International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 12 (2018) pp. 10196-10201

Ontology Development and Keyword Count Using Research Proposal Selection Frequency Distribution Algorithm

Dr. M. Balamurugan*, and E. Iyswarya**

**Professor and Head, School of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli -23, India.*

***Research scholar, School of Computer Science and Engineering, Bharathidasan University, Tiruchirappalli -23, India.*

Abstract

The research proposal selection is necessary and important task for research agencies and institutes. When the agency or research institution acknowledged lots of research proposals it is sent to reviewers to review the papers. As there is enormous amount of proposals, screening process by editor are done by manual and it is quite difficult. Generally, the research proposals based on keywords and the matching between the similarities is done manually and it is moved to reviewers based on the subject disciplinaries. In this research work, the intention is to afford a better frequency count method in research proposal domain name selection process and it is proposed by Research Proposal Selection Frequency Distribution Algorithm (RPSFDA). Ontology is defined as a knowledge repository in which ideas and articles are defined as well as relationship between these ideas. The activities of searching similar pattern of text effectively, efficiently and interactively is made by ontology. In this research work, ontology is generated for research proposal selection and it is split up into four departments - 'computer science', 'management science', 'multidisciplinary topics', and 'others'.

Keywords - ontology, document preprocessing, frequency distribution, nltk, matplotlib, protégé, proposal selection

INTRODUCTION

Selection of research projects is an important and frequent activities in many organizations such as private or government agencies. It is a collection of multiprocess starts from a challenging task, call for proposals (CFP) by a funding organizations. The CFP is distributed to appropriate communities such as research scholars, scientists from many institutions or universities. These proposals are then submitted to funding agencies. As there is a single point of contacts, for researchers from different areas, the proposals are grouped based on their similarity and are assigned to peer reviewers. After that, the review results are collected and best proposals are ranked based on accretion of experts results. These basic steps of research proposal selection process are common to all research funding agencies. [15]

The procedure followed for research proposal selection by government and private funding agencies. It starts from CFP, proposal submission, proposal grouping, proposal assignment to experts, peer review, aggregation of review results, panel evaluation, and final awarding decision [2]. These processes are very similar in other funding agencies as shown in Fig..1

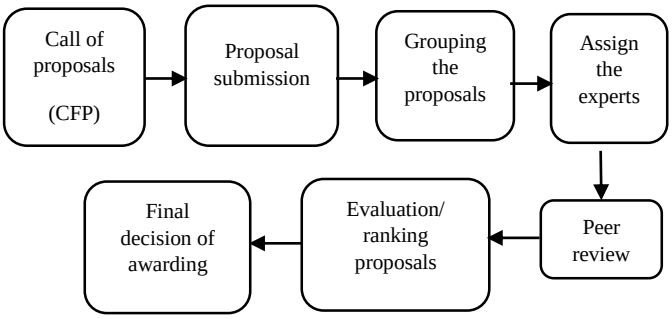

Figure 1. Traditional research proposal selection process

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 [1]. In current methods, the keywords are not often offering the widespread information about the proposals, they are just display the partial representation of proposals. So, it is not sufficient to group the proposals based on these keywords. The department responsible for grouping, faces the issues on behalf of not having adequate knowledge on areas of research proposals.

The objective of this work is to find out the frequency count of research proposal domain name and it is performed by RPSFDA and an ontology is created for these research proposals selection and that will effectively use for this purpose.

The paper is organized as follows: In section 2, related works are represented. The proposed approach on RPSFDA and constructing the research proposal ontology is given in section 3. The section 4, shows overall look of research ontology. The section 5 gives the result and discussion. The conclusion and

future work, is discussed in section 6 and followed with references and biography.

