JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Investigating Performance of N-Gram Fuzzy Keyword Search on Encrypted User Data in Cloud Using Jaccard Calculation

Sangeeta Wankhade
Designation: Lecturer
Dept. Computer Engineering
Vidyalankar Polytechnic, Mumbai, India

Dr.Prashant.P.Nitnaware
Designation: Professor
Dept. Computer Engineering
PCE, Mumbai ,India

Abstract: *Now-a-days users mostly store their personal data and professional data on the cloud. As a result, there is massive increase in the storage and computing requirements of users. Every time, the data is getting transferred to the remote server in larger chunks, without analyzing whether the server on which the data is outsourced, is a trusted server or not. But the fact is, after outsourcing the data, users are at great security risk factor that tends to lose the local possession of their large size of data. So, to maintain the privacy of personal data/documents stored in cloud environment, it should first get encrypted before outsourcing to the cloud server. After the data is placed on the cloud, retrieving the same data becomes quiet a tedious job. Thus, to retrieve the data several approaches are available in which keyword enabled search of the encrypted data is one of the outstanding techniques. Most of these approaches are only limited to handle a single keyword search with its own limitations. To enhance the searching method in terms of efficiency and speed, a fuzzy multi-key word search technique can be used to retrieve a corresponding document from cloud. The scheme of fuzzy keyword search remarkably improves the system efficiency and security over the cloud environment. The proposed scheme is convenient , manageable, and even requires less resource. The outcomes of this scheme are valid enough to get the accurate files or the closet possible match files searched by the user. Thus, we have proposed a secure search scheme supporting N-Gram fuzzy multi-keyword search over encrypted cloud data.*

***Keywords- N-Grams, Cloud Computing, Encryption, Fuzzy Keywords, Search Time, Power Consumption*.**

I. INTRODUCTION

Cloud Computing is an increasing mature model of enterprise IT infrastructure that provides high demand on quality applications and services from a shared pool of configuration computing resources. To avoid the costs of building and maintaining a private storage infrastructure the cloud customers, individuals, or enterprises can outsource their local complex data system into the cloud. The company or organization’s private and sensitive data like personal files, company records data, emails, etc which is to be shared among the selected different company employees is stored and in the centralized cloud server but with an insecure feeling that anyone can hack the data that may be very risky for the company.

Also, the data owners and cloud server should not be in the same trusted domain who put the outsourced unencrypted data, if any, at risk; the cloud server may leak data information to unauthorized entities or even be hacked[1][22]. Cloud enables large group of remote servers to be in a network to allow the centralized data repository, and access to the computer services or resources whenever required. Many users are inspired to outsource their confidential data on to the cloud. As the documents get transferred to the cloud, users do not have physical possession of that data. To make sure that the data at cloud side is safer, it must adapt to the privacy preserving storage methodology, as the cloud server is not a trusted server. To protect data confidentiality and unauthorized access to the cloud data, owners are motivated to encrypt their data before it is being outsourced to cloud[21][22].

To overcome this problem, the data stored in cloud storage database needs to be encrypted prior sending to cloud servers for storage. The number of cloud service client/users are increasing day by day because of increasing importance of storing the data effectively and computing models which are accessible within the cloud.

Subsequently, huge amount of data is being added into the cloud servers. Therefore, in such a scenario, the searching and retrieving operation on the file becomes very tedious because the data is in encrypted form and compel the user to search the data which is in encrypted form only. Hence, the data retrieval process becomes the cumbersome problem and a difficult job. It leads to unreliable way to access files by retrieving files excluding the relevance score thereby increasing the wastage of computation cost [23]. Nowadays, the efficient keyword searching technique acquires a paramount importance .

The conventional searching methodology is not so fruitful as the user keeps the data in encrypted format at cloud side.

This can be achieved by performing multi-keyword query to get the top relevant data of user interest which lead to effective searching technique on the encrypted cloud data[21][23].

