International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

Accurate Information Retrieval over Encrypted Data Storage

Leena Verma,1, Mahendra Verrma,2
[#]*Sdbct Indore A B Rd Rau Indore*

Submitted: 25-06-2021 Revised: 07-07-2021 Accepted: 10-07-2021

ABSTRACT –Search or information retrieval is a primary need of any storage system when a significant amount of stored. But when the data is stored on a cryptographic format it is much complex to extract the required data using the normal search technique. Therefore some additional technique is needed to be implementing which can help us to identify the data available on the cryptographically secured files. In this presented work we are simulating a secure search technique for cryptographic cloud storage for achieving high accuracy over data retrieval and low resource utilization. In this context two key contribution using this work is proposed first is to implement a hybrid cryptographic technique for securing the files and data over the cloud storage secondly the secure search technique which provide accurate data when user need to search the documents. The cryptographic technique is developed using the DES and SHA1 algorithm by modifying the generated key using SHA1 hash generation algorithm. The key is modified because the DES algorithm accepts only 64 bits but the SHA1 generates the 160 bits. On the other hand for designing the secure search technique the two key modules are implemented first module is used for extraction of meaningful keywords from the file. Thus here the TF-IDF concept is implemented for finding the suitable keywords. In addition of that the keywords are stored on database by processing each keyword on the basis of SHA1 algorithm. Finally a kNN based search is implemented which accept the normal keywords and search the SHA1 based keyword database. The implementation of the proposed approach is given using JAVA technology. After implementation the performance of the system is measured in terms of precision, recall and f-measure. The results demonstrate the proposed technique is efficient and accurate.

Keywords: KNN, secure search, SHA1, cloud storage, cryptographic data search, accurate search.

I. INTRODUCTION

Cloud computing is become popular day by day, additionally new and innovative techniques are also continuously added to this technology. The cloud computing technology offers the scalable and efficient resources that can easily handle the increasing work load on servers. The cloud not only includes the efficient computing resources that also involve the effective storage solutions. Due to large amount of data handling the security is an essential issue in the cloud storage, therefore most of the cloud service provides offers the cryptographic cloud services for securing the data from the other users. But the cryptographic data is secure but not readable from other users and systems therefore the data identification and correctly information retrieval is a new challenge for the cryptographic scenarios.

In this context the proposed work is focused on studying the cryptographic techniques for securing the data over cloud additionally a method is demonstrated by which the accurately the target data is identified using the user keywords. Therefore the proposed technique is named as the secure information retrieval from the cryptographic cloud. The proposed work involve the two phase of system demonstration first the cryptographic technique by which the user upload their private data to cloud with security additionally the work is also involve the technique of searching the data over the cloud using the user defined keywords.

II. PROPOSED WORK

This chapter provides the understanding of the proposed working model of cryptographic cloud based data retrieval. Therefore this chapter involves the discussion about the proposed methodology of system design and the proposed algorithm for searching and cryptographic technique utilized.

A. System Overview

Security is one of the major issues in the new generation data hosting and communication systems. Even in cloud which is much secure and efficient system the data owners are worried about their data privacy and security. Basically over the cloud, data is available in mobile form thus data is moving from one server to another data center. Therefore security is much essential in such data

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

hosting space. In order to keep secure the system the service provides are always use the cryptographic techniques and also modify these techniques time to time. The main reason behind using the cryptographic techniques is their easy to implement approach and effectively acceptable changes.

In this presented work the proposed work is focused on developing two major utilities first the hybrid cryptographic approach for secure data hosting over the cloud servers and second the secure search process for search the cryptographic data over the cloud servers. The hybrid cryptography is secure against the normal or classical cryptographic technique with small modification of cryptographic algorithm these algorithms achieves higher degree of security against the various kinds of security threads. Additionally these techniques involve the goodness of two combined algorithms or approaches. Therefore in order to implement the proposed hybrid cryptographic technique the DES algorithm and SHA1 hash function is used. Secondly for implementing the secure data search technique the keyword extraction technique, secure keyword storage and data search technique is implemented. This section provides the overview of the proposed security system for cloud data hosting. In next section the proposed system is explained in detail.

