Journal of Engineering Sciences

Vol 11, Issue 4 , April/2020
ISSN NO:0377-9254

CLASSIFICATION OF GROCERY ITEMS USING APRIORI ALGORITHM

Upputuri Srihitha[1], Kodali Harshitha[2], Mittapalli Pravallika[3], Peddi Sai Shelshe[4], V.Pavani[5]

[1,2,3,4]IV B.Tech, Department of Information Technology, Vignan's Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.

[5]Asst.Professor, Department of Information Technology, Vignan's Nirula Institute of Technology & Science for Women, Peda Palakaluru, Guntur-522009, Andhra Pradesh, India.

manojpavani81@gmail.com

ABSTRACT

Apriori algorithm is one of the major algorithm used in mining frequent item sets in a large data set. Companies ranging from small scale to large scale all store data in various forms which keeps on growing. The one thing they have in common is the fact that all data stored has common association between them and the fact that some sort of data is repeatedly entered. To determine this association, apriori algorithm is used. Apriori algorithm focuses on determining frequently occurring data and the association between them. The amount of data stored grows twice as fast as the speed of the fastest processor available to analyze it. Its main purpose is to find the association relationship among the large number of database items. It is used to describe the patterns of customers' purchase in the supermarket. This is presented in this paper.

Keywords: Grocery Items, classification, processing, clustering, machine learning.

1. INTRODUCTION

Data mining tasks can be classified into two categories: Descriptive mining and Predictive mining[1][2]. Descriptive mining refers to the method in which the essential characteristics of the data in the database are described. Clustering, Association and Sequential mining are the main tasks involved in the descriptive mining techniques tasks. Predictive mining deduces patterns from the data in a similar manner as predictions[3][4]. Predictive mining techniques include tasks like Classification, Regression and Deviation detection. Mining Frequent Item sets from transaction databases is a fundamental task for several forms of knowledge discovery such as association rules, sequential patterns, and classification[5][6].

Data Mining is the discovery of hidden information found in databases and can be viewed as a step in the knowledge discovery process[7][8]. Data mining functions include clustering,classification, prediction, and link analysis. One of the most important data mining applications is that of mining association rules. Association rules, first introduced in 1993, are used to identify relationships among a set of items in a database[9][10]. These relationships are not based on inherent properties of the data themselves, but rather based on co-occurrence of the data items.

2. LITERATURE REVIEW

Data mining finds interesting patterns from databases such as association rules, correlations, sequences, classifiers, clusters and many more of which the mining of association rules is one of the most popular problems[11][12]. Association rule mining finds interesting association or correlation relationships among a large set of data items. Association rules are derived from the frequent item sets using support and confidence as threshold levels. The sets of items which have minimum support are known as frequent item set[13][14]. The support of an item set is defined as the proportion of transactions in the data set which contain the item set. Confidence is defined as the measure of certainty or trustworthiness associated with each discovered pattern. Association rules derived depends on confidence. Frequent item set generation is done using data mining Apriori Algorithm[15][16].

Association rule mining finds interesting association or correlation relationships among a large set of data items. Association rules are derived from the frequent item sets using support and confidence as threshold levels[17][18]. The most influential algorithm for efficient association rule discovery from market databases is Apriori algorithm which is proposed by this investigation. This algorithm shows good performance with sparse datasets and hence it is considered. The Apriori algorithm[19][20] extracts a set of frequent item sets from the data and then pulls out the rules with the highest information content for different groups of customers by dividing the customers in different clusters[21].

Association rules apply to supermarket transaction data, that is, to examine the customer behavior in terms of the purchased products. Association rules describe how often the items are purchased together[34][35].

Association Rules

Association Rule Mining is defined as:

"Let I= {…} be a set of 'n' binary attributes called items. Let D= {….} be set of transaction called database[37]. Each transaction in D has a unique transaction ID and contains a subset of the items in I. A rule is defined as an implication of form X->Y where X, Y? I and X? Y=?. The set of items X and Y are called antecedent and consequent of the rule respectively[22][23]."

Learning of Association rules is used to find relationships between attributes in large databases. An association rule, A=> B, will be of the form" for a set of transactions, some value of itemset A determines the values of item set B under the condition in which minimum support and confidence are met"[24][25].

