https://doi.org/10.48047/AFJBS.6.Si4.2024.112-121

African Journal of Biological Sciences

Journal homepage: http://www.afjbs.com

Research Paper Open Access

A FAST AND ACCURATE PRIVACY-PRESERVING MULTI-KEYWORD TOP-K RETRIEVAL SCHEMA OVER ENCRYPTED CLOUD DATA

E.AISHWARYA [1] , DR.A.PRANAYANATH REDDY[2]

[1]Pg Scholar, Department of CSE, Teegala Krishna Reddy Engineering College,Meerpet, Telangana,India.

[2]Associate Professor, Department of CSE, Teegala Krishna Reddy Engineering College,Meerpet, Telangana,India.

Volume 6, Issue Si4, 2024

Received: 12 Apr 2024

Accepted: 02 May 2024

doi:10.48047/AFJBS.6.Si4.2024.112-121

ABSTRACT:

Many big data applications in industries such as research and healthcare thrive due to the widespread availability of large-scale processing capability and scalable storage via cloud computing. To enhance data management and facilitate easier mining, numerous data owners opt to outsource their data to cloud servers. However, significant privacy concerns arise when sensitive data, such as electronic health records, is shared in the cloud with partially untrusted third parties. A common approach to mitigate these risks is encrypting data before outsourcing, albeit this diminishes the data's value and renders traditional data analysis techniques, such as keyword-based top-k document retrieval, obsolete.This research addresses the challenges of multi-keyword top-k search for encrypted big data, aiming to safeguard against privacy intrusions. To provide a secure and efficient solution, we introduce a novel index structure based on trees and a random traversal method. This approach enhances query data privacy while maintaining query accuracy by ensuring that identical queries yield distinct traversal paths to the index. Additionally, we describe a group multi-keyword top-k search method that employs partitioning to construct multiple tree-based indexes for each dataset, thereby improving query efficiency. When integrated, these methods form a robust framework for handling the proposed top-k similarity search.Our methodology offers superior privacy protection, scalability, and query processing speed compared to state-of-the-art approaches. Extensive testing on real datasets has confirmed these advantages.

Index Terms: Cloud Computing, Data Privacy, Multi-Keyword Search, Top-k Retrieval, Encrypted Data, Index Structure, Query Efficiency, Data Security, Big Data Applications, Secure Search.

I.INTRODUCTION

In recent years, cloud computing has become a game-changer for IT corporations as well as academic institutions. Important features like pay-as-you-go models and high scalability let cloud users to purchase strong computing resources just when required, removing worries about resource waste and labor-intensive platform maintenance. In order to speed up data management, effective data mining, and query processing, a large number of businesses and people have started to outsource data and put services on cloud servers.
To properly reap these benefits, however, privacy issues pertaining to outsourced data must be addressed. Sensitive data, such as emails, Electronic Health Records (EHRs), and financial transaction records, are included in a large number of datasets utilized in a variety of areas. Such sensitive data is vulnerable to illegal access and analysis when it is outsourced to cloud servers that may not be reliable. These datasets' analysis may provide light on important social concerns including government services, e-research, and healthcare. Therefore, before putting their data in the cloud, data owners should look for scalable, efficient, and privacy-preserving alternatives. A popular technique for protecting sensitive data while exchanging it is data encryption, which converts plaintext into ciphertext using mathematical operations and computational techniques, making it unintelligible to outsiders. Many approaches to data encryption have been put up for data outsourcing to cloud servers. But in terms of data value, these measures often come at a high cost, making conventional data processing techniques meant for unencrypted data useless. One popular data operator used in many applications to get data from databases is keyword-based search. Unfortunately, encrypted data cannot be immediately accessed by conventional data retrieval techniques. To protect data privacy, research is presently focused on running searches over encrypted data.