RELATED WORKS

Matthew Horridge [5] states that there are no procedures for construction of ontology. The main advantage is erection of the ontology is done based on their own ideas. Selection of research projects is an important task in research and development (R&D) management. Choi and Park used text-mining approach for R&D paper screening [9]. David shotton, present his ideas on CiTO, the Citation Typing Ontology, it is an ontology labelling nature of reference citations n research articles and scholar works. The citation is described in terms of factual relationships between citing publication and cited publication. This paper describes CiTO and illustrates its usefulness both for the annotation of bibliographic reference lists and for the visualization of citation networks. CiTO is written in the Web Ontology Language OWL [6]. Ontology patterns were introduced by Blomqvist and Sandkuhl in 2005 [7]. Later the same year, Gangemi presented his work on ontology design patterns [8]. Matteo Gaeta *et al*, 2011 presented an approach for extracting relevant ontology concepts and their relationships from a knowledge base of heterogeneous text documents using e-learning perspective. Giovanni et.al, presented the ontology for historical research documents [11].

The ontology model for software engineering to represent its knowledge is shown in [12]. Its end users are software engineers sharing domain knowledge as well as instance knowledge of software engineering. An automatic algorithm to identify the topic for a textual document based on the chunks corresponding to each sentences in the document is presented [13]. Benno Stein et.al, presents the framework to specify the topic identification problem and it introduces a classification scheme for topic identification algorithms [14].

PROPOSED APPROACH

This section considers the implementation portions of proposed RPSFDA for analyzing the frequency count of domain names that are performed from the input dataset. Once the frequency distribution is performed the ontology is developed for research proposal selection. In order to create this ontology previous year papers are selected from AAAI (Association for the Advancement of Artificial Intelligence) which are in different domains. The ontology is developed by using protégé. It is a free, open- source editor for developing the ontologies that is produced by Stanford University. It is java based application and has the plug-ins like ontoViz, that visualize the ontologies [3]. The workflow of this implementation phase described as follows.

- Document Pre-processing
- Frequency distribution
- RPSFDA
- Construction of research ontology

A. Document Pre-processing

The preprocessing stage is an important process to eliminate the presence of unwanted text in a given document and select the data. In this phase, the tokenization is performed to split the input dataset into token of words. Then stopwords are removed in dataset.

B. Frequency distribution

The keyword frequency is a number of times word that appears on a piece of content i.e., it is the sum of similar words that seems in regulation for period of given time. The frequency distribution is calculated using NLTK (Natural Language Tool Kit). According to this frequency distribution the research proposal topics are identified and it also get rid of manually identifying the topics by the organization. Fig. 2 visualize the frequency distribution performed in nltk for domain name in research proposal of input dataset.

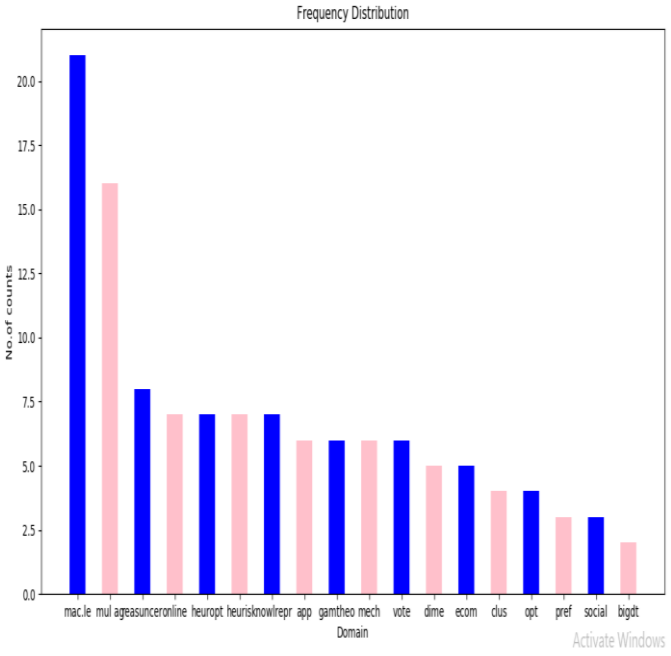

Figure 2. Frequency Distribution

C. RPSFDA

The RPSFDA is proposed in this research work for the better frequency count for words. This algorithm is based on the process of frequency distribution. It accomplishes three computations: stopwords removal, frequency distribution and graphical representation using matplotlib.