B. Methodology

The proposed work is motivated to provide two solutions for the cloud computing first the secure cryptographic technique and the second is to information retrieval technique for finding accurate data from the cryptographic data format. Therefore the entire system modeling is defined in two modules as:

a. Hybrid cryptographic technique

The proposed hybrid cryptographic technique is demonstrated using figure 2.1. In this diagram the utilized components and functions are also involved for demonstration.

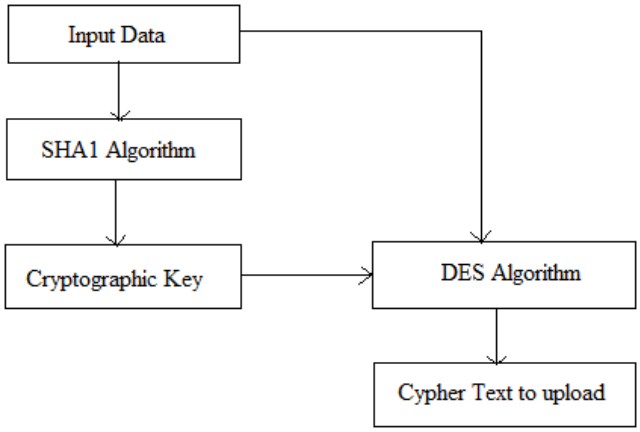

Figure 2.1 proposed cryptographic technique

Input data: that is the data which is need to be store on cloud data storage. It can be any confidential document or sensitive information for the storage.

Sha1 algorithm: SHA1 is a hash generation algorithm. That algorithm accepts any length of document or text and generates the 160 bit or 40 character hash code. The proposed technique accept the input text data as input and using the SHA1 hash key generation algorithm it generates 160 bit hash key for key generation phase.

Cryptographic key: the 160 bit hash code is now treated or processed in this phase for generation of DES key. DES encryption algorithm accepts 56 bit key and 8 bit parity bit for encryption process. Therefore the 160 bit is processed for achieving the 64 bit for DES encryption key. Therefore first 64 bit is preserved and remaining data is removed for 160 bit.

DES algorithm: the DES encryption algorithm is implemented in this phase providing cryptographic security of data. It accepts the 64 bit generated key from the last phase and initial input data for encryption. The system utilizes both the parameters and generates the cipher text for the input data.

Cipher text to upload: the generated cipher text is final outcome of the above given process. The system generated cipher text is finally uploaded to the cloud server.

b. Keyword Extraction for Search

This phase is designed for finding information from the cryptographically secured files. In this context an additional provision for secure search of data is implemented. The search keyword

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

extraction technique for the proposed system is demonstrated in figure 2.2.

Input data: it is the target file which is need to be save on server and also need to be find using the keywords when it is being encrypted. The system accept the user input or selected file name from the user local machine and the content of file is processed here before encryption of file using the below given process.

Pre-processing of data: the data input of the system is first pre-processed using this phase. Here the aim of data preprocessing is to reduce the file size or content size of file. Therefore first the unwanted characters are removed from file. Additionally the stop words are also removed from the file. In order to perform this operation a find and replace function is implemented. This function accepts a list of stop words and a character list one by one. Additionally each word or character available in the list is searched and removed for the input file.

