MAXIMAL MARGINAL RELEVANCE BASED MALAYALAM TEXT SUMMARIZATION WITH SUCCESSIVE THRESHOLDS

Ajmal E B[1] and Rosna P Haroon[2]

[1]Department of CSE, Ilahia College of Engineering and Technology, Muvattupuzha, India
[2]Department of CSE, Ilahia College of Engineering and Technology, Muvattupuzha, India

ABSTRACT

Automatic text summarization has prime importance in the area of Natural Language Processing. As we are aware a large quantity of information are there on web, it is very difficult to extract the needed information from the huge. Text summarization is the process of shorten the document, so that it retains only the important points of the original document. As the problem of information overload has grown, and the quantity of data has assumed a greater significance, the need for an instant summarization of the un-touched language -Malayalam- assumes vital importance. Lots of summarization systems have already been developed for various languages, there is no such well performing system for Malayalam. In this paper propose a Malayalam text summarization system which is based on MMR technique with successive threshold. Here the sentences are selected based on the concept of maximal marginal relevance. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the text document, which can be achieved by using successive threshold approach.

KEYWORDS

Maximum Marginal Relevance, Successive Threshold, Unit step function

1. INTRODUCTION

Automatic text summarization has prime importance in the area of Natural Language Processing. As we are aware a large quantity of information are there on web, it is very difficult to extract the needed information from the huge. Text summarization is the process of shorten the document, so that it retains only the important points of the original document. As the problem of information overload has grown, and the quantity of data has assumed a greater significance, the need for an instant summarization of the untouched language - Malayalam assumes vital importance. Lots of summarization systems have already been developed for various languages, there is no such well performing system for Malayalam. The existing systems have high computational cost, time and storage capacity. To address the issues of computational cost time and storage capacity, here

DOI: 10.5121/ijci.2016.5237

proposes a text summarization system that works on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document, which can be achieved by using successive threshold approach.

Malayalam is the official language of Kerala and there are around 33 million people who speak Malayalam. There is a vast amount of online data available in Malayalam. This warrants creating a tool that can be used to explore digital information presented in Malayalam and other native languages. In this concept, we propose the MMR based Malayalam document Summarization with Successive Thresholds.
Concept is presented in five sections. Section II reviews the related works. Section III discusses the proposed scheme. In section IV, the evaluation of the proposed scheme is presented. Section V concludes the work and future scope is discussed.

2. RELATED WORK

Attempts to automatically summarize documents started early since 1958. The method based on word frequencies by Luhn is one of the oldest but still relevant method. This method measures the importance of a sentence based on the presence of keywords (most frequently occurring words in a document otherthan the stopwords) in the sentence. Text summarization method by Ed-mundson used cue words, title words, and sentence location for determining the sentence weights [4]. Text summarization for Malayalam documents by Rajina Kabeer and Sumam Mary Idicula [1].

Graph theoretical approaches for summarization represents a document as an undirected graph, in which the nodes represent the sentences in the document. Two nodes in the graph are connected if the cosine similarity of the sentences corresponding to the nodes is above some particular threshold. The sentences corresponding to the nodes with the highest cardinality or in other words the sentences which are more similar to other sentences in the document are considered important and are included in the summary [8]. Methods based on Co-reference chains and Lexical chains are based on the semantic structure of the document.

Semantic graph based approaches extracts semantic triplets (Subject-Object-Predicate triplets) from each of the sentence in the document. These triplets are used to generate a graph of the document. A sub-graph of this graph is selected using machine learning techniques and the sentences in the sub-graph are used to form the summary [2].

Machine Learning approaches to summarization models the summarization process as a classification problem. Naïve Bayes method, Neural networks and Hidden Markov Model (HMM) are some of the machine learning approaches [4] used for text summarization.

Information extraction by abstractive text summarization for Telugu language [7], summarization of tamil document using semantic graph method [14], Text extraction for an Agglutinative Language by Sankar K, VijaySundar Ram R and Sobha Lalitha Devi which was used for summarizing Tamil documents [8], Bengalitext summarization by sentence extraction by Kamal Sarkar [6] are some text summarization works done for Indian languages.

3. PROPOSED SCHEME

In the proposed method, a single-document input is summarized based on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance and each word meaning is calculated with the help of a dictionary, finally the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document, which can be achieved by using successive threshold approach.