Algorithm 1: RPSFDA

Input: Data(i/p), nltk, tokenize, FreqDist, tick_label, matplotlib

Output: Result (word frequency count, Graph representation)

Step 1: Partition the input into two sets as training set(tr) and testing set(ts)

Step 2: Import the nltk, stopwords, matplotlib

Step 3: *w = nltk.tokenize(s)*

Step 4: *stop_words = set(nltk.corpus.stopwords.w('english')*

for word in w if word not in stop_words

Step 5: Calculate the frequency distribution of ts

fdist = nltk.FreqDist(w)

for word, frequency in fdist.most_common(n):

print(u'{} - {}'.format(word, frequency))

Step 6: Custom the matplotlib.pyplot as plt

Step 7: Mention the measures of graph.

Left = x-coordination

Height= No. of counts

Tick_label= domain

Step 8: *plt.bar (left, height, tick_label = tick_label, width = 0.5, color = ['blue','pink'])*

Step 9: From the frequency distribution perform the function to plot the graph

Plt.show()

The RPSFDA is to perform the keyword frequencies of the document. The dataset is taken as 600 research papers from AAAI and the process is splited into training dataset and testing dataset. *tr* and *ts* are considered as parameters. And by using the *ts dataset,* Import the nltk, stopwords, matplotlib. Then Perform the tokenization operation in *ts.* After this, execution and removal of stop words for *w.* Calculate the frequency distribution of *ts* as *w* words. Import the matplotlib and mention the measures of graph. Then plot the chart as *plt.bar.* At last, visualize the chart of frequency distribution.

D. Construction of research proposal ontology

In this work, the research proposal ontology is regarded according to scientific research areas that are presented in the background. Then it is developed on the basis of numerous unambiguous research areas. Next, it is further partitioned into some narrow discipline areas called as programs. Finally, it leads to research topics in terms of the feature set of discipline created in step 1. The research proposal ontology is constructed by a protégé and its visualization is shown in Fig.3. The construction is more complex than a tree-like structure. For example, '*natural language processing machine learning'* can be placed under "Machine Learning" or under "Natural Language Processing". Second, there are some synonyms used in dissimilar applicants that have the different names in different proposals but characterize the same concept. Therefore, the research ontology allows more complex relationship between concepts besides the basic tree-like structure [4].

Updating the research ontology: Once the project funding is completed each year, the research ontology is updated according to agency's policy and the change of the feature set.

COMPLETE VIEW OF RESEARCH ONTOLOGY

The ontograph tab in protégé shows the complete view of each classes, subclasses and its members associated with it. This graph also depicts the relationship between each classes. The colors used in this graph distinguish different properties. As it is more complex than tree structure Fig.3 shows the overall look of research proposal ontology for "management science" and "multidisciplinary topics".

This section, makes a clear note on selecting a research paper, like in class "Computer Science" the sub classes are 'machine learning', 'artificial intelligence', 'natural language processing', 'vision', 'knowledge representation and reasoning', 'robotics', 'planning and scheduling'. Fig.4 visualizes the "computer Science" and "others" domain. These sub-classes have their own sub-classes too. And the members are equilibrium theory, representation, natural language dialogues, brain-computer interfaces, answer set programing, etc. so that the paper proposals can easily recognize in which group the proposals are belongs to.