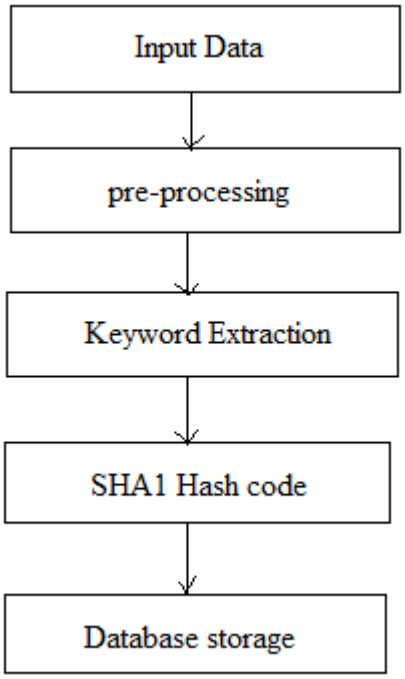

Figure 2.2 Keyword selection

Keyword extract: after removing the unwanted data from the file the file content is reduced significantly. After that it is need to find the file keywords which are essential for finding this file. In this context the word frequency of the file keywords are computed additionally the higher frequency keywords are preserved and remaining keywords from the file is removed. For finding the keyword frequency the following formula is used.

$$\text{keyword frequency} = \frac{\text{number of times a word found in file}}{\text{total count of words in file}}$$

SHA1 hash code: after identification of essential keywords on the basis of the keyword frequency these keywords are need to be store on database with the file name. but directly storage of keywords of file in a database can be create some security issues during the data discloser. Therefore for storing the files keywords into database SHA1 hash generation algorithm is used. That algorithm generates the hash code for all the selected keywords and the generated hash codes are stored on data base with the file name for security.

c. Search Process

The search process for the cryptographically secured file is described in this section. The figure 2.3 shows the search technique implemented with the system.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

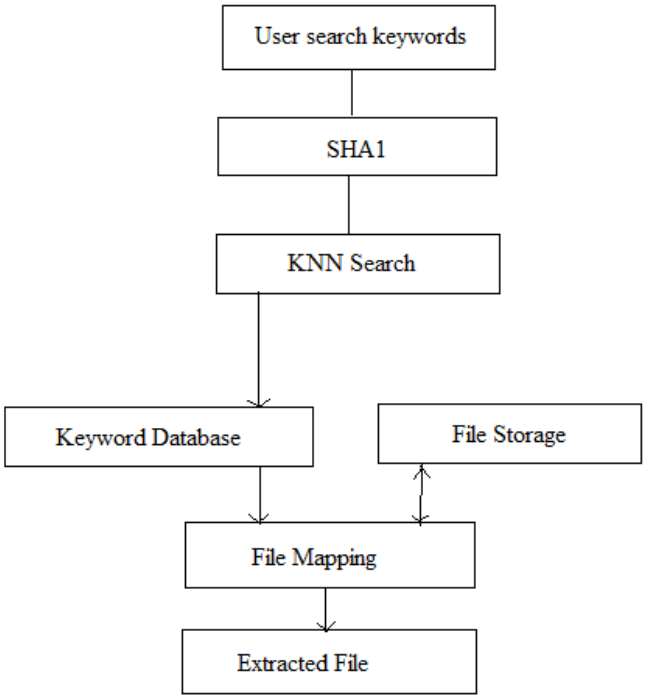

Figure 2.3 proposed search technique

User search keywords: in any search or information retrieval system it is required to provide search keywords for finding the relevant material or content form the data source. In this context here user provide some keywords into the search interface for finding the data into encrypted file storage.

SHA1: the input keywords are accepted in this phase and processed using the SHA1 hash generation algorithm. Basically the keywords of the files are stored in 160 bit hash codes therefore for identifying the target data we need to convert query keywords into the similar format. Thus the user query keywords are used with the SHA1 hash algorithm for converting the similar in format for identifying the target keywords.

KNN search: the KNN (k-nearest neighbor) algorithm is a search algorithm which is usage the distance function for finding best match of the pattern. The KNN algorithm accepts two parameters for performing search first is the query data and second is the data base. Additionally using the distance function it computes the most suitable match for the query sequence.

Keyword database: it the database of keywords and the file name. in this database structure the keywords are preserved in terms of SHA1 hash code and with these keywords the file name is also associated with the keywords. KNN accept the database sequence form this database and perform search on this database. In order to search the data from this database the Euclidean distance function is used as:

$$D(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

Where q and p are the database and user query sequence.

