Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

# Search Queries Over Encrypted Cloud Data using Hybrid Encryption

Prachi Chaudhari
Computer Science and Engineering
Sandip University , Nashik, India

V.S.Narayana Tinnaluri
Computer Science and Engineering
Sandip University, Nashik, India

***Abstract*— The Data security in the cloud has always been a big concern for both of its customers and cloud service providers. This is mainly due to threat issues from the third party audition of the data and from its internal stakeholders, that is their employees. So maintaining this security intact is always being a headache and challenging task in the cloud due to its gigantic structure. Generally to achieve this, data is always kept in encrypted format and take care that data integrity is not compromised during the audition process and other internal activities. The actual integrity problem arises when the owner of the data wants to search the stored files in the cloud based on some keywords. This process eventually needs to decrypt the data and then search for the fired query. This process actually compromises the security of the data easily, to overcome this many methodologies are existed which search the query without decrypting the stored files. Most of them uses the similarity or correlation methods to achieve this, which adds a bit of time complexity in the process. So the proposed system puts forwards an idea of searching the data over encrypted files using search tokens and Jaccard distance to enhance the time of searching more efficiently.**

***Keywords*— *Cloud computing, DropBox, Jaccard Coeffiecient, Search tokens.***

## I. INTRODUCTION

Cloud computing is one of the fastest growing technology. The Cloud is based on various different services that contain various benefits ranging from software development platforms, server, to storage and remote computing. This has all been possible due to the advent of the internet and it is how these services are linked to the customer; through the internet. There has been a large scale adoption of this technique as it is highly convenient. This has also resulted in increased research into this platform.

Due to increased research in this area, there has been a lot of development in terms of the infrastructure that supports this platform. The infrastructure has been highly augmented to support the growing number of users. A lot of organizations, as well as individual users, are adopting this platform, that is accelerating the growth of this infrastructure even further. A substantial increase in the infrastructure also facilitates lower prices that can allow this technology to enter the mainstream market.

The term cloud was coined by the developers from the early era of the internet, called the WAN (Wide Area Network). This is when the internet was in its nascent stages and was usually depicted as a cloud on many developer's flowcharts and diagrams. This representation was very ubiquitous at that time and the internet was represented as a cloud almost everywhere. This led to the representation being immortalized for the services derived from the internet collectively as the Cloud. Hence, the term cloud is used to convey any service that is reliant on the internet as its primary resource.

Cloud computing is highly useful for a number of applications, ranging from storing some data to setting up large software computational centers. This is useful for the end user as a usual set up for a server would be substantially more expensive for an organization than employing the use of the cloud as the cloud provider has already invested in the architecture and is offering it as a service for a small subscription fee each month. When compared to the substantial maintenance costs of running a big server and the initial investment, the cloud serves as a viable low-cost alternative for the majority of the users.

Majority of the cloud has been used as a storage alternative. The idea of being able to free up your local storage and still being able to access your files anywhere on any device his very convenient and highly lucrative. This has made users jump over to the cloud for their storage purposes quite readily. The user can store and access their data seamlessly and also be able to have better user experience as the local storage on your device has been freed up and the data is being available to the user anywhere in the world on any device.

There are a few drawbacks to the utility provided by the Cloud. As the data from the organizations and individuals have been migrated to the cloud, the users have to forfeit their control the data and hand it over to the cloud. Which will then separate the data into clusters and stored them at various locations and serves across its platform for achieving access and the optimum speed for the user. This is a problem as the security of the data has to be ensured as it might contain some sensitive data of the users.

Before, uploading their sensitive data on to the cloud, the organization needs to audit the security measures that are being used to safeguard the data by the cloud service provider. As the data in the cloud is decentralized, clusters of data are scattered across various servers to facilitate easy retrieval. But this technique leads to increased chances of data leakage in an event of a security lapse or an attack. Therefore, it is crucial that the users are able to evaluate and select a cloud provider that is reliable and safe.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

To increase the security of the data, there needs to be some kind of user authentication system that needs to be implemented into the cloud to reduce instances of intrusion by an unrelated person. It would also reduce the instances of Data Leakage as having an authentication system would decrease the chances of the data being exposed to unauthorized personnel. This should also be ensured that high-level clearance should not be given to a lower pay grade employee to maintain the chain of command.

