iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL
FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 7 Issue: IV Month of publication: April 2019

DOI: https://doi.org/10.22214/ijraset.2019.4530

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 7 Issue IV, Apr 2019- Available at www.ijraset.com

# Routing Queries to the Right Expert in Online CQA

M. Divya Bharathi[1], Dr. S. Christy Melwyn[2], Mr. K. Ravikumar[3]
[1, 2, 3]Rrase college of engineering, vanchuvancherry

***Abstract: The popular websites such as Stack Overflow and Yahoo provides the question and answer which is posted by the anonymous user by all over the world. In these sites, the programmers from all over the world will answer for the queries which are posted on the websites. In the existing system, if any user wants to know the exact answer they have to mine the answer between the thousands of answer posted on the website and still it was very complicated to know the exact answer and it was very time consuming. So that user has to apply all the answer to get the exact answer. So to overcome from this situation, we proposing the technique in which user can view the exact answer from the websites as it will be displayed as a first answer for that question through post voting prediction. In the voting process, the voting will be given based on the conditions. If that answer satisfies the conditions, it will be up voted. In this way, our proposed technique provides the effectiveness and efficiency, optimality, correctness and complexity.***

***Keywords: Question answering, voting prediction, non-linearity, coupling, dynamics.***

## I. INTRODUCTION

Community Question Answering (CQA) sites, such as Stack Overflow[1] and Yahoo! Answers[2], have become very popular in recent years. These sites contain rich crowd- sourcing knowledge contributed by the site users in the form of questions and answers, and these questions and answers can potentially satisfy the information needs of more users. For example, millions of programmers ask and answer questions on Stack Overflow, and even more users now use Stack Overflow to seek solutions for their programming problems. In this article, we focus on the voting score prediction of questions/answers shortly after they are posted in the CQA sites. Such a task is essential for the prosperity and sustainability of the CQA ecosystem, and it may benefit all types of users, including the information producers and consumers [2]. For example, detecting potentially high- score answers can benefit the questioners as well as the people who have similar questions; it would also be helpful to identify high-score questions in the early stage and route them to expert answerers. Generally speaking, there are three key aspects that matter with the voting prediction of a post, namely, (1) the *non-linearity* between features and output, (2) the *coupling* between questions and answers, and (3) the *dynamics* (of training data sets). First, both the contextual features (e.g., the reputation of the user who issues the question, etc.) and the content of the post (e.g., keywords, etc.) might affect its voting score, and the effect of each feature might be beyond the simple linear-relationship. Second, intuitively (which was also confirmed in our previous work [3]), the voting of a question might be correlated with that of its associated answers. Yet, the questions and answers may reside in different feature spaces. Third, CQA sites usually offer a large size of training data set, and the data may arrive in a dynamic (stream-like) way for the mining algorithms. Due to the above three aspects, it is not an easy task to comprehensively and efficiently predict the voting scores of question/answer posts. The challenges are as follows. First, while each of the above three aspects might affect the voting scores of question/answer posts, they require different treat- ments in the data mining algorithms, making any off-the- shelf data mining algorithm sub-optimal for this problem. Second, each of the above three aspects will add the extra complexity into the mining process. For example, while many machine learning algorithms (e.g., kernel regression, support vector regression, etc.) are able to capture the non- linearity aspect, they typically require at least quadratic complexity in both time and space. Moreover, when the new training examples arrive in a stream, ever-growing fashion, even a linear algorithm might be too expensive. How can we build a *comprehensive* model to capture all the above three aspects to maximally boost the prediction accuracy? How can we make our prediction algorithms *scalable* to millions of CQA posts and *adaptive* to the newly arrived training examples over time? These are the main challenges that we aim to address in this article.

## II. OBJECTIVE

The main objective of this project is to provide a exact information to the user in which the answer will be displayed as a first answer by proposing some algorithms. By using algorithm, the conditions are set, in which the answer will be up voted if and only if it satisfies the condition.

## III. SCOPE OF PROJECT

For accommodation, we utilize intense capital letters for existing frameworks/vectors at time t, and strong lower case letters for recently arrived grids/vectors at time. We utilize superscript to recognize inquiries and replies, and utilize subscript to show time. For instance, we utilize mean the component grid for inquiries at time t to indicate the element lattice of recently arrived inquiries at time t. Contains the component vector for the relating question. So also, we use to signify the vector of voting scores at time t to mean the vector of voting scores from new inquiries at time t.