File mapping: the search using KNN algorithm over the encrypted keyword database results the file name. Because the keyword database contains the file name and keywords in SHA1 hash codes. Thus the search results the file name after completing the search.

File storage: file storage is cloud based encrypted file storage database which contains the text files or documents in encrypted format. System recognize the file contents from the KNN search using the file name the file is identifying in encrypted database.

Extracted file: the resultant file is the final outcome of the system thus user can download this file when required.

C. Proposed Algorithm

The above given modules of the proposed system design is summarized in this section therefore the three algorithms are described in this section.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

Input: input file F Output: encrypted data E
Process: 1. $H_{key} = SHA1.GenrateHash(F)$ 2. $Key_{64} = SelectKeyBits(H_{key})$ 3. $E = DES.encrypt(F, K_{64})$ 4. return E

Table 2.1 proposed cryptographic algorithm

Input: Input file to upload F Output: save to keyword database KD
Process: 1. $T_n = ReadFileToken(F)$ 2. $for(i = 1; i \leq n; i++)$ a. $P_n = RemoveStopWords(T_n)$ b. $P_n = RemoveChar(P_n)$ 3. end for 4. $for(j = 1; j \leq n; j++)$ a. $W_f = count(P_j)/count(P_n)$ 5. end for 6. $KD = W_f.Select(20)$ 7. Return KD

Table 2.2 proposed keyword extraction

Input: keyword database KD, user query UK Output: file name to download F
Process: 1. $Q = ReadUserQuery(UK)$ 2. $HQ_i = SHA1.GenrateHash(Q)$ 3. $P_n = ReadDatabase(KD)$ 4. $for(i = 1; i \leq n; i++)$ a. $D(P_i, HQ_i) = \sqrt{\sum_{i=1}^{N}(HQ_i - P_i)^2}$ b. $if(D(P_i, HQ_i) \leq 0.30)$ i. $F = getFileName(P_i)$ c. End if 5. end for 6. Return F

Table 2.3 proposed search algorithm.

III RESULTS ANALYSIS

The main objective of the proposed work is to provide an efficient method for data search over the cryptographic cloud. Therefore the search system is implemented with the cryptographic cloud system. The performance of search outcomes is described in this chapter using different search parameters.

A. Precision

Precision measure is the ratio of the number of correct positive results and number of all positive results. It measures the exactness of any retrieves search process. The higher the precision means that less false positives (FP), whereas the lower precision means that more the false positives are. Here, we are showing precision rate formula.