II. PROPOSED SYSTEM

The proposed system consists of two parts. First is the admin and second is user module. Both the modules ADMIN and USER can upload, delete, and search for the data. The admin will upload some data or files to the server and the authorized user will fire a search keyword to search for the data from the cloud server. This model can enable the user to securely share their data files and information over the cloud environment or to the other users as well, i.e., the data sharing scenario. First the admin will upload his file in the encrypted form of data files along with the set of keywords which helps the user to search the particular data file to the cloud storage server, then the next thing is , an authorized and legitimate user or users' groups can search the data by the generated trapdoor by the symmetric keys using AES algorithm which is handled by the security server. Owing to the third-party settings, this model always uses symmetric public key encryption to implement the searchable encryption scheme. The user can now search the encrypted data on cloud from the storage server and gets the desired files or the most closet possible ones.

In the above figure we define and solve the problem of privacy-preserving multi-keyword ranked search over encrypted cloud data (MRSE) and establish a set of strict privacy requirements for such a secure cloud data system to become a reality. This searching scheme is used to greatly increases the usability of system by returning the matching files when users searching inputs exactly match the predefined keywords or when exact match fails, and the closest possible matching files based on keyword similarity semantics. More precisely, it uses edit distance to quantify keywords similarity and develop a novel technique that is a N-Grams and Jaccard technique, for the construction of fuzzy keyword sets. In this technique the resulted size of the fuzzy keyword sets is significantly reduced[24] by eliminating the need for enumerating all the fuzzy keywords. Encrypting the data using AES encryption before uploading to the cloud servers and by using a cloud server and employing fuzzy keyword search based on N grams.

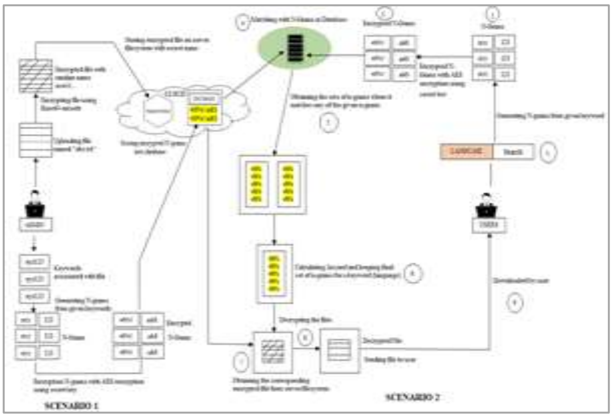


Figure 1: Block Diagram of Proposed System

III. PROPOSED METHODS USED

A. Gram-Based Technique:

This technique is the method of creating n-grams from the given keyword. The subset thus produce from the given keyword are used as the grams and the permutations of such subsets of string are used to create the grams. Therefore , to make the matching technique easier and more efficient our proposed system uses an algorithm which matches the query with the fuzzy keyword. So, for matching the n-grams which are formed are stored in the database after encryption.

B. Jaccard Index:

To check the similarity between the set of keywords in the set the Jaccard index is used. So, to calculate the similarity index the formula is the intersection of the set divided by the union of the sets. This is used in this project to calculate the similarity index between the search keyword and encrypted keyword in the stored dataset.

$$J(A,B) = |A \cap B|/|A \cup B| = |A \cap B|/(|A|+|B|-|A \cap B|)$$

In Steps, that's:

1. Count the number of members which are shared between both sets.
2. Count the total number of members in both sets (shared and un-shared).
3. Divide the number of shared members by the total number of members.
4. Multiply the number you found in by 100.
5. This percentage tells how similar the two sets are.
6. Two sets that share all members would be 100% similar. the closer to 100%, the more similarity (e.g., 90% is more similar than 89%).
7. If they share no members, they are 0% similar.
8. The midway point — 50% — means that the two sets share half of the members.

If A and B are both same, we define $J(A,B) = 1$ and 0 when they are disjoint.

$0 \leq J(A,B) \leq 1$.

E.g., if $J(A,B) > 0.8$ declare a match

C. AES Encrypt:

AES is the advanced method for encryption of the keys for safety purpose. And in the proposed system the keywords and its N-grams are encrypted and gets stored in the database. And matching will be done later with encrypted data. And files will be returned after matching the same keyword or its n-gram. AES uses bytes for its operation rather than bits, so 128 bits word will be considered as the 16-byte block.