Numerous techniques founded on searchable encryption have been created, enabling to manage Boolean searches with one or more keywords. However, single-keyword search may not be sufficient for complicated queries, and Boolean search is unfeasible owing to high transmission costs. Because multi-keyword ranked search is more suited for a pay-as-you-go cloud approach, recent research has focused on it. However, many current techniques have scalability issues since they are unable to effectively encrypt vast amounts of data while maintaining excellent data security.This study focuses on multikeyword top-k search, a subset of multi-keyword ranked search that is often used in crucial situations, in light of these difficulties. Finding the top-k papers with the greatest relevance ratings is the goal of this method. By expressing documents and queries as vectors, we propose the vector space approach to improve multi-keyword search capabilities. For ranking in top-k search, we use the TF×IDF (term frequency × inverse document frequency) model as a weighting mechanism to calculate relevance scores. Additionally, we introduce a Group Multi-keyword Top-K Search (GMTS) approach using partitioning to improve

query efficiency and user experience. The data owner uses GMTS to classify dictionary terms into numerous categories and then create a searchable index for each category. We use champion lists, where each index only keeps the top-ck documents per keyword group (where c is a positive integer) in order to efficiently manage index sizes. In order to improve data security, we also offer the Random Traversal Algorithm (RTRA). In order to allow users to provide a different key for every query, the data owner builds a binary tree searchable index and randomly assigns a switch to each node.

This randomization alters traversal patterns and query results, enhancing security without compromising query accuracy.

Finally, we propose a secure and efficient solution by integrating RTRA with GMTS. Our contributions are summarized as follows:

1. The Random Traversal Algorithm ensures exceptional query accuracy and security by randomizing server traversal on the index to provide different responses for the same query.
2. We detail a secure searchable encryption system that uses the Random Traversal Algorithm to facilitate top-k similarity search across encrypted data, allowing precise control over query unlinkability levels.
3. Our experimental results demonstrate that our methods outperform state-of-the-art technologies in terms of efficiency and user privacy protection, particularly when handling large datasets.

II.LITERATURE SURVEY

- According to Dong Jin Park, Juyoung Cha, and Pil Joong Lee, traditional single-keyword searchable encryption systems usually generate an encrypted searchable index. The content of this index is hidden from the server unless appropriate trapdoors, which are produced using secret keys, are given. While Song et al. Were the first to investigate this concept within the framework of symmetric keys. Subsequent studies by Goh, Chang et al., and Curtmola et al. Expanded upon and defined advanced security. Secure ranked keyword search, which uses phrase frequency to rank results instead of presenting homogenous results, was the focus of our initial study. Unfortunately, it's only good for searches that include a single phrase.
- In the context of public keys, Boneh et al. Outlined the first idea of searchable encryption. Everybody with the public key can update the server's data, but only authorized users with the private key may access the data via searches. In most cases, the computational cost of public key solutions is high. Keyword privacy would also be compromised with the public key option as the server might encrypt any keyword using the public key and then decrypt it using the received trapdoor.
- Lucas Ballard, Seny Kamara, and Fabian Monrose have all reported improvements to search functionality; one suggestion is to use conjunctive keywords to search across encrypted data. Such techniques incur substantial overhead as a result of

their fundamental primitives, which include computation costs associated with bilinear maps and communication costs associated with secret sharing. Recent proposals have proposed predicate encryption systems as a more general search mechanism, as they provide both disjunctive and conjunctive search. Search results for conjunction keywords are "all-or-nothing," meaning they only show results that include all of the query's keywords. A disjunctive keyword search will nonetheless return any page that has some portion of the relevant words, regardless of how few terms are of interest, since the results are not distinguished. Just so we are clear, none of the existing Boolean keyword searchable encryption methods provide anonymous, ranked search over encrypted cloud data using multiple keywords. That is what we want to explore in this study. Remember that the inner product value is concealed unless it is zero, and that predicate encryption inner product searches only find out whether two vectors are orthogonal. Because it prevents the comparison of concealed inner products, predicate encryption cannot be used with ranking search. The expensive evaluation of pairing operations on elliptic curves is also crucial to most of these schemes.