$$\text{Precision Rate} = \frac{TP}{TP + FP}$$

-TP is the number of true positives
-FN is the number of false positives

S. No.	Proposed technique	Traditional technique
1	0.81	0.79
2	0.85	0.82
3	0.83	0.80
4	0.87	0.83

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

5	0.88	0.82
6	0.83	0.76
7	0.91	0.85

Table 3.1 precision rate

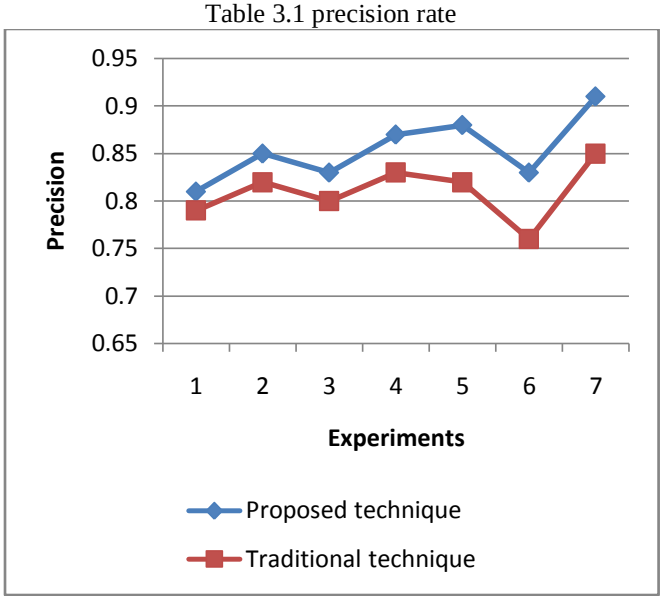

Figure 3.1 precision rate

The comparative performance in terms of precision rate of both the techniques for cryptographic data retrieval techniques is demonstrated in figure 3.1 and table 3.1. In this diagram the X axis represents the different experiments performed with the systems and observed performance results are demonstrated in Y axis. According to the obtained performance outcome the proposed technique provides more precise results as compared to the traditional approach. Therefore the proposed technique is more accurate than the traditional keyword based information extraction approach for the cryptographic data storage.

B. Recall

Recall is the ratio of the number of correct positive results and number of positive results that should have been returned. It measures the completeness or accuracy of the searching of keywords. Higher the recall means that small false negatives (FN), whereas lower the recall is more false negatives it leads to. In this, following recall rate formula used to calculate Performance of algorithm.

$$\text{Recall Rate} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

-TP is the number of true positives
-FN is the number of false negatives

S. No.	Proposed technique	Traditional technique
1	0.75	0.71
2	0.78	0.73
3	0.76	0.69
4	0.79	0.72
5	0.80	0.74
6	0.81	0.74
7	0.80	0.73

Table 3.2 recall rate

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

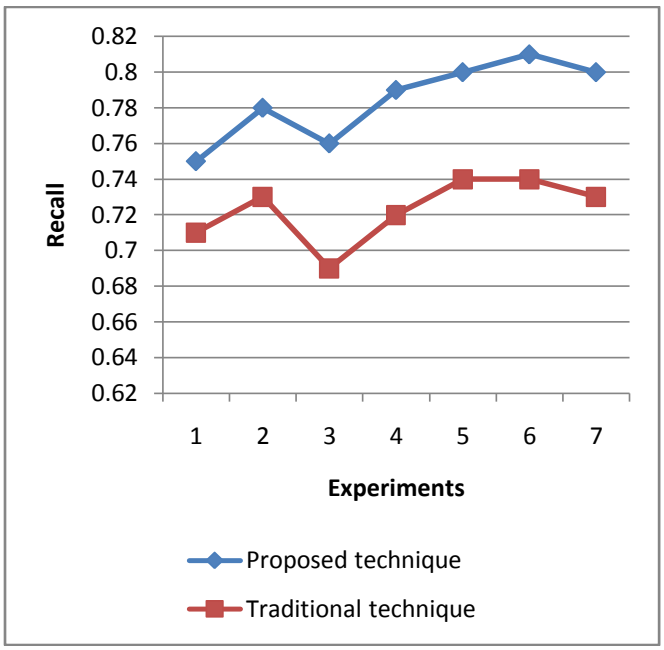

Figure 3.2 recall rate

The performance of proposed technique of cryptographic data retrieval technique in terms of recall rate is demonstrated in figure 3.2 and table 3.2. In this diagram the red line demonstrate the performance of traditional technique and performance of proposed technique is given using figure 3.2. According to the diagram the X axis of the diagram shows the different experiments conducted with the system and Y axis shows the recall rate of both the systems. The obtained experimental results show the proposed technique achieves higher outcomes as compared to the traditional information retrieval technique. Thus the proposed technique is accurate for cryptographic data retrieval as compared to the traditional technique.

