International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 151 - 155

A Novel Approach for Preserving Privacy of Content Based Information Reterival System

Miss Sayali .P Shinde
Department of Computer Engineering
R.M.D.S.S.O.E Warje , Savitribai Phule University
Pune, Maharashtra
Email:sayali.shinde610@gmail.com

Prof. Jyoti. S Raghatwan
Department of Computer Engineering
R.M.D.S.S.O.E Warje , Savitribai Phule University
Pune, Maharashtra
Email:jyotiraghatwan2@gmail.com

***Abstract*—** Content based information retrieval system (CBIR) are advanced version of retrieval systems where search is based upon specific criteria in order to get relevant items. In networking environment, as search is based on content it is easy for server to know client's interest, where client has to trust server to get relevant items. Sometimes query contains sensitive information that client does not want to reveal it, but still search should be performed. This is achieved by our proposed structure, where mainly it will deal with multimedia items such as image or audio files. In order to preserve privacy , client selects multimedia file of which hash value is generated, this value is fired towards cloud server. Cloud server contains database of stored hash values of multimedia items and based upon hamming distance and similarity search, encrypted candidate list is prepared and send it to client. Client finds best item by carrying decryption.

Keywords- *Hashing, Indexing,, PCBIR, Data privacy, Hamming distance, Audio fingerprint.*

*****

I. INTRODUCTION

Content-based information retrieval (CBIR) is used as query by content It is the application of computer vision techniques. It is used to solve retrieval problem, that is, allows the multiple users to search relevant content in large datasets. Here content means search should be done based upon specific criteria The term content refers to visual features, such as color, texture, and shape information, from which images can be extracted automatically in case of image database and melody pitch, rhythm in case of audio files.

The CBIR has become popular because of the various limitations in metadata-based systems, as well as for large range of users to have efficient retrieval. Existing systems successfully provides the textual information about multimedia files but requires humans to manually describe each object in the database. This will become very time consuming and impractical for very large databases.. Therefore there is a need of CBIR system.

Initial CBIR systems were developed to search databases based on visual features. With the invention of these systems, there was requirement to have suitable interfaces Efforts were made in the CBIR field to have feasible design that will meet needs of the user while performing search. It will have variety of methods that may allow user feedback, machine learning where all needs of user are satisfied. In networking environment usually roles of database owner, service provider is exchanged. So in case of CBIR systems client is providing server with specific criteria for retrieving results , so it is very easy for server to know what is client interest. There may be possibility that server may make misuse of client information or may build wrong profile. This issue needs to be addressed with sufficient privacy mechanisms and can be handled by privacy preserving content based information retrieval (PCBIR) which our proposed framework .

There are various application scenarios where there is need to handle privacy issue. If client has invented trademark and want to search in public database if similar exists without revealing it, here privacy concern is of client. Another application of this can be, client sends his medical images to a syndrome database for automatic similarity search. Here main motive is that the server should not see the query that is it should not be able to predict patient health condition but should perform the search for disease. There is case where database contains private information and restricted from public access or usage for example if client has audio clip and wants to know name of song so here main concern is privacy of database and also client's interest profile. So here main concept is that usually database owner and client does not trust the server. Here sensitive information should be protected from server and it is achieved through our proposed structure called (PCBIR) systems.

The main idea is to perform search without revealing the content where privacy of client must be protected while sending queries, database privacy while server is sending answers to client, and client privacy while receiving the relevant answers. Many works are done on this issue but they have non-scalability issue, degraded retrieval performance, Expensive and hard implementation and unbalanced load between client and server. The proposed work is carried under search with reduced reference approach. Here it will deal with multimedia data such image, audio. In order to provide privacy on data exchanged between client and server, robust hashing and encryption techniques on files are applied. System uses RC4 algorithm for encryption and decryption. It uses random projection which divide image into feature vector and apply hashing on that values. FDMF and fingerprint system are used for audio files.

In this paper a study about the related work and its background is done in section II, the implementation details. in section III where we see the system architecture, modules description, mathematical models, algorithms and experimental setup. In section IV we discuss about the expected results and at last we provide a conclusion in section V.

