International Journal of Advanced Trends in Engineering and Technology (IJATET)
International Peer Reviewed - Refereed Research Journal, Website: www.dvpublication.com
Impact Factor: 5.965, ISSN (Online): 2456 - 4664, Volume 10, Issue 1, January - June, 2025

EFFECTIVE PATTERN DISCOVERY FOR TEXT MINING USING HIDDEN PATTERN FILTER SORTING TECHNIQUES

A. Indhuja*, V. Alamelu Mangayarkarasi & G. R. Gnana Raja*****
* Research Scholar, Department of Computer Science, S.T.E.T Women's College, Mannargudi, Tamil Nadu, India
** Assistant Professor, Department of Computer Science, S.T.E.T Women's College, Mannargudi, Tamil Nadu, India
*** Assistant Professor, Department of English, Khadir Mohideen College, Adirampattinam, Tamil Nadu, India

Cite This Article: A. Indhuja, V. Alamelu Mangayarkarasi & G. R. Gnana Raja, "Effective Pattern Discovery for Text Mining Using Hidden Pattern Filter Sorting Techniques", International Journal of Advanced Trends in Engineering and Technology, Volume 10, Issue 1, January - June, Page Number 13-17, 2025.

Abstract:

In this paper, the emergence of the internet, billions of websites were created, which made it hard for the average user to extract useful information from the web efficiently for a specific search. For resolving a problem we can make exact information search system. In this system admin can upload the data's and also mention the important words are as keywords. If users search some information, first we can separate as individual keywords that can compare with the uploaded files keywords to finds the exact matching information. And also it have some symbolic meaning to symbols (+, - etc...) if we specify (+) between words it add the information if we specify (-) between words it eliminates that specific information. According to this we can make a search process easily and also make an accuracy results to the users.

Key Words: Text Mining, Pattern Filter, Sorting Techniques.

1. Introduction:

Text mining is the discovery of interesting knowledge in text documents. It is challenging issue to find accurate knowledge in text documents to help users to find what they want. Many applications, such as market analysis and business management, can benefit by the use of the information and knowledge extracted from a large amount of data. Knowledge discovery can be effectively use and update discovered patterns and apply it to field of text mining. Data mining is therefore an essential step in the process of knowledge discovery in databases, which means data mining is having all methods of knowledge discovery process and presenting modeling phase that is application of methods and algorithm for calculation of search pattern or models. In the past decade, a significant number of data mining techniques have been presented in order to perform different knowledge tasks. These techniques include association rule mining, frequent item set mining, sequential pattern mining, maximum pattern mining and closed pattern mining. Most of them are proposed for the purpose of developing efficient mining algorithms to find particular patterns within a reasonable and acceptable time frame. With a large number of patterns generated by using the data mining approaches, how to effectively exploit these patterns is still an open research issue. Text mining is the technique that helps users find useful information from a large amount of digital text data. It is therefore crucial that a good text mining model should retrieve the information that users require with relevant efficiency. Traditional Information Retrieval (IR) has the same objective of automatically retrieving as many relevant documents as possible whilst filtering out irrelevant documents at the same time. However, IR-based systems do not adequately provide users with what they really need. Many text mining methods have been developed in order to achieve the goal of retrieving for information for users. We focus on the development of a knowledge discovery model to effectively use and update the discovered patterns and apply it to the field of text mining.

2. Related Work:

Fast Algorithms for Mining Association Rules:

In this paper consider the problem of discovering association rules between items in a large database of sales transactions. We present two new algorithms for solving the problem that are fundamentally different from the known algorithms. Empirical evaluation shows that these algorithms outperform the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problems. We also show how the best features of the two proposed algorithms can be combined into a hybrid algorithm, called Apriori Hybrid. Scale-up experiments show that Apriori Hybrid scales linearly with the number of transactions. Apriori Hybrid also has excellent scale-up properties with respect to the transaction size and the number of items in the database

Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections:

Traditionally, texts have been analyzed using various information retrieval related methods, such as full-text analysis, and natural language processing. However, only few examples of data mining in text, particularly in full text, are available. In this paper we show that general data mining methods are applicable to text analysis tasks such as descriptive phrase extraction. Moreover, we present a general framework for text mining. The framework follows the general knowledge discovery process, thus containing steps from preprocessing to the utilization of the results. The data mining method that we apply is based on generalized episodes and episode rules. We give concrete examples of how to preprocess texts based on the intended use of the discovered results and we introduce a weighting scheme that helps in pruning out redundant or non-descriptive phrases. We also present results from real-life data experiments.