3.1. MAXIMAL MARGINAL RELEVANCE

The key idea in this technique is to use a unit step function at each step to decide the maximum marginal relevance. The automatic summarization process is explained below:

1. Input a document to be summarized
2. Now the document is traversed and eliminates the words that are not useful (stop word removal)
3. Starting with the starting position of the sentence until the document finishes
4. Identify the most important word/sentence (by meaning) with the help of a malayalam dictionary
5. Using the unit step function we can calculate the relevant information required. The unit step function used in the algorithm is given as:

$$u_k + 1 = \arg\max(Sim1(u_i, Q) - \max(sim2(u_i, u_j)))$$

   Where
   Q : User input document
   u_i : Most important word/sentence
   u_j : Remaining sentences in the document
   U : Selected list of sentences
6. The process may be terminated once an appropriate number of words or sentences are in U. Which can be achieved by using successive threshold approach

3.2. SUCCESSIVE THRESHOLD APPROACH

The concept behind this approach is that the number of sentences present in the summary would be equal to the number of paragraphs or the average number of sentences present in the input text document. That is initially count the total number of paragraphs and sentences in the given text document, if the total number of paragraphs in the input text document is meet a threshold value then take the value of ‘n’ as number of paragraphs otherwise take ‘n’ as average number of sentences in the input document. After applying all the pre-processing steps and the MMR technique to select the relevant information or the sentences from the document. Then counts the total number of sentences say it is ‘m’, if m is equal to ‘n’, then these are the sentences finally included in the summary. Else, repeat the steps of MMR technique until the ‘m’ value will be equal to ‘n’.

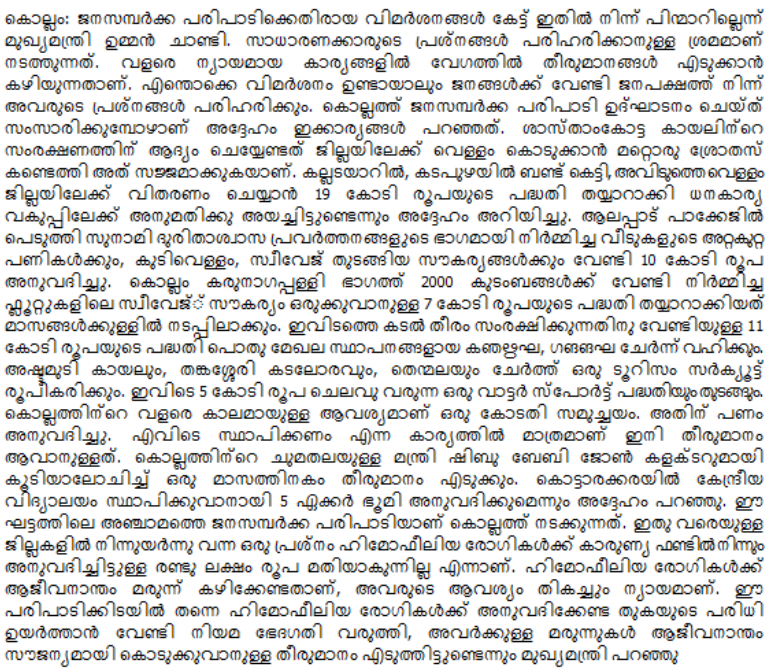
The proposed system uses the following algorithm. The algorithm consists of two sections; the first section uses a unit step function that identifies the maximum marginal relevance that is the

relevant sentences from the input document. Then the next section uses a successive threshold approach. By using this approach the total number of sentences in the final summary can be calculated. Process is explained below:

Input: Malayalam document
Output: Summarized document

1. Input a document to be summarized
2. Now the document is traversed and eliminates the words that are not useful (stop word removal).
3. Identify the most important word (by meaning) with the help of a Malayalam dictionary from the input document
4. Starting with the starting position of the sentence until the document finishes.
5. Using the unit step function and dictionary calculate the first level important sentence from the document by using the important word identified in step 3
   The unit step function used in the algorithm is given as:
   $$u_k + 1 = \arg\max(Sim1(u_i, Q) - \max(sim2(u_i, u_j)))$$
   Where
   Q : User input document
   u_i : Most important word/sentence
   u_j : Remaining sentences in the document
   U : Selected list of sentences
6. Then the second level sentence is identified using the first level sentence and the dictionary by using the unit step function
7. The next level sentence is identified using the sentence identified in step 6 and the dictionary by using the unit step function
8. Repeat the step 7 until an appropriate number of sentences is in U. Which can be achieved by using successive threshold approach
9. Stop

As an example consider the following input text shown in figure 1 and the corresponding output obtained for the text using proposed method is shown in figure 2.