## IV. CONTRIBUTIONS

The prominent sites, for example, Stack Overflow and Yahoo gives the inquiry and answer which is posted by the unknown client by everywhere throughout the world. The main advantages of proposed technique is that it provides the efficiency and correctness. In these locales, the software engineers from everywhere throughout the world will respond in due order regarding the questions which are posted on the sites. In the current framework, if any client needs to know the correct answer they have to mine the appropriate response among answer posted on the site and still it was exceptionally muddled to know the correct answer and it was extremely tedious. With the goal that client needs to apply all the response to find the correct solution. So to overcome from this circumstance, we proposing the system in which client can see the correct answer from the sites as it will be shown as a first response for that inquiry through post voting forecast. In the voting procedure, the voting will be given in view of the conditions. On the off chance that that answer fulfills the conditions, it will be up voted. Thusly, our proposed method gives the adequacy and proficiency, optimality, rightness and unpredictability.

## V. RELATED WORK

### A. *Mining CQA Sites*

There is a large body of existing work on mining CQA sites. For example, Li et al. [18] aim to predict question quality, which is defined as the combination of user attention, answer attempts and the arrival speed of the best answer. Ravi et al. [19] also study the question quality which is combined by voting score and page views. Jeon et al. [20] and Suryan to et al. [21] evaluate the usefulness of answer quality and incorporate it to improve retrieval performance. To predict the quality of both questions and answers, Agichtein et al. [22] develop a graph-based model to catch the relationships among users, Li et al. [23] adopt the co-training approach to employ both question features and answer features, and Bian et al. [24] propose to propagate the labels through user-question- answer graph so as to tackle the sparsity problem where only a small number of questions/answers are labeled. Recently, Anderson et al. [25] propose to predict the long- lasting value (i.e., the page views) of a question and its answers. Dror et al. [26] aim to predict whether a question will be answered or not. How to predict the answer that the questioner will probably choose as the accepted answer is also well studied [27], [28], [29], [30]. Overall, our work differs from these existing work at the methodology level. While most of the existing work treats the prediction problem as a single, and/or linear, and/or static problem.

### B. *Mining Stream Data*

From the dynamic aspect, our LIP problem is related to stream mining [31] and time- series mining [32]. The main focus of existing stream/time- series mining work is on pattern discovery, clustering, and classification tasks. Chen et al. [33] and Ikonomovska et al. [34] study the regression problem in data streams; however, they still focus on a single and linear prediction problem. Several researchers also consider the non-linear and dynamic aspects in regression problem [6], [35]. Different from these existing work, we consider the coupling between questions and answers, and propose approximation methods to speed-up and scale-up the computation.

### C. *Other Related Work*

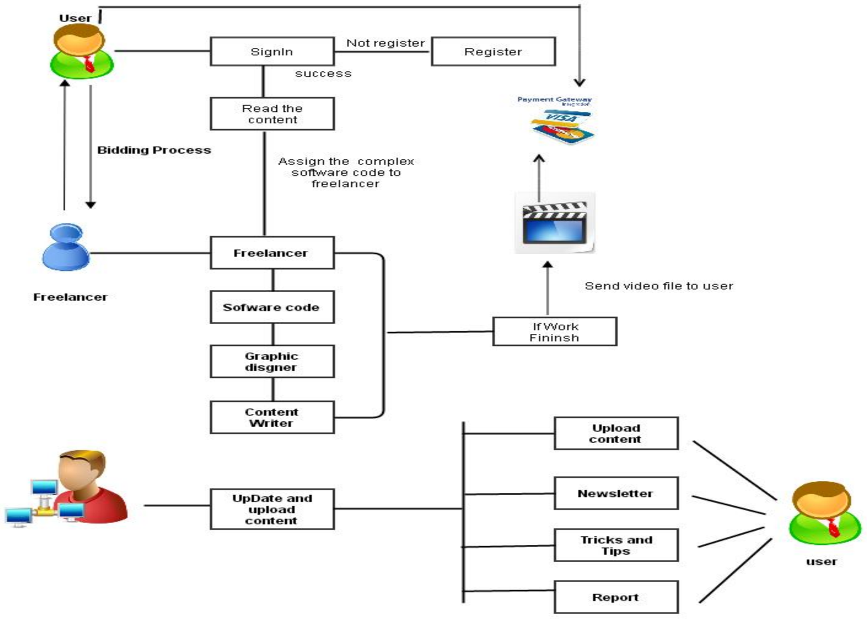
There are several pieces of in- teresting work that are remotely related to our work. Liu et al. [36] propose the problem of CQA site searcher satisfaction, i.e., whether or not the answer in a CQA site satisfies the information searcher using the search engines. Shtok et al. [37] attempt to answer certain new questions by existing answers. Zhou et al. [38] aim to find similar questions for a given query. Zhou et al. [39] propose to identify whether a user is asking a subjective question or not. Wei et al. [40] discard user biases to re-rank the voting scores of answers. Omari et al. [41] propose to provide a set of diverse and novel answers for questions. Sung et al. [42] aim to detect the potentially contributive users from recently-joined users. The question routing problem (e.g., how to route the right question to the right answerer) is also an active research area [43], [44], [45].In sun et al. used encrypted index tree structure to implement secure query results verification functionality. in this scheme, when the query ends, the cloud server returns query results along with a minimum encrypted index tree, then the data user searches this minimum index tree using the same search algorithm as the cloud server did to finish result verification zheng .constructed a verifiable secure query scheme over encrypted cloud data based on attribute-based encryption technique (abe) in the public-key setting .referred to the merkle hash tree and applied pairing operations to implement the correctness and completeness verification of query results for keyword search over large dynamic encrypted cloud data.