Support and Confidence can be represented by the following example:

Bread=> butter [support=2%, confidence-60%]

The above statement is an example of an association rule[27][28][29]. This means that there is a 2% transaction that bought bread and butter together and there are 60% of customers who bought bread as well as butter[30][31].

Support and Confidence for Item set A and B are represented by formulas[33]:

$$\text{Support (A)} = \frac{\text{Number of transaction in which A appears}}{\text{Total number of transactions}}$$

$$\text{Confidence } (A \longrightarrow B) = \frac{\text{Support(AUB)}}{\text{Support(A)}}$$

3. PROPOSED METHOD

One of the most popular data mining methods is to discover frequent item sets from a transaction dataset and develop association rules. Apriori algorithm was the first algorithm that was proposed for frequent item set mining. This algorithm uses two steps "join" and "prune" to reduce the search space. It is an iterative approach to discover the most frequent item sets.

Support is an indication of how frequently the items appear in the data.

Support=P(A U B)/N

Confidence specifies the ratio that how may instances satisfy the rule.

Confidence=P(A U B)/P(A)

Lift is a measure of the performance of a targeting model at predicting or classifying cases.

Lift=P(A ∩ B)/P(A)*P(B)

Apriori says:

The probability that item I is not frequent is if:

- P (I) < minimum support threshold, then I is not frequent.
- P (I+A) < minimum support threshold, then I+A is not frequent, where A also belongs to itemset.
- If an itemset set has value less than minimum support then all of its supersets will also fall below min support, and thus

can be ignored. This property is called the Antimonotone property.

The steps followed in the Apriori Algorithm of data mining are:

1. **Join Step**: This step generates (K+1) itemset from K-itemsets by joining each item with itself.
2. **Prune Step**: This step scans the count of each item in the database. If the candidate item does not meet minimum support, then it is regarded as infrequent and thus it is removed. This step is performed to reduce the size of the candidate itemsets [41] [42].

Steps in Apriori

Apriori algorithm is a sequence of steps to be followed to find the most frequent itemset in the given database. This data mining technique follows the join and the prune steps iteratively until the most frequent itemset is achieved. A minimum support threshold is given in the problem or it is assumed by the user.

1) In the first iteration of the algorithm, each item is taken as a 1-itemsets candidate. The algorithm will count the occurrences of each item.

2) Let there be some minimum support, min_sup. The set of 1 – itemsets whose occurrence is satisfying the min sup are determined. Only those candidates which count more than or equal to min_sup, are taken ahead for the next iteration and the others are pruned [43].

3) Next, 2-itemset frequent items with min_sup are discovered [44]. For this in the join step, the 2-itemset is generated by forming a group of 2 by combining items with itself.

4) The 2-itemset candidates are pruned using min-sup threshold value. Now the table will have 2 –itemsets with min-sup only.

5) The next iteration will form 3 –itemsets using join and prune step. This iteration will follow antimonotone property where the subsets of 3-itemsets, that is the 2 –itemset subsets of each group fall in min_sup. If all 2-itemset subsets are frequent then the superset will be frequent otherwise it is pruned [45].

6) Next step will follow making 4-itemset by joining 3-itemset with itself and pruning if its subset does not meet the min_sup criteria. The algorithm is stopped when the most frequent itemset is achieved [46].