The Data could also be encrypted before being uploaded onto the cloud to ensure that the data is completely secure. As the encryption key would be personal and not shared with anyone, even in the event of a data leakage the sensitive data would be secure as it would be encrypted and essentially useless without the private key and would remain scrambled protecting the integrity of the data intact. The cloud service provider would also have their own encryption algorithms that can also be used in conjunction with the private encryption to ensure an even greater degree of security for the sensitive data.

To ensure the safety of the sensitive data and prevent any data leakages, encryption is always been touted as the fail-safe mechanism that can be highly secure and resilient. This is a very useful feature that would protect the data and safeguard it from the intruders and attackers. This is quite an ingenious feature to be implemented to make sure the data uploaded onto the cloud is safe. But this is a redundant feature that reduces the User Experience in favor of the security on the cloud. This reduction in the Quality of Experience is attributed to the fact that Encrypted data cannot be searched over and this is quite a problem for organizations with huge amounts of data.

It is not at all feasible to manually download all the data, decrypt it and sift through the multitude of data to extract the relevant information. This quite a tedious process and almost close to impossible in certain scenarios as they have Petabytes of data to process. To solve this problem various techniques have been researched and tested. The process of searching over encrypted data is highly difficult and there have been some ingenious techniques to achieve this.

Various researchers have utilized the metadata for the purpose of enabling search over the encrypted data, by searching through the metadata. Another popular technique is to utilize various tags that define the data being encrypted, the tags are then used to achieve searching over the actual data which is encrypted, thereby maintaining the security of the data while enabling search over it.

In this paper, section 2 is dedicated for the literature review of past work ,Section 3 describes the details of the developmental procedure of the model. Section 4 evaluates the results through some experiments and finally section 5 concludes this research article with the traces of the future scope.

## II. ITERATURE REVIEW

D. Kamini [1] introduces cloud computing and the benefits it has on the economy of the place. There has been a lot of individuals and organizations that have been adopting this technology readily. This is a cause of concern as uploading the sensitive data on to the cloud needs to be done with the utmost security and it is usually encrypted to prevent data leakage. This is a good practice but it ends up with the data being unsearchable, therefore, the authors have presented a technique for enabling search over encrypted data which supports gram-based search. The only drawback is that it takes an unrealistically large amount of time.

D. Rane expresses the kind of drawbacks are faced while uploading sensitive data on to the cloud. As when the data is being uploaded by the individual or an organization, they forfeit their control over to the cloud when the data is transferred, this should only be done if the cloud is trustworthy. Therefore, the data is encrypted before being uploaded onto the cloud and this is problematic to search the encrypted data [2]. The authors introduce a technique for enabling ranked based search over the encrypted data uploaded on to the cloud while preserving the data.

L. Zhang examines the predominant use of cloud computing and the increase in the infrastructure of this technology. Due to a lot of organizations understanding the convenience of cloud computing, there has been a large scale adoption of this technology. But due to the fact that a public cloud cannot be trusted, the users encrypt their data before uploading, which becomes a nuisance in the long run [3]. Therefore, the authors have implemented a technique for performing a privacy-preserving search over the encrypted data.

Z. Xia elaborates on the topic of cloud computing as there has been a widespread increase in the adoption of cloud technology. This has increased the number of people utilizing various benefits of cloud computing and uploading their data on to the cloud. The data needs to be encrypted to ensure that the sensitive data is protected, but this makes the data unsearchable. The authors have implemented an ingenious technique to perform searches with the help of the KNN algorithm [4]. There are a few issues with the deletion and update mechanisms as they aren't that secure.

M. Shen [5] explains the novelty of a phrase search as it is capable of retrieving the files that contain only that phrase when the query is fired. It is a highly useful method to access various cloud-based IoT devices. Due to the fact that majority of the data stored on the cloud is encrypted, the authors had to develop a technique to enable search on the encrypted data with the help of Privacy-preserving search or P3 for short. The one downside to this technique is the document index that needs to be refreshed after every search to maintain the integrity of the data.

S. Lavnis introduces the realm of cloud computing and the speed at which it has been advancing and generating the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

infrastructure. There has been significant growth in the number of users opting to store their data on to the cloud. The data has to encrypted to deny any chances of foul play and leakage of any sensitive data. This is problematic as searches cannot be performed on the encrypted data. Therefore, the authors have presented Contextual Oblivious Similarity-based Search – COS2 [6], which helps the system search over the encrypted data. The only drawback is the limited implementation and the lack of any features such as ranking system, multiple keywords, etc.