Kernel Methods for Document Filtering:

This paper describes the algorithms implemented by the KerMIT consortium for its participation in the Trec 2002 Filtering track. The consortium submitted runs for the routing task using a linear SVM, for the batch task using the same SVM in combination with an innovation threshold-selection mechanism, and for the adaptive task using both a second-order perceptron and a combination of SVM and perceptron with uneven margin. Results seem to indicate that these algorithm performed relatively

well on the extensive TREC benchmark.

Statistical Phrases in Automated Text Categorization:

In this work we investigate the usefulness of n-grams for document indexing in text categorization (TC). We call n-gram a set t k of n word stems, and we say that t k occurs in a document d j when a sequence of words appears in d j that, after stop word removal and stemming, consists exactly of then stems in t k , in some order. Previous researches have investigated the use of n-grams (or some variant of them) in the context of specific learning algorithms, and thus have not obtained general answers on their usefulness for TC. In this work we investigate the usefulness of n-grams in TC independently of any specific learning algorithm. We do so by applying feature selection to the pool of all #-grams (# # n), and checking how many n-grams score high enough to be selected in the top # #-grams. We report the results of our experiments, using several feature selection functions and varying values of #, performed on the Reuters-21578 standard TC benchmark. We also report results of making actual use of the selected n-grams in the context of a linear classifier induced by means of the Rocchio method.

UnderstandingandImprovingPersonalFileRetrievalPersonalfileretrieval - the task of locating and opening files on a computer - is a common task for all computer users. A range of interfaces are available to assist users in retrieving files, such as navigation with in a file browser, search interfaces and recent items lists. This thesis examines two broad goals in file retrieval: understanding current file retrieval behaviour, and improving file retrieval by designing improved user interfaces. A thorough understanding of current file retrieval behaviour is important to the design of any improved retrieval tools, however there has been surprisingly little research about the ways in which users interact with common file retrieval tools. To address this, this thesis describes a longitudinal field study that logs participants' file retrieval behaviour across a range of methods, using a specially developed logging tool called File Monitor. Results confirm findings from previous research that search are used as a method of last resort, while providing new results characterizing file retrieval. These include analyses of revisitation behaviour, file browser window reuse, and interactions between retrieval methods, as well as detailed characterizations of the use of navigation and search.

Data Mining for Web Intelligence:

Through the billions of Web pages created with HTML and XML, or generated dynamically by underlying Web database service engines, the Web captures almost all aspects of human endeavor and provides a fertile ground for data mining. However, searching, comprehending, and using the semi structured information stored on the Web poses a significant challenge because this data is more sophisticated and dynamic than the information that commercial database systems store. To supplement keyword-based indexing, which forms the cornerstone for Web search engines; researchers have applied data mining to Web-page ranking. In this context, data mining helps Web search engines find high-quality Web pages1 and enhances Web click stream analysis. For the Web to reach its full potential, however, we must improve its services, make it more comprehensible, and increase its usability. As researchers continue to develop data mining techniques, we believe this technology will play an increasingly important role in meeting the challenges of developing the intelligent Web.

Mining Frequent Patterns without Candidate Generation:

Mining frequent patterns in transaction databases, time series databases, and many other kinds of databases has been studied popularly in data mining research. Most of the previous studies adopt an Apriori-like candidate set generation-and-test approach. However, candidate set generation is still costly, especially when there exist prolic patterns and/or long patterns. In this study, we propose a novel frequent pattern tree (FP-tree) structure, which is an extended prextree structure for storing compressed, crucial information about frequent patterns, and develop an efficient FP-tree based mining method, FP-growth, for mining the complete set of frequent patterns by pattern fragment growth. affiance of mining is achieved with three techniques: (1) a large database is compressed into a highly condensed, much smaller data structure, which avoids costly, repeated database scans, (2) our FP-tree-based mining adopts a pattern fragment growth method to avoid the costly generation of a large number of candidate sets, and (3) a partitioning-based, divide-and-conquer method is used to decompose the mining task into a set of smaller tasks for mining conned patterns in conditional databases, which dramatically reduces the search space. Our performance study shows that the FP-growth method is scalable for mining both long and short frequent patterns, and is about an order of magnitude faster than the Apriori algorithm and also faster than some recently reported new frequent pattern mining methods

