International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Survey on Privacy Preservation in Content Based Information Retrieval

Bhagyashree V. Khapli[1], Manjushri A. Mahajan[2]

[1, 2]G. H. Raisoni College of Engineering, Savitribai Phule Pune University, Dhomkhel Road, Wagholi, Pune 412207, Maharashtra

Abstract: ***Content based information retrieval is latest versions of the Information retrieval system in which content is used for information retrieval. Content can be image, audio snippet, video clip and Content based information retrieval will give related respective information. There are many applications of the Content based information retrieval. First is Google goggle which is Image search engine, flip kart's android application also launched its image searching. You-tube applied Content based information retrieval system for content identification system. Before uploading any video it checks the similar video present in database or not. CBIR system comprises of multiple parties. User sends content i.e. image or video or audio snippet as query to the CBIR system. Other parties in the system may use these queries to model the profile of the user. One more point is that if content is sensitive and user don't wants to reveal this query to CBIR server then there is need of privacy preservation. This paper addresses the privacy preservation issue of the .Content based information retrieval. This paper describes the proposed approaches in the literature to preserve the privacy of Content based information retrieval system. In recent Content based information retrieval system one important problem us that both client and server sides are loaded with the same tasks. One approach to solve this problem will be outsource the common tasks to the third party is such way that privacy should also preserved. Different types of Approximate nearest neighbor techniques proposed in the recent years. Recent of Approximate nearest neighbor techniques can be used to improve the existing privacy preserving content based information retrieval systems.***

Keywords: Content Based Information Retrieval System, CBIR, PCBIR, Approximate nearest neighbor techniques

1. Introduction

1.1 What is Content Based Information Retrieval?

Google goggle and Flip kart"s Image search engines are the examples of the Content based Information Retrieval. In Image search engines, user gives one image as a query and gets similar images with the input query as an output. System checks the content similarity of query image and images in the database. Images with higher similarity than threshold are returned as an output. Image search engine is one of the applications in which Content Based Information Retrieval is used. System which uses the content of the object for Information Retrieval is the CBIR [6].

Mostly CBIR is applied on the Multi-media data. Main problem of the CBIR was the high dimensionality. Due to high dimensionality CBIR becomes highly computation bound and Time consuming system. To reduce the computation cost and Time requirement images are compared on some features such as color (Histogram of the image), Texture, Shape present in the image.

2. Need of Privacy Preservation in CBIR

Privacy can be defined as „No one should know or interfere with whatever I am doing". In CBIR user gives some Multimedia object (e.g. Image, Video, Song snippet etc.) as query and System gives user required results. In this process system needs to find similarity between user query and objects in the DB. At this time system gets the user query and query reveals the content and interest of the user and here is the privacy leak in the system. Generally CBIR is multi-party system, in which there are minimum 2 parties, User and Information retrieval system. Sensitive information may envelop in search query or database to which the request is made. Both the parties consider unknown and untrusted to each other. If none of them want to reveal such sensitive information then here comes the need of Privacy Preservation in CBIR.

3. Literature Survey

Many times annotation or names to the multimedia objects is not available. If we want to retrieve information from such data then there is need of Content based Information Retrieval. M. S. Lew et. al. [6] described Content Based Information Retrieval (CBIR) for multimedia data, applications of the CBIR and methods proposed in the literature. Content Identification System, Images search Engine are the Applications of the CBIR. Privacy Preservation in CBIR (PCBIR) can be achieved by two approaches. First is Signal Processing in Encrypted Domain (SPEED) which used Encryption technique for PCBIR [9] [10]. Multimedia content processing and content encryption are independent and sequential operations. Sometimes it is desirable to carry out processing on encrypted signals directly. Cryptographic research and signal processing has some challenges. In [9] cryptographic primitives used in existing solutions to processing of encrypted signals are discussed. Two domains i.e. analysis and retrieval of multimedia content and multimedia content protection in which secure signal processing taken as a challenge is described. State-of-the-art algorithms are described in each domain. Heavy encryption technique like Homomorphic encryption so that encrypted data can be used for information retrieval.

