A Privacy-Preserving Framework for Large-Scale Content- Based Information Retrieval

Mayuri Sewatkar[1] Raghvendra Choudhary[2] Saurabh Bagde[3] Chetan Shahade[4]
[1,2,3,4]Savitribai Phule Pune University, Sinhgad Institute of Technology, Kusgaon(BK), Lonavala

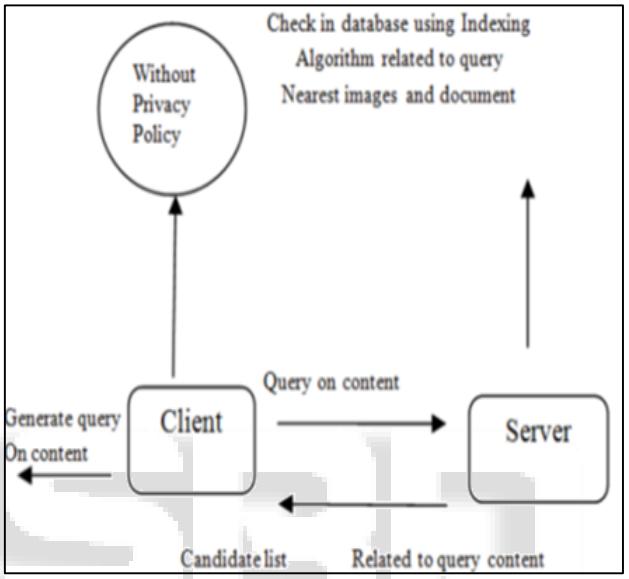
Abstract— It is very necessary to protect personal confidential data that we share or search through web. Earlier there are number of privacy preserving mechanism has been developed. In this project, we develop a new privacy protection framework for huge- content-based information retrieval. We are contributing protection in two layers. Originally, 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. Due to the suppressed information, it is computationally difficult for the server to know the client's concern. The server has to return the hash values of all desirable candidates to the client. The client executes a search within the candidate list to find the best match. Subsequently only hash values are exchanged between the client and the server, the privacy of both parties is protected. We imported the concept of tunable privacy, where the privacy protection level can be adjusted according to a policy. It is concluded through hash-based piecewise inverted indexing. The concept is to divide a feature vector into pieces and index each piece with a sub hash value. Every sub hash value is associated with an inverted index list. The framework has been majorly tested using a large image database. We have calculated both retrieval performance and privacy-preserving performance for a particular content identification application. Both algorithms illustrate satisfactory performance in comparison with state-of-the-art retrieval schemes. The results show that the privacy enhancement somewhat improves the retrieval performance. We consider the majority voting attack for reckoning the query category and identification. Experiment results show that this attack is a threat when there are near-duplicates, but the success rate reduces with the number of omitted bits and the number of distinct items.

Key words: Framework, Information Retrieval

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. "Content Based" means that the search analyses the contents of the image rather than the metadata such as keyword, tags or the description associated with the image. The term "content" in this context might refer to the colour shape and texture or any other information that can be retrieved from the image itself. With the emergence of new applications, an issue with content-based search has arisen sometimes the query or the database contains privacy-sensitive information. A privacy issue arises when an untrusted party wants to access the private information of another party. In that case, measures should be taken to protect the inter-related information. The main ultimatum is that the search has to be performed without revealing the original query or the database. This triggers the demand for privacy-preserving CBIR (PCBIR) systems.

II. EXISTING SYSTEM



1) The client will send query to the server which is based on content
2) The server develops 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 returns back all matching items
4) Among the collected 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 probable for server to know the interest of client based upon data 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 conquer this drawback a new scheme is proposed, where both privacy of client, server is preserved.

III. PROBLEM STATEMENT

A privacy issue emerges when an untrusted party wants to access the private information of another party. In that case, measures should be taken to protect the analogous information. The main ultimatum is that the search has to be performed without revealing the original query or the database. This provokes the need for privacy-preserving CBIR (PCBIR) systems. In order to protect privacy, original data cannot be used as queries. Erstwhile even features are not safe, because they still reveal information about the

original content. Rather than encryption, we generate queries from original data by robust hashing.

IV. PROPOSED FRAMEWORK

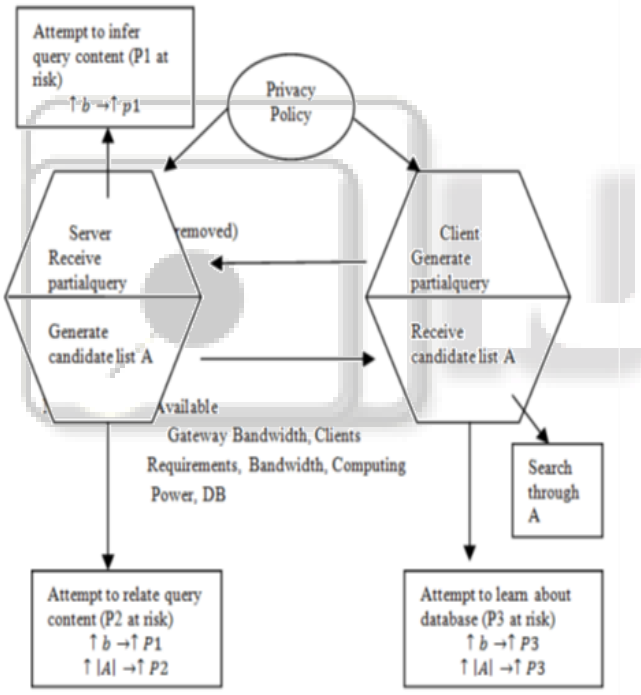
A new PCBIR structure is developed which can be used for both public and private databases

- 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.



