A Survey Paper on Secure Privacy Preserving Structure for Content Based Information Retrieval on Large Scale

Sayali P. Shinde
MEStudent
Department of ComputerEngineering
RMDSCOEWarje
Pune, India
sayali.shinde610@gmail.com

Prof. J. S. Raghatwan
AssistantProfessor
Departmentof ComputerEngineering
RMDSCOE, Warje
Pune, India
jyotiraghatwan2@gmail.com

Abstract:-It is very essential to protect personal confidential data that we share or search through web. Previously there are number of privacy preserving mechanism has been developed. Here we develop a new privacy protection framework for huge- content-based information retrieval. We are offering protection in two layers. Initially, robust hash values are taken as queries to avoid revealing of unique features or content. Then, the client has to select to skip some of the bits in a hash value for increasing the confusion for the server. Since we are reducing information it is not so easy for servers to know about interest of the client. The server needs to give back the hash values of all promising candidates to the client. The client will find the best match by searching in the candidate list. Because we are only sharing hash values between server and client the privacy of client and server will be protected. We begin the idea of tunable privacy, where we can adjust level of privacy protection according to the policy. We can realized it by hash based. It can be realized through piecewise inverted indexing based on hash. We have to divide extracted feature vector into pieces and index each and every piece with a value. Every value is linked with an inverted index list. The framework has been comprehensively tested with very huge image database. We have estimated both privacy-preserving performance and retrieval performance for those content recognition application. Couple of robust hash algorithm is being used. One is based on discrete wavelet transform; the other is based on the random projections. Both of these algorithms demonstrate acceptable recital in association with state-of-the-art retrieval schemes. We believe the bulk voting attack for guesstimate the query recognition and sort. Experiment results confirm that this attack is a peril when there are near-duplicates, but the success rate is depends upon the number of distinct item and omitted bits, success rate decrees when omitted bits are increased.

Keywords: *Content based retrieval, image hashing, indexing, multimedia database, PCBIRI.*

*****

I. INTRODUCTION

This is the era of World Wide Web, everyone is using internet for various purposes. Web is the huge source of information which is available everywhere every time. So to make retrieval of information easier and efficient content based information system had been discovered. Previously search is done on the basis of textual queries and available metadata. In CBIR system user has to provide user query which content relevant data is regarding information which user wants to search. Query can be text, image, audio, video or any type of multimedia file.

One can share and access personal, confidential information with CBIR, in network there are huge number of client who fires queries, there are number of servers so it is essential to authenticate users or server to access confidential data. If such confidential data goes into the wrong hand then there will be big trouble or risk. So it is very crucial to preserve privacy of such data. In case of biometric authentication it is not feasible for the entire user to validate their self by using biometric devices [1]. Some of the information providing websites likes quora, Facebook make user to create their profile which contains much of personal information about every user who wish to search content using their sites.

Some time it is seen that such website makes misuse or they provide this profile information to others. So it is very necessary to take care of important data when we share or access it over web [2]. Some recommendation providing system is getting into the action nowadays. This system also requires profiling which is a big concern now days [3].In public content based information retrieval system one has to deal with the authentication, sometimes database owner doesn't have trust on to the database servers, server may contains private information which is requested by unknown client in such case privacy preserving concept comes into focus. Some clarification are depends on the idea of Signal Processing in Encrypt Domain, many time called as (SPEED) [4]. They normally rely on intense cryptographic calculation, like multiparty computation and homomorphic encryption [5].

Some other clarification are there which depends on the idea of Search with Reduced Reference. The new confront is the retrieval difficulty. It like that the server should not have any information accurately which database point has been retrieved by the client. Very a small number of works have attended to this issue.

It is also necessary that privacy protection mechanism should require less cost than the cost of data retrieval. Our system allows calculation difficulty in a systematic method and bendy trade-offs between privacy. So, it can be applied in a heterogeneous network where devices have various computing bandwidths and power.