P. Kale expresses the rise of cloud computing as a ubiquitous and reliable alternative to computing. A lot of individuals and organizations have been taking advantage of this by storing most of their data on to the cloud, including the sensitive data, therefore, it is a common practice to encrypt the data before uploading. This is a hindrance for the search option as it won't be able to search over the encrypted data. To solve the problem, the authors have implemented a technique that allows ranked keyword search over the encrypted data [7]. The major drawback of this technique is that the authors have not utilized any authentication features such as IMEI which will be implemented in the future.

M. Strizhov states that a lot of individuals are utilizing the cloud platform for convenient and ubiquitous storage. This is a concerning situation due to the fact that not all of the cloud servers are trustworthy and there is a need to safeguard the data. The author presents an innovative technique for the encryption of the data such that it can still be searchable. The search mechanism proposed by the author is capable of performing a multi-keyword ranked search over the data stored on an encrypted cloud [8]. The major drawback of this technique is that substrings cannot be used to perform searches which will be implemented in the future along with genetic databases.

P. Pandiaraja [9] explains that there are a lot of problems that are faced when searching for data that is stored on the cloud that is encrypted. As most of the users are highly concerned about the privacy of their data being stored on a public cloud system there is a lack of searchability in the cloud. Therefore, to ameliorate this effect, the authors have proposed an innovative technique that enables search while preserving the privacy of the system. The search is multi-keyword and search is achieved without any leakage of data. The major drawback is the use of the Apriori algorithm that has a lot of limitation oversize and computational capacity.

P. Ponnusamyproposes a variety of techniques that can preserve the privacy of the data while allowing it to be searched in a cloud environment. This is highly useful in a public cloud scenario where the incidence of data leakage is pretty high, therefore, sensitive data needs to be encrypted before uploading on to the cloud [10]. The authors have implemented a multi-keyword search on the encrypted data in the cloud. The major drawback of this system is that the system cannot safeguard the queries that can be extracted later.

X. Shiv [11] expresses that the cloud storage is a very uncertain and highly insecure destination of the data and it needs to be safeguarded efficiently and increase its reliability of the system. But this proves to be a challenge as encrypted data cannot be searched or retrieved easily. Therefore, the authors proposed a system based on a fuzzy keyword search for encrypted data on the cloud. The technique is highly efficient, but there are a few drawbacks, such as feasibility issues while performing plaintext in the search operations.

S. Mittal states that the major part of the cloud infrastructure is cloud storage. As the majority of the users in this platform utilize this feature of the cloud. Most of the users also fall into this category, that is the users only utilize the storage features and nothing else on the cloud platform. The most common technique of protecting the data is encrypted as the data is encrypted and uploaded on the cloud. This introduces a new problem as the encrypted data cannot be searched, therefore, the authors present a technique that utilizes a fuzzy based ranked multi-keyword search. One of the drawbacks in this paper is that keywords are not used to widen the focus of the search as well as the lack of syntactic transformations. [12]

A.Jivaneexplains that the preservation of the privacy of the data on the cloud is of the maximum importance as the majority of the organizations utilize the cloud to store their sensitive documents. As encryption is one of the most widely used mechanisms to increase data security, it hampers the search. Therefore, the authors have implemented a technique for the search over encrypted data while preserving its privacy [13]. The only drawback in this technique is the lack of semantics to allow for greater precision in the ranking system.

R. Ma [14] introduces the rising concerns over the growing infrastructure of the cloud computing scenario, especially for its use for storage purposes. As most of the documents stored on the cloud are sensitive in nature, the only viable option is to encrypt the file when it is being uploaded. This hampers the retrieval process negatively, therefore, the authors have designed a framework called EnDAS. The only shortcoming of this technique is that it is specialized for its application on mobile devices and therefore, needs more work for a generalized version with wide support.