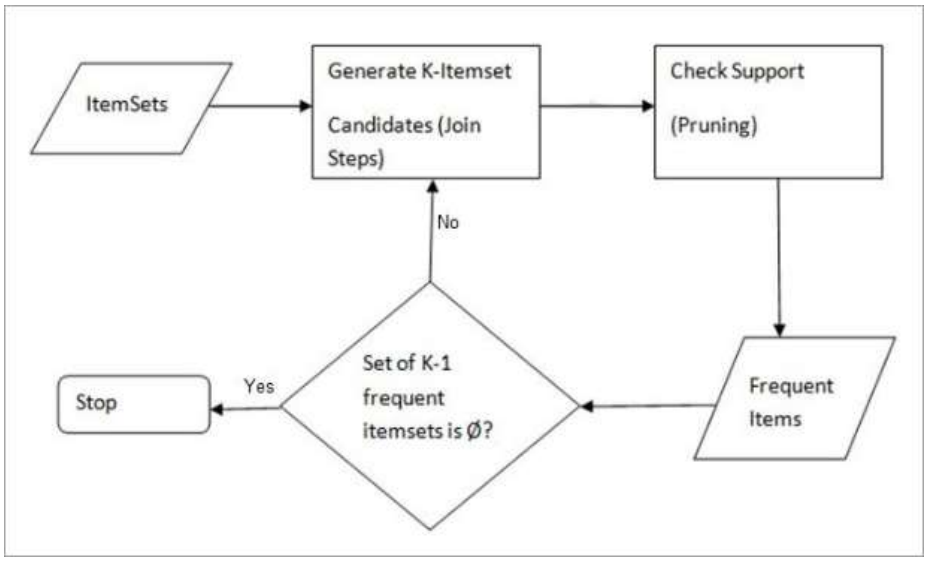

Fig 1: Apriori Algorithm Flowchart

The Apriori Algorithm: Pseudo Code

C: Candidate item set of size k

L: Frequent itemset of size k

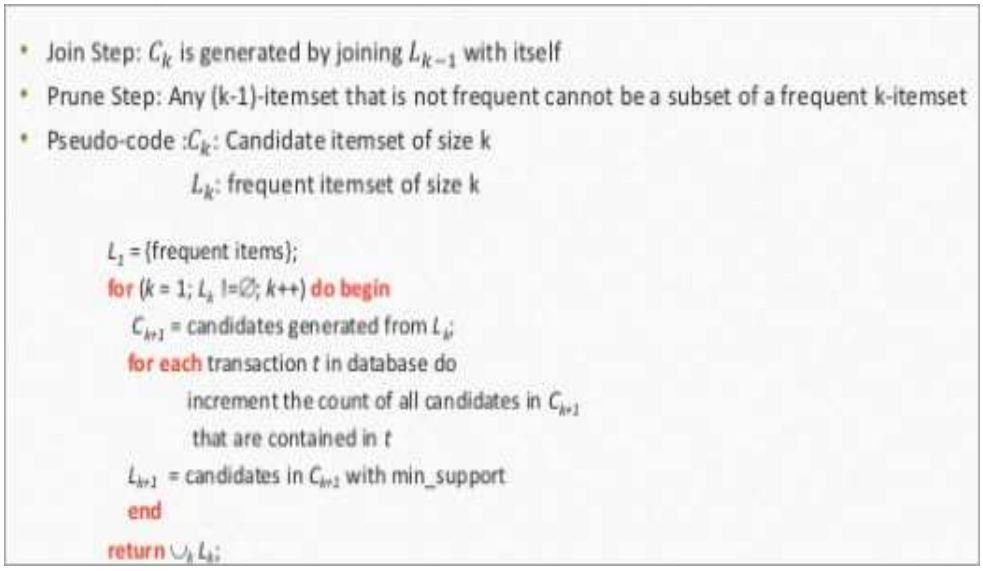

Fig 2: pseudocode for Apriori Algorithm

Methods to Improve Apriori Efficiency

Many methods are available for improving the efficiency of the algorithm.

1. **Hash-Based Technique:** This method uses a hash-based structure called a hash table for generating the k-itemsets and its corresponding count. It uses a hash function for generating the table.
2. **Transaction Reduction:** This method reduces the number of transactions scanning in iterations. The transactions which do not contain frequent items are marked or removed.
3. **Partitioning:** This method requires only two database scans to mine the frequent itemsets. It says that for any itemset to be potentially frequent in the database, it should be frequent in at least one of the partitions of the database [47] [48].
4. **Sampling:** This method picks a random sample S from Database D and then searches for frequent itemset in S. It may be possible to lose a global frequent itemset. This can be reduced by lowering the min_sup.
5. **Dynamic Itemset Counting:** This technique can add new candidate itemsets at any marked start point of the database during the scanning of the database.