The proposed framework is essentially an SRR approach. Main strategy used here is robust hashing and piece-wise inverted indexing. The flexibility is mainly embodied in which any feature is converted into hash values. In addition, we are able to control privacy by using policy . These elements work together according to a new PCBIR protocol

II. BACKGROUNDANDRELATEDWORK

H.N. Ranotkar proposed [6] the structure that improves and protects the solitude of the information and simultaneously enables the secure watermark attraction. The structure improves the solitude of data by building the newest protected information using data watermarking principle which makes it supportable contrary to the semi-honest protection conjecture. In addition, it predictable to progress the scalability, robustness and quality of a graphic data as it works on DWT coefficients of image data. It's proposed to offer greater SNR ration than the present system.

This paper [7] makes the first endeavor on content-based retrieval over a protected media database. Using picture repository for example, we give attention to building protected research indexes, which protect the solitude of picture material from server and maintain the ability of likeness comparison. Two protected indexing schemes, particularly, protected inverted index and protected min-Hash sketches, are designed by jointly exploiting techniques from cryptography, picture control, and information retrieval In this paper [8], author attempted presenting an alternative signal running on the solitude preserving search problem. The proposed approach mimics thefunction of problem modification rules with delicate details about touch reliability executedin the spread manner.

The outcome of computer simulations proves the consistency of the proposed construction for the extensive class of disturbances versions from the signal running group of distortions. It is based on signal processing, main computational work is owed to a data user, who confront an unsecure server by numerous requests, whereas the task of the server is condensed to appropriately responding to these dispute. The main advantage is it impersonate the load of error correction codes with soft data about bit dependability executed in the distributed manner

This technique need to examine the broader class of deformation under both list decoding and unique regimes. The results of computer simulations confirm the reliability of the proposed system for the broad class of deformation models from the signal processing family of deformation

An early system was proposed by Shashank et al. [9].That is an active protocol that retrieves one bit at a time. The user has to maintain the hierarchy information of the database. It directs multiple indices to the host and hides thecorrect question by exploiting the quadratic residuosity assumption.

The drawback with this system is that the burden of search is absolutely shifted to an individual, and additionally, it violates the privacyof the database. Sabbu et al.[10] proposed yet another answer predicated on homomophic encryption, where in fact the database privacy is way better protected

S.Voloshynovskiy, propose a brand new privacy amplification approach [11] based upon on data hiding principal which is dependent on data covering maxims and benefits from side information about touch consistency a.k.a. soft fingerprinting. In this paper, we investigate the identification-rate compared to privacy-leak trade-of art. The evaluation is performed for the case of a great fit between side data discussed between the encoder and decoder along with for the case of partial side information.

One can attain substantial confidentiality amplification by using even deficient side information devoid of the recognition rate loss. Here improved data hiding techniques need to be resolved. The privacy augmentation can be resolved without any publicly stored data about consistent bits opposing to the state-of-the-art technique.

L. Weng, propose a story image verification system [12] by combining perceptual hashing and sturdy watermarking. A picture is divided into blocks. Each stop is displayed by a compact hash value. The hash value is stuck in the block. The authenticity of the image may be confirmed by re-computing hash values and comparing them with those extracted from the image

It achieves a very high safety level, since they carry out verification on image block level. The method can endure a broad range of incidental alteration, and locate interfere area small as 1/64 of an image. A hash value is entrenched in a relatively tiny image block, and image excellence can be ruined due to watermark embedding

Shashank J [13] resolved the problem of private collection in image databases and deals with retrieving comparable images from an image database without illuminating the content of the query image not even to the database server.

The algorithm is found to be customizable for hierarchical information structures as well as hash-based indexing schemes. First examine has shown that the algorithm is accurate, successful and scalable.We planned algorithms which are completely private and feasible on reasonably large sources using a variety of state of heart indexing schemes.

It is fully confidential and realistic on reasonably huge databases using a range of state of the art indexing system. Here there is need to minimize the time complexity to retrieve the image. Image databases are open to considerably quicker private retrieval by developing the possessions of image data.

