INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

Online Product Quantization

LANKALAPALLI T S S SRI DURGA
PG Scholar, Department of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India.
E-Mail: taranginilankalapalli6465@gmail.com

Mr. NAGA. SRINIVASA RAO*
Asst. Professor, Dept of M.C.A,
S.K.B.R P.G College,
Amalapuram, E.G.Dt., A.P, India.
E-Mail:naagaasrinu@gmail.com

Abstract

Approximate nearest neighbor (ANN) search has achieved great success in many tasks. However, existing popular methods for ANN search, such as hashing and quantization methods, are designed for static databases only. They cannot handle well the database with data distribution evolving dynamically, due to the high computational effort for retraining the model based on the new database. This project addresses the problem by developing an online product quantization (online PQ) model and incrementally updating the quantization codebook that accommodates to the incoming streaming data. Moreover, to further alleviate the issue of large scale computation for the online PQ update, design two budget constraints for the model to update partial PQ codebook instead of all. Derive a loss bound which guarantees the performance of our online PQ model. Furthermore, develop an online PQ model over a sliding window with both data insertion and deletion supported, to reflect the real-time behaviour of the data. The experiments demonstrate that our online PQ model is both time-efficient and effective for ANN search in dynamic large scale databases compared with baseline methods and the idea of partial PQ codebook update further reduces the update cost.

Keywords : Quantization (signal), Computational modeling, Data models, Maintenance engineering, Computational efficiency,

1. INTRODUCTION

1.1 Introduction:

Online social media is one of the defining phenomena in this technology-driven era. Platforms, such as Face book and Twitter, are instrumental in enabling global connectivity. 2.46 billion Users are estimated to be now connected and by the year 2020 one third of the global population will be connected. Users of these platforms freely generate and consume information leading to unprecedented amounts of data. Several domains have already recognized the crucial role of social media analysis in improving productivity and gaining competitive advantage. Information derived from social media has been utilized in health-care to support effective service delivery, in sport to engage with fans, in the entertainment industry to complement intuition and experience in business decisions and in politics to track election processes, promote wider engagement with supporters and predict poll outcomes. However, alongside the benefits, the rapid increase in social media spam contents questions the credibility of research based on analyzing this data. A report by Nexgate estimates that on average one spam post occurs in every 200 social media posts and a more recent study reports that approximately 15% of active Twitter users are automated bots. The growing volume of spam posts and the use of autonomous accounts (social bots) to generate posts raise many concerns about the credibility and representativeness of the data for research. In this report, focus on Twitter and propose a novel, effective approach to detect and filter unwanted tweets, complementing earlier approaches in this direction. Previous studies rely on historical

ISSN: 2057-5688

features of tweets that are often unavailable on Twitter after a short period of time, hence not suitable for real-time use. Our approach utilizes an optimized set of readily available features, independent of historical textual features on Twitter. The employed features are categorized as related to the Twitter account, the user or referring to the pair wise engagement between users. A number of machine learning models have been trained. Recursive feature elimination has been employed in order to ascertain the robustness and the discriminative power of each feature. In comparison to an earlier study, the proposed features exhibit stronger discriminative power with more consistent performance across the different learning models. Spam posting users exhibit some evasive tactics, such as posting on average of 4 tweets per day, and tricks to balance the follower–followee relationship. Our analysis shows that an average automated spam posting account posts at least 12 tweets per day within well-defined activity periods. The activity pattern resembles the staircase function exhibiting surges of intermittent activities. Our study contributes (a) a new set of lightweight features suitable for real-time detection of spammers on Twitter and (b) an additional dataset source1 offering an insight into the behavior of spam users on Twitter to support further studies.

