International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 4 Issue: 6 314 - 317

A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval Using K-Secure Sum Protocol

Supriya G, More[1]
ME Student in Computer Engg., ACEM
Pune, India
supriyamore161@gmail.com

Ismail Mohammed[2]
Prof. In Computer Engg Dept., ACEM
Pune, India
ismail_009@yahoo.com

***Abstract*—**We propose a privacy protection framework for large-scale content-based information retrieval. It offers two layers of protection. To begin with, robust hash values are utilized as quiries to avoid uncovering unique content or features. Second, the customer can choose to exclude certain bits in a hash values to further expand the ambiguity for the server. Due to the reduced information, it is computationally difficult for the server to know the customer's interest. The server needs to give back the hash values of every single possible to the customer. The customer performs a search within the candidate list to locate the best match. Since just hash values are exchanged between the client and the server, the privacy of both sides is ensured. We present the idea of tunable privacy, where the privacy protection level can be balanced by policy. It is acknowledged through hash-based piecewise inverted indexing. The thought is to gap a highlight vector into pieces and list every piece with a sub hash value. Each sub hash value is connected with an inverted index list. The framework has been broadly tested using a large scale image database. We have assessed both retrieval performance and privacy-preserving performance for a specific content identification application. Two unique developments of robust hash algorithms are utilized. One depends on random projections; the other depends on the discrete wavelet transform. Both algorithm exhibit satisfactory performances in comparison with state-of-the-art retrieval performances. The outcomes demonstrate that the privacy upgrade somewhat enhances the retrieval performance. We consider the majority voting attack for evaluating the query category and identification. The test results demonstrate that this attack is a threat when there are close duplicities, yet the achievement rate diminishes with the quantity of discarded bits and the number of distinct items.

Keywords*-Multimedia database, image hashing, indexing,content-based retrieval, data privacy.*

*****

I. INTRODUCTION

In the Internet era, multimedia content is massively produced and distributed. In order to efficiently locate content in a large-scale database, content-based search techniques have been developed. They are used by content based information retrieval (CBIR) systems to complement conventional keyword-based techniques in applications such as near-duplicate detection, automatic annotation, recommendation, etc. In such a typical scenario, a user could provide a retrieval system with a set of criteria or examples as a query; the system returns relevant information from the database as an answer. Recently, with the emergence of new applications, an issue with content-based search has arisen sometimes the query or the database contains privacy-sensitive information. In a networked environment, the roles of the database owner, the database user, and the database service provider can be taken by different parties, who do not necessarily trust each other. 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 corresponding information.

The main challenge is that the search has to be performed without revealing the original query or the database. This motivates the need for privacy-preserving CBIR (PCBIR) systems. Privacy raised early attention in biometric systems, where the query and the database contain biometric identifiers. Biometric systems rarely keep data in the clear, fearing thefts of such highly valuable data. Similarly, a user is reluctant in sending his biometric template in the clear. Conventionally, biometric systems rely on cryptographic primitives to protect the database of templates. In the multimedia domain, privacy issues recently emerged in content recommendation. With recommendation systems, users are typically profiled. Profiles are sent to service providers, which send back personalized content. Users are today forced to trust the service providers for the use of their profiles. Although CBIR systems have not been widely deployed yet, similar threats exist. Recently, the one-way privacy model for CBIR was investigated. The one-way privacy setting assumes that only the user wants to keep his information secret because the database is public. Public databases against which users may wish to run private queries have become commonplace nowadays. Some of them already integrate similarity search mechanisms, such as Google Images or Google Goggles. It is likely that others will soon follow that path, turning Flickr, YouTube, Facebook into content-based searchable collections (in addition to already being tag searchable). Put in a larger picture, PCBIR is one of many aspects on privacy protection in the big data era where profiling becomes ubiquitous. For example, recent news claims that advertisers and Facebook can generate user profiles of political opinions and behaviors. Latest research discovers that websites are actually fingerprinting users on the Internet by their system (e.g. browser) configurations. There is already some initiatives in web search privacy. The trend shows that privacy protection will become an indispensable part of future content-based search systems.

II. PROBLEM DEFINITION

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 corresponding information. The main challenge is that the search has to be performed without revealing the original query or the database. This motivates the need for privacy-preserving CBIR (PCBIR) systems. 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.

III. OBJECTIVES

We propose a privacy protection framework for large-scale content-based information retrieval. It offers two layers of protection. First, robust hash values are used as queries to prevent revealing original content or features. Second,the client can choose to omit certain bits in a hash value to further increase the ambiguity for the server.