II. RELATED WORK

In paper [1] a scheme is proposed for privacy preservation of CBIR system ,in which client generates partial query from which certain bits are omitted from hash value which will create ambiguity to server. Server will return list of all matching items belonging to hash value from which client will search required item. Here only hash values are exchanged so privacy is better protected. But drawback of system is that client cannot retrieve whole multimedia object only it can see whether the required item is present in database or not. Also both client and server are using same hashing techniques, indexing schemes. It deals with particularly image database

A work in [2] is proposed the isolated information and similarly enables the secure watermark attraction that protects information Data watermarking principle is supportable opposite to the semi-honest protection conjecture and also improves the isolated data by building the newest protected information. In addition, graphic data works on DCT coefficients of image data and it is able to predict the progress of scalability, robustness and quality, also proposed to offer greater SNR ratio than the present system.

One of the study on content-based retrieval on a protected media database is carried out in [3] which maintains the ability of likeness comparison and the isolation of picture material from the server is protected by using repository, here attention to building protected research indexes is given. Using exploiting techniques from cryptography, various protected indexing schemes are designed such as protected inverted index, protected min-Hash sketches. Author in [4] tries to present an alternative signal running to see on solitude preserving search problem. The results of computer simulations proves that the consistency of the proposed construction for the extensive class, upset versions from the signal running group of distortions.

A work in [5] is proposed for an active protocol that retrieves one bit at a time. User's responsibility is to maintain the hierarchy information of the database. By exploiting the quadratic residuosity assumption, it directs multiple indices to the host and hides the correct question. The limitation of this system is that load of searching is completely shifted towards an individual and also it violates the privacy of the database. Database privacy is way better protected using homomorphic encryption presented by Sabbu et al.

This paper [6] presents a new privacy amplification approach which is dependent on data covering maxims and has flexibility benefits of soft fingerprinting and investigation of the identification-rate compared to privacy-leak trade-of art. Side data is discussed between the encoder and decoder and the evaluation is performed for the great fit along with case of partial side information.

This paper [7] combines perceptual hashing and sturdy watermarking. A picture is divided into blocks and each stop is displayed by a compact hash value stuck in the block. The authenticity is confirmed by extracted image by comparing and re-computing hash values of the image.

III. PROBLEM DEFINITION

To perform search without revealing the original query which contains sensitive information

The main aim is to protect

1) Client's privacy when sending queries
2) Client's privacy when receiving answers
3) Server's privacy when sending answers

This is made possible by (PCBIR) systems

IV EXISTING SYSTEM

In existing system a scheme is proposed for privacy preservation of CBIR system ,in which client generates query as hash value and fired towards server. Server will return list of all matching items belonging to hash value from which client will search required item. Here only hash values are exchanged so privacy is better protected. But drawback of system is that client cannot retrieve whole multimedia object only it can see whether the required item is present in database or not. Also both client and server are using same hashing techniques, indexing schemes. It deals with particularly image database

V PROPOSED SYSTEM ARCHITECTURE

A. System Overview

System is consisting of multiple clients and a cloud server. Server stores the multimedia objects such as image or audio files with their respected hash values. On the other hand client wants to retrieve particular multimedia files which can be either image or audio. To hide the identity of original file contents, hash function is used to generate its hash value. Finally client sends hash value of multimedia file to the cloud server as an input query.. After receiving input query, server carries search for related items in database. Actually server compares the hash value of input query with hash values of multimedia files stored in database on the basis of hamming distance. From this based upon similarity search server generates candidate list. This candidate list is encrypted using RC4 algorithm and sends it to client as an output. At client side, encrypted candidate list is received and performs final search within the candidate list for best suitable item.

System consisting of following modules:

*1) **Query Generation:***

In order to provide privacy, contents that represent client interest that is fired by query are not safe Sometimes even features are not safe, because they reveal information about the original content. To provide the privacy, input file is compressed preserving the core parts and then generates its hash and finally sends hash value of multimedia file to server.