## VI. BACKGROUND

To clarify our proposed problems, in this section, we present our system model, threat model, used to implement our scheme.

### *A. System Model*

The system model of the secure search over encrypted cloud data usually includes three entities: data owners, data users, and the cloud server, which describes the following scenario: data owners encrypt their private data and upload them to cloud server for enjoying the abundant benefits brought by the cloud computing as well as guaranteeing data security. Meanwhile, the secure searchable indexes are also constructed to support effective keyword search over encrypted outsourced data. An authorized data user obtains interested data files from the cloud server by submitting query trapdoors (encrypted query keywords) to the cloud server, trapdoors and sends the query results to the data user. The above application scenario is based on an ideal assumption that the cloud server is considered as an honest entity and always honestly returns all qualified query results. In this paper, we consider a more challenging model, where the query results would be maliciously deleted or tampered by the dishonest cloud server. When the query results face the risks that are deleted or tampered, a well-functioning secure query system should provide a mechanism that allows the data user to verify the correctness and completeness of query results. To achieve the results verification goal, we propose to construct secure verification objects for data files that are outsourced to the cloud with encrypted data and secure indexes together. The query results along with corresponding data verification object are returned to the data user when a query ends. The improved system model of verifiable secure search over encrypted cloud data is illustrated in Fig. 1.



### *B. Threat Model*

In this paper, compared with the previous works, an important distinction about the threat model is that the cloud is considered to be an un trusted entity. More specifically, first of all, the cloud server tries to gain some valuable information from encrypted data files, secure indexes, and verification objects (e.g., a misbehaving cloud administrator aims at obtaining these information for possible monetary profits). Then, the cloud server would intentionally return false search results for saving computation resource or communication cost. Further, if the cloud server knows a query results verification mechanism is embedded, he may tamper or forge verification objects to escape responsibilities of misbehavior. Similar to the previous works, both data owners and authorized data users are considered to be trusted in our threat model.

## VII.PROPOSED METHOD

In this section, we propose our solutions for the LIP problem. We start with presenting two algorithms for Problem 1 (subsection 3.1) and Problem 2, respectively; and then address the computational challenges.

### *A. LIP-KM Algorithm for Problem 1*

Here, we address the static LIP problem (Problem 1). We propose an algorithm (LIP-KM) to capture both the non- linearity and the coupling aspects. For the non-linear aspect, a natural choice is to kernelize a linear prediction model (e.g., linear ridge regression). Recall that kernel method aims to produce non-linear versions of linear learning algorithms by mapping the data points into a high-dimensional Hilbert space with a non- linear function φ [7]. The key idea behind kernel methods is to use the kernel functions to replace the inner-product operations in the high-dimensional Hilbert space , and such replacement can be ensured by Mercer's Condition [8]. In other words, for two data points $\mathbf{F}(i, :)$ and $\mathbf{F}( j, :)$, the inner product of $\varphi(\mathbf{F}(i, :))$ and $\varphi(\mathbf{F}( j, :))$ in the Hilbert space can be directly computed by a Mercer kernel $\kappa(\mathbf{F}(i,:),\mathbf{F}(j,:))$

$$\kappa(\mathbf{F}(i, :), \mathbf{F}( j, :)) = < \varphi(\mathbf{F}(i, :)), \varphi(\mathbf{F}( j, :)) >$$
$$= \varphi(\mathbf{F}(i,:))\varphi(\mathbf{F}(j,:))^J$$

Where <, > indicates the inner product in. As we can see from Eq. (3), we can derive the non-linear models without any explicit knowledge of either φ or. Com- mon kernel functions include Gaussian kernel, polynomial kernel, cosine kernel.

