IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 2, No. 4
November 2013

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 2, No. 4, November 2013
© 2013 IJERST. All Rights Reserved

Research Paper

AUTOMATIC ONTOLOGY CREATION FOR RESEARCH PAPER CLASSIFICATION

Jay Prakash Pandey[1]*, Shrikant Lade[1], Manish Kumar Suman[1]

*Corresponding Author: **Jay Prakash Pandey,** ✉ jp.pandey@gmail.com

As a large number of research, educational, institutes are opened day-by-day, so research project selection is an important task for different government and private research funding agencies, Journals, etc. As a large number of research proposals are received, 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. As one person not have the whole knowledge of the different research paper, so rich information in the proposals' full text can be used effectively. By Implementing Text-mining methods to solve the problem by automatically classifying text documents, mainly in English. This paper presents a complete automatic ontology-based text-mining approach where one put paper and year of submission, then it automatically cluster research proposals based on their similarities in research areas. The method is based on use of keywords for creating ontology, then for similarities whole paper is scan based on similarities with the ontology that paper can be classify. It can be efficient and effective for clustering research proposals with English texts as most of research paper are in English language.

Keywords: Information Extraction, Text Analysis, Ontology, Feature extraction, Text categorization, Clustering

INTRODUCTION

The problem of clustering has been studied widely in the database and statistics literature in the context of a wide variety of data mining tasks (Everitt B S, 1978). The clustering problem is defined to be that of finding groups of similar objects in the data. The similarity between the mining text data objects is measured with the use of a similarity function. The problem of clustering can be very useful in the text domain, where the objects to be clusters can be of different granularities such as documents, paragraphs, sentences or terms.

For many research funding agencies, international journals, national journals, such as either government or private agencies, the

[1] Information and Technology Branch of RKDF Institute of Science and Technology Bhopal.

selection of research project proposals is an important and challenging task, when large numbers of research proposals are collected by the organization (Figure 1). 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 (Young *et al.,* 2009). 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.

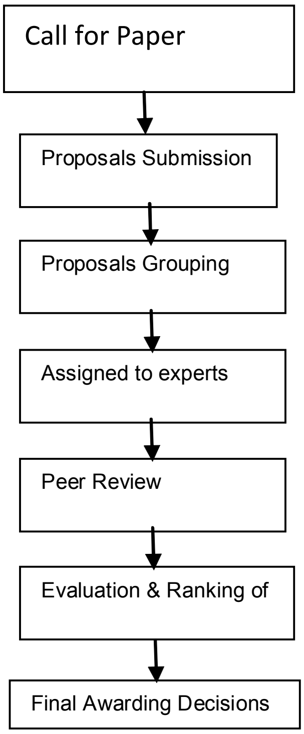
RELATED WORK

Ontology patterns were introduced by Blomqvist and Sandkuhl in 2005. Later the same year, Gangemi (2005) presented his work on ontology design patterns. Such patterns, encodings of best practices, were intended to reduce the need of extensive experience when developing ontologies.

Rainer Malik *et al.,* (2006). 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. Boosting classifier was used for performing supervised learning and used on the test data set. Henriksen and Traynor (1999) presented a scoring tool for project evaluation and selection. Ghasemzadeh and Archer (2000) offered a decision support approach to project portfolio selection. Machacha and Bhattacharya, proposed a fuzzy logic approach to project selection. Butler *et al.(1997)* used a multiple attribute utility theory for project ranking and selection. Loch and Kavadias, established a dynamic programming model for project selection, while Meade and Presley, developed an analytic network process model. Greiner *et al.* proposed a hybrid AHP and integer programming approach to support project selection.

Methods have been developed to group proposals for peer review tasks. For example, Hettich and Pazzani proposed a text-mining

Figure 1: Paper Selection Procedure



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.

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.

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 (Everitt B S, 1978) to requirements management (Pedrycz, 2005) to general knowledge fusion architectures (L Kaufman and P J Rousseeuw, 1990).

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 (Gangemi A, 2005). 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 a 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.

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.

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 B S, 1978):

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.

THE PROPOSED APPROACH

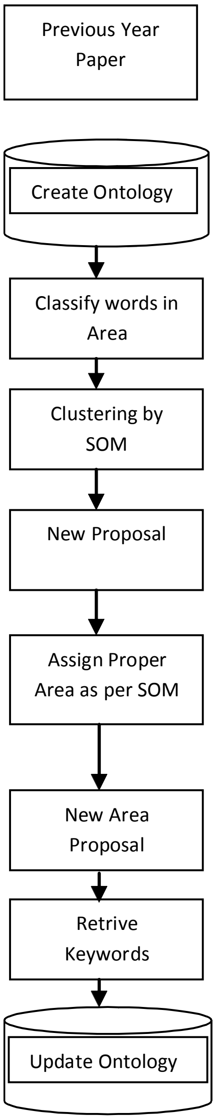
In this paper research project/paper are clustered into specific area using ontology of the different areas. So following are the modules of approach. From raw paper collections to classified as per area (Figure 2).