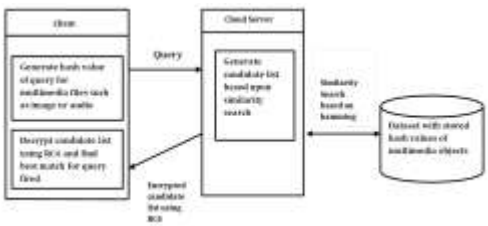

Fig. 1: Proposed System Architecture

*2) **Database Indexing**:*

The database indexing is based on the concept of piece-wise inverted indexing. There is a general component which extracts features of multimedia file or object. From the extracted features a vector is formed which is capable of

characterizing the underlying content. Every feature is associated with sub hash value and whole file is concatenation of sub hash values and represented by single hash value. Also sub hash value is associated with inverted index list. The list contains id of multimedia objects corresponding to sub hash value.

3) ***Database Search***

In this based upon hamming distance, nearest neighbor search is done within radial distance. Multimedia objects having smaller radial distance in hamming sphere are retrieved and placed in candidate list.

VI. IMPLEMENTATION DETAILS

A. Algorithms

Algorithm 1: Random Projection Algorithm

1. Set the hash code V = 0;
2. For every feature f do create feature vector C using feature f.
3. for every image k = 1 to C do
4. Generate a hashing function using component f_k;
5. Build a hashing able where each bucket store image.
6. Names corresponding to the same hash code.
7. add the hash code to the set V = V [f_k]
8. end for
9. end for
10. Design the a graph consisting of V nodes corresponding to all hash codes
11. Measure the similarity between each pair of the codes.
12. Select m hash codes in a greedy way using. The set of selected hash codes is denoted as A.
13. for every query q do
14. Initialize the retrieve set as R = 0;
15. For code m in A do R = R[samples in the same bucket of q using code m].
16. end for

Algorithm 2: RC4 Algorithm

Rc4 is presumably the most generally utilized stream cipher as a part of the world because of its straightforwardness and proficiency. RC4 generates a pseudorandom stream of bits called as key stream. As with any stream cipher, these can be used for encryption concatenation of content using bit-wise exclusive-or; decryption is performed in the same way Pseudo-random generation algorithm (PRGA) the PRGA modifies the state and outputs a byte of the key stream. PRGA (P)

Initial conditions
i←0
j← 0
Formation of loop:
i ←(i+1)mod256
j← (j+S[i])mod256
P[i] ↔ P[j]
Output z ←+P[P[i]+P[j]mod256]
PRGA(P,i,j)
Generation loop:
P[i]↔ P[j]
j← (j-P[i]+256)mod256
i ←(i-1+256)mod256
Output z ←+P [(P[i] + P[j]) mod256]

Algorithm 3: FDMF - Find Duplicate Music Files and Fingerprint

FDMF is an acronym for 'Find Duplicate Music Files' and fingerprint system provided by Kurt Rosenfeld It has been designed to detect equal versions of the same song title even if the meta-data is not the same

1. Decode / decompress the input file to raw audio data (PCM)
2. Apply the Fast Fourier Transformation (FFT)
3. Divide the frequency spectrum into 4 non-overlapping frequency bands (B1 -B4)
4. Calculate the energy of the bands for each 250 milliseconds chunks
5. Define a result list (FP . . . fingerprint) consisting of regions and calculate the following values:
a) FP[0..255] = b) FP[256..511] = (B2 + B3)/(B0 + B1)-ratio for each chunk
b) FP[512..767] = (B0 + B2)/(B1 + B3)-ratio for each chunk
The sum of the energy bands for each chunk 6. Calculate the power spectra for each array
7. Apply spline fit smoothing operation
8. Apply a one-bit median quantization on these values
9. Concatenate the bit string to get the full fingerprint out of 3 times 256 bit, a 768 bit signature 2
10. Store the representation to the database

For the evaluation, the following steps have to be performed:
1. Extraction of the fingerprint as described
2. Comparison with database entries
3. If the results (i.e. distance values) exceed the given thresholds (one for each region), a proper identification is confirmed and printed.