Mining sequential patterns using graph search techniques:

Sequential patterns discovery had emerged as an important problem in data mining. In this paper, we propose an effective GST algorithm for mining sequential patterns in a large transaction database. Different from the apriori-like algorithms, the GST algorithm can out of order find large k-sequences (k >= 3);i.e., we can find large k-sequences not directly through large (k-1)-sequences. This leads to that our algorithm has much better performance than the Apriori-like algorithms. Besides, we also propose the method to find new sequential patterns by scanning only new transactions since the database was increased. Through several comprehensive experiments, the GST algorithm gains a significant performance improvement over the Apriori-like algorithms. Also we found as long as the ratio of the items purchased in new transactions is always much better than scanning the entire database.

Identifying Comparative Sentences in Text Documents:

This paper studies the problem of identifying comparative sentences in text documents. The problem is related to but quite different from sentiment/opinion sentence identification or classification. Sentiment classification studies the problem of classifying a document or a sentence based on the subjective opinion of the author. An important application area of sentiment/opinion identification is business intelligence as a product manufacturer always wants to know consumers' opinions on its products. Comparisons on the other hand can be subjective or objective. Furthermore, a comparison is not concerned with an object in isolation. Instead, it compares the object with others. An example opinion sentence is "the sound quality of CD player X is poor". An example comparative sentence is "the sound quality of CD player X is not as good as that of CD player Y". Clearly, these two sentences give different information. Their language constructs are quite different too. Identifying comparative sentences is also useful in practice because direct comparisons are perhaps one of the most convincing ways of evaluation, which may even be more important than opinions on each individual object. This paper proposes to study the comparative sentence

identification problem. It first categorizes comparative sentences into different types, and then presents a novel integrated pattern discovery and supervised learning approach to identifying comparative sentences from text documents. Experiment results using three types of documents, news articles, consumer reviews of products, and Internet forum postings, show a precision of 79% and recall of 81%. More detailed results are given in the paper.

A Probabilistic Analysis of the ROCCHIO Algorithm with TFIDF for Text Categorization:

The Rocchio relevance feedback algorithm is one of the most popular and widely applied learning methods from information retrieval. Here, a probabilistic analysis of this algorithm is presented in a text categorization framework. The analysis gives theoretical insight into the heuristics used in the Rocchio algorithm, particularly the word weighting scheme and the similarity metric. It also suggests improvements which lead to a probabilistic variant of the Rocchio classier. The Rocchio classier, its probabilistic variant, and a naive Ba yes classier are compared on six text categorization tasks. The results show that the probabilistic algorithms are preferable to the heuristic Rocchio classier not only because they are more well-founded, but also because they achieve better performance

3. Effective Pattern Discovery for Text Mining:

String Identification Techniques form a structured data on data mining techniques. This system introduces a notion of search pattern privacy, which gives a measure of security against the leakage from trapdoor. We have shown that our scheme is secure under search pattern ineffective definition.

Login Module:

Here admin has to login by using their unique username and password. Admin is the only authorized person to access this module for security purpose. So other users don't get rights to access this module for their purpose.

Registration Module:

In this module, user can register the details and get unique username and password. Using this user can login to the system each time. And also have facility if the user forgot the password we can change the password using any personal queries.

Admin Module:

Here admin can upload the data for user. Admin has the authority to upload the useful information and mention some important words as keywords. These keywords helps user to get exact information from the server.

Search Module:

In this module user can search data which admin has uploaded. The search module helps user to extract useful information from web efficiently for the specific search. The data's will be available relevant to the user's search.

Separate Module:

This module helps user to search some information and separate as individual keywords that can compare with the uploaded files keywords to finds the exact information. This module contains some symbolic meaning, if we specify (+) symbol between words, it add the information. If we specify (-), it eliminates specific information.