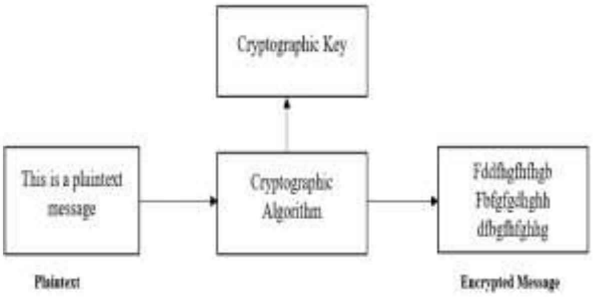


Figure 2: Encryption Process

IV. METHODOLOGY:

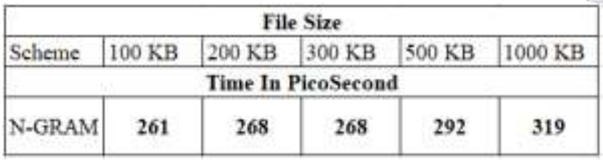
In recent years the energy consumed by computers and its components have caught major attention. However, a recent survey shows that majority of power consumption has occurred in their display and some heating components by about 100TWh/year. Among this, 65 TWh/year is consumed in the field of enterprises & corporates. This is the reason incorporating in energy wastage in computers. Therefore, by using the proposed system the searching time can be minimized as it is able to efficiently handle spelling mistakes, typos and any morphological variants and simultaneously joulemeter app has been used to measure the total power consumed by the proposed system to reduce the carbon footprints.

On the other hand, the downloading of redundant file is also another trend of wasting energy. There also come a lot of similar redundant files existing caused by repeated downloading especially in LAN where the users may have common interest. The types of files also lead to unnecessary energy wastage on the Internet making the computers energy-consuming monster. The Microsoft Joulemeter is a software tool that estimates the power consumption of computer. It tracks computer resources, such as CPU utilization and screen brightness, and estimates power usage. Joulemeter can be used for gaining visibility into energy use and for making several power management and provisioning decisions in data centers, client computing, and software de-sign. Joulemeter estimates the energy usage of a VM, computer, or software by measuring the hardware resources (CPU, disk, memory, screen, etc.) being used and converting the resource usage to actual power usage based on automatically learned realistic power models. Joulemeter provides a software tool to estimate the energy usage of a virtual machine, a computer, server, or software application. It also allows modeling the impact of power management of various components such as the CPU, screen, memory, and storage on total power use. Hence, the power consumed without using proposed system and with using proposed system is measured and analyzed.

V. RESULT ANAYLSIS

- **Search Time Analysis:**

The proposed system thus implemented is enables users to achieve efficient searching of data with reduced time.



	File Size				
Scheme	100 KB	200 KB	300 KB	500 KB	1000 KB
	Time In PicoSecond				
N-GRAM	261	268	268	292	319

Table 1: Search Time

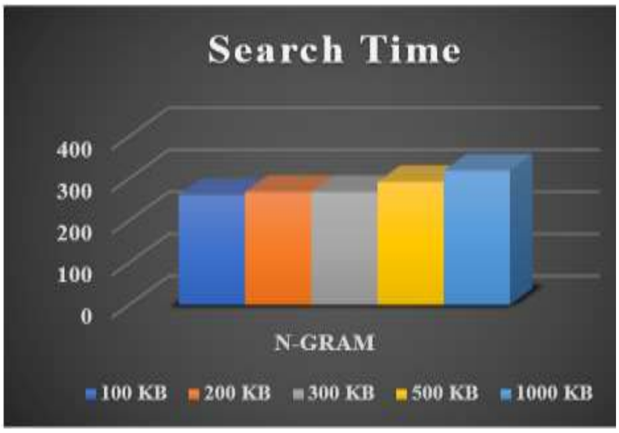
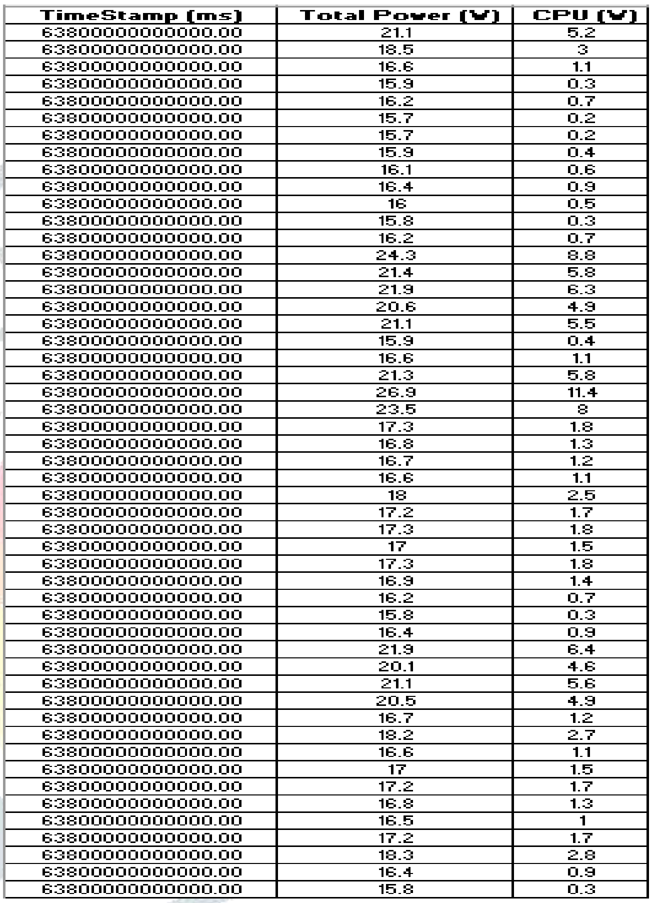


