., Tennakoon, L., Vismer, D., Astakhov, V., Gupta, A., Grethe, J.S.,Martone, M.E., Das, A.K., McAuliffe, M.J.: Ontology Driven Data Integration for Autism Research. 22nd IEEE International Symposium on Computer Based Medical Systems, pp. 1–7, Albuquerque, NM (2009)

[3]. List of English stop words, http://members.unine.ch/jacques.savoy/clef

[4]. Joho, H., Sanderson, M., Retrieving Descriptive Phrases from Large Amounts of Free Text. 9th ACM Conference on Information and Knowledge Management, pp. 180--186, McLean, VA (2000)

[5]. M. W. Berry, Survey of Text Mining: Clustering, Classi_cation, and Retrieval, New York: Springer, 2003, pp. 1-122.

[6]. M. Nagy and M. Vargas-Vera, “Multiagent ontology mapping ramework for the semantic web,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 693–704, Jul. 2011.

[7]. G. H. Lim, I. H. Suh, and H. Suh, “Ontology-based unified robot knowledge for service robots in indoor environments,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 3, pp. 492–509, May 2011.

[8]. Y. Liu, X. Wang, and C. Wu, “ConSOM: A conceptional self-organizing map model for text clustering,” Neurocomputing, vol. 71, no. 4–6, pp. 857–862, Jan. 2008.

[9]. 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.

[10]. W. D. Cook, B. Golany, M. Kress, M. Penn, and T. Raviv, “Optimal allocation of proposals to reviewers to facilitate effective ranking,” Manage. Sci., vol. 51, no. 4, pp. 655–661, Apr. 2005.

[11]. F. Ghasemzadeh and N. P. Archer, “Project portfolio selection through decision support,” Decis. Support Syst., vol. 29, no. 1, pp. 73–88, Jul. 2000.

[12]. H. C. Yang, C. H. Lee, and D. W. Chen, “A method for multilingual text mining and retrieval using growing hierarchical self-organizing maps,” J. Inf. Sci., vol. 35, no. 1, pp. 3–23, Feb. 2009.

[13]. Q. Liang, X. Wu, E. K. Park, T. M. Khoshgoftaar, and C. H. Chi, “Ontology-based business process customization for composite web services,” IEEE Trans. Syst., Man, Cybern. A, Syst., Humans, vol. 41, no. 4, pp. 717–729, Jul. 2011.