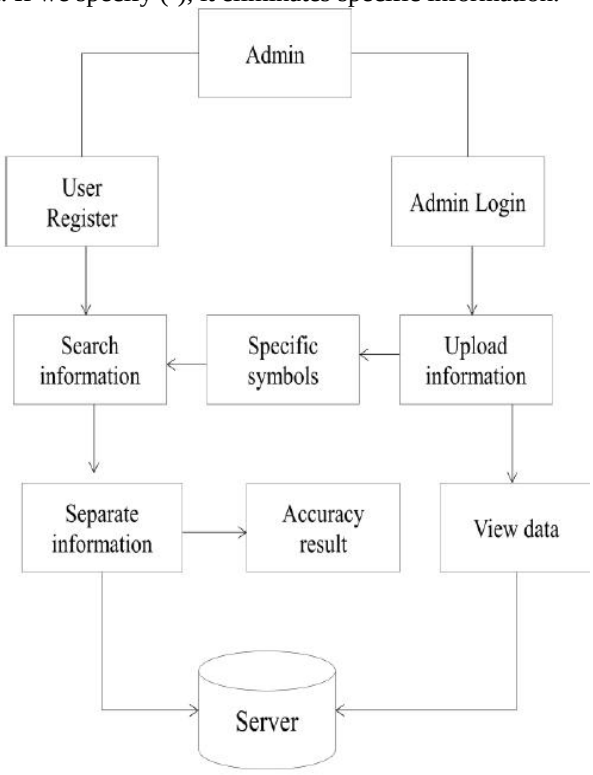

Figure 1: Workflow

4. Experimental Results:

Text mining process can work with unstructured or semi-structured data to convert numerical values which can be better solution for structured data on data mining techniques, to improve the effectiveness of using and updating discovered patterns for finding relevant and filtering information.

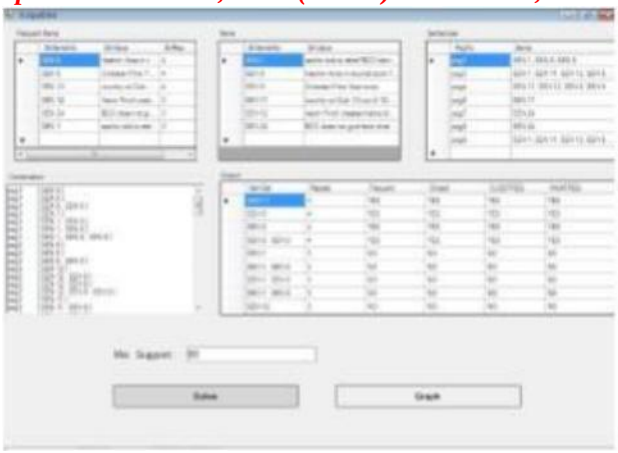

Figure 2: Data Set

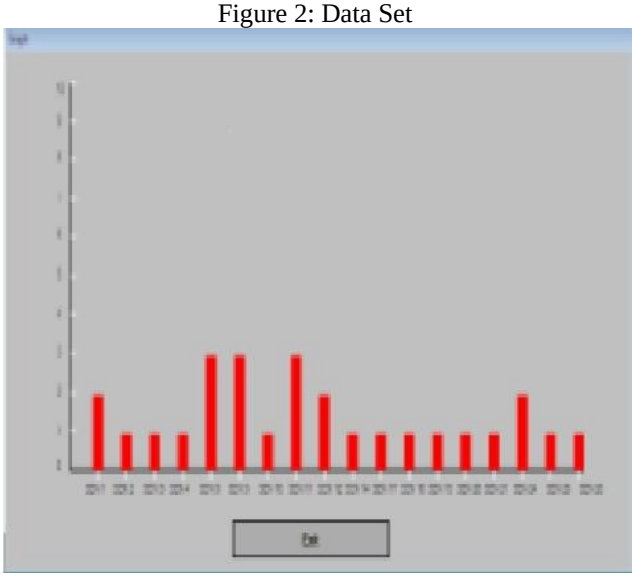

Figure 3: Efficiency

5. Conclusion:

In this paper, we proposed the Hidden pattern filter sorting techniques uses the context variable in pattern mining (feature extraction) and to eliminate the candidate generation is proposed. This paper mainly focused on developing the effective mining algorithms for discovering patterns from large volume of data. To improve the effectiveness of using and updating discovered patterns for finding relevant and filtering information. The experimental results improve the performance and specific file filtering through user indexing search method. In our future work we will investigate better means of exploration of long patterns and look at more diverse kinds of texts, especially large collections of text where a two level hierarchy may not be sufficient. We will also support the filtering of patterns by their usage trend over time. Metrics can be defined to characterize frequency distributions associated with each pattern and identify that are increasing, decreasing, showing spikes or gaps, etc. Finally, we have focused here on patterns of repetitions; other features can be extracted from the text (e.g. name entities, part of speech patterns) and explored in a similar fashion.