Figure 3: Search Time Comparison

- **Energy Consumption Analysis:**

A desktop uses an average of 200 W/hour when it is being used (loudspeakers and printer included). A computer that is on for eight hours a day uses almost 600 kWh and emits 175 kg of CO2 per year. A laptop uses between 50 and 100 W/hour when it is being used, depending on the model. A laptop that is on for eight hours a day uses between 150 and 300 kWh and emits between 44 and 88 kg of CO2 per year.The internet is a virtual space, using it still requires power and results in CO2 emissions.

Total Power Saved= Power Consumed without using the proposed system- Power Consumed with using proposed system =21.1- 15.8= 5.3 Watt



TimeStamp (ms)	Total Power (W)	CPU (W)
6380000000000.00	21.1	5.2
6380000000000.00	18.5	3
6380000000000.00	16.6	1.1
6380000000000.00	15.9	0.3
6380000000000.00	16.2	0.7
6380000000000.00	15.7	0.2
6380000000000.00	15.7	0.2
6380000000000.00	15.9	0.4
6380000000000.00	16.1	0.6
6380000000000.00	16.4	0.9
6380000000000.00	16	0.5
6380000000000.00	15.8	0.3
6380000000000.00	16.2	0.7
6380000000000.00	24.3	8.8
6380000000000.00	21.4	5.8
6380000000000.00	21.9	6.3
6380000000000.00	20.6	4.9
6380000000000.00	21.1	5.5
6380000000000.00	15.9	0.4
6380000000000.00	16.6	1.1
6380000000000.00	21.3	5.8
6380000000000.00	26.9	11.4
6380000000000.00	23.5	8
6380000000000.00	17.3	1.8
6380000000000.00	16.8	1.3
6380000000000.00	16.7	1.2
6380000000000.00	16.6	1.1
6380000000000.00	18	2.5
6380000000000.00	17.2	1.7
6380000000000.00	17.3	1.8
6380000000000.00	17	1.5
6380000000000.00	17.3	1.8
6380000000000.00	16.9	1.4
6380000000000.00	16.2	0.7
6380000000000.00	15.8	0.3
6380000000000.00	16.4	0.9
6380000000000.00	21.9	6.4
6380000000000.00	20.1	4.6
6380000000000.00	21.1	5.6
6380000000000.00	20.5	4.9
6380000000000.00	16.7	1.2
6380000000000.00	18.2	2.7
6380000000000.00	16.6	1.1
6380000000000.00	17	1.5
6380000000000.00	17.2	1.7
6380000000000.00	16.8	1.3
6380000000000.00	16.5	1
6380000000000.00	17.2	1.7
6380000000000.00	18.3	2.8
6380000000000.00	16.4	0.9
6380000000000.00	15.8	0.3