This Approach provides good privacy preservation [11] but computation cost is high therefore cannot be used with low

configuration devices. Second approach is Search with Reduced Reference (SRR) in which small representative (secure index) of the actual content is generated and used for information retrieval. As actual content is not used for the information retrieval privacy is preserved automatically. PCBIR contains the two parts: 1) Find nearest neighbor and 2) Oblivious Retrieval.

In Oblivious retrieval constrain is that, server should not know which items were accessed by user. Paper [12] describes main three types of privacy preserving nearest neighbor solutions: 1) Computational methods which corresponds to the SPEED approaches. 2) Information-theoretic methods which requires third party for distance calculation. 3) Randomized embedding which corresponds to SRR approaches.

For computer vision and machine learning problems many algorithms are available. However, Cost overhead part of many algorithms consists of finding nearest neighbor matches in high dimensional space to present training data. The paper [1] worked two algorithms i.e. randomized k-d forest and the priority search k-means tree which is newly proposed in [1] addresses the problems of fast nearest neighbor search in high dimensional space and fast approximate matching of binary features.

Searching large databases using descriptors like local binary patterns proved inefficient because cost of the linear search and inappropriate performance of existing indexing methods. Paper [2] proposes Discrete Cosine Transform hashing for creating index structures for face descriptors. Hash suppression used to improve accuracy and retrieval efficiency. Hashing plays a key role. Created Index is queried to find the desired images most similar to the query image. Object retrieval efficiency and accuracy achieved using common hash suppression.

Nearest Neighbor searching (NN) has high importance for various purposes such as feature matching, objet recognition, image retrieval etc. in high dimensional similarity searching; exact NN search is computationally not sufficient. Approximate nearest neighbor search algorithm solves this problem by providing high probability to find nearest neighbor.

Paper [3] considers the projection errors in quantization process which addresses the problem of large scale ANN search in high dimensional space by proposing method of projected residual vector quantization for ANN. Approach in [4] is based on the dictionary learning for sparse coding, proposes new Nearest Neighbor retrieval framework known as Robust Sparse Hashing (RSH).

When data is noisy, direct application of sparse code to NN retrieval has technical difficulties. Dictionary learning formulation problem was introduced considering uncertainty of ellipsoidal data. Sparse coding framework and Novel robust dictionary learning called RSH addresses the problem by learning dictionaries on the robust counterpart of the perturbed data points. Algorithm is scalable and effective to solve the resulting robust objective.

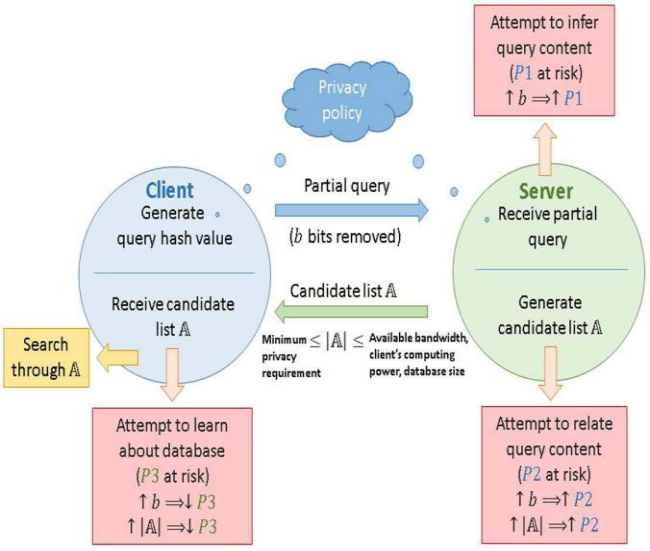

Figure 1 Existing approach in PCBIR [5]

In paper [5] an approach is proposed for PCBIR with facility to adjust the level of privacy accordingly. As in fig.1 user queries the database for similar content retrieval. For e.g. user provides an image to database for searching. Prior to this for Privacy Preservation Secure index of the database is generated using robust hash algorithm. When user querying the object, secure index is generated by removing some bits from the secure index. Then this query and position of removed bits is sent to server. Server finds the n nearest neighbor of the query. In return, Hash i.e. Secure Index of the n nearest neighbor is sent to the user. On user side required results are searched using original query and hash list received from server. In existing approach the problem is that Searching and indexing on both sides i.e. client and server leads to computation overhead, load on both the sides.