### *B. LIP-KIMAA Algorithm*

Compared with LIP-KIM, LIP-KIMA is much more scalable, being *linear* in terms of both time and space complexity. However, if the new training examples arrive in a stream-like, ever-growing fashion, a linear algorithm might be still too expensive.

## VIII. EXPERIMENTS

### *A. Experiment Setup*

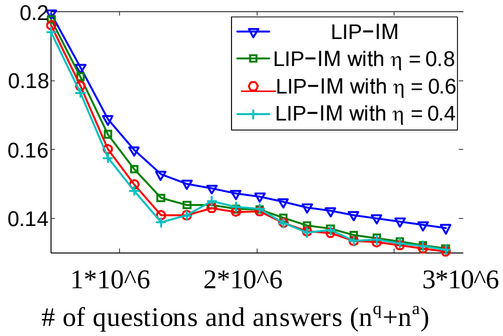
We use the data from two real CQA sites, i.e., Stack Over- flow (SO) and Mathematics Stack Exchange (Math). SO and Math are two popular CQA sites for programming and math, respectively. Both data sets are officially published and publicly available[3].

For SO and Math data, we use both content and con- textual features. For content features, we adopt the "bag of words" model to extract content features after removing the infrequent words. This model is widely used in natural language processing where the frequency of each word is used as a feature for training.

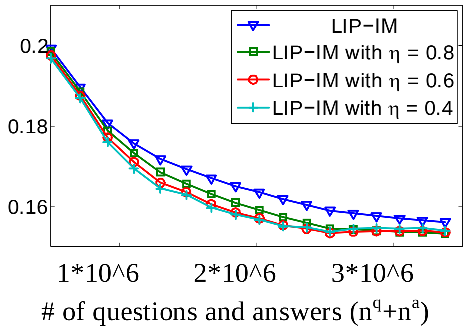
## IX. EFFECTIVENESS RESULTS

### *A. The Effectiveness Comparisons*

We first compare the effectiveness of the proposed algorithms (i.e., LIP-KIM, LIP-KIMA, and LIP-KIMAA) with two state-of-the-art non-linear regression methods, i.e., kernel ridge regression (KRR) [4] and support vector regression (SVR).The prediction results of questions and answers on the SO and Math data sets are shown in Fig. 2. On SO data, we only report the first few points because some of the algorithms(e.g., KRR) cannot finish training within 1 hour. We do not report the results by linear models (e.g., linear ridge regression) since their performance (RMSE) is much worse than non-linear models.



(a)Question voting prediction on SO data



(b)Answer voting prediction on SO data

Fig. (a) The effectiveness results of LIP IMF. Lower is better improve the accuracy of voting prediction.

Fig. (b) Fading the effects of old examples can help to lower is better improve the accuracy of voting prediction.

## X. CONCLUSION

In this article, we have proposed a family of algorithms to comprehensively and efficiently predict the voting scores of questions/answers in CQA sites. In particular, some of the proposed algorithms can capture three key aspects (non-linearity, coupling, and dynamics) that matter with the voting score of a post, while others can handle the special cases when only a fraction of the three aspects are prominent. In terms of computation efficiency, some algorithms enjoy linear, sub-linear, or even constant scalability. The proposed algorithms are also able to fade the effects of old examples and select a subset of features/examples. We analyze our algorithms in terms of optimality, correctness, and complexity, and reveal the intrinsic relationships among different algorithms. We conduct extensive experimental evaluations on two real datasets to demonstrate the effectiveness and efficiency of our approaches.