1.2 Purpose:

On online hashing show that hashing based ANN approaches can be adapted to the dynamic database environment by updating hash functions accommodating new streaming data and then updating the hash codes of the exiting stored data via the new hash functions. Searching is performed in the Hamming space which is efficient and has low computational cost. However, an important problem that these works have not addressed is the computation of the hash code maintenance. To handle the streaming fashion of the data, the hash functions are required to be frequently updated, which will result in constant hash code recomputation of all the existing data in the reference database. This will inevitably in cur an increasing amount of update time as the data volume increases. In addition, these online hashing approaches require the system to keep the old data so that the new hash code of the old data can be updated each time, leading to in efficiency in memory and computational load. Therefore, computational complexity and storage cost are still our major concerns in developing an online indexing model.

1.3 Scope:

ANN search in a dynamic database has a widespread applications in the real world. For example, a large number of news articles are generated and updated on hourly/daily basis, so a news searching system requires to support news topic tracking and retrieval in a frequently changing news database. For object detection in video surveillance, video data is continuously recorded, so that the distances between/among similar or dissimilar objects are continuously changing. For image retrieval in dynamic databases, relevant images are retrieved from a constantly changing image collection, and the retrieved images could therefore be different over time given the same image query. In such an environment, real-time query needs to be answered based on all the data collected to the database so far.

1.4 Motivation:

In many real-world applications, data is continuously generated everyday and the database needs to get updated dynamically by the newly available data. For example, news articles can be posted any time and it is important to enhance user experience in news topic tracking and related news retrieval. New images with new animal species may be inserted to the large scale image database. Index update needs to be supported to allow users to retrieve images with expected animal in a dynamically changing database. A live video or a surveillance video may generate several frame per second, which makes the real-time object tracking or face recognition a crucial task to

ISSN: 2057-5688

solve. In this experiment, we evaluate our model on how it handles dynamic updates in both time efficiency and search accuracy in three different types of data: text, image and video.

1.5 Overview:

In recent years, there has been an increasing concern over the computational cost and memory requirement dealing with continuously growing large scale databases, and therefore there are many online learning algorithm works proposed to update the model each time streaming data coming in. Therefore, we consider the following problem. Given a dynamic database environment, develop an online learning model accommodating the new streaming data with low computational cost for ANN search.

2. LITERATURE SURVEY

Huang et al.,proposes an online hash model to accommodate data coming in stream for online learning. Specifically, a new loss function is proposed to measure the similarity loss between a pair of data samples in hamming space. Then, a structured hash model is derived and optimized in a passive-aggressive way. Theoretical analysis on the upper bound of the cumulative loss for the proposed online hash model is provided. They extend online hashing from a single-model to a multi-model online hashing that trains multiple models so as to retain diverse online hashing models in order to avoid biased update.

Big data is becoming ever more ubiquitous, ranging over massive video repositories, document corpuses, image sets and Internet routing history. Proximity search and clustering are two algorithmic primitives fundamental to data analysis, but suffer from the "curse of dimensionality" on these gigantic datasets. A popular attack for this problem is to convert object representations into short binary codewords, while approximately preserving near neighbor structure. However, there has been limited research on constructing codewords in the "streaming" or "online" settings often applicable to this scale of data, where one may only make a single pass over data too massive to fit in local memory. Ghashami A. Abdullah [2] applies recent advances in matrix sketching techniques to construct binary code words in both streaming and online setting.

Extensive new algorithms have been developed and successfully applied for hashing based approximate nearest neighbor (ANN) search. However, two critical problems are rarely mentioned. First, in real-world applications, the data often comes in a streaming fashion but most of existing hashing methods are batch based models. Second, when the dataset becomes huge, it is almost impossible to load all the data into memory to train hashing models. Leng et al., [3], propose a novel approach to handle these two problems simultaneously based on the idea of data sketching. A sketch of one dataset preserves its major characters but with significantly smaller size. With a small size sketch, our method can learn hash functions in an online fashion, while needs rather low computational complexity and storage space.

Hashing methods that map the data into Hamming space have shown promise, however, many of these methods employ a batch-learning strategy in which the computational cost and memory requirements may become intractable and infeasible with larger and larger datasets. To overcome these challenges, Cakir and Sclaroff [4] propose an online learning algorithm based on stochastic gradient descent in which the hash functions are updated iteratively with streaming data.