A good system should keep P1, P2, and P3 sufficiently large. Specifically, partial query has some constraints that need to be followed:

1) It is difficult to infer the original query,
2) It is feasible to generate and perform search with the extended query list,
3) The properties of A, e.g. the size and the diversity, can be controlled by the partial query and
4) It is easy to estimate P1.

A. Query Generation:

In order to protect privacy, original content cannot be used as queries. Sometimes even features are not safe, because they still reveal information about the original content. Instead of encryption, we generate queries from original content by robust hashing.

Robust hashing is also called perceptual hashing or robust fingerprinting (for multimedia data), or locality-sensitive hashing (LSH) (for generic data). It is a framework that maps multimedia data to compact hash values.

Elegantly, a robust hash value is a short string of equally probable and independent bits. It can be used to carefully identify or authenticate the underlying content, just like a "fingerprint". The basic property of robust hashing is that similar data should result in similar hash values. More significantly, hash algorithms have the one-way property that it is computationally difficult to infer the input from the output, because hashing is essentially a many-to-one mapping.

The leverage of robust hashing in this application is mainly two-fold: 1) the compact size can facilitate fast search (in the Hamming space if binary); 2) due to the one-way property, the privacy requirements P1 and P3 can be achieved by using the hash value instead of the original content (or features) for the search and return of answers. Using hash values for privacy protection is also called generic privacy amplifications. A conventional system can be enhanced by converting feature vectors into hash values. Another advantage of robust hashing is the possibility to overcome the semantic gap by supervised learning.

Robust hashing typically involves feature extraction, orthogonal transformation, dimension reduction, and quantization.

B. Database Indexing:

The database indexing is based on the concept of piece-wise inverted indexing. We assume there is a general feature extraction component. The extracted feature vectors are efficient of characterizing the underlying content. They first go through an orthogonal transform and dimension reduction. Only significant features are preserved. The elements of a feature vector are splitted into n groups. A robust hash value hi (i = 0, 1, ・ ・ ・, n − 1) is computed from the ith group. We call it a sub-hash value. The above step creates a new coordinate system, with every coordinate represented by a sub hash value. Lastly, a multimedia object in the database is indexed by the overall hash value i.e., the concatenation of sub-hash values.

Each sub-hash value is associated with an inverted index list (also called a hash bucket). The list contains the IDs (identification information) of multimedia objects corresponding to the sub-hash value. The size of a sub-hash value l depends on the significance of its corresponding feature elements.

C. Database Search:

When privacy protection is not required, the proposed framework can work as efficiently as a normal CBIR scheme. In general, there are several possibilities to perform database search. They mainly differ in the domain for distance computation, which can be the feature space, the quantized feature space, or the hash space. In order to facilitate the explanation, we assume that an original query is a hash value. It can be generated by the client, or the server. In the former case, P1 is still preserved, but there is no guarantee for P2. While in the latter case, since the client sends the original content to the server, no privacy is preserved for the client.

1) Approximate Nearest Neighbour Search: When the server receives a hash value, it checks the table for each sub hash value and optionally performs a nearest neighbour search within a Hamming sphere. For each binary sub-hash value, the multimedia objects IDs within a small Hamming radius r are retrieved. When $r \geq 1$, we call it multi-probing, because this is similar to the concept of multi-probe LSH. Additionally, when side information is available, different policies can be applied to prioritize sub-hash values in the neighbourhood. The retrieved objects for all sub-hash values are put into a list A. This list of candidates is sorted according to the hash distance from the query. The hash distance can be defined similarly as the L1 distance

$$D(H1, H2)|L1 = \sum_{i=0}^{n-1} \llbracket |dH(h1i, h2i)| \rrbracket , \quad (1)$$

where dH denotes e.g. the Hamming distance between two sub-hash values. In general, we assume that similar multimedia objects should have similar hash values. Therefore, the nearest neighbours can be obtained from the sorted list. If not specified otherwise, we use D(H1, H2)|L1 in the later experiments.

Distance computation can also be performed with feature vectors or quantization indices. In that case, we just need to replace dH in the above equations with the distance in the feature space or the quantization space. It is also possible to use other similarity metrics, such as the L2 distance or the cosine similarity.

V. CONCLUSION

In this project, we have presented a privacy preserving framework for large scale content-based information retrieval. It can be utilized for any CBIR framework based on features and similarity. This framework is mainly light of robust hashing and piece-wise inverted indexing.

The framework has been implemented and broadly assessed in different situations. We demonstrate that the security level, e.g., the number and the diversity of candidates can be tuned by the privacy policy. A few guidelines are given on how to choose the omitted bits. We have exhibited both retrieval performance and privacy-preserving performance for a specific content identification application. Experiment results show that query items with near-duplicates are likely to be vulnerable to majority voting. The chance of success is equivalent to the chance that a query item has more near-duplicates than other irrelevant items in the candidate list. The results also show that the success rate decreases with the number of omitted bits and the number of distinct items.

ACKNOWLEDGEMENT

We would like to thank to thank to our guide Prof. P. P. Ahire, our project co-ordinator Prof. R. S. Badodekar, our HOD sir Prof. N. A. Dhawas and all the people who were so much helpful.

REFERENCES

[1] Sayali P. Shinde, "A Survey Paper on Secure Privacy Preserving Structure for Content Based Information Retrieval on Large Scale".
[2] Cha Zhang and Tsuhan Chen, "An Active Learning Framework for Content Based Information Retrieval".
[3] Supriya G. More and Ismail Mohammed, "Survey on CBIR using K-Secure Sum Protocol in Privacy Preserving Framework".
[4] Ritendra Datta, Jia Li James and Z. Wang, "Content-Based Image Retrieval - Approaches and Trends of the New Age".