B. Mathematical Model

System S = {I, H, Q, D, En(L), Dn(L), O }

Input: I= Multimedia file image or audio
I = $\{I_1, I_2, \ldots I_n\}$ where I is a set of input
Output: O =Candidate list
O= $\{O_1, O_2, \ldots O_n\}$ where O= Required item.
Process:

1. Hash values generation:-
H=$\{H_1, H_2, \ldots H_n\}$ where H is set hash values generated

2. Query acceptance at server side:-
Q= $\{ H_1, H_2, \ldots H_n \}$ where Q is a set of queries accepted by cloud server.
3. Hamming Distance Calculation:-
$D(H_1,H_2)\ |L_1=\sum_{i=0}^{n-1} |\ d_H(h_1,h_2)\ |$

where $L_{1=}$ Hash distance, d_H =Hamming distance between two sub hash values

4. Encryption:-
   En(L)= {I_1, I_2, … I_n} where En(L) is encrypted list of all similar items to the input query

5. Decryption:-
   Dn(L)= {I_1, I_2, … I_n}where Dn(L) is decrypted list and finding the best match.

6. Output O=Required item

C. Experimental Setup

The proposed system is using Java (jdk 1.8 version) on Windows platform. The Net beans (version 8.2) is used as a development tool. There no any specific hardware required to run, so any standard machine is capable of running the application. The system analysis is carried out on datasets consisting of files.

VII RESULT AND DISCUSSION

Following gives idea about results obtained from the scheme proposed

Fig 2: Selection of file to be fired as input query

Fig 3: File selected whose hash value should be generated

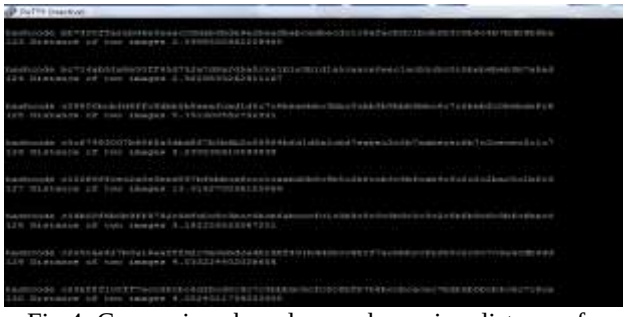

Fig 4: Comparison based upon hamming distance for similarity search by cloud server

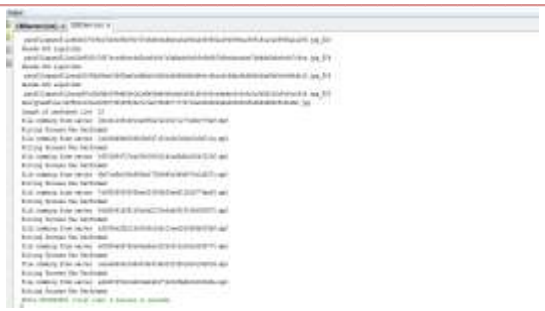

Fig 5: Comparison based upon hamming distance for similarity search by Normal server

Fig 5: Encrypted Candidate list recevied at client side

Fig 6:Decryption of Candidate list at client side

Fig 7. Secure retrieval of audio files matching query as hash value

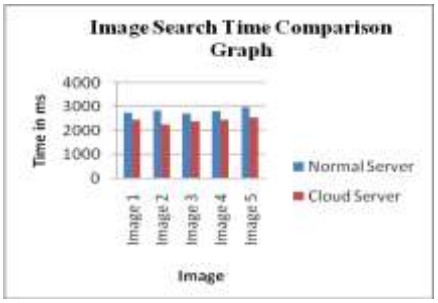

Fig 8: Image Retrieval Time comparison between existing and proposed System

Image	Existing System with Normal Server	Proposed System with Cloud Server
Image 1	2745 ms	2436 ms
Image 2	2839 ms	2250 ms
Image 3	2690 ms	2380 ms
Image 4	2785 ms	2432 ms
Image 5	2950 ms	2538 ms

Table 1: Time Comparison for retrieval of image

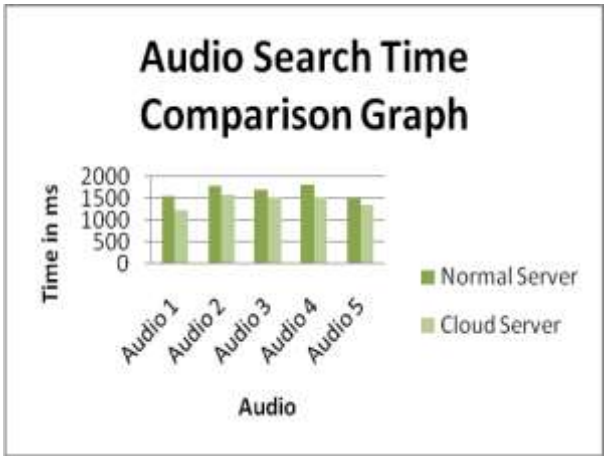

Fig 9: Audio Retrieval Time comparison between existing and proposed System

Audio	Existing System with Normal Server	Proposed System with Cloud Server
Audio 1	1550 ms	1230 ms
Audio 2	1760 ms	1570 ms
Audio 3	1667 ms	1495 ms
Audio 4	1788 ms	1500 ms
Audio 5	1482 ms	1330 ms