Yang et al., [5] consider updating a hashing mod-el upon gradually increased labelled data in a fast response to users, called smart hashing update (SHU). In order to get a fast response to users, SHU aims to select a small set of hash functions to re-learn and only updates the corresponding hash bits of all data points. More specifically, put forward two selection methods for performing efficient and effective update. In order to reduce the response

ISSN: 2057-5688

time for acquiring a stable hashing code, also propose an accelerated method to further reduce interactions between users and the computer

Most state-of-the-art supervised hashing approaches employ batch-learners. Unfortunately, batch-learning strategies may be inefficient when confronted with large datasets. Moreover, with batch-learners, it is unclear how to adapt the hash functions as the dataset continues to grow and new variations appear over time. To handle these issues, Cakir et al., [6], propose OSH: an Online Supervised Hashing technique that is based on Error Correcting Output Codes. They consider a stochastic setting where the data arrives sequentially and our method learns and adapts its hashing functions in a discriminative manner. This method makes no assumption about the number of possible class labels, and accommodates new classes as they are presented in the incoming data stream.

J´egou et al., [7] introduces a product quantization-based approach for approximate nearest neighbor search. The idea is to decompose the space into a Cartesian product of low-dimensional subspaces and to quantize each subspace separately. A vector is represented by a short code composed of its subspace quantization indices. The euclidean distance between two vectors can be efficiently estimated from their codes. An asymmetric version increases precision, as it computes the approximate distance between a vector and a code.

Gong and Lazebnik [8] addresses the problem of learning similarity-preserving binary codes for efficient similarity search in large-scale image collections. They formulate this problem in terms of finding a rotation of zero-centered data so as to minimize the quantization error of mapping this data to the vertices of a zero-centered binary hypercube, and propose a simple and efficient alternating minimization algorithm to accomplish this task. This algorithm, dubbed iterative quantization (ITQ), has connections to multiclass spectral clustering and to the orthogonal Procrustes problem, and it can be used both with unsupervised data embeddings such as PCA and supervised embeddings such as canonical correlation analysis (CCA).

Zhang et al., [9] presents a novel compact coding approach, composite quantization, for approximate nearest neighbor search. The idea is to use the composition of several elements selected from the dictionaries to accurately approximate a vector and to represent the vector by a short code composed of the indices of the selected elements. To efficiently compute the approximate distance of a query to a database vector using the short code, introduce an extra constraint, constant inter-dictionary-element-product, resulting in that approximating the distance only using the distance of the query to each selected element is enough for nearest neighbor search.

Babenko and Lempitsky [10] propose a new vector encoding scheme (tree quantization) that obtains lossy compact codes for high-dimensional vectors via tree-based dynamic programming. Similarly to several previous schemes such as product quantization, these codes correspond to codeword numbers within multiple codebooks. Also propose an integer programming-based optimization that jointly recovers the coding tree structure and the codebooks by minimizing the compression error on a training dataset.

He et al., [11] propose a novel Affinity-Preserving K-means algorithm which simultaneously performs k-means clustering and learns the binary indices of the quantized cells. The distance between the cells is approximated by the Hamming distance of the cell indices. Li et al., [12] propose an adaptive binary quantization method that learns a discriminative hash function with prototypes correspondingly associated with small unique binary codes. This alternating optimization adaptively discovers the prototype set and the code set of a varying size in an efficient way, which together robustly approximate the data relations. This method can

ISSN: 2057-5688

be naturally generalized to the product space for long hash codes.

Liu et al., [13] propose a structure sensitive hashing based on cluster prototypes, which explicitly exploits both global and local structures. An alternating optimization algorithm, respectively, minimizing the quantization loss and spectral embedding loss, is presented to simultaneously discover the cluster prototypes for each hash function, and optimally assign unique binary codes to them satisfying the affinity alignment between them. For hash codes of a desired length, an adaptive bit assignment is further appended to the product quantization of the subspaces, approximating the Hamming distances and meanwhile balancing the variance among hash functions.