## REFERENCES

[1] T. Osbourn, “Getting the most out of the web,” Software, IEEE, vol. 28, no. 1, pp. 96–96, 2011.
[2] Y. Yao, H. Tong, T. Xie, L. Akoglu, F. Xu, and J. Lu, “Joint voting prediction for questions and answers in cqa,” in ASONAM, 2014.
[3] “Want a good answer? ask a good question first!” arXiv preprint arXiv:1311.6876, 2013.
[4] C. Saunders, A. Gammerman, and V. Vovk, “Ridge regression learning algorithm in dual variables,” in ICML, 1998, pp. 515–521.
[5] S. Haykin, Adaptive filter theory, 2005.
[6] Y. Engel, S. Mannor, and R. Meir, “The kernel recursive least-squares algorithm,” IEEE Transactions on Signal Processing, vol. 52, no. 8, pp. 2275–2285, 2004.
[7] N. Aronszajn, “Theory of reproducing kernels,” Transactions of the American mathematical society, vol. 68, no. 3, pp. 337–404, 1950.
[8] C. J. Burges, “A tutorial on support vector machines for pattern recognition,” Data mining and knowledge discovery, vol. 2, no. 2, pp. 121–167, 1998.
[9] P. Drineas and M. W. Mahoney, “On the nystr¨om method for approximating a gram matrix for improved kernel-based learning,” The Journal of Machine Learning Research, vol. 6, pp. 2153–2175, 2005.
[10] G. Golub and C. Van Loan, “Matrix computations,” 1996.
[11] J.-L. Lin and J.Y.-C. Liu, "Privacy preserving item set mining through fake transactions," in Proceedings of the 2007 ACM symposium on applied computing. Seoul, Korea: ACM Press, 2007, pp. 375-379.
[12] Z.-Y. Chen and G.-H.Liu, "Quantitative Association Rules Mining Methods with Privacy-preserving," Sixth International Conference on Parallel and Distributed Computing, Applications and Technologies, 2005.PDCAT 2005, pp. 910-912, 2005.
[13] A. Evfimevski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), pp. 217 -228, 2002.
[14] S. Rizvi and J. R. Haritsa, "Maintaining Data Privacy in Association Rule Mining," VLDB, pp. 682-693, 2002.
[15] D. W. Cheung, S. D. Lee, and B. Kao, "A general incremental technique for maintaining discovered association rules," Proc. 5th Int. Conf. Database Systems Advanced Applications, pp. 1-4, 1997.
[16] J. S. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," 8th ACM SIGKDD Inter-national Conference on Knowledge Discovery and Data Mining, pp. 639 -644, 2002.
[17] M. Naveed, M. Prabhakaran, and C. A. Gunter, “Dynamic searchable encryption via blind storage,” in IEEE S&P, May 2014, pp. 639–654.
[18] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, “Secure ranked keyword search over encrypted cloud data,” in IEEE ICDCS, 2010, pp. 253–262.
[19] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, “Privacy-preserving multi-keyword ranked search over encrypted cloud data,” in IEEE INFOCOM, 2011, pp. 829–837.
[20] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, Y. T. Hou, and H. Li, “Privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking,” in ACM ASIACCS, 2013.
[21] B. Wang, S. Yu, W. Lou, and Y. T. Hou, “Privacy-preserving multi keyword fuzzy search over encrypted data in the cloud,” in IEEE INFOCOM, 2014, pp. 2112–2120.

[22] W. Zhang, S.Xiao, Y. Lin, J. Wu, and S. Zhou, “Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing,” IEEE Transactions on Computers, vol. 65, no. 5, pp. 1566–1577, May 2016.

[23] Z. Xia, X. Wang, X. Sun, and Q. Wang, “A secure and dynamic multi-keyword ranked search scheme over encrypted cloud data,” IEEE Transactions on Parallel and Distributed System, vol. 27, no. 2, pp. 340–352, 2015.

[24] Z. Fu, X. Sun, Q. Liu, L. Zhou, and J. Shu, “Achieving efficient cloud search services: Multi-keyword ranked search over encrypted cloud data supporting parallel computing,” IEICE Transactions on Communications, vol. E98-B, no. 1, pp. 190–200, 2015.

[25] H. Yin, Z. Qin, L. Ou, and K. Li, “A query privacy enhanced and secure search scheme over encrypted data in cloud computing,” Journal of Computer and System Sciences, http://dx.doi.org/10.1016/j.jcss.2016.12.003.