III.PROBLEM STATEMENT

One common technique for safeguarding private information when exchanging data is data encryption. It combines algorithmic systems and mathematical computations to transform plaintext into cipher-text, which is unintelligible to other people. Encrypting sensitive data using one of the numerous suggested methods is standard procedure before transmitting it to distant servers in the cloud. The issue lies in the fact that standard data processing methods designed for unencrypted data function poorly when used to encrypted data since they often result in a significant loss of data usefulness. Keyword-based searches are the foundation of many database and information retrieval applications; however, encrypted data is incompatible with the common processing methods used by these searches. Therefore, figuring out how to run such searches on encrypted data while maintaining data privacy becomes a crucial field of research. Fortunately, several methods using searchable encryption have been studied. Handle searches with a single phrase and ensure that multiple keyword boolean searches are supported. However, single-keyword searches aren't intelligent enough to address complicated issues, and boolean searches are unfeasible because of the enormous communication cost they cause. For this reason, newer systems such priority multi-keyword ranked search operate on the pay-as-you-go cloud paradigm. However, these methods have serious scalability and performance problems when it comes to huge data encryption, and they are unable to provide both robust data security and great search efficiency at the same time.

IV.METHODOLOGY

Our primary focus is on multi-keyword top-k search, a subset of multi-keyword ranked searches that specifically returns the top-k items with the highest relevance scores. This approach has proven crucial in numerous applications as a preferred database operation. To facilitate multi-keyword search, we adopt the vector space model, utilizing vectors to represent both documents and queries. To compute relevance scores for ranking purposes, we employ the TF×IDF (term frequency × inverse document frequency) model as a weighting mechanism. This helps determine how pertinent documents are to specific queries.

Additionally, we introduce GMTS (Group Multi-keyword Top-k Search), a partition-based strategy that enhances query efficiency and user experience by enabling top-k similarity search across encrypted data. Implemented by the data owner, GMTS categorizes keywords into multiple groups and creates an index for each category, assuming the dictionary encompasses all possible keywords derived from all texts. To manage index sizes effectively, we utilize champion lists, ensuring each index contains only the top-ck documents per keyword group, where c is a positive integer and k is the number of documents.

Furthermore, our proposed Random Traversal Algorithm (RTRA) enhances data security. By constructing a searchable binary tree and assigning a random switch to each node, the data owner allows data users to apply a unique key to each query. This randomization enables users to alter query paths and results using different keys while maintaining high query accuracy.Finally, we address our research problem by integrating GMTS with RTRA, resulting in a secure and efficient solution. Our contributions can be summarized as follows:

- **Random Traversal Method:** Enhances security by causing the cloud server to randomly navigate the index, providing varied results for the same query while preserving query accuracy.
- **Secure Searchable Encryption Method:** Built on the Random Traversal Algorithm, this method supports top-k similarity search over encrypted data. The data owner can control query unlinkability levels, ensuring accuracy.
- **Performance:** Our experimental results demonstrate that our methods outperform state-of-the-art techniques in terms of efficiency and privacy protection. Particularly, our approach excels in scalability when handling massive datasets.

In conclusion, our study addresses the challenges of multi-keyword top-k search over encrypted data, presenting novel methodologies that effectively balance security and performance, thereby advancing the state of the art in searchable encryption.

V.SYSTEM ARCHITECTURE DIAGRAM

In consequence of the aforementioned problem, we provide a technique for ranking search terms over encrypted cloud data that is both fast and accurate, allowing for the efficient and accurate recovery of the top k most relevant documents. With FHOPE, the FASE system encrypts the query and index vectors. Order comparison, homomorphism multiplication, and homomorphism addition may be implemented by the FHOPE over encrypted data. So, the FHOPE may use

encrypted data to generate the relevance score, and the score is cipher text that will stay in order. The cloud server may use the relevance score to rank things without disclosing any information. Furthermore, the search results are returned by accurately computing the query vector and document vector, ensuring that no artificial keywords are included. The vectors for the document mark and the query mark are both created by us. A large number of irrelevant documents are effectively filtered by matching the document mark vector and query mark vector, and the time needed to compute the relevance score and ranking may be significantly reduced. When the cloud server ranks the results, it does so based on how closely the keywords match. If numerous sites match the same keyword degree, we rerank the search results using relevance scores. This enhances the precision of ranking.