Product quantization (PQ) is an effective vector quantization method. A product quantizer can generate an exponentially large codebook at very low memory/time cost. The essence of PQ is to decompose the high-dimensional vector space into the Cartesian product of subspaces and then quantize these subspaces separately. The optimal space decomposition is important for the PQ performance, but still remains an unaddressed issue. Ge et al., [14] optimize PQ by minimizing quantization distortions with respect to the space decomposition and the quantization codebooks. We present two novel solutions to this challenging optimization problem. The first solution iteratively solves two simpler sub-problems. The second solution is based on a Gaussian assumption and provides theoretical analysis of the optimality.

3. PROBLEM STATEMENT

Existing hashing methods are grouped in data-independent hashing and data-dependent hashing. One of the most representative work for data-independent hashing is Locality Sensitive Hashing (LSH), where its hashing functions are randomly generated. Data-independent hashing methods are independent from the input data, they can be easily adopted in an online fashion.

Data-dependent hashing learns the hash functions from the given data, which can achieve better performance than data independent hashing methods. Its representative works are Spectral Hashing (SH), which uses spectral method to encode similarity graph of the input into hash functions, IsoH which finds a rotation matrix for equal variance in the projected dimensions and ITQ which learns an orthogonal rotation matrix for minimizing the quantization error of data items to their hash codes.

Online Hashing, AdaptHash and Online Supervised Hashing are online supervised hashing methods, requiring label information, which might not be commonly available in many real-world applications. Stream Spectral Binary Coding (SSBC) and Online Sketching Hashing (OSH) are the only two existing online unsupervised hashing methods which do not require labels, where both of them are matrix sketch-based methods to learn to represent the data seen so far by a small sketch.

Composite Quantization (CQ), Sparse Composite Quantization (SQ), Additive Quantization (AQ) and Tree Quantization (TQ) require the codeword maintenance of the old data to update the code words due to the constraints or structure of their models. Product Quantization (PQ), as one of the most classical Multi-codebook quantization (MCQ) method for fast nearest neighbor search, decomposes the input space into a Cartesian product of subspaces. The codeword of a data instance is represented by the concatenation of the subcodeword of the data in all subspaces.

Extension works of Q such as Optimized Product Quantization (OPQ), Kmeans Hashing (KMH), Adaptive Binary Quantization (ABQ) and Structure Sensitive Hashing (SSH) can also be developed to an online fashion. Specifically, KMH, ABQ and SSH approximate the distance between codewords by their Hamming distance, so they will require codewords maintenance in the online setting.

INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688

3.1 Drawbacks : These existing works provides poor caching ratio and less hit ratio**.** These existing works provides poor caching ratio and less hit ratio

4. PROPOSED SYSTEM

This project proposes a novel online paradigm for PQ. The codebook at each iteration gets updated by the streaming data without retraining all the collected data. ANN search can be conducted against the latest codebook in terms of user queries. Unlike online hashing methods which update hashing functions and hash codes of the existing data, online PQ updates codebooks only and the codeword index of the existing data remains the same.

4.1 Advantages:

The proposed wildcard-rule caching algorithm could have better caching ability than the other existing algorithms. Furthermore, the proposed cache replacement algorithm could have higher hit ratio than the other existing algorithms.

5. IMPLEMENTATION

5.1 Code Book Generation

Code book generation is the initial step of quantization. Codebook can be generated in spatial domain or transform domain using clustering algorithms. Codebook generation is the key component of VQ, where the performance of the total VQ depends mainly on the quality of the codebook that is generated. K-means clustering algorithm is used to generate the code book.