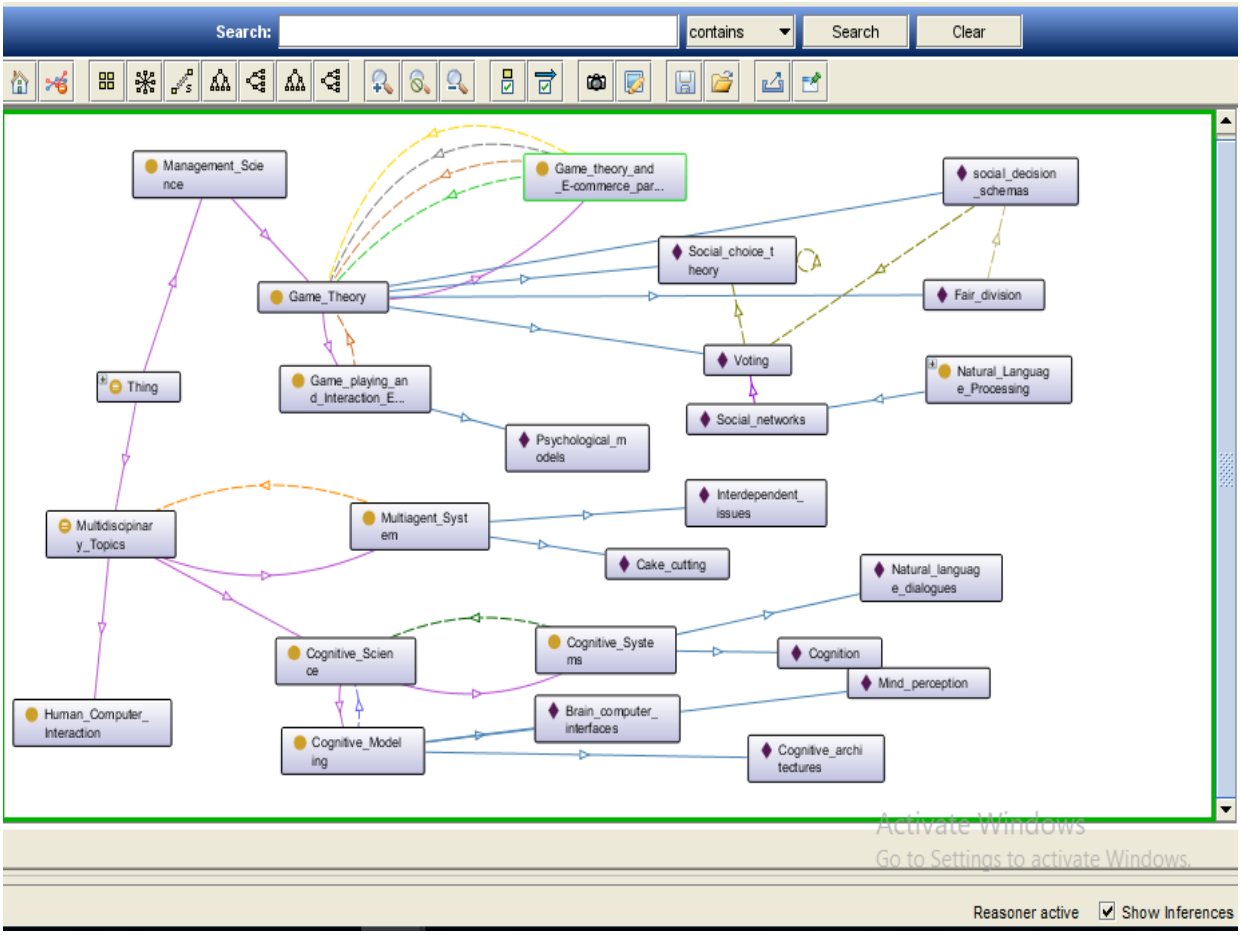

Figure 3. Ontology for management science and multidisciplinary

RESULT AND DISCUSSION

This section discusses the process of keyword frequency count. The proposed method of frequency distribution technique is compared with existing systems. As in our work, the frequency count of words is done by RPSFDA as in existing systems the frequency count for document is obtained by feature set (Nok, IDk, year, {(keyword1, frequency1), (keyword2, frequency2), . . ., (keywordk, frequencyk)}) [4]. The consumption of time to find the frequency is more because the feature set for each domain are created and based on this specific research area are obtained and then narrow line of research topics are attained. In our proposed approach, the frequency distribution is gained for overall input set at a single time.

CONCLUSION AND FUTURE WORK

From this implementation work, the RPSFDA is utilized for the keyword frequency counts. Here using the dataset from AAAI the frequency distribution and ontology are also erected for research proposal selection. This proposed technique used text from the datasets to perform the training operations and testing operations. Then the obtained discussion is compared with existing manual process and feature set operation of keyword frequencies. The next phase of this work will be classification to the research proposals and grouping them. And the optimization technique is also undertaken in clustering phase as a future work.

