International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-5, March 2020

Development of Research Proposal Selection Based on Domain Ontology using K-Means Categorical Clustering

Iyswarya E., Balamurugan M., Vinoth Kumar N.J.

Abstract: ***With the prompt improvement in research progress of various zones, selection of research proposals became a remarkable methodology in many research funding agencies and organizations. When a less number of research proposals are received, then it is ease to cluster the research proposals and the selection process became as non-problematic way. If a number of research proposals elevated, then the clustering and selecting the proposals became complicated. In current system, proposals grouping is done in manual-based or along with their similarities in subject disciplinaries which yield irrelevant results in some cases. The main goal of this research work is to develop an enhanced system in selection of research proposals based on Domain ontology, where the ontology acts as a searching criteria for the topics of research proposals. This proposed system will help to select the topics of research proposals in well-systematic way without the interference of manual progression. In this paper, an algorithm is proposed as Scikit-learn K-means Multiclass Document Clustering(SKMDC) to group each subject discipline according to their sub-topics and sub-domains. Here, the k-means clustering technique is implemented on categorical data to implement the clustering process. As, the categorical data are not able to applied directly in K-means clustering algorithm, the LabelEncoder method is implemented to encode the text data to numerical values and the dimensions of a dataset are reduced using Principal Component Analysis. This paper also overwhelms the weaknesses of k-means technique in specification of cluster number in initial stage. It is done through the determination of optimal number of clusters by using Elbow Curve method and it is cross-validated through Silhouette Score analysis.***

Keywords: ***k-means clustering, Principal Component Analysis, Elbow Curve, Silhouette Score, LabelEncoder, Research Proposal Selection, Domain Ontology***

I. INTRODUCTION

For any research funding agencies or organizations, for instance government or non-government funding agencies, research proposal selection remains as predominant task when huge number of research proposals are received.

Revised Manuscript Received on March 30, 2020.
* Correspondence Author
Iyswarya E.*, Research Scholar, School of Computer Science, Engineering and Applications, Khajamalai Campus, Bharathidasan University, Tiruchirappalli, India. Email id: eiyswarya@gmail.com
Balamurugan M., Professor and Head School of Computer Science, Engineering and Applications, Khajamalai Campus, Bharathidasan University, Tiruchirappalli, India. Email id: mbala@gmail.com
Vinoth Kumar N.J., Assistant Professor, Department of Electrical and Electronics Engineering, Government Polytechnic College, Nagercoil, India. Email: vinothkumarnj@gmail.com

It is a multi-process task starts from call-for-proposals to germane communities. The collected proposals are categorized and grouped by appropriate authorities and then the proposals are assigned to peer review to select the praiseworthy proposals. [1] Figure 1. presents the traditional method of selecting a proposal. These steps are followed mostly in all research funding agencies. [2].

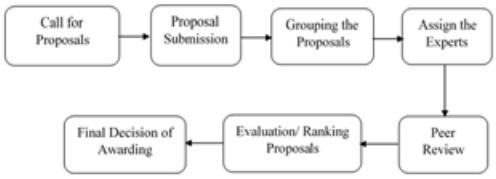

Fig 1. Research Proposal Selection Process

In current method, once the proposals are submitted from various sources, categorizing the proposals along with their subject domains and grouping the research proposals along their subject domains and then assigning the proposals to experts for review process are done manually or keyword similarities. [3] In manual progression, the authorities in grouping departments may not have sufficient ideas in all subject views which leads to misinteruptions in subject concepts [4].

And also in some cases, the research agencies follow some private software or program to do the selection process, that too not help while the proposals are huge in number [5]. On the other hand, the grouping is done according to similarities of keywords in subject domains but not in semantic manner which results in wrong set of grouping proposals.

To conquer the demerits of current research proposal selection system, Domain Ontology based K-means clustering algorithm is proposed to group the research proposals for each subject disciplinaries. Ontology is defined as a semantic technique of knowledge symbolization by using a set of concepts and properties for a specific domain. With the help of this ontology based clustering, the research proposals are grouped in well-defined manner as well as in semantic way.

Contribution of the Work

In this paper, a new approach is presented to select the research proposals without the interference of manual progress and to group the research proposals of each subject areas by using K-means clustering algorithm and overwhelm the manual progression in research proposal selection system. the main contributions of this research work are stated as follows:

- We present a new solution for keyword occurrences of research proposals in a single set of time by using a frequency distribution method.
- A novel approach is modeled to search a domains and sub-domains of a subjects with the help of domain ontology.
- An enhanced method is introduced to cluster the research proposals with a K-means clustering algorithm.
- A traditional process of K-means algorithm are surmounted by determining an optimal number of clusters.