കൊല്ലം: ജനസമ്പർക്ക പരിപാടിക്കെതിരായ വിമർശനങ്ങൾ കേട്ട് ഇതിൽ നിന്ന് പിന്മാറില്ലെന്ന് മുഖ്യമന്ത്രി ഉമ്മൻ ചാണ്ടി. സാധാരണക്കാരുടെ പ്രശ്നങ്ങൾ പരിഹരിക്കാനുള്ള ശ്രമമാണ് നടത്തുന്നത്. വളരെ ന്യായമായ കാര്യങ്ങളിൽ വേഗത്തിൽ തീരുമാനങ്ങൾ എടുക്കാൻ കഴിയുന്നതാണ്. എന്തൊക്കെ വിമർശനം ഉണ്ടായാലും ജനങ്ങൾക്ക് വേണ്ടി ജനപക്ഷത്ത് നിന്ന് അവരുടെ പ്രശ്നങ്ങൾ പരിഹരിക്കും. കൊല്ലത്ത് ജനസമ്പർക്ക പരിപാടി ഉദ്ഘാടനം ചെയ്ത് സംസാരിക്കുമ്പോഴാണ് അദ്ദേഹം ഇക്കാര്യങ്ങൾ പറഞ്ഞത്. ശാസ്താംകോട്ട കായലിന്റെ സംരക്ഷണത്തിന് ആദ്യം ചെയ്യേണ്ടത് ജില്ലയിലേക്ക് വെള്ളം കൊടുക്കാൻ മറ്റൊരു ശ്രോതസ് കണ്ടെത്തി അത് സജ്ജമാക്കുകയാണ്. കല്ലടയാറിൽ, കടപുഴയിൽ ബണ്ട് കെട്ടി, അവിടുത്തെവെള്ളം ജില്ലയിലേക്ക് വിതരണം ചെയ്യാൻ 19 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കി ധനകാര്യ വകുപ്പിലേക്ക് അനുമതിക്കു അയച്ചിട്ടുണ്ടെന്നും അദ്ദേഹം അറിയിച്ചു. ആലപ്പാട് പാക്കേജിൽ പെടുത്തി സുനാമി ദുരിതാശ്വാസ പ്രവർത്തനങ്ങളുടെ ഭാഗമായി നിർമ്മിച്ച വീടുകളുടെ അറ്റകുറ്റ പണികൾക്കും, കുടിവെള്ളം, സ്വീവേജ് തുടങ്ങിയ സൗകര്യങ്ങൾക്കും വേണ്ടി 10 കോടി രൂപ അനുവദിച്ചു. കൊല്ലം കരുനാഗപ്പള്ളി ഭാഗത്ത് 2000 കുടംബങ്ങൾക്ക് വേണ്ടി നിർമ്മിച്ച ഫ്ലൂറ്റുകളിലെ സ്വീവേജ് സൗകര്യം ഒരുക്കുവാനുള്ള 7 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കിയത് മാസങ്ങൾക്കുള്ളിൽ നടപ്പിലാക്കും. ഇവിടത്തെ കടൽ തീരം സംരക്ഷിക്കുന്നതിനു വേണ്ടിയുള്ള 11 കോടി രൂപയുടെ പദ്ധതി പൊതു മേഖല സ്ഥാപനങ്ങളായ കഞ്ഞിഘ, ഗങ്ങഘ ചേർന്ന് വഹിക്കും. അഷ്ടമുടി കായലും, തങ്കശ്ശേരി കടലോരവും, തെന്മലയും ചേർത്ത് ഒരു ടൂറിസം സർക്യൂട്ട് രൂപീകരിക്കും. ഇവിടെ 5 കോടി രൂപ ചെലവു വരുന്ന ഒരു വാട്ടർ സ്പോർട്ട് പദ്ധതിയും തുടങ്ങും. കൊല്ലത്തിന്റെ വളരെ കാലമായുള്ള ആവശ്യമാണ് ഒരു കോടതി സമുച്ചയം. അതിന് പണം അനുവദിച്ചു. എവിടെ സ്ഥാപിക്കണം എന്ന കാര്യത്തിൽ മാത്രമാണ് ഇനി തീരുമാനം ആവാനുള്ളത്. കൊല്ലത്തിന്റെ ചുമതലയുള്ള മന്ത്രി ഷിബു ബേബി ജോൺ കളക്ടറുമായി കൂടിയാലോചിച്ച് ഒരു മാസത്തിനകം തീരുമാനം എടുക്കും. കൊട്ടാരക്കരയിൽ കേന്ദ്രിയ വിദ്യാലയം സ്ഥാപിക്കുവാനായി 5 ഏക്കർ ഭൂമി അനുവദിക്കുമെന്നും അദ്ദേഹം പറഞ്ഞു. ഈ ഘട്ടത്തിലെ അഞ്ചാമത്തെ ജനസമ്പർക്ക പരിപാടിയാണ് കൊല്ലത്ത് നടക്കുന്നത്. ഇതു വരെയുള്ള ജില്ലകളിൽ നിന്നുയർന്നു വന്ന ഒരു പ്രശ്നം ഹിമോഫീലിയ രോഗികൾക്ക് കാരുണ്യ ഫണ്ടിൽനിന്നും അനുവദിച്ചിട്ടുള്ള രണ്ടു ലക്ഷം രൂപ മതിയാകുന്നില്ല എന്നാണ്. ഹിമോഫീലിയ രോഗികൾക്ക് ആജീവനാന്തം മരുന്ന് കഴിക്കേണ്ടതാണ്, അവരുടെ ആവശ്യം തികച്ചും ന്യായമാണ്. ഈ പരിപാടിക്കിടയിൽ തന്നെ ഹിമോഫീലിയ രോഗികൾക്ക് അനുവദിക്കേണ്ട തുകയുടെ പരിധി ഉയർത്താൻ വേണ്ടി നിയമ ഭേദഗതി വരുത്തി, അവർക്കുള്ള മരുന്നുകൾ ആജീവനാന്തം സൗജന്യമായി കൊടുക്കുവാനുള്ള തീരുമാനം എടുത്തിട്ടുണ്ടെന്നും മുഖ്യമന്ത്രി പറഞ്ഞു