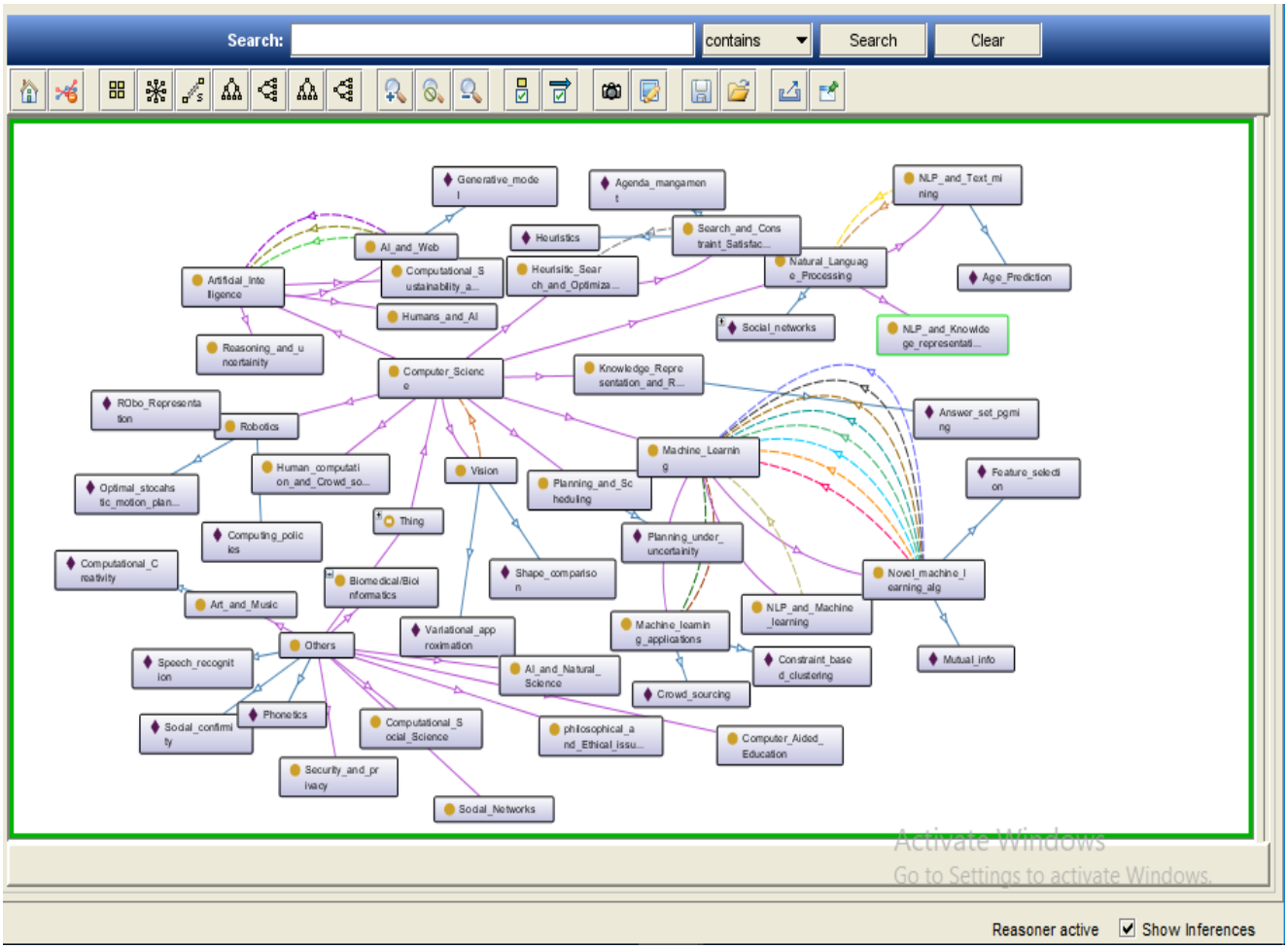

Figure 4. Ontology for computer science and others

REFERENCES

[1] Ankita, "Automatic Ontology Creation for Research Paper Classification", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727, Volume 18, Issue 2, Ver. II, Mar-Apr. 2016.