C. F-Measure

The F-measure is also known as f-score of the search system, which represent the harmonic mean of the search capability. To compute the f-measure the precision and recall values of the search outcome is computed in the following manners.

$$\mathrm{F-measure} = 2 * \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

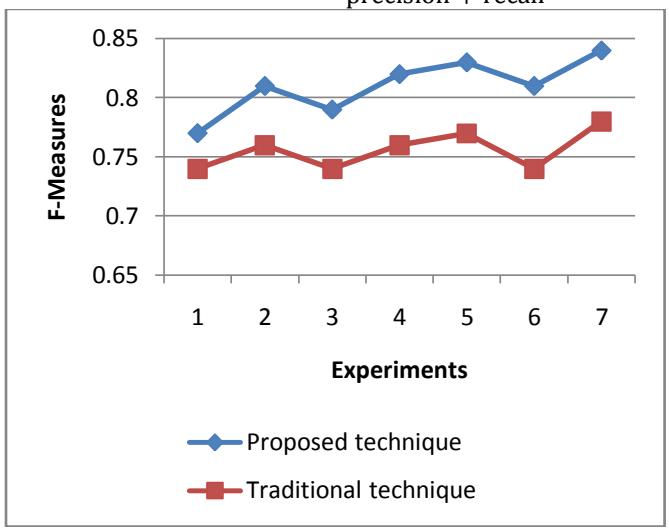

Figure 3.3 F-measure

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

S. No.	Proposed technique	Traditional technique
1	0.77	0.74
2	0.81	0.76
3	0.79	0.74
4	0.82	0.76
5	0.83	0.77
6	0.81	0.74
7	0.84	0.78

Table 3.3 F-measure

The performance of the proposed cryptographic search technique in terms of f-measures is demonstrated in figure 3.3 and table 3.3. In order to demonstrate the performance of the approach the figure includes in X axis as the different experiments performed and the Y axis contains the computed F-measures or f-score. According to the given results the f-measures of the proposed technique is higher as compared to the traditional approach therefore the proposed technique omit high accurate results as compared to the traditional techniques.

D. Memory Usages

The memory usages of the proposed work are demonstrated in this section. The memory usages of the algorithm are also known as the space complexity of the algorithm. The memory usage of the algorithm is the amount of main memory required to execute the algorithm is known as the memory usage or space complexity of algorithm. That is computed using the following formula:

$$\text{memory usage} = \text{total memory space} - \text{free memory space}$$

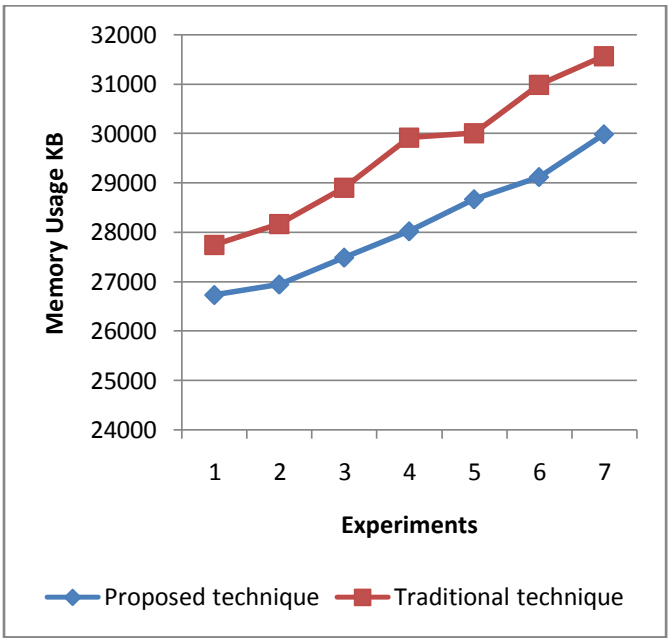

Figure 3.4 memory usage

S. No.	Proposed technique	Traditional technique
1	26736	27748
2	26948	28173
3	27493	28911
4	28025	29918
5	28673	30013
6	29123	30997
7	29987	31572

Table 3.4 memory usages

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

The memory usages in terms of kilobytes (KB) of the proposed method and traditional technique are given using figure 3.4 and table 3.4. For demonstration of the performance of both the technique X axis include the different experiments conducted with the system and Y axis contains the memory consumed in terms of KB. According to the given results the proposed approach consumes less amount of memory as compared to the traditional technique therefore the proposed technique is more efficient as compared to traditional information retrieval technique.