4. RESULTS

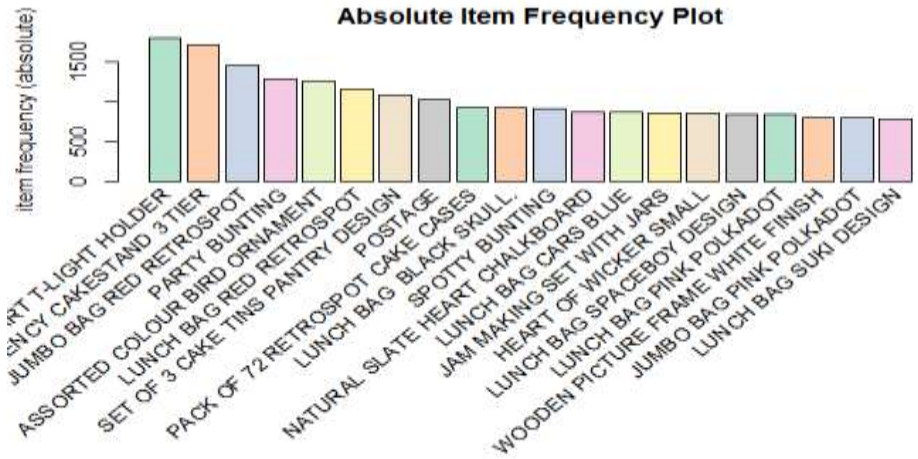

Fig 3:-Absolute Item Frequency Plot

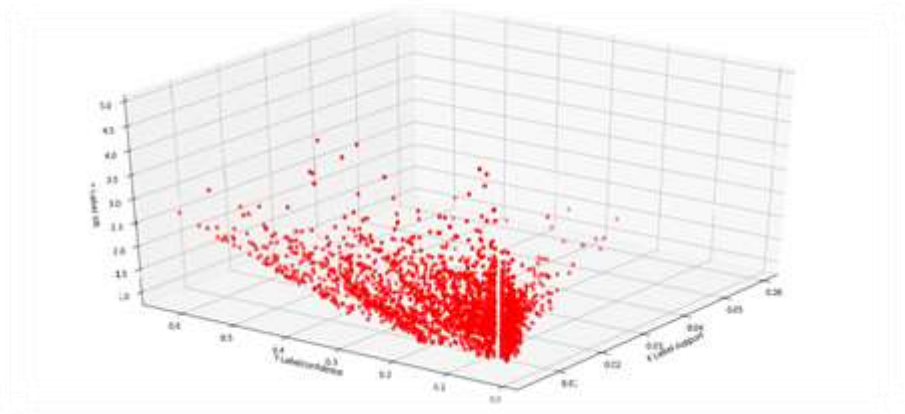

Fig 4 : Apriori algorithm generates highly informative frequent item sets

5. CONCLUSION

Apriori algorithm effectively generates highly informative frequent item sets and association rules for the Stores. Stores widely used the market basket analyses to manage the placement of goods in their store layout. Related products are placed together in such a manner that customers can logically find items he/she might buy which increases the customer satisfaction and hence the profit. Customers are segmented and association rules are separately generated to satisfy their specific needs in a cost effective manner using some special promotions for the common groups. From the results it is shown that the market basket analysis using Apriori algorithm for stores improves its overall revenue.

References

[1]. Lakshman Narayana Vejendla and A Peda Gopi, (2019),” Avoiding Interoperability and Delay in Healthcare Monitoring System Using Block Chain Technology”, Revue d'Intelligence Artificielle , Vol. 33, No. 1, 2019,pp.45-48.

[2]. Gopi, A.P., Jyothi, R.N.S., Narayana, V.L. et al. (2020), “Classification of tweets data based on polarity using improved RBF kernel of SVM” . Int. j. inf. tecnol. (2020). https://doi.org/10.1007/s41870-019-00409-4.