Figure 1. Input Text

കല്ലടയാറിൽ, കടപുഴയിൽ ബണ്ട് കെട്ടി, അവിടുത്തെ വെള്ളം ജില്ലയിലേക്ക് വിതരണംചെയ്യാൻ 19 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കി ധനകാര്യ വകുപ്പിലേക്ക് അനുമതിക്കു അയച്ചിട്ടുണ്ടെന്നും അദ്ദേഹം അറിയിച്ചു. ആലപ്പാട് പാക്കേജിൽ പെടുത്തി സുനാമി ദുരിതാശ്വാസപ്രവർത്തനങ്ങളുടെ ഭാഗമായി നിർമ്മിച്ച വീടുകളുടെ അറ്റകുറ്റ പണികൾക്കും, കുടിവെള്ളം, സ്വീവേജ് തുടങ്ങിയ സൗകര്യങ്ങൾക്കും വേണ്ടി 10 കോടി രൂപ അനുവദിച്ചു. കൊല്ലം കരുനാഗപ്പള്ളി ഭാഗത്ത് 2000 കുടംബങ്ങൾക്ക് വേണ്ടി നിർമ്മിച്ച ഫ്ലൂറ്റുകളിലെ സ്വീവേജ് സൗകര്യം ഒരുക്കുവാനുള്ള 7 കോടി രൂപയുടെ പദ്ധതി തയ്യാറാക്കിയത് മാസങ്ങൾക്കുള്ളിൽ നടപ്പിലാക്കും. ഇവിടത്തെ കടൽ തീരം സംരക്ഷിക്കുന്നതിനു വേണ്ടിയുള്ള 11 കോടി രൂപയുടെ പദ്ധതി പൊതു മേഖല സ്ഥാപനങ്ങളായ കഞ്ഞിഘ, ഗങ്ങഘ ചേർന്ന് വഹിക്കും.

Figure 2. Output Text

2. EVALUATION

As shown in the below table is the different parameter evaluation of the existing method. The method is implemented on 6 different dataset of different sizes and various parameters such as precision, recall and F-measure is calculated.