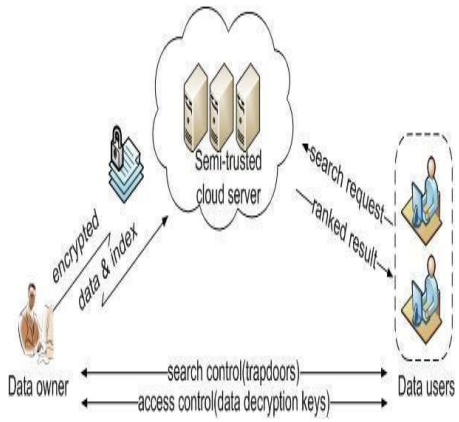


Figure.1 SYSTEM ARCHITECTURE

VI.MODEL DESCRIPTION:

- **System Model**

  To use our proposed system into action, we build the System Model in the first module. This model comprises the Admin, Users, Data Owners, and Cloud Servers. Granting access to Data Owners is the responsibility of the Administrator. The initial step for a Data Owner is to register, and the Admin must approve all registration requests. Following acceptance, the Data Owner will get an email with the necessary access credentials.

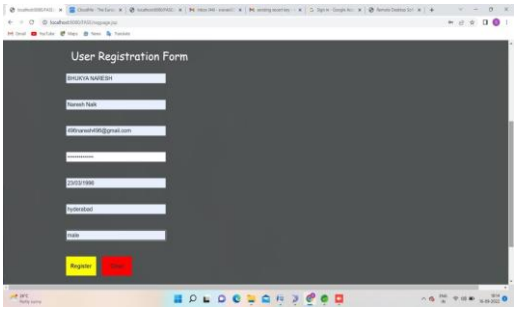
  A user's global identity and the collection of characteristics bestowed upon them by various attribute authorities make up the system's core functionality. A private key associated with these attributes is the user's right. A scalable plug-and-play architecture is possible with the help of the proposed method, which allows for the addition of new Data Owners to the system without negatively impacting existing users or Data Owners.

  Data Owners may send encrypted data to the Cloud Server using the Admin. Access control is implemented via encryption instead than relying on the server. In order to decipher the ciphertext, a user's attributes must match the access policy given in the cipher text; nevertheless, users' characteristics determine the quantity of information that may be accessed.

- **Data User Authentication**

  The administration server re-encrypts trapdoors for data users only after they authenticate. This prevents attackers from statistically attacking authorized users by assuming their identities. In the standard model, there are three stages to an authentication process: first, the requester and authenticator both have access to the

same secret key (k0); second, the requester encrypts their personal data (d0) using the secret key (k0); and finally, the authenticator verifies the requester's identity by decrypting the received data (d0)k0. For authentication to work, you need secret keys that change



over time as well as the user's related data from the past.

- **Illegal Search Detection**
  To ensure the authentication method is secure, dynamic secret keys and historical data are used. Even if an attacker gets their hands on the secret key, they'll still have to come up with reliable authentication data. The administration server becomes aware of them because they are unable to provide trustworthy authentication data without access to historical data, such as the request counter and the last request time. No amount of listening in on Uj's data or fabricating authentication data will be able to hide the secret key mismatch between the administration server and Uj, which will reveal any illegal behavior if Uj does a search.
- **Search over Multi-owner**
  Files encrypted for several data owners using distinct keys may be searched across by the suggested system utilizing multiple keywords. The cloud server will sort the search results by different data owners and then provide the top k results. It is there that all the encrypted files and keywords belonging to various data owners are kept. In response to a query, the cloud server examines all data owners' databases. Part one of this method involves the cloud matching the query keywords with all stored keywords to build a candidate file set. Part two involves the cloud ranking the files in this set to find the top k most relevant files. By encoding the relevance ratings, the proposed technique efficiently obtains the top k search results.