[3]. A Peda Gopi and Lakshman Narayana Vejendla, (2019),” Certified Node Frequency in Social Network Using Parallel Diffusion Methods”, Ingénierie des Systèmes d' Information, Vol. 24, No. 1, 2019,pp.113-117.. DOI: 10.18280/isi.240117

[4]. Lakshman Narayana Vejendla and Bharathi C R ,(2018),"Multi-mode Routing Algorithm with Cryptographic Techniques and Reduction of Packet Drop using 2ACK scheme in MANETs", Smart Intelligent Computing and Applications, Vo1.1, pp.649-658. DOI: 10.1007/978-981-13-1921-1_63 DOI: 10.1007/978-981-13-1921-1_63

[5]. Lakshman Narayana Vejendla and Bharathi C R, (2018), "Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS", Modelling, Measurement and Control A, Vol.91, Issue.2, pp.73-76. DOI: 10.18280/mmc_a.910207

[6]. Lakshman Narayana Vejendla , A Peda Gopi and N.Ashok Kumar,(2018)," Different techniques for hiding the text information using text steganography techniques: A survey", Ingénierie des Systèmes d'Information, Vol.23, Issue.6,pp.115-125.DOI: 10.3166/ISI.23.6.115-125

[7]. A Peda Gopi and Lakshman Narayana Vejendla (2018), "Dynamic load balancing for client server assignment in distributed system using genetic algorithm", Ingénierie des Systèmes d'Information, Vol.23, Issue.6, pp. 87-98. DOI: 10.3166/ISI.23.6.87-98

[8]. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Using customized Active Resource Routing and Tenable Association using Licentious Method Algorithm for secured mobile ad hoc network Management", Advances in Modeling and Analysis B, Vol.60, Issue.1, pp.270-282. DOI: 10.18280/ama_b.600117

[9]. Lakshman Narayana Vejendla and Bharathi C R,(2017),"Identity Based Cryptography for Mobile ad hoc Networks", Journal of Theoretical and Applied Information Technology, Vol.95, Issue.5, pp.1173-1181. EID: 2-s2.0-85015373447

[10]. Lakshman Narayana Vejendla and A Peda Gopi, (2017)," Visual cryptography for gray scale images with enhanced security mechanisms", Traitement du Signal,Vol.35, No.3-4,pp.197-208. DOI: 10.3166/ts.34.197-208

[11]. A Peda Gopi and Lakshman Narayana Vejendla, (2017)," Protected strength approach for image steganography", Traitement du Signal, Vol.35, No.3-4,pp.175-181. DOI: 10.3166/TS.34.175-181

[12]. Lakshman Narayana Vejendla and A Peda Gopi, (2020)," Design and Analysis of CMOS LNA with Extended Bandwidth For RF Applications", Journal of Xi'an University of Architecture & Technology, Vol. 12, Issue. 3,pp.3759-3765. https://doi.org/10.37896/JXAT12.03/319.

[13]. Chaitanya, K., and S. Venkateswarlu,(2016),"DETECTION OF BLACKHOLE & GREYHOLE ATTACKS IN MANETs BASED ON ACKNOWLEDGEMENT BASED APPROACH." Journal of Theoretical and Applied Information Technology 89.1: 228.

[14]. Patibandla R.S.M.L., Kurra S.S., Mundukur N.B. (2012), "A Study on Scalability of Services and Privacy Issues in Cloud Computing". In: Ramanujam R., Ramaswamy S. (eds) Distributed Computing and Internet Technology. ICDCIT 2012. Lecture Notes in Computer Science, vol 7154. Springer, Berlin, Heidelberg

[15]. Patibandla R.S.M.L., Veeranjaneyulu N. (2018), "Survey on Clustering Algorithms for Unstructured Data". In: Bhateja V., Coello Coello C., Satapathy S., Pattnaik P. (eds) Intelligent Engineering Informatics. Advances in Intelligent Systems and Computing, vol 695. Springer, Singapore

[16]. Patibandla, R.S.M.L., Veeranjaneyulu, N. (2018), "Performance Analysis of Partition and Evolutionary Clustering Methods on Various Cluster Validation Criteria", Arab J Sci Eng ,Vol.43, pp.4379–4390.

[17]. R S M Lakshmi Patibandla, Santhi Sri Kurra and N.Veeranjaneyulu, (2015), "A Study on Real-Time Business Intelligence and Big Data",Information Engineering, Vol.4,pp.1-6.