P. Sreekumari proposes a framework for the efficient and privacy-preserving technique for the retrieval and multi-keyword search over the encrypted data [15]. The authors utilize a fuzzy technique to search the user's private documents without any data leakage whatsoever. There is a need for a system like this as the landscape for cloud computing has been rapidly expanding. The major drawback with this technique is that there are various parameters that have not been considered by the authors, such as, efficiency, verifiability, and security.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

## III PROPOSED METHODOLOGY

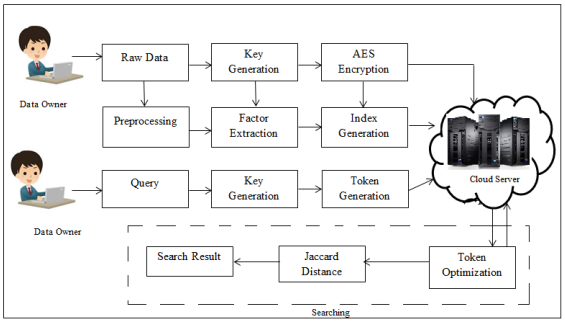


Figure 1: Proposed System overview Diagram

The proposed methodology of search over encrypted data is depicted in Figure no1. And it is detailed in the below mentioned steps.

*Step 1: Key generation and Encryption* - This is the very first step of the proposed model, where a user is registered with the developed software by feeding his credentials. After registration is done user is allowed to login into the system to upload the data in the cloud for storage. The model is designed to deal with textual files and once the file is fed, then a key is being generated using the key space entity in Java, and the key is named as $K_{SPEC.}$ As the key is generated, that is then used to encrypt the file using AES scheme. The whole file encrypted and then it is kept aside for uploading to the cloud.

The proposed model deploys the system using the DropBox public cloud. This is done based on the creation of a user account and then integrating the access token of that user account in the developed software.

*Step 2: Index File Generation* - This is very crucial and important step of the proposed model which runs simultaneously with the previous step of key generation and file encryption. Here first the plain text of the input file is read and then it subjected to the process of preprocessing. This processing step involves four steps as described below.

*Special Symbol removal*- The input file plain text is used here to remove all the special symbols except space.

*Tokenization*- Once the special symbols are removed then the string is divided into the words to store in a list.

*Stopword Removal*- The stopwords are the conjunction words in English which hardly adds any semantics to the sentence or the document. So the removal of these stopwords makes the text more lighter for the further processing. Some of the stop words are like is, of , and , the , from , to etc..

*Stemming*- After removal of the stopwords, the plain text is now free from special symbol and conjunction words. Now these words are subjected to bring to their base form by replacing the suffix words in them by the deserved other words. For example, studying becomes study, and also studied becomes study.

*Factor extraction* - Once the preprocessing is over, then this preprocessed textual word is stored in a list. Then these words tend to the form the factor or bucket list. For example a factor for a word " search" can be denoted as [ sea, sear, searc, search]. Then these factors are encrypted using the AES encryption scheme to concatenate with a space which forms an index for searching. This concatenated string is stored in a factor file, which is being uploaded along with the complete encrypted file of the past step. This process of factor creation can be shown in the below algorithm 1.

Algorithm 1: Factor File creation

// Input : Preprocessed List $P_L$
// Output : Factor File $F_F$
**Function** : factorFileCreation(**$P_L$** )
Step 0: Start
Step 1: **$F_{STR}$** =Ø
[$F_{STR:}$ Factor String ]
Step 2: ***for*** i=0 **to** size of $P_L$
Step 3: W= $P_{Li}$
Step 4: ***if W length > 3, THEN***
Step 5: ***for*** j=3 **to** length of **W**
Step 6: $S_{SUB}$ =Substring_of_W ( 0, j)
Step 7: $S_{SUB=}$ aesEncrypt($S_{SUB}$)
Step 8: $F_{STR=}$ $F_{STR+}$ " "+$S_{SUB}$
Step 9: **end *for***
Step 10: **end *if***
Step 11: ***ELSE***
Step 12: W = aesEncrypt(W)
Step 13: $F_{STR=}$ $F_{STR+}$ " "+W
Step 14: **end *for***
Step 15: write **$F_{STR}$** to **$F_F$**
Step 16: return **$F_F$**

*Step 3: Searching* - Here the user enters a multiword search query. The complete query is preprocessed and a factor list is created as mentioned in the last step. This factor list is encrypted using the AES encryption scheme to call this as search token or Trap door.

All the file contents of the factor file in DropBox cloud is downloaded and then they are read in a string. These factor file contents are loaded in the double dimension list, which contains the File name and also the factor set of that file.