Figure 4: Energy Consumption Readings from Joulemeter

1. A laptop consumes about 21.1Watts(0.0211KWh) without using the proposed system.
   Energy Usage for a working day (9am-5pm) =0.0211*8 =0.1688 Kwh
   Energy Usage over the week(considering 6 days as a working days)=0.1688 x 6 =1.0128 Kwh
   Energy usage over one year = 1.0128*52 = 52.66 Kwh
   The average price people in India pay for electricity is about Rs.6.09/Kwh
   Annual Energy Cost for the Desktop CPU = 52.66 * 6.09 =approx. Rs.320.
2. A laptop consumes about 15.8 Watts(0.0158 Kwh) with using the proposed system.
   Energy Usage for a working day (9am-5pm)=0.0158*8=0.1264 Kwh
   Energy Usage over the week(considering 6 days as a working days) =0.1264x 6 =0.758 Kwh Energy usage over one year = 0.758*52 = 39.41 Kwh
   The average price people in India pay for electricity

is about Rs.6.09/Kwh.
Annual Energy Cost for the Desktop CPU = 39.41 x 6.09= approx. Rs.240

Therefore, Total Savings = 320-240= Rs.80 Per Single Laptop

Thus, the proposed scheme consumes low power on computation and communication.

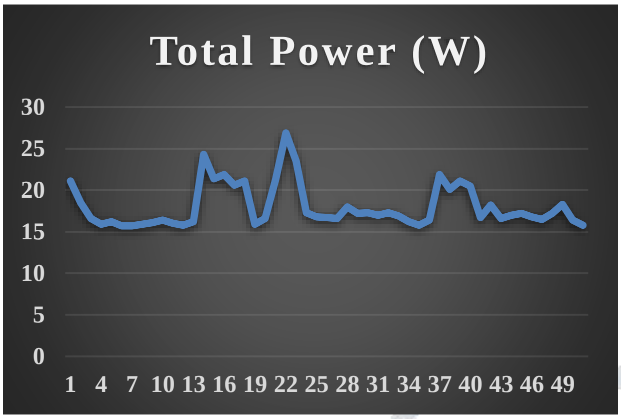


Figure 5: Power Consumption

The proposed scheme introduces low overhead on computation and communication.

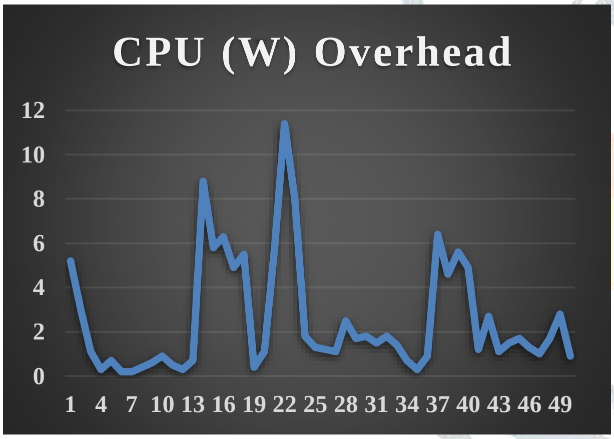


Figure 6: CPU Overhead

VI. CONCLUSION

The proposed system mainly focuses on reducing energy consumption and reducing emission of Carbon Dioxide. The system uses less time to search a file by using N-Gram methodology. Using this proposed system file downloading and uploading will not cause more overheard on CPU. The N-Gram fuzzy keyword search improves the file downloading speed while keeps the merit of energy saving for computer. As per the analysis and the experimental results, we can hereby conclude that if the proposed system mechanism is implemented for optimizing the resource usage in institutes or organizations, the huge amount of energy will be saved and the major trouble of energy wastage will be overcome.

REFERENCES

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

[2] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, "Secure ranked keyword search over encrypted cloud data," in Proc. of ICDCS'10, 2010.

[3] Y.-C. Chang and M. Mitzenmacher, "Privacy preserving keyword searches on remote encrypted data," in Proc. of ACNS, 2005.

[4] D. Song, D. Wagner, and A. Perrig, "Practical techniques for searches on encrypted data," in Proc. of S&P, 2000.

[5]Zhihua Xia, Member, IEEE, Xinhui Wang, Xingming Sun, Senior Member, IEEE, and Qian Wang, Member, IEEE "A Secure and Dynamic Multi-keyword Ranked Search Scheme over Encrypted Cloud Data" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL., NO.1,2015.