5.2 Sub-Quantization Learning

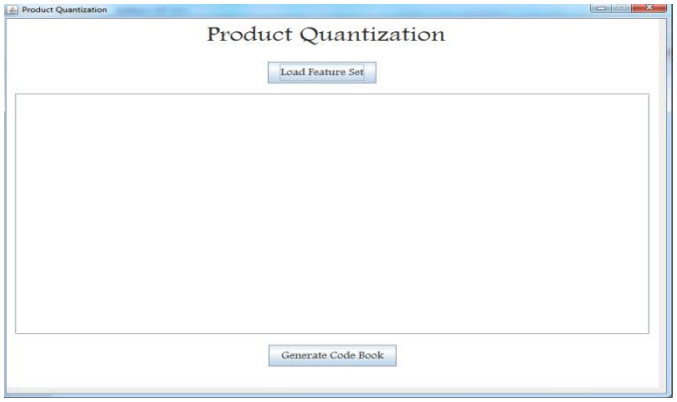
This module generates sub-quantizer based on the generated codebook. Codebook is then composed of M sub-codebooks and each of the sub-codebook contains K sub-code words quantized from a distinct sub-quantizer. Any codeword belongs to the Cartesian product of the sub-codewords in each sub-codebook. The codeword of x is constructed by the concatenation of M sub-codewords $z = [z_{1,k1}, \ldots, z_{m,km}, \ldots, z_{M,kM}]$, where $z_{m,km}$ is the sub-codeword of x_m.

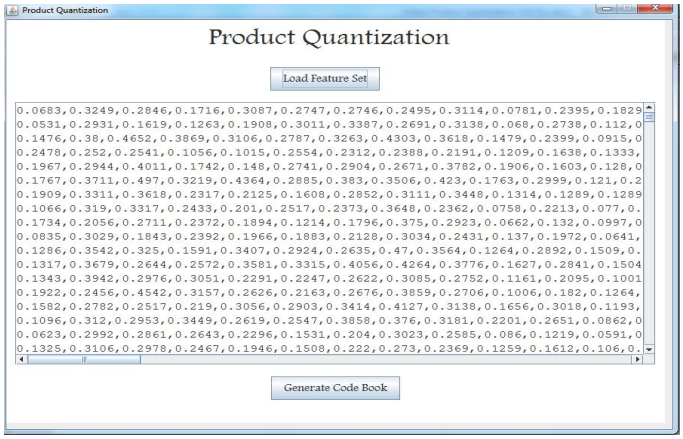
Each sub-codeword is represented by an index. Thus the codeword of a data instance is the concatenation of the sub-codeword indices of all subspaces.

5.3 Searching

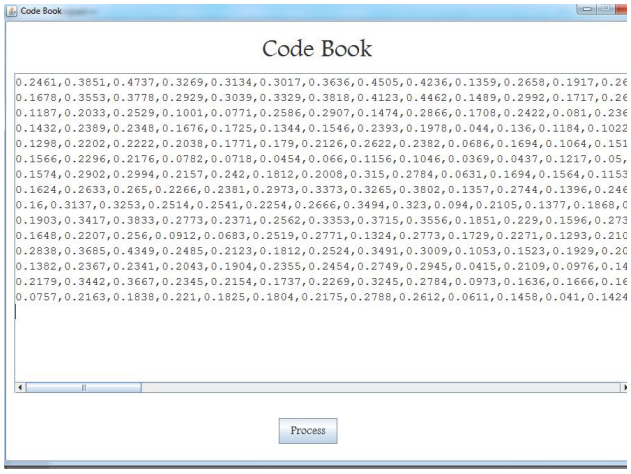
This module use approximate nearest neighbor (ANN) search. An approximate nearest neighbor search algorithm is allowed to return points, whose distance from the query is at most *c* times the distance from the query to its nearest points.

