ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 12, December 2015

A Methodology for the Implementation of Research Proposal Selection

Yathiraj GR[#1], KC Thangamma[#2], Bharath BK[*3], Bopanna KN[#4]

[#1]Assistant Professor, Dept. of Computer Science, Coorg Institute of Technology, Ponnampet, Kodagu, Karnataka, India

[*2]Assistant Professor, Computer Science Department, Coorg Institute of Technology, Ponnampet, Kodagu, Karnataka, India

[*3]Assistant Professor, Computer Science Department, Coorg Institute of Technology, Ponnampet, Kodagu, Karnataka, India

[#4]Assistant Professor, Electronics and Communication Department, Coorg Institute of Technology, Ponnampet, Kodagu, Karnataka, India

ABSTRACT : As a large number of research centers, educational, institutes are opened day by day, the research project selection has became an important tasks for different government and private research funding agencies. As a large number of research papers are received, next step is to group them according to their similarities in the research disciplines. By implementing the text mining approach, classification of project proposals can be done automatically. And ranking of proposals can be done based on the value of feature vector which gives the effective proposals in sorted fashion. The outcome will conclude us the effective way of selecting the best research proposals among them.

KEYWORDS : Ontology based text mining,, Classification, Clustering

I. Introduction

In computer science, ontology can be said as set of concepts that is knowledge within domain and relation between the pairs of concepts. Ontology is used in various domains as a form of knowledge representation about the world. In this project, ontology is a model for describing the world that gives the mapping between the properties and relationship types which gives the close relationship between the ontology and real world.

Research project selection is an important task in government as well as private funding agencies. It is a challenge for multi process task that begins with a call for proposals by the funding agencies. Earlier it was a manual method for classifying but this method has extended from manual to automatically to be done based on feature vector value. After submission of the proposals, need to apply preprocessing step like data cleaning to remove all the stop words from proposals.

The web technology has defined many stop words. By applying the preprocessing step like data cleaning, can remove all the stop words from all the submitted proposals. Obtained clean data words can be considered as tokens, assigned with unique id to each tokens. Later calculate the number of times the token has been repeated, which gives the frequency of tokens. Now apply the frequency tokenized algorithm for calculating the inverse documents frequency text which gives number of documents or proposals. By multiplying the frequency of text to the obtained, idft value, will get the feature vector value. Finally the proposal, whose feature vector value is highest that will appear at the top of browser and then in descending order based on is value. Following figure 1 shows the system architecture of this paper.

DOI: 10.15680/IJIRCCE.2015. 0312043

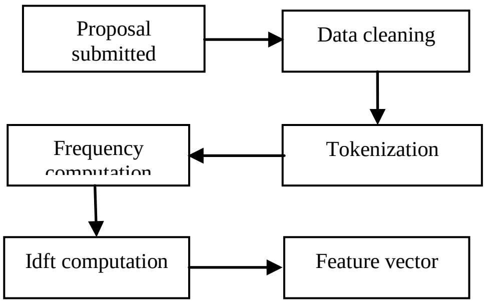


Fig 1: System Architecture

II. LITERATURE SURVEY

Ontology patterns were introduced by Blomqvist and Sandkuhl in 2005. Later the same year, Gangemi (2005) presented his work on ontology design patterns. Rainer Malik et al., (2006). have used a combination of algorithms of text mining to extract the keywords relevant for their study from various databases.

The selection of research proposals in existing system is done manually. Here proposals needs to submit to funding agency and according to name and keywords used in proposals are classified into groups. These all things has done by manually means by human beings.

But this is not suitable for large data because it might make misplacements of proposals to wrong groups due to manual process. Misplacement of proposals can be happen for following reasons. First, keywords might give an incomplete meaning about whole proposals. Second, keywords which are provided by applicants may have misconception and also we can say keywords will give only partial representation of proposals. Third, manual grouping which is done by area expert.

III. BACKGROUND

This project uses the concept of ontology based text mining approach such as classification and clustering algorithms. The proposed system builds the research ontology and applies the decision tree algorithm to classify the data into the disciplines using created ontology and then the resultant of classification is helps to make clusters of similar data.

A. Ontology

Ontology has several technical advantages like flexibility and easily accommodates heterogeneous data. Nowadays, ontology has become a prominent in the research field especially in computer science. Ontology is knowledge repository which defines the terms and concepts. And also represent the relationship between various concepts. It’s a tree like structure defined by author Gangemi A, 2005. Ontology in this paper is created by submitting proposals which containing the keywords, which are the representation of overall project. Creating a list of keywords from specific area itself is an area of ontology. By creating this it will be easy to classify the proposals into their respective area by checking number of times words have been appeared in paper.

B. Classification

Based on the data, input text data can be classified into number of classes in classification. 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.

C. Clustering

Number of similar objects collected and grouped together is called a cluster. Following are few of the definitions of the cluster.