Search tokens are used to search the matched tokens in the double dimension list using the Jaccard Coefficient. Jaccard Coefficient can be shown in the below mentioned equation 1. The matched token file names are collected and then they are displayed to the user as the output for his fired query.

$$J(String\ 1, String\ 2) = \frac{|S1 \cap S2|}{|S1 \cup S2|}$$

## IV RESULT AND DISCUSSIONS

The proposed model of search over encrypted data in cloud is deployed in real time public cloud such as DropBox. Initially a DropBox account is created and then the access token provided by the public cloud DropBox is integrated with the developed application. The proposed application is

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

developed using a windows machine and Java as the programming language. Machine is equipped with the Core i5 processor and Primary memory of 6GB.

To develop the proposed idea Netbeans is used as the standard IDE along with the Mysql as the database to store the user credentials and search results. Some experiments are conducted to prove the effectiveness of the system as described below.

Precision and Recall- Precision and Recall are considered as the one of the best measuring parameters for the searching techniques. The proposed model is tried for different number of query keywords and recorded the result as shown in Table 1. And when the precision and recall are measured and found that the proposed model provides almost 90.1 % average precision and 98.75% of Average Recall.

As the precision and recall of the proposed model is compared with that of [16] which is working on the basis of K nearest neighbor clustering technique, the proposed model yields better results than that of [16]. This is mainly because in [16] the K nearest neighbor algorithm adds some complexities because of its iterations. Whereas the proposed model searches the files based on the Jaccard Coefficient model, which anyhow takes less complexity and provides better results than that of [16]. The comparison table is tabulated in Table 2.

| Experiment No | Relevant File extracted ( A) | Irrelevant File extracted ( B) | Relevant Files not extracted ( C) | Precision in % ( A/ (A+B))* 100 | Recall in % ( A/ (A+C))* 100 |
|---|---|---|---|---|---|
| 1 | 5 | 2 | 0 | | 100 |
| 2 | 7 | 0 | 1 | 100 | 87.5 |
| 3 | 5 | 1 | 0 | 83.33333333 | 100 |
| 4 | 12 | 1 | 0 | 92.30769231 | 100 |
| 5 | 10 | 0 | 0 | 100 | 100 |
| 6 | 6 | 0 | 0 | 100 | 100 |
| 7 | 2 | 0 | 0 | 100 | 100 |
| 8 | 4 | 1 | 0 | 80 | 100 |
| 9 | 6 | 1 | 0 | 85.71428571 | 100 |
| 10 | 8 | 1 | 0 | 88.88888889 | 100 |

Table 1: Precision and Recall experiment Results of the proposed model

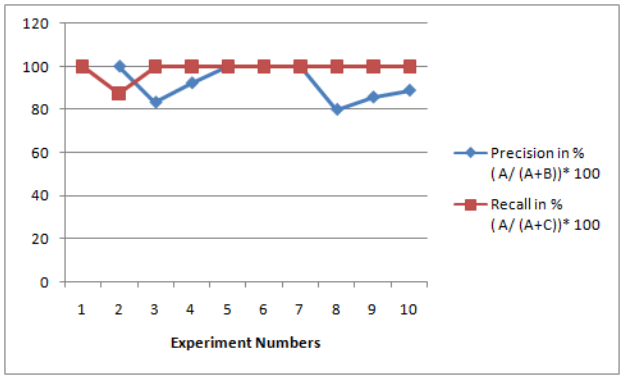


Figure 2: Precision and Recall experiment results of the proposed model

| Methodology | Average Precision | Average Recall |
|---|---|---|
| KNN Search | 84.5 | 97 |
| Jaccard Distance Search | 90.1 | 98.75 |

Table 2: Comparative Results of Precision and Recall between KNN Search and Jaccard Search

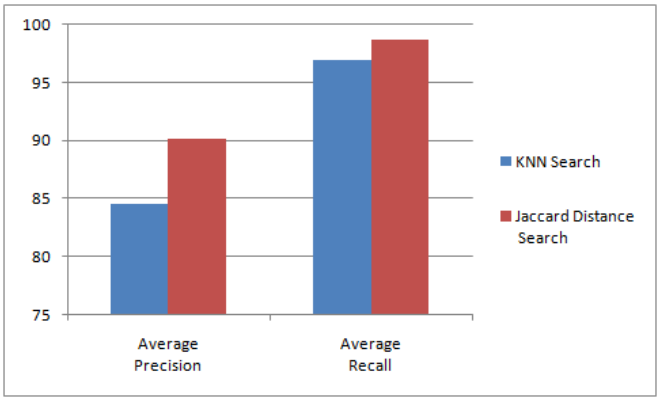


Figure 2: Comparative Results of Precision and Recall, KNN Search V/s Jaccard Search