Table 1. Parameter evaluation – Existing method

Dataset	**Precision**	**Recall**	**F- Measure**
Dataset1	0.485	0.525	0.530
Dataset2	0.5309	0.5765	0.5665
Dataset3	0.5807	0.6635	0.5978
Dataset4	0.6679	0.7814	0.7449
Dataset5	0.656	0.7756	0.7645
Dataset6	0.772	0.7901	0.7801

Table 2 shows parameter evaluation of the proposed method. The method is implemented on 6 different dataset of different sizes and various parameters such as precision, recall and F-measure is calculated.

Table 2. Parameter evaluation – Proposed method

Dataset	**Precision**	**Recall**	**F- Measure**
Dataset1	0.535	0.565	0.543
Dataset2	0.5407	0.5805	0.5785
Dataset3	0.5917	0.6743	0.6537
Dataset4	0.6779	0.7896	0.7549
Dataset5	0.673	0.7826	0.7775
Dataset6	0.852	0.8910	0.8018

The following chart shows the comparison of proposed MMR method with existing Sentence scoring method.

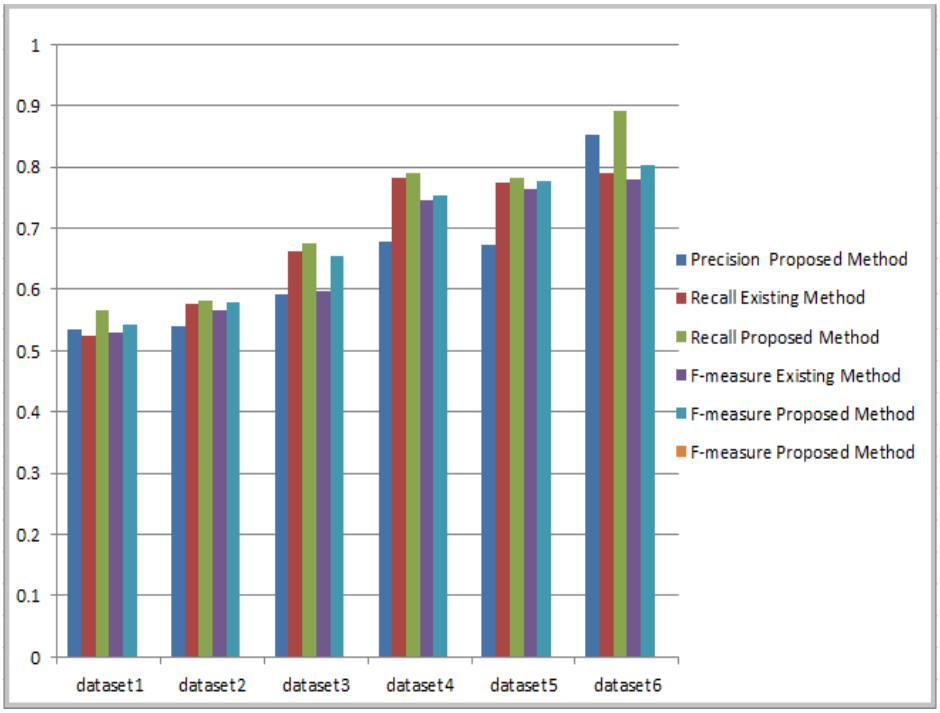


Figure 3. Graphical evaluation

3. CONCLUSIONS AND FUTURE WORK

The text summarization provides the summary of the input document. Here in this concept an efficient technique of document summarization is proposed. The proposed method is works on the concept of maximal marginal relevance between the sentences or the words. The key idea is to use a unit step function at each step to decide the maximum marginal relevance, and the number of sentences present in the summary would be equal to the number of paragraphs

or the average number of sentences in the input text document, which can be achieved by using successive threshold approach. Analysis shows that proposed method is more accurate. More quality parameters are generated by incorporate another methods is future work

ACKNOWLEDGEMENTS

The authors would like to thank, Prof. Rosna P Haroon, Head of the Department, Department of Computer Science and Engineering, Ilahia College of Engineering And Technology,Muvattupuzha, kerala, for her timely advices and suggestions.

REFERENCES

[1] Kabeer, Rajina, and Sumam Mary Idicula (2014) "Text summarization for Malayalam documents—An experience." *2014 IEEE International Conference On Data Science & Engineering (ICDSE).*

[2] Leskovec, J., Milic-Frayling, N., Grobelnik, M. and Leskovec, J., (2005) "Extracting summary sentences based on the document semantic graph", *Microsoft Research, Microsoft Corporation*.