E. Time Usages

The time usages or time requirement of the proposed search system is described in this section. The time consumed is also known as the time complexity of the algorithms. The consumed time is estimated using the following formula.

$$\text{time consumed} = \text{start time} - \text{end time}$$

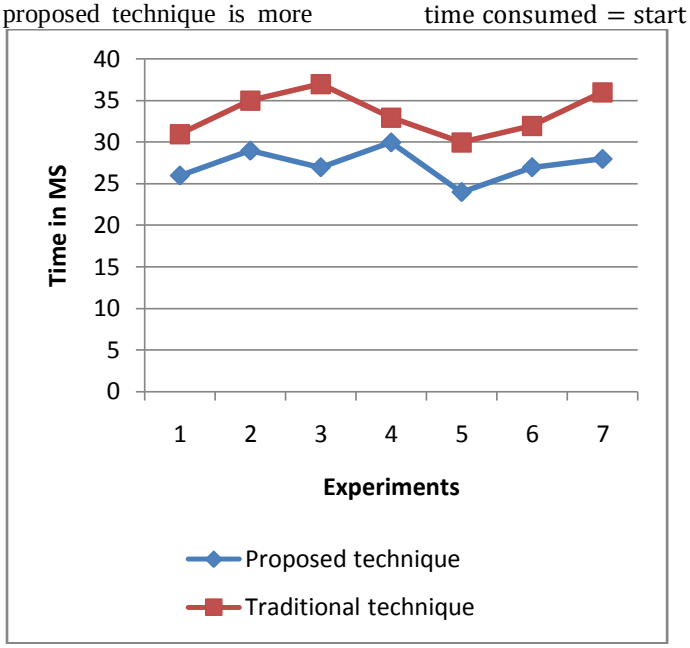

Figure 3.5 time consumption

S. No.	Proposed technique	Traditional technique
1	26	31
2	29	35
3	27	37
4	30	33
5	24	30
6	27	32
7	28	36

Table 3.5 time consumption

The time consumption of the proposed cryptographic data retrieval technique is reported using figure 3.5 and table 3.5. Both the table and figure demonstrate the comparative performance study among the proposed technique and traditionally available technique. In this diagram the X axis shows the experiments performed with the proposed system and the Y axis shows the time consumed in terms of milliseconds. According to the obtained time requirements the proposed technique requires less amount of time to find the data as compared to the traditional technique.

IV. CONCLUSION & FUTURE WORK

This chapter provides the conclusion of the implemented system and the experimental observations. Therefore that involve the conclusion based on the observations and the future extension of the proposed work is also involved.

A. Conclusion

The cloud computing is one of the popular computing and storage technology. The cloud computing offers the solutions for scalable storage and computational needs therefore that is suitable for large scale application and data handling. In this presented work the cloud data storage is key area of

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 7 July 2021, pp: 1576-1586 www.ijaem.net ISSN: 2395-5252

study and system design. The basic issue is the data format and the security concern on parking the sensitive and private data over the cloud storage. Therefore by the aim of storing data on third party storage becomes safe and easy to identify when required by using the data keywords. The traditional approach requires additional encoding and decoding schemes for managing such kind of system which time consuming and as well as the resource consuming techniques. Therefore an efficient and accurate technique is required for performing the proposed task.

The proposed technique is divided in two major parts first is responsible for providing the cryptographic security for data hosting on cloud server and second module is used for demonstration of the cryptographic data search. In this context the first module usages the DES algorithm and SHA1 algorithm for designing the hybrid cryptographic technique and in second module the KNN algorithm is applied for searching the cryptographic data. The proposed cryptographic technique is a symmetric key cryptographic technique which is first generates the cryptographic key using the SHA1 hash generation algorithm. This key is used with the DES algorithm and data for encryption of data. On the other hand for searching of data the SHA1 based hash code is generated and compared with the database keywords and the target data is extracted from data storage.