[6] Bing Wang, Wei Song, Wenjing Lou, and Y. Thomas Hou "Inverted Index Based Multi-Keyword Public-key Searchable Encryption with Strong Privacy Guarantee" IEEE Conference on Computer Communications (INFOCOM), 2015.

[7] Yanzhi Ren, Yingying Chen, Jie Yang, Bin Xie "Privacy-preserving Ranked Multi-Keyword Search Leveraging Polynomial Function in Cloud Computing" Globecom Communication and Information System Security Symposium 2014.

[8] Hongwei Li, Dongxiao Liu,Yuanshun Dai, Tom H. Luan, And Xuemin (Sherman) Shen "Enabling Efficient Multi-Keyword Ranked Search Over Encrypted Mobile Cloud Data Through Blind Storage", December 2014.

[9] Mikhail Strizhov and Indrajit Ray "Multi-keyword Similarity Search Over Encrypted Cloud Data" International Conference on Pervasive Computing (ICPC), 2012.

[10] E.-J. Goh, "Bloom filters in order to construct the indexes for the data files" IEEE Conference on Computerer Communications 2016.

[11] Jun Zhou, Zhenfu Cao, Xiaolei Dong and Xiaodong Lin "More Efficient Verifiable Outsourced Computation from Any Oneway Trapdoor Function" IEEE ICC - Communication and Information Systems Security Symposium, 2015.

[12] Fanyu Bu, Yu Ma, Zhikui Chen and Han Xu "Privacy Preserving Back-Propagation Based on BGV on Cloud" 2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC), 2015 IEEE 7th International Symposium on Cyberspace Safety and Security (CSS), and 2015 IEEE 12th International Conf on Embedded Software and Systems (ICESS).

[13] Joseph K, "Secure Sharing and Searching for Real-Time Video Data in Mobile Cloud" 2015. [14] Zhangjie Fu, Member, IEEE, Jiangang Shu, Xingming Sun, and Nigel Linge "Verifiable Keyword-based Semantic Search over Encrypted Cloud Data" IEEE Transactions on Consumer Electronics, Vol. 60, No. 4, November 2014.

[15] J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, and W. Lou, "Fuzzy keyword search over encrypted data in cloud computing," in Proc. of IEEE INFOCOM'10 Mini-Conference, San Diego, CA, USA, March 2010.

[16] D. Boneh, G. D. Crescenzo, R. Ostrovsky, and G. Persiano, "Public key encryption with keyword search," in Proc. of EUROCRYPT, IEEE Conference on Computer Communications 2004.

[17] Utkarsh Joshi, Neeraj Vishwakarma , A.Murugan "Fuzzy Keyword Search over Encrypted Data" in International Journal of Innovative Research in Science, Engineering and Technology Vol. 6, Issue 4, April 2017.

[18] Dr.Narendra Shekokar , Kunjita Sampat, Chandni

Chandawalli ,Janvhi Shah ”Implementation of fuzzy keyword search over encrypted data in cloud computing” in ICACTA-2015.

[19] Manish Kumar Yadav, Drishti Gugal, Shivani Matkar, Sanket Waghmare “Encrypted Keyword Search in Cloud Computing using Fuzzy Logic” in IEEE Xplore 2019.

[20] Saumya Sharma, Amrita Bhagtani, Parth Agarwal, Ankit Mohite[20] “N-Gram Fuzzy Keyword Search On Encrypted User Data in Cloud” in International Journal of Open Information Technologies 2017.

[21] C. Wang, N. Cao, J. Li, K. Ren, and W. Lou, “Secure ranked keyword search over encrypted cloud data,” in Proc. of ICDCS’10, 2010.

[22] Ms. Jabeen Akkalkot , Ms. S. Shanmug Priya, “A survey on keyword based search mechanism for data stored in cloud,” in Proceedings of International Journal of Computer Science and Mobile Computing. ACM, May- 2016, pg. 235-240.

[23] Y.-C. Chang and M. Mitzenmacher, “Privacy preserving keyword searches on remote encrypted data,” in Proc. of ACNS, 2005.

[24] JianLi,Ruhui Ma, HaibingGaun, “TEES: An Efficient Search Scheme Over Encrypted data on Mobile Cloud” IEEE Transactions on Cloud Computing, TCC 2015.

[25] 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.