VII.RESULTS

Figure.2 User Registration

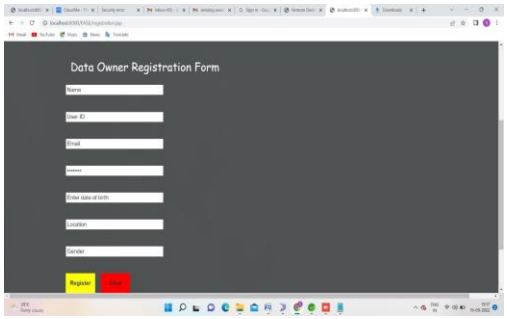


Figure.3 Data Owner Registration

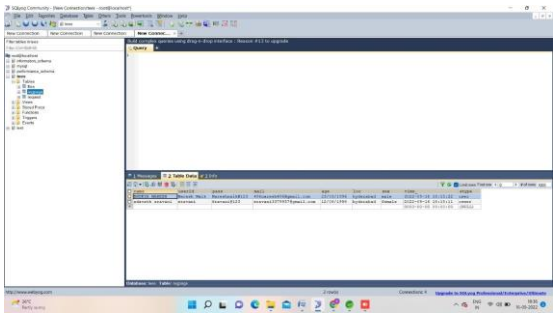


Figure.4 Database

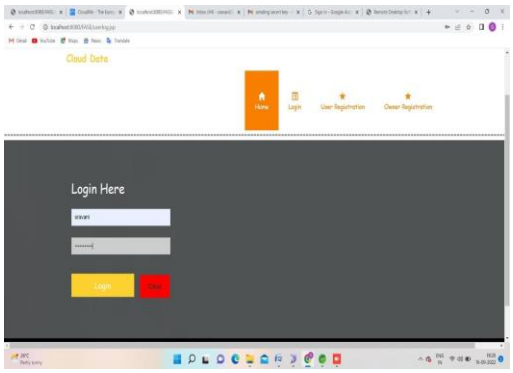


Figure.5 Data Owner Login

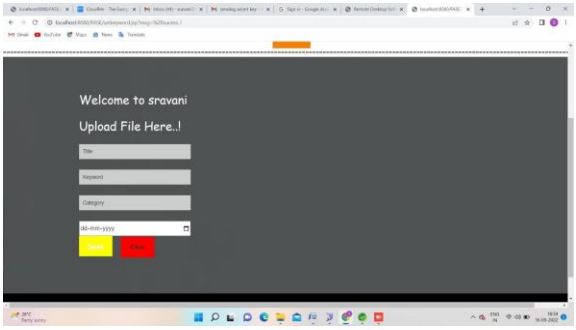
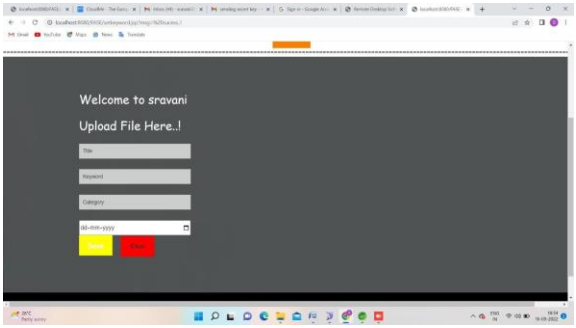