1. A cluster is a set of entities which are like and entities from different clusters are not alike.

2. A cluster is an aggregation of points in the test space such that distance between any two points in the cluster is less than the distance between any point in cluster.
3. Clusters are connected regions of multi dimensional space containing high density of points separated by low density of points.
4. Clustering means grouping of similar types of objects into one cluster

Clustering is a technique used to make group of documents having similar features. Documents within cluster have similar objects and dissimilar objects as compared to any other cluster. Clustering algorithms creates a vector of topics for document and measures the weights of how well the documents file into each cluster.

IV. PROPOSED SYSTEM

The proposed system is based on the ontology based text mining. It includes four phases. Ontology based text mining cluster the research proposals according to their domain. Unstructured text is processed and extracted interesting information and knowledge by applying text mining.

1. *Construction of ontology:* The project which is funded in last five years are used to construct the ontology according to keywords and it gets updated annually. Research ontology expressed the topics of different disciplines more clearly to understand.

2. *Classification of new proposal:* According to the keyword of paper which is match with started keywords of specific research domain, using this proposals are classified

3. *Clustering of research proposals according to similarities:* text mining technique is used to cluster the proposals in each discipline once the classification is done according to domain. Five steps performed to cluster the proposals. Following figure 2 shows how exactly clustering process will takes place.

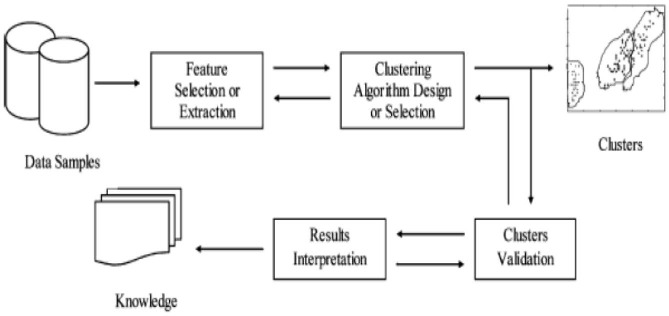

Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Knowledge
Results Interpretation
Clusters Validation

Fig 2: Process Of Clustering

4. *Balancing of research proposals and regrouping:* if the each cluster contains more than 20 proposals in its single group then that can be divided into smaller groups. This is called balancing of proposals and regrouping.

5. *Ranking the proposals:* the number of times tokens has been appeared in document will gives the frequency of text. Inverse document frequency text can be a ratio of number of documents to the frequency. So finally we can get the feature vector value can be calculated

V. METHODOLOGY

In this paper research projects are clustered into specific area using ontology of different areas. So following are the modules of approach from new paper collections to classify as per area.

Module 1: In the first module users have to submit proposals. In this project, five proposals at a time can be submitted. Proposals along with their abstract will be sent and those will be stored on ontology.

Module 2: By applying preprocessing step like data cleaning, we can remove all the stop words from proposals. Then obtained data will be cleaned data.

Module 3: The cleaned data words are called tokens which assigns with unique id. Number of times token has been repeated is called frequency computation.

Module 4: The ratio of number of documents to the frequency of text is called inverse document frequency text. And the product of inverse document frequency text and frequency of text will give the feature vector of proposals. This gives feature vector value based on this value, papers can be ranked in sorted fashion.

Classification is the main steps involved are

1. Document preprocessing
2. Feature extraction/selection
3. Model selection
4. Training and testing the classifier

Pre-processing: Data pre-processing reduces the size of the input text documents significantly. It involves the boundary like sentence boundary, natural language stop words elimination and stemming. Stop words are functional words which occur frequently in language of text so that they are not useful for classification.

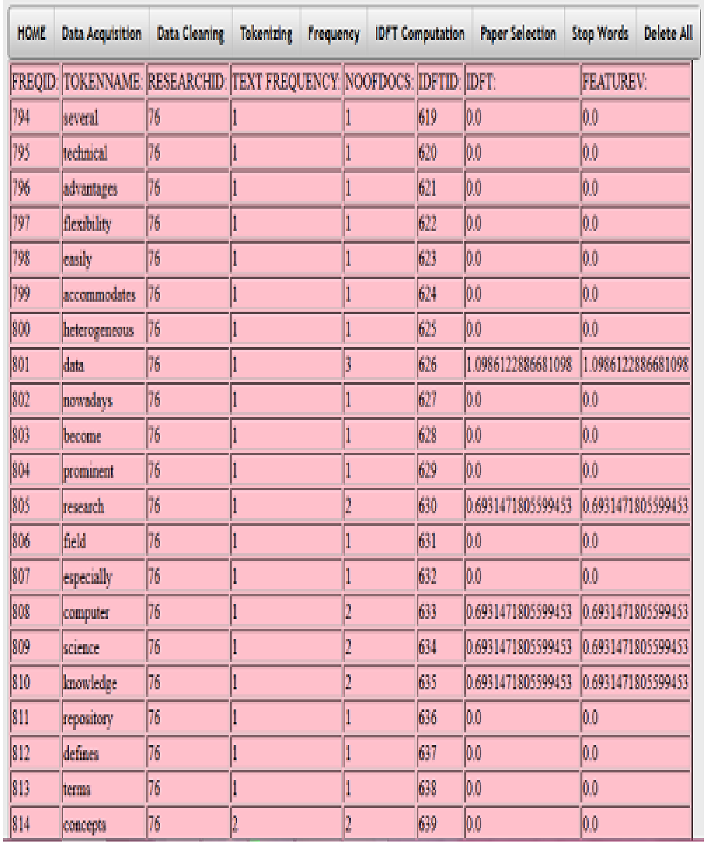
Feature extraction: The linked list which contains the pre-processed data is use for collecting feature of that document. This is done by comparing the linked list with keywords of ontology of different area. So the refined vector will act as feature vector for that proposal.