Module 1

Previous year papers are select for creating the ontology which may be of different field as per the required paper. Here from each paper keywords are fetch as 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 $k(k = 1, 2, . . . , K)$. A research ontology can be constructed in the following steps to represent the topics of the disciplines.

Here feature vector of the different vector is create in this way which is the collection of one

Figure 2: Different Module of the Proposed Work



identification number, then keyword, then frequency of keyword that how many time that keyword appear in the different project/proposal submit. It look like a vector {Id, Area, Keyword1, frequency1, ... keyword2, frequency2,... keyword-n, frequency-n}. One may use either TFIDF in place of the number of time it appear. 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. Once the ontology is create then it required that to prune the keywords for the sufficient time of occurrence as the upper limit and lower limit of the frequency in the feature vector. As there may be a chance that some keywords are very few in number then existing, this can be omit as they may misguide the clustering process.

Finally all the keywords 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.

Module 2

Here a more general name of the area is assign as Digital image watermarking, Object tracking, image recongnization, etc., are some of the research area where one can work now the more general word for all above research area is the Image Processing as all above topic come under that broad field. So assigning such kind of field name is done in second phase where user select the areas.

In this module areas are still separate but the keywords of different area are come under same area with same field name and different area name.

Module 3

In this module Clustering of new Research Proposals is 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 {a1, f1, s1, a2, s2, a3, a4, f2…………..an} and let the stop words collection is {a1,a2,a3,………….am}. Then the vector obtain after the Pre-Processing is {f1, s1, s2, f2,……….fx}.

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 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 {f1, s1, s2, f2,……….fx}. Now let features are f1, f2,….fx then comparing those from those from the ontology we will find a feature vector of the inserted proposal/paper.

Model Selection: 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).

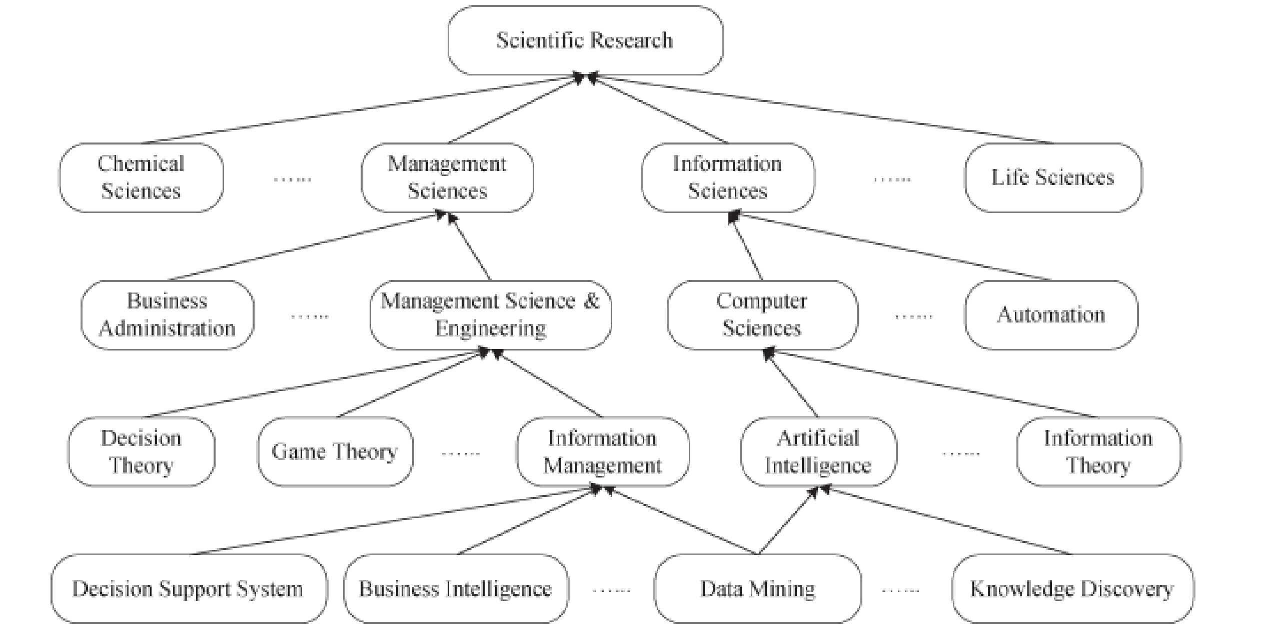
Here the created ontology is used for training the SOM neural network (Self Organized Mapping). Here the feature vector is pass into the network in form of vector of the keywords frequency. Here we pass the created ontology and feature vector both so it will train then specify the corresponding research area.

Training and Testing: Here the created research projects feature vectors are transfer in form of input as the training 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. At the end research project is classified in one of the area as shown in Figure 3.