Paper Organization:

Rest of a paper is organized as follows: Section II, presents a brief work that are related to cluster the research proposals, frequency distribution method and domain ontology framework. Section III, discuss the problem statement of this proposed research methodology. In section IV, presents a proposed approach for research proposals selection system. Section V, shows a proposed algorithm on research proposals and in section VI, the experimental results are evaluated. Finally, in section VII, the conclusion of this research work is presented.

II. RELATED WORK

Okko Rasanen, Shreyas Seshadri et. al [6] presents an Automatic Word Count Estimation (AWCE) to estimate a number of words spoken in an audio recording. But this method is used only for a words that are in an audio format. Xiuwen Chen et.al [7] proposed a co-word analysis based on keywords from the funded project. The relationship between the research topics are studied by cluster analysis and social network analysis. The keywords with the frequencies under 8 are removed from the list because only few researchers pay attention on them.

In general, there are no proper "methodology" for developing the ontology. It can be constructed based on the needs and purpose of an application and user i.e. it is mostly depending upon user of ontology. Domain Ontologies are constructed for Safety Risk identification to formalize the safety risk knowledge in metro construction [8], for historical documents [9], for university purpose [10].

Qijia Tian et.al [11] proposed a decision support system for a Research and Development (R&D) project selection. the object-orient method is used to design a decision support system. but the decision models are used only in structured documents. Henriksen A.D and Traynor A.J presented a scoring tool for R&D project evaluation and selection in which they rank the project alternatives by characteristics of relevance, risk and return [12]. O. Liu et.al [13] established multilingual ontology framework for R&D system. M. Nagy and M. Vargas-Vera [14] presented multi-agent ontology mapping skeleton for a heterogeneous data integration on semantic web. Ontology matching has been progressed to assign proposals to experts by using two ontologies [15]. Hossein Shahsav and Baghdadi et al, 2011 developed an Automatic Topic Identification Algorithm to identify the topic for a textual document, by this method, they achieved 86% of matching for both total and partial matching among 200 random documents from the Wikipedia [16].

III. PROBLEM STATEMENT

A selection of research proposals is an important and difficult task for many research funding agencies. Mostly, in some organizations and agencies they follow a system named as traditional system to select the research proposals in which they face misinteruptions in terms of wrong grouping and time consuming. The proposed research methodology develops a novel approach for a research proposal selection system by developing a domain ontology for research proposals which performs as a background knowledge to search research proposal topics. In order to collect the keywords from a research proposal an algorithm is proposed as Research Proposal Selection Frequency Distribution Algorithm (RPSFDA).

Then these keywords are classified based on their subject disciplinaries by using a Multinomial Logistic Regression. Once a research proposals are classified an unsupervised approach is performed to group the proposals for each subject discipline because it is not able to directly delineate a set of techniques and methods adopted for each subject domain or discipline and then it is assigned to expert reviewers for a review the proposals. Some subject discipline techniques like medical science are not used in other domains, are directly assigned to their appropriate reviewers so that all the received proposals are under the control of a system which does not leads to any neglecting process.

IV. PROPOSED APPROACH

The proposed research progress is amalgamated with five processing phases starts from the collection of a research proposals. They are Document preprocessing, frequency distribution, Domain Ontology framework, research proposal classification and clustering the proposals for each subject discipline. The figure 2. presents the workflow for a proposed research progress to develop a research proposal selection system. The implementation of a grouping progress is accomplished as:

- RPSFDA
- Domain Ontology framework
- SKL-MLR
- SKMDC

The data taken for the proposed work is in the form of research proposals from the year of 2013 to 2016 which consist of 1630 proposals with various subject disciplinaries from Association for the Advancement of Artificial Intelligence (AAAI) repository. The research progress starts from preprocessing the research proposals by using tokenization, stopwords removal and stemming process [17]. The AAAI dataset applied for preprocessing and for normalization (stemming) is splited into training and testing phase. In preprocessing phase, the total number of tokens gained are 6350 which are reduced into 5306 using stopwords removal. Porter stemmer algorithm is applied in the refined tokens to normalize into their root words which is very accessible for the further progress of a proposed research methodology. A frequency distribution is calculated from this stemmed words.