Figure.6 Data Owner File Upload



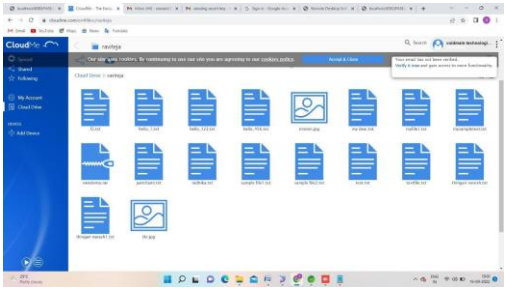


Figure.7 Upload File

Figure.8 Data Store in Cloud

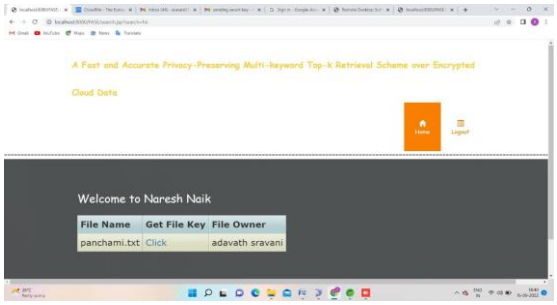


Figure.9 Data Search Keyword

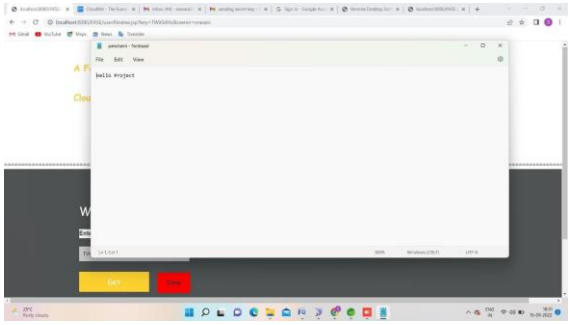


Figure.10 Download File

VIII.CONCLUSION

Improving the safety and effectiveness of top-k similarity searches using multiple keywords on encrypted data is our primary objective in this research. We begin with the random traversal strategy, which can do the following: for two identical queries with different keys, the cloud server may take different paths throughout the index, giving the data user diverse results while keeping the query accuracy high over time. The next step is to develop a group multi-keyword top-k search method. This method creates an index by saving only the top-ck documents of each word group after splitting the dictionary into several groups. As a result, search efficiency is enhanced. To ensure the query is unlinkable, we use the random traversal procedure to get the RGMTS. Because of this, it may be harder for servers in the cloud to conduct linkage attacks on two similar requests. Data owners also have some leeway in deciding how much query unlinkability they want by adjusting the value of E. In the

end, our methods are more efficient and secure than the state-of-the-art methods, according to the experimental results.

IX.FUTURE ENHANCEMENT

The FASE system isn't perfect, however. The ability to add, delete, or update documents kept on the server, as well as add new documents to the initial data collection, makes a technique that allows dynamic operations vital. It may be a rewarding but difficult endeavor to develop a searchable and dynamically operable encryption system; we want to do so in future work.

X.REFERENCES

[1] S.S.M. Chow, Y.J. He, L.C.K. Hui, and S.-M. Yiu, "SPICE – Simple Privacy-Preserving Identity- Management for Cloud Environment," Proc. 10th Int'l Conf. Applied Cryptography and Network Security (ACNS), vol. 7341, pp. 526-543, 2012.

[2] L. Hardesty, Secure Computers Aren't so Secure. MIT press, http://www.physorg.com/news176107396.html, 2009.

[3] C. Wang, S.S.M. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage," IEEE Trans. Computers, vol. 62, no. 2, pp. 362-375, Feb. 2013.

[4] B. Wang, S.S.M. Chow, M. Li, and H. Li, "Storing Shared Data on the Cloud via Security- Mediator," Proc. IEEE 33rd Int'l Conf. Distributed Computing Systems (ICDCS), 2013.

[5] S.S.M. Chow, C.-K. Chu, X. Huang, J. Zhou, and R.H. Deng, "Dynamic Secure Cloud Storage with Provenance," Cryptography and Security, pp. 442-464, Springer, 2012.

[6] D. Boneh, C. Gentry, B. Lynn, and H. Shacham, "Aggregate and Verifiably Encrypted Signatures from Bilinear Maps," Proc. 22nd Int'l Conf. Theory and Applications of Cryptographic Techniques (EUROCRYPT '03), pp. 416-432, 2003.

[7] M.J. Atallah, M. Blanton, N. Fazio, and K.B. Frikken, "Dynamic and Efficient Key Management for Access Hierarchies," ACM Trans. Information and System Security, vol. 12, no. 3, pp. 18:1-18:43,2009.

[8] S. Tang, X. Li, X. Huang, Y. Xiang, and L. Xu, "Achieving simple, secure and efficient hierarchical access control in cloud computing," IEEE transactions on computers, vol. 65, no. 7, pp. 2325–2331, 2015.

[9] J. Ning, Z. Cao, X. Dong, K. Liang, H. Ma, and L. Wei, "Auditable σ-time outsourced attribute-based encryption for access control in cloud computing," IEEE Trans. Information Forensics and Security, vol. 13, no. 1, pp. 94–105, 2018.

[10] M. J. Atallah, M. Blanton, N. Fazio, and K. B. Frikken, "Dynamic and efficient key management for access hierarchies," ACM Trans. Inf. Syst. Secur., vol. 12, no. 3, pp. 18:1–18:43, 2009.

[11] E. S. V. Freire, K. G. Paterson, and B. Poettering, "Simple, efficient and strongly ki-secure hierarchical key assignment schemes," in Topics in Cryptology - CT-RSA 2013 - The Cryptographers' Track at the RSA Conference 2013, San Francisco,CA, USA, February 25-March 1, 2013. Proceedings, pp. 101–114, 2013.

[12] A. Castiglione, A. D. Santis, B. Masucci, F. Palmieri, A. Castiglione, and X. Huang, "Cryptographic hierarchical access

control for dynamic structures," IEEE Trans. Information Forensics and Security, vol. 11, no. 10, pp. 2349–2364, 2016.

[13] Q. Jiang, J. Ma, and F. Wei, "On the security of a privacy-aware authentication scheme for distributed mobile cloud computing services," IEEE Systems Journal, vol. 12, no. 2, pp. 2039–2042, 2018.

[14] M. Hwang, W. Tzeng, and W. Yang, "An access control scheme based on chinese remainder theorem and time stamp concept," Computers & Security, vol. 15, no. 1, pp. 73–81, 1996.

[15] X. Zou, B. Ramamurthy, and S. S. Magliveras, "Chinese remainder theorem based hierarchical access control for secure group communication," in Information and Communications Security, Third International Conference, ICICS 2001, Xian, China, November 13-16, 2001, pp. 381–385, 2001.

[16] D. He, S. Zeadally, N. Kumar, and W. Wu, "Efficient and anonymous mobile user authentication protocol using self-certified public key cryptography for multi-server architectures," IEEE Transactions on Information Forensics and Security, vol. 11, no. 9, pp. 2052– 2064, 2016.

[17] H. Petersenl and P. Horster, "Self-certified keysconcepts and applications," in Communications and Multimedia Security, pp. 102–116, Springer, 1997.

[18] D. He and D. Wang, "Robust biometrics-based authentication scheme for multiserver environment," IEEE Systems Journal, vol. 9, no. 3, pp. 816–823, 2015.

[19] V. Odelu, A. K. Das, and A. Goswami, "A secure biometricsbased multi-server authentication protocol using smart cards," IEEE Transactions on Information Forensics and Security, vol. 10, no. 9, pp. 1953–1966, 2015.

[20] Q. Feng, D. He, S. Zeadally, and H. Wang, "Anonymous biometrics-based authentication scheme with key distribution for mobile multi-server environment," Future Generation Comp. Syst., vol. 84, pp. 239–251, 2018.