EXPERIMENT AND RESULT

To implement this paper start with first module-1 and for this creating ontology will be done by 100 proposal of four fields are used then these are selected one by one with their field name and year. It will automatically read the keywords and store them in a separate file for further modules. It is a simple Bag of Words (BOW) with specific area, year. Example {Image Processing, 2012, watermarking, DCT, 3, DFT, 5……..,etc.}

Figure 3: Formal Example of Paper Classification



Then second Module is simple and it just group different area in a specific field. These grouping is done by checking the keywords of the different area which have mostly common words. Here this module is just a generalization of the areas. Like keywords of some area may be common which is totally depend on the type of paper we put in it. For this some threshold is set so for areas whose keywords ratio that cross that threshold is consider as a similar field. Example let two area keywords are {f1, f2, f3,……….fn} and {s1, s2, s3,…….sn} and if common ration is above 0.6 then it is consider as similar field area.

For Third module it will preprocess the testing proposal by removing the stop words and then extract features from the vector, once the feature vector is created of the new proposals. Then pass the vector to the trained network, which is trained by the ontology of the different area. This passing of feature vector to the trained network will specify the type of class it belong, its like winner take all.

To test our result we use following measures the accuracy of the text mining approach, that is to say Precision, Recall and F-score.

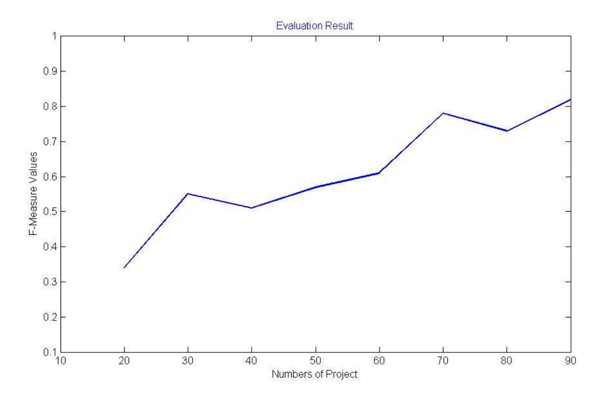
$$\text{Precision} = \frac{\text{true positives}}{(\text{true positives} + \text{false positives})}$$

$$\text{Recall} = \frac{\text{true positives}}{(\text{true positives} + \text{false negatives})}$$

$$\text{F-score} = \frac{2 * \text{Precision} * \text{Recall}}{(\text{Precision} + \text{Recall})}$$

From the Table 1 and Figure 5, it was found that proposed completely automatic 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 testing data increases then the number of f-measure score is also increases, because

Figure 5: Graph of the F-Score for Different Number of Papers



the accuracy is high so the number of test sample increase then value of f-measure raises.

Table 1: Results of the Different Measure for Different Area

Measuring Parameter	Area			
	A1	A2	A3	A4
Precision	0.82	0.94	0.79	0.88
Recall	0.78	0.65	0.84	0.91
F-Score	0.799	0.768	0.814	0.873

CONCLUSION

Exploiting knowledge present in textual documents is an important issue in building systems for knowledge management and related tasks. In this paper Automatic, 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 0.91. With the combination of both text mining and neural network approach new bridge of learning is developed for paper classification. This same approach can be used for story, article, topic, classification without any manual interference.

REFERENCES

1. Everitt B S (1978), "Graphical Techniques for Multivariate Data", *Elsevier North-Holland Inc.*, New York, USA.
2. Pedrycz W (2005), "Knowledge-Based Clustering", *From Data to Information Granules*, John Wiley & Sons Inc., Hoboken, New Jersey.
3. Kaufman L and Rousseeuw P J (1990), "Finding Groups in Data: An Introduction to Cluster Analysis", *Wiley Interscience*.
4. Blomqvist E and Sandkuhl K (2005), "Patterns in ontology engineering: Classication of ontology patterns", In *Proceedings of the 7th International Conference on Enterprise Information Systems*.
5. Gangemi A (2005), "Ontology design patterns for semantic web content", In *The Semantic Web* {ISWC 2005, *Springer.*
6. Rainer Malik, Lude Franke and Arno Siebes (2006), "Combination of text-mining algorithms increases the Performance", *Bioinformatics.*
7. Henriksen A D and Traynor A J (1999), "A practical R&D project-selection scoring tool," *IEEE Trans. Eng. Manag.*, Vol. 46, No. 2, pp. 158-170.
8. Ghasemzadeh and Archer (2000), "Project Portfolio selection through decision support", *Decision Support Systems*, pp. 73-88.
9. Butler J C, Jia J and Dyer J S (1997), "Simulation techniques for the sensitivity analysis of multi-criteria decision models", *Eur. J. Oper. Res.*, pp. 531-545.

IJERST

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com

ISSN 2319-5991
04
9 772319 599001