Model selection: Now, the way by which that paper is categorized into research area is clustering of proposal. This is done by many approaches but this paper use the k-means algorithm for it. Here created ontology is used for training the network. Here created ontology and feature vector both so it will train then specify the corresponding research area.

Training and Testing: Here created research projects feature vector are transfer in the form of input as the training data to network for training. And this trained network is test with different proposal’s feature vector so one can obtain belonging class of proposal.

VI. RESULT AND DISCUSSION

Following section will discuss about the overall result of this paper. Any government and private funding agency will call for proposals. Then users can submit their research proposals and applying pre-processing steps like data cleaning, all the stop words will be removed. Each tokens have their own unique id, then calculate number of times token has appeared which gives frequency. By using appropriate formula we can calculate idft and feature vector. Based on the feature vector value proposals can be ranked.



FREQID:	TOKENNAME:	RESEARCHID:	TEXT FREQUENCY:	NOOFDOCS:	IDFTID:	IDFT:	FEATUREV:
794	several	76	1	1	619	0.0	0.0
795	technical	76	1	1	620	0.0	0.0
796	advantages	76	1	1	621	0.0	0.0
797	flexibility	76	1	1	622	0.0	0.0
798	easily	76	1	1	623	0.0	0.0
799	accommodates	76	1	1	624	0.0	0.0
800	heterogeneous	76	1	1	625	0.0	0.0
801	data	76	1	3	626	1.0986122886681098	1.0986122886681098
802	nowadays	76	1	1	627	0.0	0.0
803	become	76	1	1	628	0.0	0.0
804	prominent	76	1	1	629	0.0	0.0
805	research	76	1	2	630	0.6931471805599453	0.6931471805599453
806	field	76	1	1	631	0.0	0.0
807	especially	76	1	1	632	0.0	0.0
808	computer	76	1	2	633	0.6931471805599453	0.6931471805599453
809	science	76	1	2	634	0.6931471805599453	0.6931471805599453
810	knowledge	76	1	2	635	0.6931471805599453	0.6931471805599453
811	repository	76	1	1	636	0.0	0.0
812	defines	76	1	1	637	0.0	0.0
813	terms	76	1	1	638	0.0	0.0
814	concepts	76	2	2	639	0.0	0.0

Fig 3: Result Analysis

In the above screen shot few of feature vector value is zero, it means that few of tokens are unique, not repeated in any of the proposals leads the frequency value to 1. Hence idft is log of n to f where n is number of documents and f is frequency then multiply idft value with frequency will give the feature vector value. So like this, the proposal whose feature vector value is highest that will come at top of list. Like this we can rank the papers.

VII. CONCLUSION

This paper has presented the ontology in text mining for grouping of proposals. Research ontology is constructed to categorize the concepts in different discipline area. And also form a relationship among them. The text mining technique provides different methods like classification, clustering etc. for extracting important information from unstructured text document. Feature vector value will be calculated based on number of times token has been repeated in proposals with product of inverse document frequency text value. Finally proposals are ranked based on the feature vector value. Highest feature vector valued proposal will appear at top this can give effective research proposal.

REFERENCES

[1] 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*.
[2] Gangemi A (2005), “Ontology design patterns for semantic web content”, In *The Semantic Web* ISWC 2005, *Springer*.
[3] Rainer Malik, Lude Franke and Arno Siebes (2006), “Combination of text-mining algorithms increases the Performance”, *Bioinformatics.*
[4] Henriksen A D and Traynor A J (1999), “A practical R&D project-selection scoring tool,” *IEEE Trans. Eng. Manag*., Vol. 46, No.
[5] 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.
[6] S. Bechhofer et al., OWL Web Ontology Language Reference, W3C recommendation, vol.10, p. 2006-01, 2004.
[7] 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
[8] jay prakash oandey et al,automatic ontology creation for research paper classification vol. 2, no 4 Nov 2013.