III. EXISTINGSYSTEM

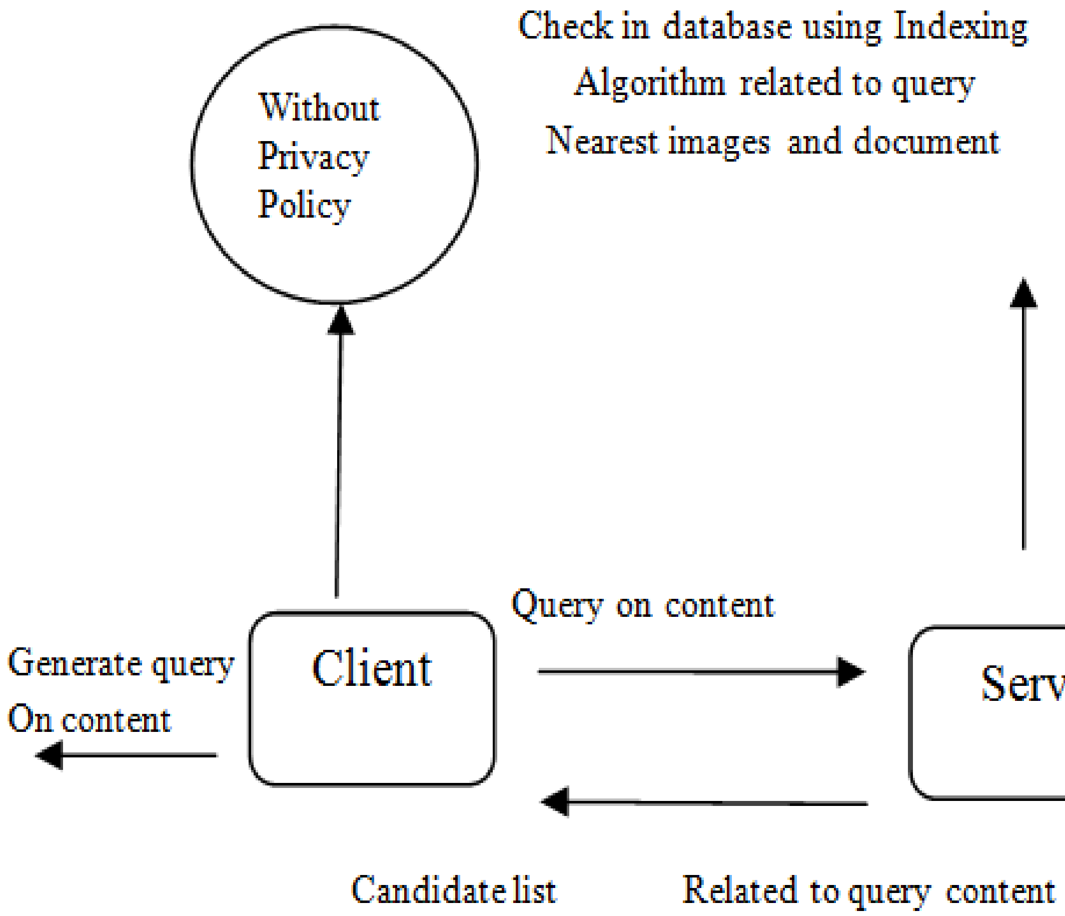

Fig 1. Basic CBIR Scheme

1.The client will send query to the server which is based on content
2. The server generates an candidate list of items which is based on the query content
3. The server will do a search with the extended query list, and sends back all matching items
4. Within the received set of items client will perform match based upon query fired

A drawback of this proposal is that privacy is not maintained in which client has to trust the server to get required item, where server can cause threat to the client as it is possible for server to know the interest of client based upon content query fired by client. Also there is possibility that client is able to know database contents based upon candidate list returned by server
In order to overcome this drawback a new scheme is proposed, where both privacy of client, server is preserved

IV. PROBLEM STATEMENT

We propose a security insurance system for huge scale substance based data recovery. It offers two layers of security. In the first place, strong hash qualities are utilized as inquiries to avoid uncovering unique substance or elements. Second, the customer can decide to discard certain bits in a hash worth to further build the equivocalness for the server. CBIR systems face the common problem that the server is not trusted by the database owner or the user. The query has some secrete information. The CBIR technique involves some private information, e.g. proprietary technology.
Signal Processing in Encrypted Domain (SPEED) rely on heavy cryptographic computation, such as homomorphic encryption and multiparty computation, cryptographic level protection of private data. Concepts of Search with Reduced Reference (SRR) use a secure index (the reduced reference) as the query. The scrutinized information will protect the original content and accelerate the search using PCBIR system.