IV. SYSTEM ARCHITECTURE

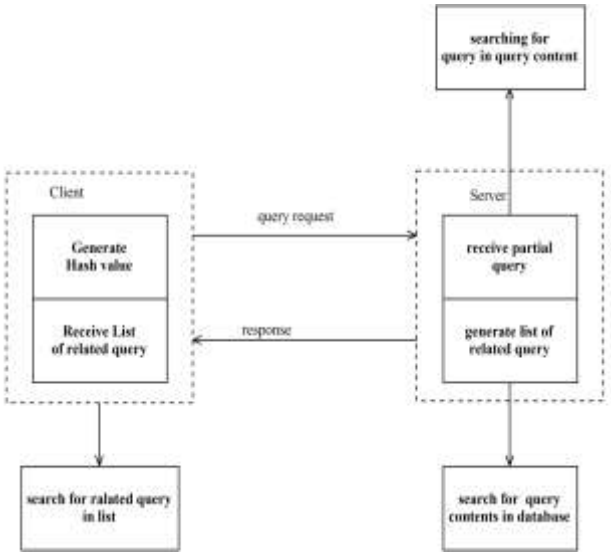


Fig: system architecture

In our system user register and login. After login user put query. User input query compare with database and mining will be perform. Data mining is the extraction of data. Optimization performs when mining process will be done. Optimization reduces our cost function *J* (W, b), adopt the gradient descent to update the parameters. Our model holds three sparsely hooked hidden layers and they are pre-trained sequentially. Final result will be Different links related to user entered keyword (Query).

V. DESIGN MODULES

A. Query Generation.
B. Database Indexing.
C. Database Search.

Module Description:

A. Query Generation:

Instead of encryption, we generate queries from original content by robust hashing. It is a framework that maps multimedia data to compact hash values. Ideally, a robust hash value is a short string of equally probable and independent bits. It can be used to persistently identify or authenticate the underlying content, just like a "fingerprint". The basic property of robust hashing is that similar content should result in similar hash values. Robust hashing typically involves feature extraction, orthogonal transformation, dimension reduction, and quantization.

More importantly, 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. By robust hashing we can accomplish two-folds that are

1) The compact size can facilitate fast search (in the Hamming space if binary)

2) Due to the one-way property, the privacy requirements can be achieved by using the hash value instead of the original content (or features) for the search and return of answers. 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.

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 capable of characterizing the underlying content. They first undergo an orthogonal transform and dimension reduction. Only significant features are preserved. The elements of a feature vector are divided into n groups. A robust hash value hi where, i = 0, 1, ... , n – 1 is computed from the i th group. We call it a sub-hash value. The above step creates a new coordinate system, with each coordinate represented by a sub hash value. Finally, a multimedia object in the database is indexed by the overall hash value H = h0||h1|| . . . . ||hn−1. 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.

1) Approximate Nearest Neighbor Search:

When the server receives a hash value, it checks the table for each sub hash value and optionally performs a nearest neighbor search within a Hamming sphere. For each binary sub-hash value, the multimedia object 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 neighborhood. The retrieved objects for all sub-hash values are put into a list. This list of candidates is sorted according to the hash distance from the query.

In general, we assume that similar multimedia objects should have similar hash values. Therefore, the nearest neighbors can be obtained from the sorted list.

2) Approximate Nearest Neighbor Search With Privacy:

When privacy protection is "turned on", the hash value of the query content must be generated by the client. A partial query is then formed by omitting some bits in one or more sub- hash values according to a privacy policy. In general, the more bits are missing, the more client privacy is preserved. The partial hash value is sent to the server along with the privacy policy, i.e., positions of the absent bits. The entire candidate IDs are sent back to the client, together with the corresponding hash values. The client eventually performs a search by comparing the hash values in the list with the original one.

VI. MATHEMATICAL MODEL

Let W be the whole system which consists:
W= {IP, PRO, OP}
Where,
A. IP is input given to system
IP= {C, S, Q, PQ}.
Where,
C is the set of number of clients in the system.
C= {c1, c2 …cn}.
S is the server.
Q is the query given by the client to server.
PQ is the partial query given by client to server.
B. PRO is procedure of system.

Server sends the whole database to the client hence client get privacy problem because client is not feasible for a limited bandwidth and violates the server privacy. Even if the client obtains the whole database from server for his query, it may not be able to store or process the database for his query. Then client removes some details from database for his query which creates some ambiguity for the server.

The client creates a partial query PQ, and sends it to the server S.

The server generates an extended query list based on the partial query.

The client performs a search within the received set of matching results using the original query.

C. OP is the Output of the system.
Output of the system will be: client will get only matching query content for his query instead of large scale content.

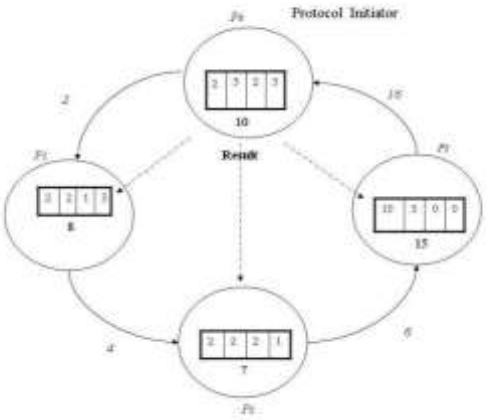
VII Algorithms Used

1. Robust Hash for Query Generation A robust hash value is a short string of independent bits. In robust hash similar contents have similar hash values. It is computationally difficult to affect the input from the output, because hashing is essentially a many-to-one mapping.

2. Piecewise Inverted Index Piecewise Inverted Index is used for database indexing. The Sub-hash values of queries are associated with an inverted index list. This index consists of IDs of multimedia objects corresponding to sub-hash value.

3. Approximate Nearest Neighbor Search: Whenever server get hash values from user then it first checks the tables for each hash values and execute nearest neighbor search when the server receives a hash value, it checks the table for each sub-hash value and optionally performs a nearest neighbor search within a Hamming sphere and when privacy protection is on then client generate hash values. The partial query discards some of bits in one or more sub-hash values according to the privacy policy.

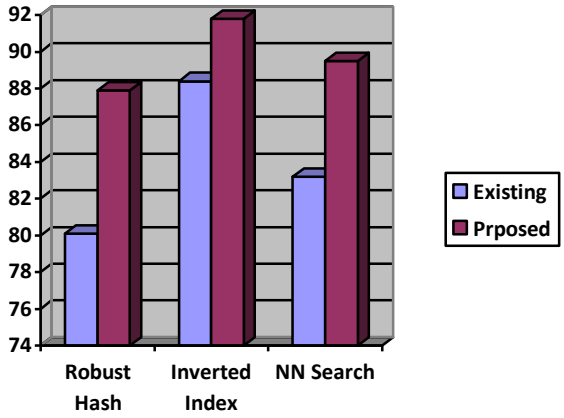
4. Secure K sum protocol: The information piece of every gathering is apportioned into a settled number of portions. Assume party P is chosen as convention initiator then this sending so as to gather will begin the convention the first section of its information piece. The stream of fractional aggregate will take after a unidirectional ring. The subsequent aggregate is reported by the convention initiator.



VIII. RESULT ANALYSIS:

TABLE I. Performance of CBIR

Algorithms	Existing System	Proposed System
Robust Hash	80.1%	87.9%
Piecewise Inverted Index	88.4%	91.8%
Approximate Nearest Neighbor search	83.2%	89.5%



IX. CONCLUSION:

In this project, a privacy preserving framework is provided 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. In this 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. 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.

ACKNOWLEDGMENT

It gives me a great pleasure and immense satisfaction to present this special topic on A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval, The success of this topic has throughout depended upon an exact blend of hard work and unending co-operation and guidance, extended to me.

REFERENCES:

[1] Li Weng, *Member, IEEE*, Laurent Amsaleg, April Morton, and Stéphane Marchand-Maillet, "A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval" at IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 10, NO. 1, JANUARY 2015.

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

[3] C.-Y. Hsu, C.-S. Lu, and S.-C. Pei, "Image feature extraction in encrypted domain with privacy-preserving SIFT," *IEEE Trans. Image Process.*, vol. 21, no. 11, pp. 4593–4607, Nov. 2012.

[4] W. Zhang, K. Gao, Y.-D. Zhang, and J.-T. Li, "Data-oriented locality sensitive hashing," in *Proc. ACM Int. Conf. Multimedia*, 2010, pp. 1131–1134.

[5] M. Diephuis, S. Voloshynovskiy, O. Koval, and F. Beekhof, "DCT sign based robust privacy preserving image copy detection for cloud-based systems," in *Proc. 10th Workshop Content-Based Multimedia Indexing (CBMI)*, Jun. 2012, pp. 1–6.

[6] 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.

[7] [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. 18P. 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. 349354.

[8] Z. Erkin, M. Franz, J. Guajardo, S. Katzenbeisser, I. Lagendijk, and T. Toft,Privacy-preserving face recognition, in Proc. 9th Int. Symp. Privacy Enhancing Technol. (PETS), 2009, pp. 235253.

[9] A.-R. Sadeghi, T. Schneider, and I. Wehrenberg, Efcient privacy- preserving face recognition, in Proc. 12th Int. Conf. Inf. Secur. Cryptol. (ICISC), 2009, pp. 229244.