The implementation of the proposed technique is performed using the JAVA technology more specifically using the JSP (java server pages). Additionally for storing the data MySql database server is used. After implementation of the proposed technique the performance of the implemented cryptographic system is computed and compared with the similar functioning technique. According to obtained performance results the performance summary is demonstrated in table 4.1.

S. No.	Parameters	Proposed technique	Traditional technique
1	Precision	0.81 – 0.91	0.76 - 0.85
2	Recall	0.75 – 0.81	0.69 – 0.74
3	F-measure	0.77 – 0.84	0.74 – 0.78
4	Memory usages	26736 – 29987 KB	27748 – 31572 KB
5	Time consumption	24 – 30 MS	30 – 37 MS

Table 4.1 performance summary

According to the obtained performance the proposed technique found acceptable, accurate and efficient. Therefore the proposed technique can be used with the real world applications for finding or retrieving the accurate data from the cryptographic data storage.

B. Future Work

The main aim of the proposed work is to design and develop an efficient and accurate information retrieval technique for cryptographic cloud data. The made effort demonstrate the proposed technique achieves the required goal. Therefore for future extension the following work is proposed.

1. The current technique only focused on keywords based search in near future the work is extended for obtaining the semantically similar keyword extraction technique
2. The proposed technique is provide the higher accurate search results but need some improvement on keyword selection phase for finding more accurate results therefore in near future the keywords is modified.

REFERENCES

[1] Victor Chang, Muthu Ramachandran, “Towards achieving Data Security with the Cloud Computing Adoption Framework”, IEEE TRANSACTIONS on Services Computing, Volume: 9, Issue: 1, Jan.-Feb. 1 2016

[2] Neha Rawat and Ratnesh Srivastava, “Data Security Issues in Cloud Computing”, Open Journal of Mobile Computing and Cloud Computing, Volume 1, Number 1, August 2014

[3] Venkata Sravan Kumar Maddineni an Shivashanker Ragi, Security Techniques for Protecting Data in Cloud Computing, Master Thesis, Electrical Engineering November 2011

[4] John Harauz and Lori M. Kaufman, “Data Security in the World of Cloud Computing”, IEEE Security & Privacy (Volume: 7 , Issue: 4 , July-Aug. 2009)

[5] Vaquero, Luis M., et al. "A break in the clouds: towards a cloud definition." ACM SIGCOMM Computer Communication Review 39.1 (2008): 50-55.

[6] Armbrust, Michael, et al. "A view of cloud computing." Communications of the ACM 53.4 (2010): 50-58.
[7] Jinesh varia," AWS Cloud Security Best Practices", "White Paper", November 2013
[8] Luit Infotech Private Limited, "What is Cloud Computing", available online at: http://www.luitinfotech.com/kc/what-is-cloud-computing.pdf
[9] B. Grobauer, T. Walloschek, and E. Stöcker, "Understanding Cloud Computing Vulnerabilities". 2011 IEEE Security and Privacy, pp. 50-57.
[10] S. Zhang, S. F. Zhang, X. B. Chen, and X. Z. Huo, "Cloud Computing Research and Development Trend," In Proceedings of the 2010 Second International Conference on Future Networks (ICFN '10). IEEE Computer Society, Washington, DC, USA, pp. 93-970.
[11] Obrutsky, S. "Cloud Storage: Advantages, Disadvantages and Enterprise Solutions for Business", (2016).
[12] Wu, Jiyi, et al. "Cloud storage as the infrastructure of cloud computing", 2010 International Conference on Intelligent Computing and Cognitive Informatics (ICICCI), IEEE, 2010.
[13] "Cloud Storage", Nonprofit Technology Collaboration, available online at: http://www.baylor.edu/business/mis/nonprofits/doc.php/197132.pdf
[14] Akingbade L.O, "Cloud Storage problems, benefits and solutions provided by Data De-duplication", International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 5, Issue 6, November 2016.