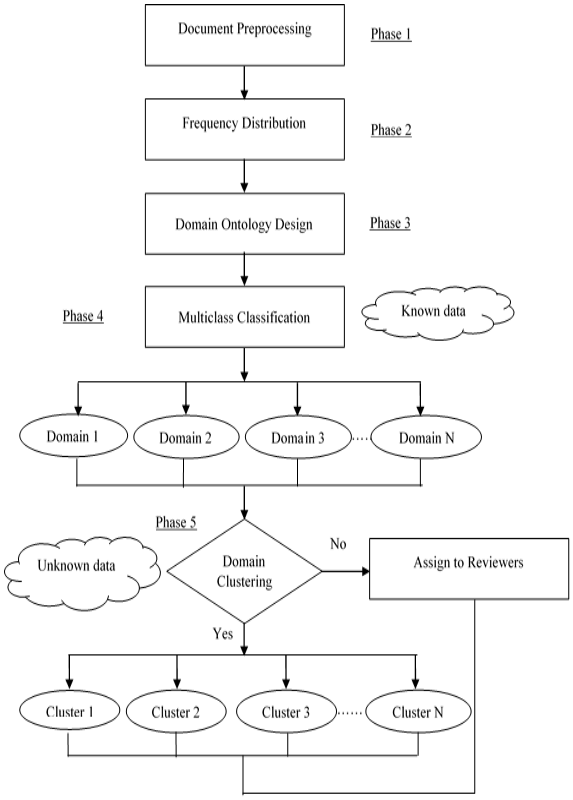

Fig 2. Workflow of Research Process

A. RPSFDA

The frequency distribution is defined as a number of occurrences of a word in a document. The frequency distribution is calculated for a stemmed token with Research Proposal Selection Frequency Distribution Algorithm (RPSFSA) [18]. It is proposed to count the topics of a research proposals and it also overcomes the existing system of keyword count based on feature set [19]. The consumption of time is more in existing system because a feature set has to be created for each domain then the research topics are attained based on this feature set. The RPSFDA acquired a list of keyword count for overall refined data in single set of time. Figure 3. displays a word frequencies for a sample set of 50 classes (subjects).

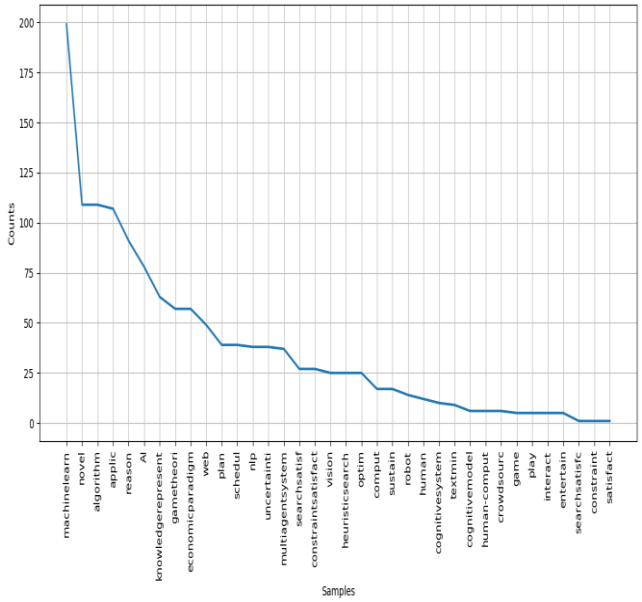

Fig 3. RPSFDA

B. Domain Ontology Framework

The domain ontology is developed according to research areas that are presented as a background for a searching basis. The ontology editor, Protégé [20] is used to develop this proposed ontology. The development of this ontology is more complex than a normal tree-like structure representation. The keywords gained from RPSFDA is taken to develop it which are narrow down it specific discipline areas called as programs. For example, the keyword as 'machine learn nlp' can be placed under the domain 'Machine Learning' and 'Natural Language Processing'. There are some terms with different names in different proposals which characterize the same concept must also be identified. This leads to a complex relationship between the concepts in ontology comparing to a basic tree-like structure. [19]

Once funding the research proposals completed for each year, the domain ontology is updated according to a subject techniques and funding agencies policy.

C. SKL-MLR

The Scikit-Learn Multinomial Logistic Regression algorithm (SKL-MLR) is proposed to achieve a better classification and analysis in the research topics. The performance of this proposed algorithm is evaluated by precision, recall, confusion matrix, intercepts and coefficients. The accuracy achieved for this proposed classification algorithm is 0.97%.