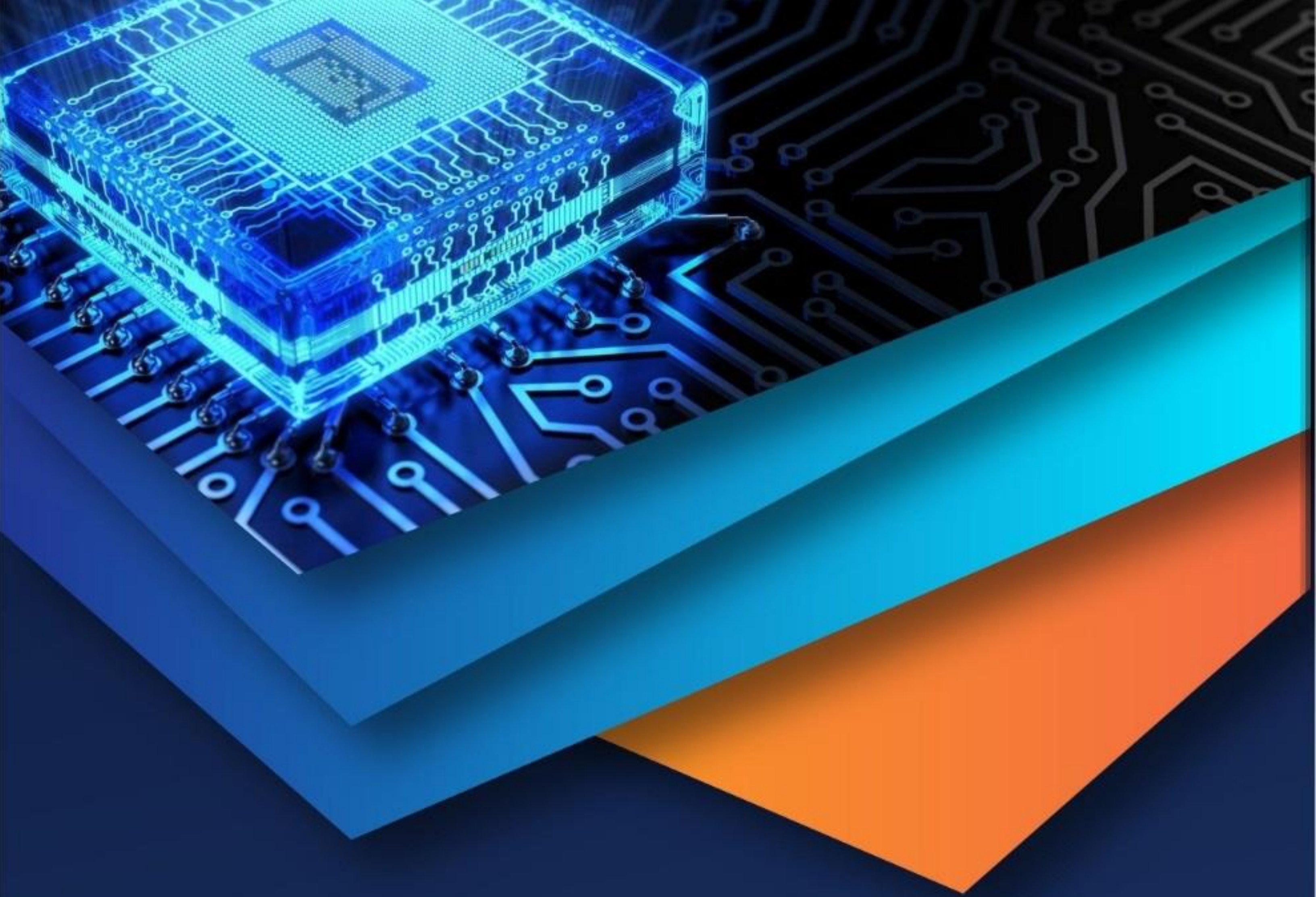
[26] B. Wang, B. Li, and H. Li, “Oruta” Privacy-preserving public auditing for shared data in the cloud,” IEEE Transactions on Cloud Computing, vol. 2, no. 1, pp. 43–56, 2014.

[27] C. Wang, N. Cao, K. Ren, and W. Lou, “Enabling secure and efficient ranked keyword search over outsourced cloud data,” IEEE Transactions on Parallel and Distributed Systems, vol. 23, no. 8, pp. 1467–1479, 2012.

[28] W. Sun, B. Wang, N. Cao, M. Li, W. Lou, and Y. T. Hou, “Verifiable privacy-preserving multi-keyword text search in the cloud supporting similarity-based ranking,” IEEE Transactions on Parallel and Distributed Systems, vol. 25, no. 11, pp. 3025–3035, 2014.

[29] Q. Zheng, S. Xu, and G. Ateniese, “Vabks: Verifiable attribute based keyword search over outsourced encrypted data,” in IEEE INFOCOM, May 2014, pp. 522–530.

[30] W. Sun, X. Liu, W. Lou, Y. T. Hou, and H. Li, “Catch you if you lie to me: Efficient verifiable conjunctive keyword search over large dynamic encrypted cloud data,” in IEEE INFOCOM, April 2015, pp. 2110–2118.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY