Journal of
Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7766108

EFFICIENT N-GRAM BASED FUZZY KEYWORD SEARCHOVER ENCRYPTED DATA IN CLOUD

Panchal Mital Nikunj [1],Dr. Dushyantsinh B. Rathod [2]
[1]PhD Scholar,Faculty of Engineering and Technology,
Sankalchand Patel University,Visnagar,Gujarat
[2] Professor & HOD,Computer Engineering Department,
Ahmedabad Institute of Technology,Ahmedabad
mital@ldce.ac.in

Abstract

Since approximately two decades ago, the term "cloud computing" has become widely used among IT professionals. The advantages of cloud services have no boundaries. Users can gain from cloud computing in several ways, including the flexibility to pay as they go and the availability of computer resources as needed. For customers, it is quite convenient to receive data and information whenever they need it. Only what they utilise must be paid for. The costs for cloud computing and storage services should be transparent to users and enterprises. These services provide a range of data processing and storage options. Besides this, customers have a very tough time locating the best cloud services for keeping various options of usage. In this research, we have considered the problem of security and privacy while receiving data from the data owner. The process includes multiple algorithms for securing data, accessing keys and transferring data to the data user.Security of data is enhanced with the AES encryption model for sets of n-gram words. The set is maintained with an index to obtain the sequence of data along with its key. The doc is encrypted to maintain privacy while sending and storing it on the cloud server. Parameters of comparison like a list of fuzzy keywords, multi keywords and asingle keyword are demonstrated.

Keywords — cryptography, encryption schemes, fuzzy keywords,Keyword Search,privacy, trapdoor

1. INTRODUCTION

Users who use cloud storage can store data online and view it on any device. Local data is increasingly being stored on cloud servers because they are practical, affordable, and simple to use. The concept of public-key encryption with keyword search (PEKS) was introduced by Boneh et al[2][3]. to provide flexible usage of the encrypted data. Unfortunately, most of the PEKS schemes are not secure against inside keyword guessing attacks (IKGA), so the keyword information of the trapdoor may be leaked to the adversary.To preserve your privacy, data transmitted over the cloud is encrypted[1]. Messages can be encoded using encryption so that only those with the proper access can read them. The data has been encrypted using a variety of encryption methods. This results in a significant issue for encrypted data search. The symmetric encryption algorithm, also known as the key algorithm, and the asymmetric encryption algorithm are the two key-based encryption methods (or called public key algorithm). The fundamental difference is that symmetric encryption algorithms may readily extract the decryption key from the encryption by using the same key for both encryption and decryption.

The possibility of attacks is diversified with the use of cloud computing for multiple services and multiple platforms. Lack of user control, potential illegal secondary usage, data proliferation, cross-border data flow, dynamic provisioning, retention of data, ensuring data has been destroyed, and privacy breaches are the main privacy challenges in public clouds [7]. Access, control over the data lifecycle, availability and backup, a lack of standards, multi-tenancy, and audit are examples of security problems. Weak trust ties and low consumer trust are two trust difficulties [7]. The many security vulnerabilities that exist in cloud environments were likewise divided into levels [8]. Depending on the level of protection offered, various solutions were offered to address the security challenges in the cloud.The comparison of advantages and disadvantages of cloud computing services based on the type of network is mentioned in the table below.

Type of cloud	Advantage	Disadvantage
Public cloud	• Ease of use • Unlimited scalability • Predictable operating expenses • Service provider manages the services • No upfront cost	• Some services are hidden behind firewalls • Cost is higher • Services are managed by the user •
Private cloud	• Control over data • More secure • Provide services at low cost • Responsibility is bear by staff	• Limited visibility • Less services provided in comparison to public and hybrid • User is solely responsible to manage
Hybrid cloud	• More control over cloud environment • Better administration with simpler setup • Useful for monitoring activities in secured platform	• Configuration is difficult • Cost can be high • Needs to setup on premises software.

Table -1Comparison of services on different cloud

Software rules that let users utilise the security tools they require at the levels where the problems occur. The security over cloud is categorized as legal issues, governance, data security , virtualization, network security, interfaces and compliance.

The threats over Domain no. Cloud computing architectural framework is shared technology issues. In governance and enterprise risk management face account hijacking, malicious insiders, and abusive use of cloud services. In legal issues, the problems of insufficient due diligence while information management and data security map the data breaches, data loss, audit hijacking and malicious insiders. Insecure APIs can be the problem faced by interoperability and portability. Identity, entitlement and access management can catch malicious insiders attack and shared technology can be an issue. Virtualization needs to handle problems of data breaches, data loss and shared technology issues. Among all D11 – Encryption and key management suffer the account hijacking, denial of service, insecure APIs and shared technology issues. To manage the attacks in encryption and key management there are various mitigation strategies proposed. Identity based encryption attribute based encryption, fully homomorphic encryption, signature based encryption and structured encryption helps in reducing the complexity of the key generation process and strengthens the key. AES, DES, RSA, and Blowfish are modern key generation algorithms to build an efficient cryptography system.

2. RELATED WORK

Techniques that help in maintaining privacy and data security are derived from mathematical concepts and predefined algorithms. Some of the most used techniques are:

Phases in searchable encryption is inevitable. Multiple phases that deals with trapdoor generation, key generation, encryption, index generation and decryption of documents are categorised in different phases. The major advantage is to use them for separate stages on platforms[1].Privacy leakage is not an issue that provide meaningful information to trespasser as its generated with probabilistic based trapdoor.

Cryptography - Cryptography is the process of generating secret codes, enabling the confidentiality of communication through an insecure channel [1]. It protects against unauthorized parties by preventing unauthorized alteration of use. Generally speaking, it uses a cryptographic system to transform a plaintext into a cipher text. using most of the time a key. It has different Encryption and Decryption algorithms to do so.

Cipher Text - This is the scrambled message produced as output from Encryption algorithm. It depends on the plaintext and the secret key. For a given message, two different keys will produce two different cipher texts.

Encryption - Encryption is the process of converting data, in plain text format into a meaningless cipher text by means of a suitable algorithm. The algorithm takes secret key and plain text as input and produces cipher text.

Decryption - Decryption is converting the meaningless cipher text into the original information using decryption algorithms. The decryption algorithm is inverse of encryption algorithm. This takes key and cipher text as input and produces original plain text[2].

Symmetric Key Cryptography - It uses the same secret (private) key to encrypt and decrypt its data. It requires that the secret key be known by the party encrypting the data and the party decrypting the data.

Asymmetric Key Cryptography - Asymmetric uses both a public and private key. This allows for distribution of your public key to anyone with which they can encrypt the data they want to send securely and then it can only be decoded by the person having the private key[3].

To search over encryption data multiple search techniques are invented which are described below.

In the paper of [7] they proposed an efficient verifiable keyword-based semantic search scheme. Comparing to most of the existing searchable encryption schemes, the proposed scheme is more practical and flexible, better suiting users different search intensions. Moreover, the proposed scheme protects data privacy and supports verifiable searchability, in the presence of the semi honest server in the cloud computing environment.

In the paper of [6] proposed an effective approach to solve the problem of synonym based multi-keyword ranked search over encrypted cloud data. The main contributions are summarized in two aspects; synonymbased search and similarity ranked search. The search results can be synonyms of the predefined keywords, not the exact or fuzzy matching keywords, due to the possible synonym substitution and/or her lack of exact knowledge about the data. The vector space model is adopted combined with cosine measure, which is popular in information retrieval field, to evaluate the similarity between search request and document.

In the paper [8] solved the problem of multi-keyword ranked search over encrypted cloud data and establish a variety of privacy requirements. Among various multikeyword semantics, we choose the efficient similarity measure of coordinate matching i.e. as many matches as possible, to effectively capture the relevance of outsourced documents to the query keywords and use inner product similarity to quantitatively evaluate such similarity measure. For meeting the challenge of supporting multikeyword semantic without privacy breaches, they proposed a basic idea of MRSE using secure inner product computation. Then they have given two significantly improved MRSE schemes to achieve various stringent privacy requirements in two different threat models.

In the paper [8] proposed the first verifiable SSE scheme, which offers data privacy. Verifiable searchability and efficiency, in the presence of an unusually strong adversarial server in cloud computing environment. The rigorous security analysis together with through experimental evaluations on a resource-constrained device using real data sets confirms that the proposed VSSE realizes our design goals and is a promising solution to mediate the conflicts between data usability and data privacy in such scenario.

I reviewed related work and illustrate the difference of different keyword-based search techniques. Li at al [6] firstly proposes a fuzzy keyword search scheme over encrypted cloud data. Wang et al [18] proposed a secure ranked search scheme. This scheme supports only single keyword search while fuzzy keyword search scheme tacks the problem of minor typos and format inconsistencies. Yang et al [4] propose a verifiable search scheme which can prove correctness and completeness of result efficiently. Based on VSSE and fuzzy keyword search, Wang et al [18] proposes a scheme supporting both verification and fuzzy search, but the scheme ignores result ranking.

One is that most of these schemes supports only exact keyword search. That means returned result is completely dependent on whether query terms users enter match pre-set keywords. The other one is that most existing searchable schemes assume that the cloud server is honest-but-curious. However, Yang et al [4] notice that the cloud may be selfish to save its computation or download bandwidth. That is, the cloud server might conduct only a fraction of search of search operation or return a part of result honestly.

Besides, Fu et al [7] recently proposed a multi-keyword search scheme in encrypted cloud environment which can achieve synonym query. The main contribution of the scheme is that it solves the problem of synonym search.

3. Methodology

Flow of Diagram:

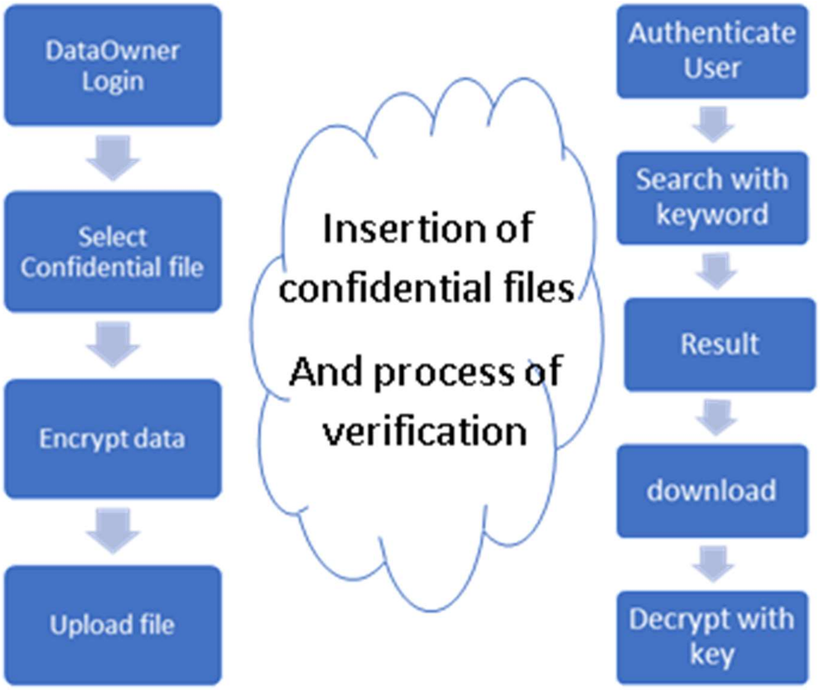

Figure -1Processing secure data sharing over cloud

The project's process flow is depicted in the work flow diagram. All key words pertaining to our project are included in the diagram. The Data owner login, which is the authorised user with full access to view and edit all confidential data stored in the cloud, is provided in the diagram above. In the diagram that is provided, we have displayed the key steps and how they are carried out in proposed research. uploading a file and using a keyword to search for it. The first is that in order to upload a file to the cloud, an administrator must first log in. After choosing the file he wants to upload, the system performs file encryption before uploading the file to the cloud and storing it there.

The detail architecture of system is presented with all the methodologies and algorithms used to fine grain the process to enhance security perspective.

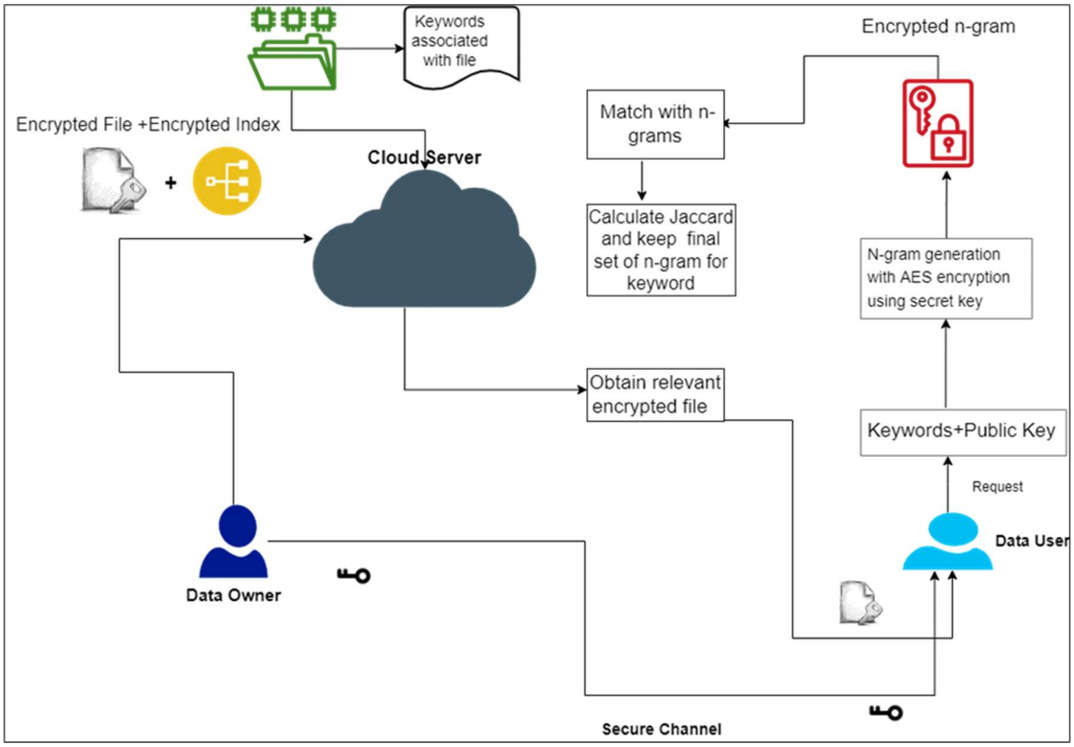
Keywords associated with file
Encrypted n-gram
Match with n-grams
Encrypted File +Encrypted Index
Cloud Server
Calculate Jaccard and keep final set of n-gram for keyword
N-gram generation with AES encryption using secret key
Obtain relevant encrypted file
Keywords+Public Key
Request
Data User
Data Owner
Secure Channel

Figure-2 Proposed architecture

The architecture explains that in today's cloud storage commodities, fuzzy search is crucial to retrieving information. For example, suppose some cloud users submit their queries with a few typographical errors or without having a clear understanding of the underlying keywords of stored data on the cloud. Before outsourcing to the cloud, the data is encrypted to protect user privacy. This may lead to confusion or misunderstandings about the data, its flexibility and its efficiency in usage. This work claims that, given the similarity of the keywords, n-gram with Jaccard similarityequation can produce a top ranked fuzzy keyword search. After this process, the encrypted files and index is searched for the appropriate word match that accumulates the files in the result data. In the meanwhile, the authenticity of the user is verified. Authenticated user is allowed to download the file and provided the key , user can decrypt the document. The process is divided into three phases for which the steps are as below.

Phase 1:

Step-1 Generate encrypted file ids with ABE.(at server if used public cloud service)

Step-2 Produce secret keys with AES

Step-3 Create encrypted file ids.

Step-4 Add secret keys with file.

Step-5 Upload encrypted files on server.

Phase 2:

Requests secret key on any attributes set.

Returns the secret key for Request R

Store secret key to S .

Challenge :

Receives two messages m1,m2(1. Request id and 2. Files) then returns Ciphertext CT (in the form of file from cloud server).

Phase 3:

Repeat as Phase 2.

Functions used in the proposed hybrid algorithm are AES , ABE and n-gram. Jaccard similarity is performed to match the keyword with entered fuzzy keyword. Trapdoors are unaffected by the keyword size for the same number of query keywords. In the process INDEX ENCRYPTION processassume that each index's keyword dictionary and file numbers are defined as u and d, respectively. During the creation of the index tree, O(d) nodes are produced, and the encryption process takes O(u) time for each node. The construction of an index has O time complexity (ud). Then, MERGE INDEX Assume that m is an n-gram-based keyword set to be merged. It is the same as combining their roots to combine m indexes. To merge these encrypted indexes, O(um) is required. Consequently, the merging index's time complexity is O. (um). At last, its presume that n is the total number of leaf nodes that include at least one search term. The max range of the merged index is a log and the time required to calculate the relevance score is O(u). The token process is by choosing the keyword u randomly and generating a new number of higher lengths that is encrypted by merging the timestamp. The token is generated by extracting 8-bit digits as max length. For public key k in {0,1}n to map n (number of keywords) then f(x,y) is a one way hash function. Trapdoor is generated by prime function and Rabin carp algorithms combination.

Proposed Flow:

We design a new dynamic secure index structure based encryption. Benefiting from the property of the special n-gram words with encrypted index simultaneously supports efficient search and dynamic update operations like insertion and searching of files flexibly. The index includes the encrypted keyword-file relevance scores, which are computed and encrypted by AES encryption. To perform search, server only needs to conduct simple vector multiplication between search token and index to judge whether the search keywords appears, then compute the sum of the multi-keyword weight scores on the encrypted file.

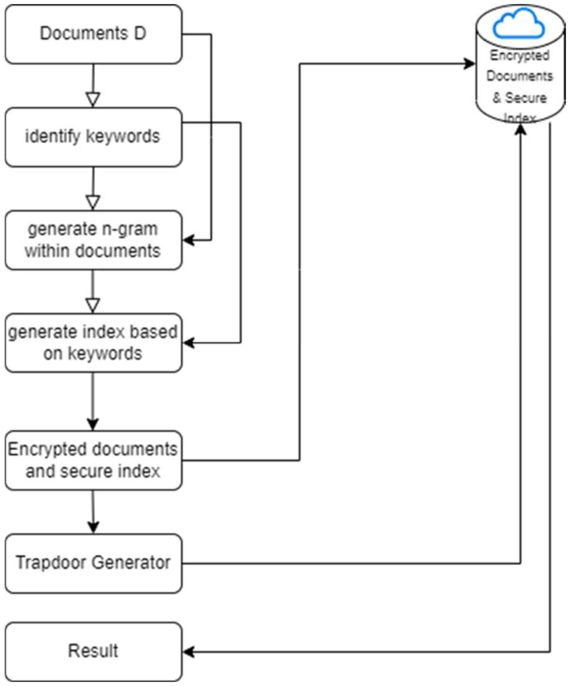
Documents D
identify keywords
generate n-gram within documents
generate index based on keywords
Encrypted documents and secure index
Trapdoor Generator
Result
Encrypted Documents & Secure Index

Figure 3 : Proposed Flow

4. Results and Analysis

We have implemented our proposed research on java platform with java's library pairing with bcprov and cryptography supported jars.This application require that the client not add a group of documents to the server but also add related keywords. It means that users can add the encrypted indices and files securely. The below stated table is demonstrating the results of comparison of time complexity of process of searching a document by entering keyword with documents from Enron dataset. The time taken to process is mentioned.

Enron dataset						
path	**File name**	**File size**	**Keyword type**	**Keywords**	**Time complexity**	**Time taken to process**
allen-p/all documents	4.txt	7 kb	multi keyword	photo, holiday, photographs	O(n)	0.243 sec
allen-p/all documents	1.txt	10 kb	fuzzy keyword	addision, peces	O(n)	0.450 sec
4168rnold-j/2000 conference	1.txt	1kb	single keyword	check	O(n/2)	0.122 sec

Table-2 : Time taken for different documents of enron email dataset

Given below table is comparing the time taken to process user owned documents which are created for personal use. The results show that the complexity remains the same searching with fuzzy keywords.

	Local files					
Sr.no	**File name**	**File size**	**Keyword type**	**Keywords**	**Time complexity**	**Timetaken to process**
1	laptop.txt	5kb	multi keyword	data, laptop, check	O(n)	0.124 sec
2	mobile.txt	10kb	fuzzy keyword	infor, teknologi	O(n)	0.200 sec
3	test.txt	7kb	single keyword	programming	O(n/2)	0.87sec

Table-3 : Time taken for different documents of local document set

Accuracy calculation with private hosting with VPSAccess is provided with specific boundaries, like firewall settings and VPN of the organization to access it. Its demonstrated for three types of keywords search, multi , fuzzy and single. The results are promising and does not vary on the size of the documents set.

Multi keywords

no. of keywords	Size of documents (KB)	Total indexed keywords	TP	TN	ACCURACY
3	121	23	2	1	94%
5	139	49	3	2	92%
7	168	10	5	2	95%
6	154	20	5	1	90%

Table-4 : Accuracy of results for multiple keywords

Fuzzy keywords + multi keywords

no. of keywords	Size of documents (KB)	Total indexed keywords	TP	TN	ACCURACY
2	121	23	2	1	97%
3	139	49	2	1	87%
4	168	10	3	1	78%
1	154	20	1	1	98%

Table – 5 Accuracy of results for multiple Fuzzy keywords

Single keyword

no. of keywords	size of documents (KB)	total indexed keywords	TP	TN	ACCURACY
1	121	23	1	0	98%
1	139	49	1	0	98%
1	168	10	1	0	97%

1	154	20	1	0	98%

Table – 6 : Accuracy of results for single keywords

Space complexity comparison stated in table below is comparatively promising as its occupying less space as compared to existing algorithms.

Comparison of computational space complexity

	single keyword search	multi keyword search	multi+fuzzy keyword search
scheme function[11]	θ(mn)	θ(mn + m2 + en2)	θ(mn) | O(n + 1)
ranked keyword search[11]	θ(mn)	θ(mn)	θ(2Qm + mn)
unranked keyword search[11]	θ(2m)	θ(2Qm)	θ(2Qm + n+1)
proposed research	θ(n)	θ(n+k)	θ(n+1)

Table-7 : Space complexity calculated for proposed research after implementation

Time complexity comparison

size of documents collection	1000	1500	2000	3000
bloom filter[11]	18	40	58	80
tree-based index[7]	40	75	105	148
btcs search[18]	320	600	800	1000
proposed algorithm	16.52	34.7	49.2	65.901

Table-8 : Comparison of time complexity over different approach

In response to the search request, the total time computation cost in is O(N)). And the server needs to do search matching keywords to perform search. for generating fuzzy keyword search result with different keywords size, the results displayed in table above are noted.

5. Conclusions

To provide searchable encryption based on the type of keywords, numerous solutions are developed and examined. Along with the words formed, fuzzy keywords that are ambiguously related to the words entered yet present in the document are also encrypted. It is advised to use prime inner product encoding to match words to vectors [13]. One of the most effective methods that outperform word direct matching algorithms is keyword vectorization. Word pattern function is used to generate fuzzy keyword searchable encryption approaches, and the character appearing technique suggests a maximum error tolerance value in cases where a secret key component is utilised to decrypt a cypher text component.

6. References

[1] Zi-Yuan Liu, Yi-Fan Tseng, Raylin Tso, Masahiro Mambo, and Yu-Chi Chen,“Public-key Authenticated Encryption with Keyword Search: Cryptanalysis, Enhanced Security, and Quantum-resistant Instantiation∗”IEEE, 2021.

[2] Shweta Agrawal, Dan Boneh, and Xavier Boyen. 2010. Efficient Lattice (H)IBE in the Standard Model. In EUROCRYPT

[3] Shweta Agrawal, Dan Boneh, and Xavier Boyen. 2010. Lattice Basis Delegation in Fixed Dimension and Shorter-ciphertext Hierarchical IBE. In CRYPTO.

[4] Lu Yang, Wang Gang, Li Jiguo, “Keyword guessing attacks on a public key encryption with keyword search scheme without random oracle and its improvement”,ScienceDirect, 2021.

[5] Yuan Ping 1, Wei Song 2, Zhili Zhang 1 , Weiping Wang 3 and Baocang Wang 2,4,, " A Multi-Keyword Searchable Encryption Scheme Based on Probability Trapdoor over Encryption Cloud Data," MDPI, 2020.

[6] Meng Meng Li, Guijuan Wang, Suhui Liu, and Jiguo Yu, “ Multi-keyword Fuzzy Search over Encrypted Cloud Storage Data ,” Elsevier, school of computer science,P.R.china, 2021.

[7] Jing Chen, Kun He, Lan Deng, Quan Yuan, Ruiying Du, Yang Xiang, and Jie Wu, “EliMFS: Achieving Efficient, Leakage-resilient, and Multi-keyword Fuzzy Search on Encrypted Cloud Data,” Proceedings of IEEE 2017.

[8] Qin. Liu and Yu. Peng, Shuyu Pei, Jie Wu, Tao Peng, Guojun Wang, “Prime Inner Product Encoding for EffectiveWildcard-based Multi-Keyword Fuzzy Search,” Proceedings of IEEE 2020.

[9] Siani Pearson and Azzedine Benameur, "Privacy, Security and Trust Issues Arising from Cloud Computing," in 2nd IEEE International Conference on Cloud Computing Technology and Science, USA, 2010.

[10] Mhammed Chraibi, Hamid Harroud and Abdelilah Maach, "Classification of Security Issues and Solutions in Cloud Environments," in IIWAS '13 Proceedings of International Conference on Information Integration and Web-based Applications & Services, New York, 2013.

[11] Xinrui Ge1, Jia Yu 1,2,3, Chengyu Hu 4, Hanlin Zhang1, And Rong Hao1, “Enabling Efficient Verifiable Fuzzy Keyword Search Over Encrypted Data in Cloud Computing,” Proceedings of IEEE-ACCESS,2018.

[12] J Hong Zhu1, Zhuolin Mei1(B), Bing Wu1, Hongbo Li1, and Zongmin Cui2, “Fuzzy Keyword Search and Access Control over Ciphertexts in Cloud Computing,” Proceedings of SPRINGER 2017, School of computer science and technology, wuhan, china.

[13] Shaojing Fu1,2 · Qi Zhang1 · Nan Jia1 · Ming Xu1, “A Privacy-preserving Fuzzy Search Scheme Supporting Logic Query over Encrypted Cloud Data,” SPRINGER-2020.

[14] Jin Li, Xiaofeng Chen, “Efficient multi-user keyword search over encrypted data in cloud computing,” Computing and Informatics, Vol. 32, 2013, 723-738,.

[15] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of IEEE Symposium on Security and Privacy'00,2000.

[16] https://eprint.iacr.org/2019/639.pdf

[17] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, ‘‘Secure ranked keyword search over encrypted cloud data,’’ in Proc. IEEE 30th Int. Conf. Distrib. Comput. Syst., Genova, Italy, Jun. 2010, pp. 253–262.

[18] Prasanna, B. T., and C. B. Akki. "A comparative study of homomorphic and searchable encryption schemes for cloud computing”arXiv preprint arXiv:1505.03263(2015).