D. SKMDC

Once the proposals are classified according to their subject areas. The clustering approach is performed on each subject areas. Each subject discipline has 1 to N number of clusters which are grouped according to their specific techniques and sub-domains. A Scikit-learn K-means Multiclass Document Clustering (SKMDC) algorithm is proposed to perform a clustering process which are presented in upcoming section.

V. PROPOSED ALGORITHM FOR GROUPING RESEARCH PROPOSALS

The novel Scikit-learn K-means Multiclass Document Clustering (SKMDC) algorithm is to cluster the research proposals by using a python tool with sklearn library. The Elbow Curve and Silhouette Score analysis is performed to determine an optimal number of clusters for k- means algorithm. The step-by-step procedure of this proposed algorithm is described as follows:

Algorithm: Novel SKMDC Algorithm

Input: Data, Numpy (np), Pandas (pd), PCA.

Output: Result (PCA variables, Elbow Curve, Silhouette Score, kmeans clustering report)

Step 1: Import linear algebra(numpy), data analysis (pandas), dimensionality reduction(PCA) and metrics to compute a score.

Step 2: Read a preprocessed data by using a data analysis(pd).
Step 3: Encode a categorical data to numerical variables using a LabelEncoder.
Step 4: Perform Principal Component Analysis to simplify a dimension of data in order to identify a better cluster.

For N samples,

$$mean\ vector(\bar{S}) = (S_1 + S_2 + \cdots + S_N)/N$$
$$S_{mean} = [\bar{S_1} \ldots \bar{S_N}]$$
$$cov_{(x,y)} = (S_X - \bar{S}).(S_Y - \bar{S}) \quad // X, Y \text{ are document topic vectors }$$
$$\lambda\iota - \mathrm{C} = 0$$
$$(\lambda - \mathrm{C}) * V_k = 0$$
$$S_B = S_{mean} * EV$$
$$PCA = S_B * S_{mean}$$

Step 5: The above process is computed for a encoded data and stored a result as $npca$**.**
Step 6: The optimal number of clusters is obtained initially for a k- means algorithm.
Step 7: First, the Elbow Curve method is plotted as $avg\ within\ cluster\ sum\ of\ square\ vs\ No.of\ clusters$.

$$wss = \sum_{r=1}^{k} \frac{1}{n_r} D_r$$

Where,

k = No. of clusters.
n_r = No. of points in r.
D_r = Sum of distance between all points in clusters

Whereas,

$$D_r = \sum_{i=1}^{n_r - 1} \sum_{j=1}^{n_r} \|d_i - d_j\|^2$$

Step 8: Then, the Silhouette Score analysis is calculated.

$$score = \frac{(p - q)}{\max(p, q)}$$

Where,

p = mean distance to points in nearest cluster
q = mean intra-cluster distance to all points

Step 9: From the above two methods, the optimal number of clusters is determined.
Step 10: The Euclidean distance measure is used as a metric for the determination of optimal cluster and for $k$means clustering process.
Step 11: Fit the $k$means

$$kmeans(n\_clusters = optimal\ no.of\ clusters).fit\ (npca)$$

Step 12: Plot the result

$$(kmeans.label, kmeans.cluster\_center)$$

The proposed algorithm explains a categorical clustering of research proposals for a selection of funding agencies. In initial stage, the PYTHON tool imports a package for linear algebra, data analysis, dimensionality reduction and a metrics to explore a score analysis. The pandas package is used to read the input dataset file which contains a refined set of tokens. As, an input dataset consists of categorical data, for a best performance and to identify a better clusters the categorical data is encoded into numerical variables.

Then, the Principal Component Analysis (PCA) is performed for N samples of a input data with following steps: (i) Compute a mean vector($\bar{S}$) for N samples and then mean adjusted matrix (S_{mean}) is assembled. Once this standardization is done, all variables are transformed into same scale. The Covariance Matrix ($cov_{(x,y)}$) is computed to identify the correlations, (i.e.,) how the variables of input data varies from the mean. The Covariance Matrix is a summaries of correlations between all pairs of variables.

Next, the Eigen Value (λ) and Eigen Vector (EV) of a Covariance Matrix($cov_{(x,y)}$) is performed to recognize a principal components. The principal components are new variables that are constructed as a linear combination of initial variables. Then, the basis vector (S_B) is executed to choose whether to keep all the components or to abandon the components that have a less importance and to build a matrix for a remaining ones. It keeps a EV as column which initiate a first step for a dimensionality reduction. Then, the final dimensionality reduction data (PCA) is obtained by multiplying a S_B and S_{mean}. Once a result is acquired the PCA result is stored as npca.