[3] Bijalwan, Vishwanath, Pinki Kumari, Jordan Pascual, and Vijay Bhaskar Semwal (2014) "Machine learning approach for text and document mining." *arXiv preprint arXiv:1406.1580*

[4] Gupta, Vishal, and Gurpreet Singh Lehal. (2010) "A survey of text summarization extractive techniques." *Journal of Emerging Technologies in Web Intelligence*2, no. 3: 258-268..

[5] Dipanjan Das, Andre F.T. Martins, (2007) "A Survey on Automatic Text Summarization, Language Technologies Institute", Carnegie Mellon University.

[6] Sarkar, K., August. (2012) "An approach to summarizing Bengali news documents." In *proceedings of the International Conference on Advances in Computing, Communications and Informatics* (pp. 857-862). ACM..

[7] Jagadish S Kallimani, Srinivasa KG, Eswara Reddy B, (2011) "Information Extraction by an Abstractive Text Summarization for an Indian Rgional Language Natural Language Processing and Knowledge Engineering (NLP-KE)", *2011 7th International Conference on* Natural Language Processing and Knowledge Engineering.

[8] Sankar K, VijaySundar Ram R and Sobha Lalitha Devi (2011) "Text Extraction for an Agglutinative Language"

[9] Martin Hassel, (2004) "Evaluation of AutomaticText Summarization: A practical implementation".

[10] Bindu.M.S, Sumam Mary Idicula, (2011) "A Hybrid Model For Phrase Chunking Employing Artificial Immunity System And Rule Based Methods", *International Journal of Artificial Intelligence Applications(IJAIA)*, Vol.2, No.4.

[11] Rajeev RR, Rajendran N, Elizabeth Sherly, (2005) "A Suffix Stripping Based Morph Analyser For Malayalam Language", *Proceedings of 20th Kerala Science Congress*.

[12] Jayashree.R, SrikantaMurthy.K,Sunny, (2011) , " Keyword extraction based summarizationof categorised Kannada text documents*", International Journal on Soft Computing (IJSC)* , Vol.2, No.4.

[13] Aysun Guran, Eren Bekar, Selim Akyokus, (2010) "A Comparison of Feature and Semantic-Based Summarization Algorithms for Turkish", *International Symposium on Innovations in Intelligent Systems and Applications*, Kayseri Cappadocia,TURKEY.

[14] Banu, M., Karthika, C., Sudarmani, P. and Geetha, T.V., (2007) "Tamil Document Summarization Using Semantic Graph Method", *IEEEConference on Computational Intelligence and Multimedia Applications, 2007.* (Vol. 2, pp. 128-134).

[15] Lin, Hai, Lusheng Wang, and Ruoshan Kong. (2015) "Energy Efficient Clustering Protocol for Large-Scale Sensor Networks." *Sensors Journal, IEEE* 15, no. 12 7150-7160.

[16] Liu, Dawei, Saifeng Cai, and Xiaohong Guo. "Incremental sequential pattern mining algorithms of Web site access in grid structure database." *Neural Computing and Applications*: 1-9.

[17] Demertzis, Kostantinos, Lazaros Iliadis, Stavros Avramidis, and Yousry A. El-Kassaby. "Machine learning use in predicting interior spruce wood density utilizing progeny test information." *Neural Computing and Applications*: 1-15.

[18] Sahoo, G. "A two-step artificial bee colony algorithm for clustering." *Neural Computing and Applications*: 1-15.

[19] Stein, Procópio, Anne Spalanzani, Vitor Santos, and Christian Laugier (2014) "Leader following: A study on classification and selection." *Robotics and Autonomous Systems*

[20] Acampora, Giovanni, Matteo Gaeta, and Vincenzo Loia (2011), "Hierarchical optimization of personalized experiences for e-Learning systems through evolutionary models", *Neural Computing and Applications*,Vol.20, No.5, PP 641-657.

Authors

Ajmal E B received his **B Tech degree** in Computer Science and Engineering from Ilahia College of Engineering and Technology, Muvattupuzha, Kerala, India in the year 2013 and received **M Tech degree** in Computer science and engineering from the same college in the year 2015. His research interests include Natural Language Processing, Data mining and Evolutionary Algorithms, and Digital Image Processing.