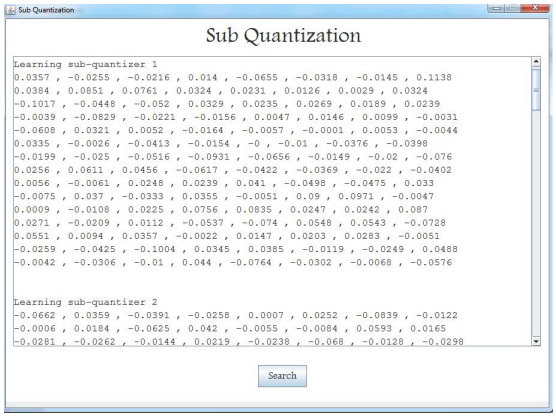
6. OUTPUT RESULTS

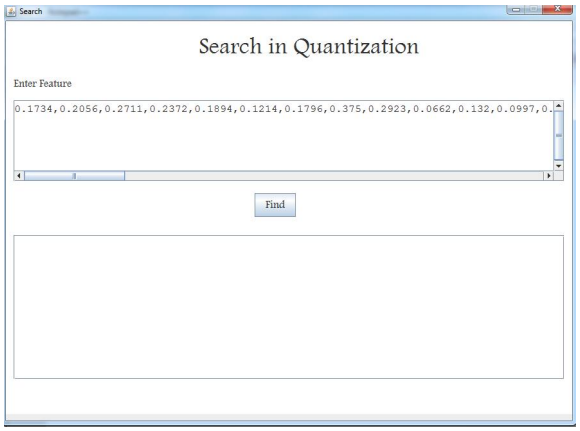


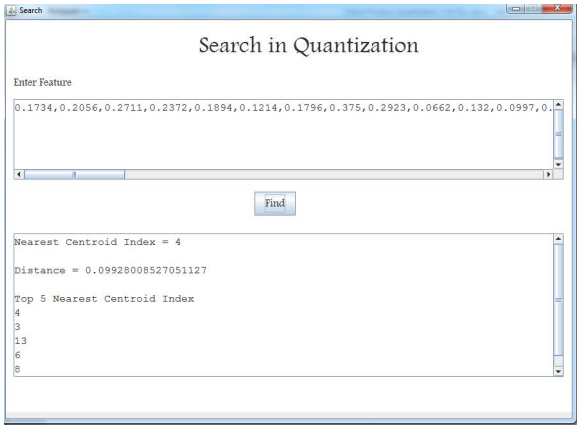


INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE
ISSN: 2057-5688









7. CONCLUSION

This project presents online PQ method to accommodate streaming data. In addition, employ two budget constraints to facilitate partial codebook update to further alleviate the update time cost. A relative loss bound has been derived to guarantee the performance of our model. In addition, we propose an online PQ over sliding window approach, to emphasize on the real-time data. Experimental results show that our method is significantly faster in accommodating the streaming data, outperforms the competing online hashing methods and unsupervised batch mode hashing method in terms of search accuracy and update time cost, and attains comparable search quality with batch mode PQ.

8. FUTURE ENHANCEMENT

Analysis of spammers network to unearth different types of coordinated spam campaigns run by the spambots seems one of the promising future directions of research. Moreover, analyzing the temporal evolution of spammers' followers may reveal some interesting patterns that can be utilized for spammers characterization at different levels of granularity.

9. BIBLIOGRAPHY

ISSN: 2057-5688

1. Improving product marketing by predicting early reviewers on E-Commerce websites
S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, "Improving product marketing by predicting early reviewers on E-Commerce websites," Deleted Journal, no. 43, pp. 17–25, Apr. 2024, doi: 10.55529/ijrise.43.17.25.

2. Kodati, Dr Sarangam, et al. "Classification of SARS Cov-2 and Non-SARS Cov-2 Pneumonia Using CNN." Journal of Prevention, Diagnosis and Management of Human Diseases (JPDMHD) 2799-1202, vol. 3, no. 06, 23 Nov. 2023, pp. 32–40, journal.hmjournals.com/index.php/JPDMHD/article/view/3406/2798, https://doi.org/10.55529/jpdmhd.36.32.40. Accessed 2 May 2024.