Next, to perform k − means clustering algorithm an optimal number of clusters are determined initially. To do so, the Elbow Curve method and Silhouette Score analysis is performed. In Elbow Curve method, the optimal number of clusters obtained is 3. Though, both the process (i.e., Elbow Curve method and Silhouette Score) are used to determine an optimal number of clusters, but in some cases, the Elbow in the arm shape curve is not displayed clearly. So, to provide more confident decision, the Silhouette Score analysis is performed and results achieved as a 3 which are same as the number of optimal cluster gained by Elbow Curve method. Then, the kmeans clustering process is performed using a Euclidean Distance as similarity measure for both optimal number and clustering process.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed experiment evaluates the progress of clustering research proposals in algorithm based system for a funding agencies. The proposed Novel SKMDC algorithm makes research proposals to group according to their similarities in subject disciplinaries. Grouping or clustering the proposals after a classification progress became more advantage for agencies to allocate them to appropriate expertise to review the proposals. The determination of optimal number of clusters and clustering quality are evaluated as an important factor for a clustering process. The Elbow method and the cluster report are visualized. The Silhouette Score and PCA are also performed and analyzed.

A. Principal Component analysis

The PCA is a mathematical tool from applied linear algebra. It is one of the important backbone of a data analysis and it is assumed as a black box because it is highly efficient but poorly understood. In general, PCA are used to reduce a dimensionality of a dataset which consists of many variables that are correlated with each other.

The transformation of a variable in same scale of data to a new set is called as a Principal Components. The PCA is applied before a k − means clustering algorithm in order to improve the cluster results as it reduce a noise in data. In PCA, it represents all n data vectors as linear combination of eigenvectors which are done to minimize reconstruction error, in other hand, in k-means it represent all n data vectors as linear combination of a small number of cluster centroid vectors. This the main connection between this PCA and k − means clustering algorithm.

Figure 4, shows the PCA results for some sample components from the results obtained for overall refined dataset.

```
[[-5.00594935e+01 -1.84572551e+00]
 [ 1.95463480e+00  8.98441679e-01]
 [ 5.29638561e+01  2.64754013e+00]
 [ 2.69395748e+01 -2.22450112e+00]
 [ 3.94969144e+00 - 1.11384550e-01]
 [-1.90598686e+01 -1.99821958e+00]
 [ 1.79396837e+01 -2.18022864e+00]
 [ 2.79543202e+01  7.70543420e-01]
 [ 4.09443246e+01 -1.29338151e+00]
 [-2.50352002e+01  3.03123490e+00]
 [-3.00597355e+01 -1.94410878e+00]
 [-3.90497883e+01  1.00139493e-01]
 [-5.30446997e+01  1.16899569e+00]
 [ 3.69394538e+01 -2.27369276e+00]
 [ 1.96447313e+00  2.89841748e+00]
 [-2.20549131e+01 -9.83474192e-01]
 [ 3.95461060e+00  8.88603351e-01]
 [ 9.64485229e-01  2.90333664e+00]
 [ 3.99492559e+01 -2.88474446e-01]
 ..... ..... ..... .....

[-1.00403008e+01  1.95745954e+00]
 [-3.90399500e+01  2.10011529e+00]
 [-2.20499940e+01  1.65137085e-02]
 [ 4.89491470e+01 -3.32746920e-01]
 [-5.00299786e+01  4.15420190e+00]
 [ 4.19393933e+01 -2.29828858e+00]]
```

Fig 4. Principal Component Analysis Results

B. Optimal Number of Clusters

In k −means clustering algorithm the right number of cluster should be specified initially and it is considered as demerit for this efficient algorithm. It is overwhelmed in this proposed work by determining the optimal number of clusters. In general, getting the optimal number of clusters is a significant in the analysis. If k is too high, then each datapoint acts as a cluster and if k is low, then the points are not clustered correctly. So, forming an optimal number of cluster leads to granularity in cluster process.