6. References:

1. K. Aas and L. Eikvil, “Text Categorisation: A Survey,” Technical Report NR 941, Norwegian Computing Center, 1999.
2. R. Agrawal and R. Srikant, “Fast Algorithms for Mining Association Rules in Large Databases,” Proc. 20th Int’l Conf. Very Large Data Bases (VLDB ’94), pp. 478-499, 1994.
3. H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, “Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections,” Proc. IEEE Int’l Forum on Research and Technology Advances in Digital Libraries (ADL ’98), pp. 2-11, 1998.
4. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.
5. N. Cancedda, N. Cesa-Bianchi, A. Conconi, and C. Gentile, “Kernel Methods for Document Filtering,” TREC, 2002.
6. N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders, “Word- Sequence Kernels,” J. Machine Learning Research, vol. 3, pp. 1059- 1082, 2003.
7. M.F. Caropreso, S. Matwin, and F. Sebastiani, “Statistical Phrases in Automated Text Categorization,” Technical Report IEI-B4-07- 2000, Instituto di Elaborazione dell’ Informazione, 2000.
8. C. Cortes and V. Vapnik, “Support-Vector Networks,” Machine Learning, vol. 20, no. 3, pp. 273-297, 1995.
9. S.T. Dumais, “Improving the Retrieval of Information from External Sources,” Behavior Research Methods, Instruments, and Computers, vol. 23, no. 2, pp. 229-236, 1991.
10. J. Han and K.C.-C. Chang, “Data Mining for Web Intelligence,” Computer, vol. 35, no. 11, pp. 64-70, Nov. 2002.
11. J. Han, J. Pei, and Y. Yin, “Mining Frequent Patterns without Candidate Generation,” Proc. ACM SIGMOD Int’l Conf.