3. V. Srikanth, "CHRONIC KIDNEY DISEASE PREDICTION USING MACHINE LEARNING ALGORITHMS," IJTE, pp. 106–109, Jan. 2023, [Online]. Available: http://ijte.uk/archive/2023/CHRONIC-KIDNEY-DISEASE-PREDICTION-USING-MACHINE-LEARNING-ALGORITHMS.pdf

4. V. SRIKANTH, "DETECTION OF PLAGIARISM USING ARTIFICIAL NEURAL NETWORKS," International Journal of Technology and Engineering, vol. XV, no. I, pp. 201–204, Feb. 2023, [Online]. Available: http://ijte.uk/archive/2023/DETECTION-OF-PLAGIARISM-USING-ARTIFICIAL-NEURAL-NETWORKS.pdf

5. V. SRIKANTH, "A REVIEW ON MODELING AND PREDICTING OF CYBER HACKING BREACHES," IJTE, vol. XV, no. I, pp. 300–302, Mar. 2023, [Online]. Available: http://ijte.uk/archive/2023/A-REVIEW-ON-MODELING-AND-PREDICTING-OF-CYBER-HACKING-BREACHES.pdf

6. S. Kodati, M. Dhasaratham, V. V. S. S. Srikanth, and K. M. Reddy, "Detection of fake currency using machine learning models," Deleted Journal, no. 41, pp. 31–38, Dec. 2023, doi: 10.55529/ijrise.41.31.38.

7. "Cyberspace and the Law: Cyber Security." IOK STORE, iokstore.inkofknowledge.com/product-page/cyberspace-and-the-law. Accessed 2 May 2024.

8. "Data Structures Laboratory Manual." IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-structures-laboratory-manual. Accessed 2 May 2024.

9. Data Analytics Using R Programming Lab." IOK STORE, www.iokstore.inkofknowledge.com/product-page/data-analytics-using-r-programming-lab. Accessed 2 May 2024.

10. V. Srikanth, Dr. I. Reddy, and Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, 501301, India,

ISSN: 2057-5688

"WIRELESS SECURITY PROTOCOLS (WEP,WPA,WPA2 & WPA3)," journal-article, 2019. [Online]. Available: https://www.jetir.org/papers/JETIRDA06001.pdf

10. V. SRIKANTH, "Secured ranked keyword search over encrypted data on cloud," IJIEMR Transactions, vol. 07, no. 02, pp. 111–119, Feb. 2018, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/1121_approvedpaper.pdf

11. V. SRIKANTH, "A NOVEL METHOD FOR BUG DETECTION TECHNIQUES USING INSTANCE SELECTION AND FEATURE SELECTION," IJIEMR Transactions, vol. 06, no. 12, pp. 337–344, Dec. 2017, [Online]. Available: https://www.ijiemr.org/public/uploads/paper/976_approvedpaper.pdf

12 . SRIKANTH MCA, MTECH, MBA, "ANALYZING THE TWEETS AND DETECT TRAFFIC FROM TWITTER ANALYSIS," Feb. 2017. [Online]. Available: http://ijmtarc.in/Papers/Current%20Papers/IJMTARC-170309.pdf

14 Srikanth, V. 2018. "Secret Sharing Algorithm Implementation on Single to Multi Cloud." International Journal of Research 5 (01): 1036–41. https://journals.pen2print.org/index.php/ijr/article/view/11641/11021.

5. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023.

16. Babu, Dr P. Sankar, et al. "Intelligents Traffic Light Controller for Ambulance." Journal of Image Processing and Intelligent Remote Sensing(JIPIRS) ISSN 2815-0953, vol. 3, no. 04, 19 July 2023, pp. 19–26, journal.hmjournals.com/index.php/JIPIRS/article/view/2425/2316, https://doi.org/10.55529/jipirs.34.19.26. Accessed 24 Aug. 2023.

17. S. Maddilety, et al. "Grid Synchronization Failure Detection on Sensing the Frequency and Voltage beyond the Ranges." Journal of Energy Engineering and Thermodynamics, no. 35, 4 Aug. 2023, pp. 1–7, https://doi.org/10.55529/jeet.35.1.7. Accessed 2 May 2024.

18. K. Meenendranath Reddy, et al. Design and Implementation of Robotic Arm for Pick and Place by Using Bluetooth Technology. No. 34, 16 June 2023, pp. 16–21, https://doi.org/10.55529/jeet.34.16.21. Accessed 20 Aug. 2023