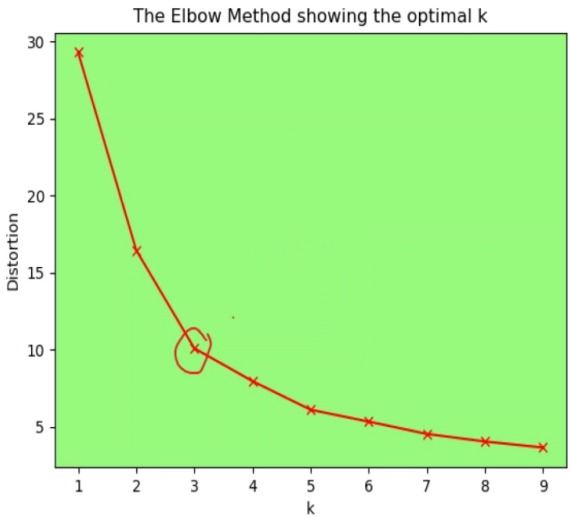

Fig 5. Elbow Curve point

The Elbow Curve is visible by plotting the k values which is like arm shape point. In some cases, the elbow point is not clearly visible so it is cross validated by silhouette score analysis. Figure 5, shows the elbow point for optimal number of clusters.

C. Silhouette Score Analysis

The performance of a clustering is attained by number of metrics. One of the main performance measure that are used to quantify the quality of clustering process is a Silhouette Score or Silhouette Coefficient. It refers a consistency and validation within a cluster of data. The value acquired by a Silhouette measure shows of how similar an object is to its cluster comparing to other clusters. The values ranges from -1 to $+1$. The Silhouette Score, $s(x)$, for a data point, x is defined as:

$$s(x) = \begin{cases} 1 - \dfrac{a(x)}{b(x)}, & \text{if } a(x) < b(x) \\ 0 & \text{if } a(x) = b(x) \\ \dfrac{b(x)}{a(x)} - 1 & \text{if } a(x) > b(x) \end{cases}$$

The above equation makes a clear note on the Silhouette value range as $-1 \leq s(x) \leq 1$. The Silhouette Score indicates that high values are well matched to its own clusters and the clustering configuration is more appropriate, on the other hand, the low values are poorly matched to neighboring clusters and it may have too many or too less clusters.

Table 1. Silhouette Score for a Cluster range (2, 5)

Number of Cluster	Silhouette Score Values
2	0.58953
3	0.59717
4	0.58272

The silhouette score is measured by performing two iterations from the range of cluster range (2,5) and (2,10). In both iterations of cluster range it is noted that cluster number,3, has a high value of 0.597 ± 2 and is demonstrated that optimal number, k, of proposed k − means algorithm is 3 and it also proved that the Elbow Curve method also results the same. Table 1 and 2 shows the Silhouette Score for two range of cluster iterations as (2,5) and (2, 10).

Table 2. Silhouette Score for a Cluster range (2, 10)

Number of Cluster	Silhouette Score Values
2	0.58953
3	0.59717
4	0.58272
5	0.57855
6	0.55976
7	0.54725
8	0.54906
9	0.53767

The number of cluster for cluster range (2,5) are 2,3 and 4, whereas for the cluster range (2, 10) are 2,3,4,5,6,7,8 and 9. The visual representations for both iterations are displayed in figure 6 and 7.