Management of Data (SIGMOD '00), pp. 1-12, 2000.
12. VA Mangayarkarasi, M. V. Srinath, A survey on agile testing mechanism with directed acyclic graph(DAG) based model in various platform" in international journal of Australian journal of basic applied sciences issue 13 ,volume 8, pages 266-273 , august 2014,ISSN 1991-8178, impact factor :0.425.
13. VA Mangayarkarasi, M. V. Srinath, An Efficient DAGbM-KSJS Algorithm for Agile Software Testing, International Review on Computers and Software (I.RE.CO.S.), Vol. 11, N. 10 ISSN 1828-6003 October 2016
14. VA Mangayarkarasi, M. V. Srinath. " A Novel Prioritized Deciding Factor(PDF) Approach for Directed Acyclic Graph(DAG) Based Test Case Prioritization using Agile Testing Methodology", International Journal of Computing Algorithm Volume: 05 Issue: 02 December 2016 Page No.72-78 ISSN: 2278-2397.
15. VA Mangayarkarasi, M.V.Srinath. "Big data management using NOSQL", International Journal of scientific transactions in environment and Technovation, ISSN: 0973-9157, Vol. 10(1),July-sep 2016,Page 37-42
16. K Vinayakan, M V Srinath, A Secured On-Demand Routing Protocol for Mobile Ad-Hoc Network, A Literature Survey, Vol 6, No 6, 2015, 598-604
17. K Vinayakan, M V Srinath, Reinforcing Secure on-Demand Routing Protocol in Mobile AD-Hoc Network Using Dual Cipher based Cryptography, International Journal of Control Theory and Applications, Vol. 10, No 23, 2017, 103-109
18. K Vinayakan, M V Srinath, Security Mandated Analytics based Route Processing with Digital Signature [SMARPDS] - Pseudonymous Mobile Ad Hoc Routing Protocol, Indonesian Journal of Electrical Engineering and Computer Science, Vol 10, No 2, 2018, 763-769
19. K Vinayakan, M V Srinath, A Adhiselvam, Security for Multipath Routing Protocol using Trust based AOMDV in MANETs, Vol. 2 No. 43, 2022, 1640-1654
20. K Vinayakan, M V Srinath, A Adhiselvam, Reinforced Securing of Data Leakage in Mobile Ad hoc Network (MANET) by Hybrid Mechanism of Identity Based Encryption (IBE), International Journal of Health Sciences, Volume 6. No S8, 2022, 3622-3635
21. S Sujatha, K Vinayakan, The Role of Collaborative Learning in Mathematics Education: A Review of Research and Practice, Indo American Journal of Multidisciplinary Research and Review, Vol 6, No. 2, 2022, 200-206
22. S Sujatha, K Vinayakan, Mathematical Literacy for the Future: A Review of Emerging Curriculum and Instructional Trends, International Journal of Applied and Advanced Scientific Research, Vol 7, No. 2, 2022, 65-71
23. S Sujatha, K Vinayakan, Assessing the Impact of Math Competitions and Challenges on Student Learning: A Review, International Journal of Advanced Trends in Engineering and Technology, Vol 8, No 2, 2023, 62-67
24. S Sujatha, K Vinayakan, Integrating Math and Real-World Applications: A Review of Practical Approaches to Teaching, International Journal of Computational Research and Development, Vol 8, No. 2, 2023, 55-60
25. S Sujatha, K Vinayakan, Engaging Students with Mathematics: A Review of Motivation and Engagement Strategies, International Journal of Interdisciplinary Research in Arts and Humanities, Vol 8, No. 2, 2023, 55-60
26. K Vinayakan, VA Mangayarkarasi, Review: Data Analytics Problems, Unanswered Research Challenges and Big Data Technologies, International Journal of Computational Research and Development, Vol 8, No. 2, 2023, 61-69
27. K Vinayakan, AD Kumar, Classification of Defective Product for Smart Factory through Deep Learning Method, International Journal of Scientific Research and Modern Education, Vol 9, No. 2, 2024, 10-15
28. VA Mangayarkarasi, K Vinayakan, AD Kumar, Secure Cloud Data Storage with a Zero Trust Security Foundational Deep Learning Algorithm, International Journal of Advanced Trends in Engineering and Technology, Vol 9, No. 2, 2024, 87-93
29. K Vinayakan, VA Mangayarkarasi, An Analysis of the Distinctions among IoT Network Cloud, Edge, and Fog Computing, International Journal of Applied and Advanced Scientific Research, Vol 9, No. 2, 2024, 130-134
30. R Raja, AS Reddy, AD Kumar, G Malleswari, An Integrated Model for Storage Analysis Using Blockchain and IoT Services, Second International Conference on Intelligent Cyber Physical Systems and Internet of Things, IEEE, 2024, 285-292
31. AS Reddy, G Malleswari, R Raja, AD Kumar, Health Monitoring System using Ensemble based Learning Mechanism, 8th International Conference on Electronics, Communication and Aerospace Technology, IEEE, 2024, 990-994
32. AS Reddy, R Raja, G Malleswari, AD Kumar, Reinforcement Learning with Fuzzy Neural Network for Medical Data, 8th International Conference on Electronics, Communication and Aerospace Technology, IEEE, 2024, 995-999
33. MS Kumar, K Vinayakan, Cardio Vs Strength Training: Which is Better for Overall Health?, International Journal of Computational Research and Development, Vol 9, No. 2, 2024, 99-105
34. MS Kumar, K Vinayakan, The Science of Strength: Understanding the Principles of Effective Weight Training, Indo American Journal of Multidisciplinary Research and Review, Vol 8, No. 2, 2024, 149-159
35. MS Kumar, K Vinayakan, Building a Sustainable Fitness Routine: Balancing Exercise, Rest, and Nutrition, International Journal of Interdisciplinary Research in Arts and Humanities, Vol 9, No. 2, 2024, 164-169
36. VA Mangayarkarasi, An Capable Re-Cluster Based Panel Collection Using Mst And Heuristic System, International Journal of Research and Analytical Reviews (IJRAR), October 2020,vol 7 (4) 94-100.
37. VA Mangayarkarasi, A Real Time Big Data Analysis Using R" International Journal of Research and Analytical Reviews (IJRAR) February 2021 vol8 (1), 384-389.
38. VA Mangayarkarasi, A. Adhiselvam, Ardubot Path Finder: Road Obstacles Finder through Auto Navigation Using Artificial Intelligence and Internet of Things (Iot)" Tuijin Jishu/Journal of Propulsion Technology,44(6), 6449-6459