V. PROPOSED SYSTEM

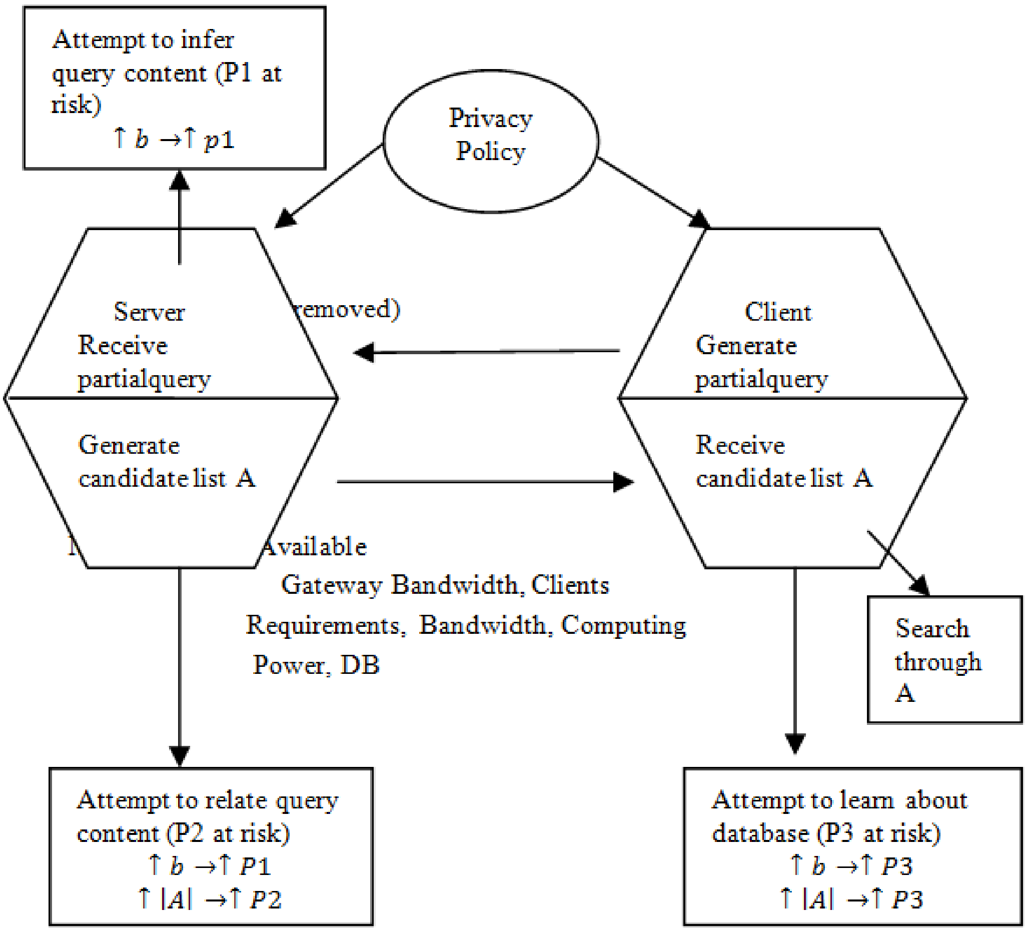
Attempt to infer query content (P1 at risk)
↑ b →↑ p1
Privacy Policy
Server Receive partialquery
emoved)
Client Generate partialquery
Generate candidate list A
Receive candidate list A
N
Available
Gateway Bandwidth, Clients Requirements, Bandwidth, Computing Power, DB
Search through A
Attempt to relate query content (P2 at risk)
↑ b →↑ P1
↑ |A| →↑ P2
Attempt to learn about database (P3 at risk)
↑ b →↑ P3
↑ |A| →↑ P3

Fig 2. Proposed PCBIR model

A new PCBIR structure is developed which can be used for both public and private databases

- It is designed for large-scale databases;
- Different levels of protection is provided
- It is easy to use and generalize.