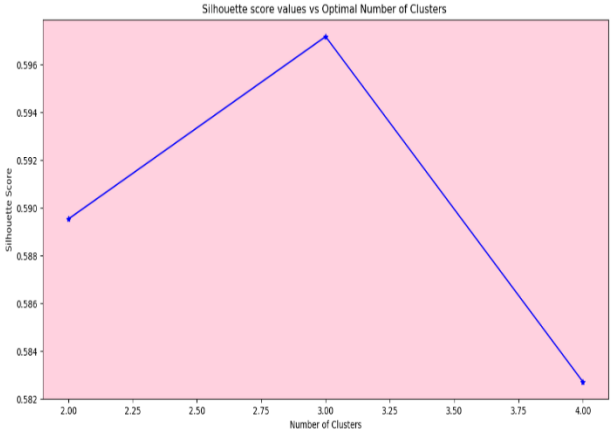

Fig 6. Graphical Representation of Silhouette Score for Cluster Range $(2, 5)$

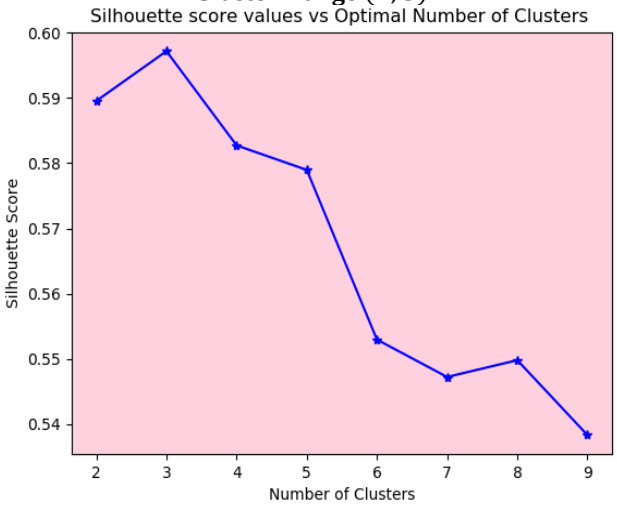

Fig 7. Graphical Representation of Silhouette Score for Cluster Range $(2, 10)$

D. SKMDC Graphical Representation

Instead of mentioning the number of clusters randomly as an input the proposed work obtained the right number of cluster. Once the optimal number of cluster is obtained then it is fit to proposed SKMDC algorithm. Figure 10. shows the report of proposed clustering algorithm with the centroids that are placed correctly in all the clusters.

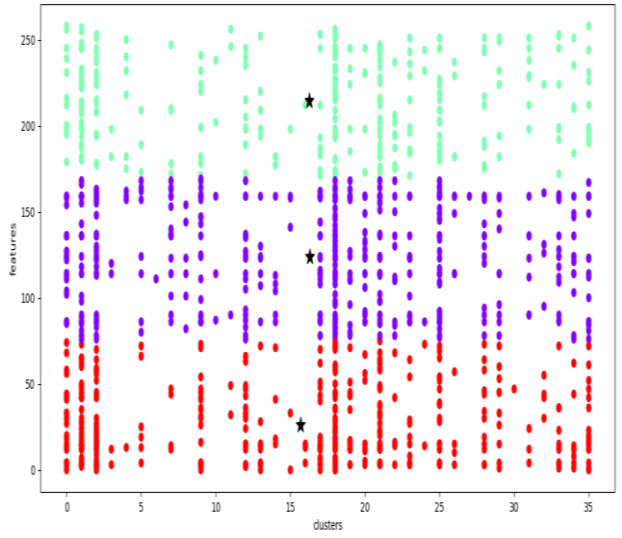

Fig 8. SKMDC Graphical Representation

VII. CONCLUSION

The research proposals for each subject discipline according to their sub-domains and sub-topics are grouped with the proposed Scikit-learn Multiclass Document Clustering (SKMDC) algorithm. It clusters each subject domain from 1 to N number of clusters according to their individual subject domain's techniques and methods. The proposed algorithm overcomes the randomly selection of cluster number in traditional k-means clustering algorithm by determining the optimal number of cluster. The Elbow Curve method is used to detect the right number of cluster. The optimal number of cluster for the proposed algorithm is, $k = 3$. Next, the quality of a cluster performance and to check the consistency of Elbow Curve results, the Silhouette Score analysis is performed and the results obtained in the two range of clusters as $(2,5)$ and $(2, 10)$. In this two iterations the optimal number of cluster is $k = 3$. So, it is proved that the optimal number of cluster obtained by both Elbow Curve method and Silhouette Score remains same. Then, the optimal number of cluster is used to fit in the proposed Scikit-learn Multiclass Document Clustering (SKMDC) algorithm. Once the research proposals are clustered it is assigned to the peer reviewers. On the other hand, some miscellaneous subject domain which cannot adopt by other subject areas are directly sent appropriate reviewers, so that all the proposals received are under the control of selection system. It is concluded that with the help of proposed SKMDC algorithm the research proposals are clustered in well-systematic way and the intervention of manual progress is also avoided.