[2] D. Saravana Priya, M. Karthikeyan, "An Ontology Based Text-Mining to Clustering the Research Projects Based on Fuzzy Technique", The International Journal of Science & Technoledge, Vol 3 Issue 7 July, 2015.

[3] DR. M. Balamurugan, E. Iyswarya, "CONSTRUCTION FOR RESEARCH PAPER SELECTION ONTOLOGY USING PROTÉGÉ", International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 13, Number 9 (2018) pp. 6989-6993.

[4] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu, "An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, VOL. 42, NO. 3, MAY 2012.

[5] Matthew Horridge, "A Practical Guide to Building OWL Ontologies Using Protégé and CO-ODE Tools Edition", The University of Manchester, March 13, 2009.

[6] David shottan, "CiTO, the citation typing ontology", Journal of Biomedical Semantics 2010, 2009.

[7] 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.

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

[9] C. Choi and Y. Park, "R&D paper screening system based on text mining approach, Int. J.Technol. Intell. Plan., vol. 2, no. 1, pp. 61–72,2006.

[10] Matteo Gaeta, "Ontology extraction for knowledge reuse the e-learning perspective", IEEE Trans on systems, man, and cybernetics—part A: systems and humans, vol. 41, no. 4, 2011.

[11] Giovanni Adorni1, Marco Maratea1, Laura Pandolfo1, and Luca Pulina, "An Ontology for Historical Research Documents", Springer international Publishing, Switzerland, 2015.

[12] Pornpit Wongthongtham, Elizabeth Chang, Tharam Dillon and Ian Sommerville, "Development of a Software Engineering Ontology for Multi-Site Software Development", IEEE Transactions On Knowledge and Data Engineering, Manuscript Id.

[13] Hossein Shahsavand Baghdadi and Bali Ranaivo-Malançon, "An Automatic Topic Identification Algorithm", Journal of Computer Science 7 (9): 1363-1367, 2011, ISSN 1549-3636, 2011 Science Publications.

[14] Benno Stein, Sven Meyer zu Eissen, "Topic Identification: Framework and Application", 4th International Conference on Knowledge Management, Journal of Universal Computer Science, pp. 353-360, ISSN 0948-6968.

[15] SaravanaPriya, Dr. M. Karthikeyan, "MLTP GROUPING FOR RESEARCH PROJECT SELECTION", International Journal of Advanced Engineering Technology, E-ISSN 0 976-3945, Vol. VII/Issue II/April-June,2016/524-531.

[16] Sunil Datir, Arpit Solanki, "OTMM for search proposal classification", International Journal of Computer Science and Mobile Computing", Vol. 4, Issue. 9, September 2015, pg.154 – 165.

[17] Almudena Ruiz-Iniesta and Oscar Corcho, "A review of ontologies for describing scholarly and scienific documents", 2008.

[18] Suganya, Shanmugakani, "An ontology Based Text Mining Approach Cluster Proposals and Expert Reviewers for Selecting R&D Project". International journal of Engineering Research and Technology, Vol.5, Issue 6, June 2016.

[19] A. Nancy, Dr. M. Balamurugan, and S. Vijaykumar, "Alzheimer's Disease Diagnosis by using Likelihood Lattice Classification Algorithm", International Journal of Pure and Applied Mathematics, Vol.118, No.7, 2018, pp.563-571, ISSN: 1314-3395.

[20] Dr.M. Balamurugan and A. Nancy, and S. Vijaykumar, "Alzheimer's Disease Diagnosis by using Dimensionality Reduction Based on KNN Classifier", Biomedical & Pharmacology Journal, Vol.10, Issue 4, December 2017, pp. 1823-1830, http://dx.doi.org/10.13005/bpj/1299.

[21] Vijaykumar Selvam, Dr. M. Balamurugan, A. Nancy, Saravanakumar S. G., "Unique Sense: A Smart Computing Prototype 2", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 3, Issue 3, pp.2024-2031, March-April.2018.

[22] S Vijaykumar, M. Balamurugan, K Ranjani, "Big Data: Hadoop Cluster Deployment on ARM Architecture", International Journal of Advanced Research in Computer and Communication Engineering (IJARCCE), vol. 4, no. 1, June 2015