## V CONCLUSION AND FUTURESCOPE

The Importance of cloud is growing day by day as the data at the user end grows. So storing all the data of the user at the cloud end and retrieve the same on the go is now a day's requirement. But data security at the cloud service provider is always under stake. So to enhance the privacy of the data on searching process by the user himself need to revise. For this many systems have introduced the concept of search over encrypted data. Proposed model also incorporates the concept of search over encrypted data using the Jaccard distance and Trapdoor or Search Token technique on the data stored in the public cloud like DropBox.

The Evaluation of the results clearly indicates that the proposed model's results are better than that of KNN search mentioned in [16] regarding precision and Recall. In the future this system can enhance to handle all kinds of the files like audio, video and images too.

## REFERENCES

[1] D. Kamini, M. Suresh, and S. Neduncheliyan, “Encrypted Multi-Keyword Ranked Search SupportingGram Based Search Technique”, International Conference on Information Communication and Embedded System, ICICES 2016.

[2] Deepali D.Rane andDr.V.R.Ghorpade, “Multi-User Multi-Keyword Privacy Preserving Ranked Based Search Over Encrypted Cloud Data”, International Conference on Pervasive Computing (ICPC), 2015.

[3] Lili Zhang, Yuqing Zhang and Hua Ma, “Privacy-preserving and Dynamic Multi-attribute Conjunctive Keyword Search over Encrypted Cloud Data”, IEEE Access, 2018.

[4] 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 Systems, Vol. 27, No. 2, February 2016.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 8 Issue 05, May-2019

[5] M. Shen, B. Ma, L. Zhu, X. Du and K. Xu, "Secure Phrase Search for Intelligent Processing of Encrypted Data in Cloud-Based IoT", IEEE Internet of Things Journal, 2018.
[6] S. Lavnis, D. Elango, and H. Velez, "Contextual Oblivious Similarity Searching for Encrypted Data on Cloud Storage Services", IEEE 8th International Symposium on Cloud and Service Computing, 2018.
[7] P. Kale and R. Wadekar, "A Survey on Different Techniques for Encrypted Cloud Data", International Conference on Intelligent Computing and Control Systems, 2017.
[8] Mikhail Strizhov, "Towards a Practical and Efficient Search over Encrypted Data in the Cloud", IEEE International Conference on Cloud Engineering, 2015.
[9] P. Pandiaraja and P. Kumar, "Efficient Multi-keyword Search Over Encrypted Data in Untrusted Cloud Environment" Second International Conference on Recent Trends and Challenges in Computational Models, 2017.
[10] P. Ponnusamy and R. Vidhyapriya, "A Survey on Multi-Keyword Ranked Search Manipulations over Encrypted Cloud Data", International Conference on Computer Communication and Informatics, 2017.
[11] X. Shi and S. Hu, "Fuzzy Multi-Keyword Query on Encrypted Data in the Cloud", 4th Intl Conf on Applied Computing and Information Technology/3rd Intl Conf on Computational Science/Intelligence and Applied Informatics/1st Intl Conf on Big Data, Cloud Computing, Data Science & Engineering, 2016.
[12] S. Mittal and R. Krishna, "Privacy-Preserving Synonym Based Fuzzy Multi-Keyword Ranked Search Over Encrypted Cloud Data", International Conference on Computing, Communication and Automation, 2016.
[13] Anjali BaburaoJivane, "Time Efficient Privacy-Preserving Multi-keyword Ranked Search over Encrypted Cloud Data", IEEE International Conference on Power, Control, Signals,and Instrumentation Engineering, 2017.
[14] R. Ma, J. Li, H. Huan, M. Xia and X. Liu, "EnDAS: Efficient Encrypted Data Search as a Mobile Cloud Service", IEEE Transactions on Emerging Topics in Computing, 2015.
[15] P. Sreekumari, "Privacy-Preserving Keyword Search Schemes over Encrypted Cloud Data: An Extensive Analysis", 4th IEEE International Conference on Big Data Security on Cloud, 2018.
[16] Cengiz Orencik, Erkay Savasy and Mahmoud Alewiwiz, " A United Framework for Secure Search Over Encrypted Cloud Data ", IACR Cryptology ePrint Archive 2017.

****