Table 2: Time Comparison for retrieval of audio

VII CONCLUSION

The proposed system improves content based information retrieval system by providing better privacy than the existing one. The existing framework is applicable on client server system, but this proposed system is applicable for cloud server and multiple clients. The proposed system deals with image and audio files. This system uses various techniques to improve performance by which privacy is better protected. Various techniques such as random projection, RC4 , FDMF are used. They exhibits satisfactory performance in terms of privacy and information retrieval.

ACKNOWLEDGMENT

I take this opportunity to express my gratitude to my guide Prof J.S .Raghatwan and head of department, Prof. V. M. Lomte, Department of Computer Engineering, RMDSSOE, Pune University, for their kind cooperation and guidance during the entire research work.

REFERENCES

[1] Li Weng, Laurent Amsaleg , “A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval”, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 1, JANUARY 2015.

[2] H.N.Ranotkar, Prof.M.S. Deshmukh, “Privacy Enhancement of data with safe watermark extraction using signal processing”, International Journal of Application or Innovation in Engineering Management (IJAIEM), Volume 3, Issue 11, November 2014.

[3] W. Lu, A. Swaminathan, A. L. Varna, and M. Wu, “Enabling search over encrypted multimedia databases,” Proc. SPIE, Media Forensics Secure., vol. 7254, pp. 725418-1-725418-11, Feb. 2009.

[4] S. Voloshynovskiy, F. Beekhof, O. Koval, and T. Holotyak, “On privacy preserving search in large scale distributed systems: A signal processing view on searchable encryption,” in Proc. Int. Workshop Signal Process. Encrypted Domain, Lausanne, Switzerland, 2009.

[5] Ms. Archana Chemate, Prof. S. P. Pingat, “Reliable data delivery in low-power and lossy networks using trust based link selection.” June 2015.

[6] S. Voloshynovskiy, T. Holotyak, O. Koval, F. Beekhof, and F. Farhadzadeh, “Private content identification based on soft fingerprinting,” Proc. SPIE, Media Watermarking, Secur., Forensics III, vol. 7880, pp. 78800M-1-78800M-13, Feb. 2011.

[7] L. Weng, G. Braeckman, A. Dooms, B. Preneel, and P. Schelkens, “Robust image content authentication with tamper location,” in Proc. IEEE Int. Conf. Multimedia Expo, Jul. 2012, pp. 380-385.

[8] J. Shashank, P. Kowshik, K. Srinathan, and C. V. Jawahar, “Private content based image retrieval,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2008, pp. 1-8.

[9] B. Mathon, T. Furon, L. Amsaleg, and J. Bringer, “Secure and efficient approximate nearest neighbors search,” in Proc. 1st ACM Workshop Inf. Hiding Multimedia Secur. (IH & MMSec), Jun. 2013, pp. 175–180.

[10] Dawson, E., and Nielsen, L. "Automated Cryptoanalysis of XOR Plaintext Strings." Cryptologia, April 1996.

[11] Knudsen, L., et al. "Analysis Method for Alleged RC4." Proceedings, ASIACRYPT'98, 1998.

[12] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, “A review of audio fingerprinting,” J. VLSI Signal Process. Syst., vol. 41, no. 3,pp. 271–284, Nov. 2005.

[13] L. Cao, Z. Li, Y. Mu, and S.-F. Chang, “Submodular video hashing:A unified framework towards video pooling and indexing,” in Proc. 20th ACM Int. Conf. Multimedia, 2012, pp. 299–308.

[14] H. Özer, B. Sankur, N. Memon, and E. Anarim, Perceptual audio hashing functions,” EURASIP J. Appl. Signal Process., vol. 2005, pp. 1780–1793, Jan. 2005