REFERENCES

1. Thorat Madhuri, Naikwadi Chaitali, "Project Paper Selection Using Ontology Based Text-Mining", Global journal of Advanced Research, Vol. 2, No. 3, Pp. 595 – 602, 2015.
2. Http://Www.Nsfc.Gov.Cn/English/Site_1/Index.Html.
3. Lavanya, Rajkumar, "Ontology Based Clustering in Research Project Selection & Assign Proposals to Experts by Ontology Matching", International Journal of Research in Engineering and Technology, Vol. 3, No. 7, 2014.
4. Ankita, "Automatic Ontology Creation for Research Paper Classification", Iosr Journal of Computer Engineering (Iosr-Jce), Vol. 18, No. 2, April 2016.
5. S. Bechhofer Et Al., "Owl Web Ontology Language Reference, W3c Recommendation", Vol.10, 2004.
6. Okko Räsänen, Shreyas Seshadri, Marisa Casillas, "Comparison of Syllabification Algorithms and Training Strategies for Robust Word Count Estimation Across Different Languages and Recording Conditions", ResearchGate, August 2018.
7. Xiuwen Chen, Jianming Chen, Dengsheng Wu, Yongjia Xie, Jing Lic, "Mapping the research trends by co-word analysis based on keywords from funded project", Information Technology and Quantitative Management (ITQM 2016), Elsevier, Procedia Computer Science, Vol. 91, Pp. 547 – 555, 2016.
8. Xuejiao Xinga Botao Zhonga, Hanbin Luoa, Heng Lic, Haitao Wu, "Ontology for Safety Risk Identification in Metro Construction", Elsevier, Computers in Industry, Vol. 109, Pg. 14–30, 2019.
9. Giovanni Adorni, Marco Maratea, Laura Pandolfo, Luca Pulina, "An Ontology for Historical Research Documents", Springer International Publishing, 2015.
10. Naveen Malviya, Nishchol Mishra, Santosh Sahu, "Developing University Ontology Using Protégé Owl Tool: Process and Reasoning", International Journal of Scientific & Engineering Research Vol. 2, No. 9, 2011.
11. Qijia Tiana, Jian Ma, Jiazhi Liang, Ron C.W. Kwok, Ou Liu, "An organizational decision support system for effective R&D project selection", Elsevier, Decision Support Systems, Vol. 39, Pp. 403–413, 2005.

12. A. D. Henriksen and A. J. Traynor, “A practical R&D project-selection scoring tool,” IEEE Trans. Eng. Management, Vol. 46, No. 2, Pp. 158–170, May 1999.
13. Liu. O and Ma. J, “A multilingual ontology framework for R&D project management systems”, Expert Systems, Elsevier, Vol. 37, No. 6, Pp. 4626–4631, 2010.
14. M. Nagy and M. Vargas-Vera, "Multiagent Ontology Mapping Framework for the Semantic Web," in IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, Vol. 41, No. 4, Pp. 693-704, 2011.
15. Pavel Shvaiko and Jerome Euzenat, “Ontology Matching: State of the Art and Future Challenges”, IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 1, 2013.
16. Hossein Shahsav and Baghdadi and Bali Ranaivo-Malançon, "An Automatic Topic Identification Algorithm”, Journal of Computer Science, Pp. 1363-1367, 2011
17. Kannan, Vairaprakash Gurusamy, “Preprocessing Techniques for Text Mining”, Conference: RTRICS, ResearchGate, March 2015.
18. M. Balamurugan, E. Iyswarya, “Ontology Development and Keyword Count Using Research Proposal Selection Frequency Distribution Algorithm”, International Journal of Applied Engineering Research, ISSN 0973-4562 Vol. 13, No. 12, Pp. 10196-10201, 2018.
19. Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wangl,” An Ontology-Based Text- Mining Method to Cluster Papers for Research Project Selection”, IEEE Transactions on systems and humans, Vol.42, No.3, 2012
20. http://www.Protégé.Standford.edu

AUTHORS PROFILE

E. Iyswarya is a Research Scholar in the area of Ontology in the Department of Computer Science in Bharathidasan University, Trichy, India, under the guidance of Dr. M. Balamurugan. She has five years of teaching experience in the department of computer science at Ragavendra College, Chidambaram. She has published more than 5 international journals. She obtained her MCA degree at Annamalai University and M.Phil. degree at Cauvery College for Women, Trichy. Her area of interest is Ontology, Semantic Web and Web Mining.

Dr. M. Balamurugan is currently working as Professor and Head in the Department of Computer Science of Bharathidasan University, Trichy, India. He has the teaching experience of more than 18 years. He has credits of 20 + international and national conferences Publications. He has published 35+ research papers in national and international journals. His research interests are mainly focused on the area of Data Science, Machine Learning and Data Mining. He has supervised several research scholars in these areas.

Dr. N.J. Vinoth Kumar is currently working as an Assistant Professor in the Department of Electrical and Electronics Engineering at Government Polytechnic College, Nagercoil. He has 15 years of experience in teaching and research academy. His main area of research is Fuzzy Logic and Load Frequency Control. Other than that he has keen interest and proficiency knowledge in Data Analysis, Big Data and Geographic Technology. He has published more than 10 International and 5 National journals. He also attended and organized more than 8 conferences.