[18]. K. Santhisri and P.R.S.M. Lakshmi,(2015), " Comparative Study on Various Security Algorithms in Cloud Computing", Recent Trends in Programming Languages ,Vol.2,No.1,pp.1-6.

[19]. K.Santhi Sri and PRSM Lakshmi,(2017), "DDoS Attacks, Detection Parameters and Mitigation in Cloud Environment", IJMTST,Vol.3,No.1,pp.79-82.

[20]. P.R.S.M.Lakshmi,K.Santhi Sri and Dr.N. Veeranjaneyulu,(2017), "A Study on

Deployment of Web Applications Require Strong Consistency using Multiple Clouds", IJMTST,Vol.3,No.1,pp.14-17.

[21]. P.R.S.M.Lakshmi,K.Santhi Sri and M.V.Bhujanga Ra0,(2017), "Workload Management through Load Balancing Algorithm in Scalable Cloud", IJASTEMS,Vol.3,No.1,pp.239-242.

[22]. K.Santhi Sri, P.R.S.M.Lakshmi, and M.V.Bhujanga Ra0,(2017), "A Study of Security and Privacy Attacks in Cloud Computing Environment", IJASTEMS,Vol.3,No.1,pp. 235-238.

[23]. R S M Lakshmi Patibandla and N. Veeranjaneyulu, (2018), "Explanatory & Complex Analysis of Structured Data to Enrich Data in Analytical Appliance", International Journal for Modern Trends in Science and Technology, Vol. 04, Special Issue 01, pp. 147-151.

[24]. R S M Lakshmi Patibandla, Santhi Sri Kurra, Ande Prasad and N.Veeranjaneyulu, (2015), "Unstructured Data: Qualitative Analysis", J. of Computation In Biosciences And Engineering, Vol. 2,No.3,pp.1-4.

[25]. R S M Lakshmi Patibandla, Santhi Sri Kurra and H.-J. Kim,(2014), "Electronic resource management using cloud computing for libraries", International Journal of Applied Engineering Research, Vol.9,pp. 18141-18147.

[26]. Ms.R.S.M.Lakshmi Patibandla Dr.Ande Prasad and Mr.Y.R.P.Shankar,(2013), "SECURE ZONE IN CLOUD", International Journal of Advances in Computer Networks and its Security, Vol.3,No.2,pp.153-157.

[27]. Patibandla, R. S. M. Lakshmi et al., (2016), "Significance of Embedded Systems to IoT.", International Journal of Computer Science and Business Informatics, Vol.16,No.2,pp.15-23.

[28]. AnveshiniDumala and S. PallamSetty. (2020),"LANMAR routing protocol to support real-time communications in MANETs using Soft computing technique", 3rd International Conference on Data Engineering and Communication Technology (ICDECT-2019), Springer, Vol. 1079, pp. 231-243.

[29]. AnveshiniDumala and S. PallamSetty. (2019),"Investigating the Impact of Network Size on LANMAR Routing Protocol in a Multi-Hop Ad hoc Network", i-manager's Journal on Wireless Communication Networks (JWCN), Volume 7, No. 4, pp.19-26.

[30]. AnveshiniDumala and S. PallamSetty. (2019),"Performance analysis of LANMAR routing protocol in SANET and MANET", International Journal of Computer Science and Engineering (IJCSE) – Vol. 7,No. 5, pp.1237-1242.

[31]. AnveshiniDumala and S. PallamSetty. (2018), "A Comparative Study of Various Mobility Speeds of Nodes on the Performance of LANMAR in Mobile Ad hoc Network", International Journal of Computer Science and Engineering (IJCSE) – Vol. 6, No. 9, pp. 192-198.

[32]. AnveshiniDumala and S. PallamSetty. (2018),"Investigating the Impact of IEEE 802.11 Power Saving Mode on the Performance of LANMAR Routing Protocol in MANETs", International Journal of Scientific Research in Computer Science and Management Studies (IJSRCSMS) – Vol.7, No. 4.

[33]. AnveshiniDumala and S. PallamSetty. (2016),"Analyzing the steady state behavior of RIP and OSPF routing protocols in the context of link failure and link recovery in Wide Area Network", International Journal of Computer Science Organization Trends (IJCOT) – Vol. 34 No 2, pp.19-22.