As far as we know, the granularity of privacy protection is a factor which is not taken into consideration in existing PCBIR solutions. In client removes bits from the query to create some confusion for the server. Proposed PCBIR procedure works as follows:

Client will produce a partial query which is hash value and will send it to server

- The server generates an extended query list based on the partial query.
- The server performs a search with the extended query list, and sends back all matching items
- Within the received set client will do search for matching results using the original query.

Specifically, partial query has some constraints that need to be followed:

- It is difficult to infer the original query;
- It is feasible to produce and have search with the extended query list;
- The properties of set ‘A’, e.g. the size and the varied items, are controlled by the partial query;
- It is easy to estimate privacy of client

There are the various constraints on the matching set ‘A’:

- ‘A’ should be compact enough to save bandwidth;
- ‘A’ contains the best answers
- The number of varied of elements in ‘A’ is sufficiently large;
- The server cannot predict are the answers best by analysing ‘A’. ‘A’ should not reveal too much information about the database

In above model P1 represents privacy of client while it is sending query ,P2 privacy of client while receiving answers, and P3 is server privacy when sending answers.

VI. CONCLUSION

We have implemented a framework for enhancing privacy for huge content-based retrieval of information which can be used for system depends on similarity search and features. It focuses on piece wise inverted indexing and robust hashing. We can also adjust the privacy level of protection by our privacy policy. Bandwidth and local processing depends upon the strangeness of the policy.

We use two different hashing algorithms which are based on sign bits of DWT coefficient and sign bits of random projection. Both this shows best performance than existing state of art reference scheme. Also the success rate depends upon the number of omitted bits and number of distinct item. If number of omitted bits increases success rate will be degrade. We observed that if server only gives one (accurate) answer, it only has knowledge about the user query. So if privacy is required it is to be expected to send more than one multimedia objects.

Here we can use this framework with several features. Sub hash value is associates with several different features. We can divide feature in number of segments for long feature vector by which number of hash table will be increased but hashing table can condense duplicate features. It seems like just the once hash supported search and indexing become broadly used, our system takes fewer exertion to execute.

ACKNOWLEDGMENT

I like to acknowledge my vigorous thanks to Prof. J.S. Raghatwan for giving her valuable suggestions. I also express my special thanks to my friends ,for supporting me in my research work.

REFERENCES

[1] J. Bringer, H. Chabanne, and A. Patey, “Privacy-preserving biometric identification using secure multiparty computation: An overview and recent trends,” *IEEE Signal Process. Mag.*, vol. 30, no. 2, pp. 42–52, Mar. 2013

[2] G. Fanti, M. Finiasz, and K. Ramchandran, “One-way private media search on public databases: The role of signal processing,” *IEEE SignalProcess. Mag.*, vol. 30, no. 2, pp. 53–61, Mar. 2013.

[3] A. Aghasaryan, M. Bouzid, D. Kostadinov, M. Kothari, and A. Nandi, “On the use of LSH for privacy preserving personalization,” in *Proc. 12th IEEE Int. Conf. Trust, Secure., Privacy Comput.Commun. (TrustCom)*, Jul. 2013, pp. 362–371

[4] Z. Erkin*et al.*, “Protection and retrieval of encrypted multimedia content: When cryptography meets signal processing,” *EURASIP J. Inf.Secure.*, vol. 2007, p. 20, Dec. 2007.

[5] R. L. Lagendijk, Z. Erkin, and M. Barni, “Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation,” *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 82–105, Jan. 2013

[6] 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

[7] 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.

[8] 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

[9] Ms.Archana chemate, Prof. S. P. Pingat, reliable data delivery in low-power and lossy networks using trust based link selection. June 2015.

[10] P. R. Sabbu, U. Ganugula, S. Kannan, and B. Bezawada, “An oblivious image retrieval protocol,” in *Proc. IEEE Int. Workshop Adv. Inf. Netw.Appl. (WAINA)*, Mar. 2011, pp. 349–354.

[11] 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.

[12] 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

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