[34]. AnveshiniDumala and S. PallamSetty. (2016),"Investigating the Impact of Simulation Time on Convergence Activity & Duration of EIGRP, OSPF Routing Protocols under Link Failure and Link Recovery in WAN Using OPNET Modeler", International Journal of Computer Science Trends and Technology (IJCST) – Vol. 4 No. 5, pp. 38-42.

[35]. VellalacheruvuPavani and I. Ramesh Babu (2019) ,"Three Level Cloud Storage Scheme for Providing Privacy Preserving using Edge Computing",International Journal of Advanced Science and Technology Vol. 28, No. 16, pp. 1929 – 1940.

[36]. VellalacheruvuPavani and I. Ramesh Babu,"A Novel Method to Optimize the Computation Overhead in Cloud Computing by Using Linear Programming" ,International Journal of Research

and Analytical Reviews May 2019, Volume 6, Issue 2,PP.820-830..

[37]. Anusha Papasani and Nagaraju Devarakonda,(2016),"Improvement of Aomdv Routing Protocol in Manet and Performance Analysis of Security Attacks", International Journal Of Research in Computer Science & Engineering ,Vol.6,No.5, pp.4674-4685.

[38]. Sk.Reshmi Khadherbhi,K.Suresh Babu , Big Data Search Space Reduction Based On User Perspective Using Map Reduce ,International Journal of Advanced Technology and Innovative Research Volume.07, IssueNo.18, December-2015, Pages: 3642-3647

[39]. B.V.Suresh kumar,Sk.Reshmi Khadherbhi ,BIG-IOT Framework Applications and Challenges: A Survey Volume 7, Issue VII, JULY/2018 pg.no 1257-1264

[40]. P.Sandhya Krishna,Sk.Reshmi Khadherbhi,V.Pavani, Unsupervised or Supervised Feature Finding For Study of Products Sentiment ,International Journal of Advanced Science and Technology, Vol 28 No 16 (2019).

[41]. K.Santhi Sri, Dr.Ande Prasad (2013), "A Review of Cloud Computing and Security Issues at Different Levels in Cloud Computing" , International Journal on Advanced Computer Theory and Engineering Vol. 2,pp 67-73.

[42]. K.Santhi Sri, N.Veeranjaneyulu(2018), "A Novel Key Management Using Elliptic and Diffie-Hellman for Managing users in Cloud Environment", Advances in Modelling and Analysis B,Vol.61,No.2,pp 106-112.

[43]. K.Santhi Sri, N.Veeranjaneyulu(2019), "Decentralized Key Management Using Alternating Multilinear Forms for Cloud Data Sharing with Dynamic Multiprivileged Groups", Mathematical Modelling of Engineering Problems,Vol.6,No.4,pp511-518.

[44]. S.Sasikala, P.Sudhakar, "interpolation of CFA color Images with Hybrid image denoising", 2014 Sixth International Conference on Computational Intelligence and Communication Networks, DOI 10.1109/.53 193 DOI 10.1109/CICN.2014.53, pp. 193-197.

[45]. Me. Jakeera Begum and M.Venkata Rao, (2015), "Collaborative Tagging Using CAPTCHA" International Journal of Innovative Technology And Research, Volume No.3, Issue No.5,pp,2436 – 2439.

[46]. L.Jagajeevan Rao, M. Venkata Rao, T.Vijaya Saradhi (2016), "How The Smartcard Makes the Certification Verification Easy" Journal of Theoretical and Applied Information Technology, Vol.83. No.2, pp. 180-186.

[47]. Venkata Rao Maddumala, R. Arunkumar, and S. Arivalagan (2018)"An Empirical Review on Data Feature Selection and Big Data Clustering" Asian Journal of Computer Science and Technology Vol.7 No.S1, pp. 96-100.

[48]. Singamaneni Kranthi Kumar, Pallela Dileep Kumar Reddy, Gajula Ramesh, Venkata Rao Maddumala, (2019), "Image Transformation Technique Using Steganography Methods Using LWT Technique" ,Traitement du Signalvol 36, No 3, pp. 233-237.

Journal of Engineering Sciences