4. Application of the CBIR

Content Identification System this is used by the You tube to avoid duplicate videos, Content Based recommendation also used by many Music and Video websites. As mentioned above Image search engines are common now like Google goggle. E- Commerce websites such as flip kart also considered the power of the CBIR and launched the Image search engine for the products in the Database in Jul 2015 [8].

Using Google image search engine when user wants to search any image, the image is provided as an input to the content retrieval system. System process the query which contains the image and related information and it extract the patterns to match with content present in the search engine database. Most similar matched content are returned to user as a result.

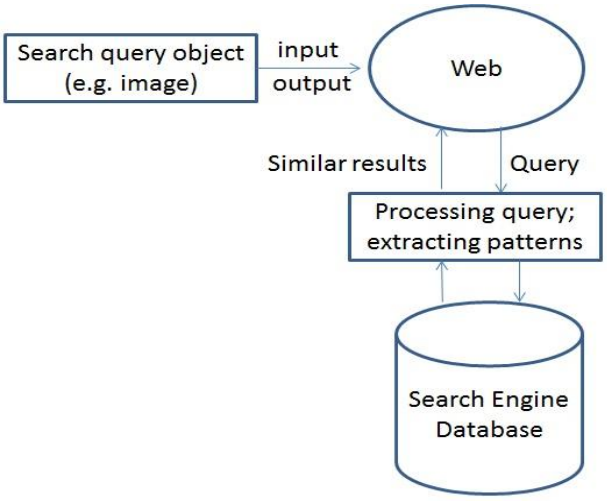

Figure 2: Example of Image Search Engine using CBIR

5. Comparison between different Hashing Techniques

Table 1 gives comparison between different hashing techniques such that viewer can easily classify the difference between those techniques. Comparison truly based on some parameters like Accuracy, Speedup, Precision etc.

Table 1: Comparison between Hashing Techniques

Algorithm Used	*Compared with*	*Datasets Used*	*Accuracy/Results/Precision*
Priority Search k-means tree [1].	Approximate Nearest Neighbor(ANN) and Locality Sensitive Hashing(LSH)	100K SIFT	Precision and Speed up parameters are inversely proportional. PS k-means outperforms ANN and LSH. Gives Up to 99% precision.
Discrete Cosine Transform Hashing [2]	LSH,E2LSH,K-means codebooks, KLSH, k means codebooks, KSH	(FERET), FEI , BioID database, and Labeled Faces in the Wild (LFW).	On the basis of retrieval accuracy DCT hashing is superior than compared algorithms. 88% retrieval accuracy.
Projected Residual Vector Quantization [3]	Residual Vector Quantization (RVQ), Product Quantization (PQ)	GIST dataset, VLAD dataset	Up to 0.9 recall , 30 ms searching time per vector , Outperforms RVQ and PQ
Robust Sparse Hashing [4]	KLSH, KDT,DL,PQ	SIFT dataset, MNIST Dataset	92% accuracy on MNIST dataset , 100% accuracy for wall type data in SIFT dataset

6. Conclusion

Privacy preservation in content based retrieval system is a must because of privacy leak issues. We have studied different approaches to address the issues in CBIR. Various Secure indexing techniques and different algorithms used to solve the problems. In recent work there are two approaches i.e. Projected Residual Vector quantization for ANN search and Efficient Nearest Neighbor via Robust Sparse Hashing proves to be more efficient for accuracy and performance in PCBIR. Existing system has the client and the server that generate the secure index i.e. hash generation on both the side. We can introduce the third party to do this job i.e. there is a scope to alleviate the load of generating secure index and search operation to the third party. As the load is reduced on both the sides, time and computation efficiency improves.

References

[1] Marius Muja, Member, IEEE and David G. Lowe, Member, IEEE, "Scalable Nearest Neighbor Algorithms for High Dimensional Data", IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 36, No. 11, November 2014

[2] Mehran Kafai, Member, IEEE, Kave Eshghi, Bir Bhanu, Fellow, IEEE, "Discrete Cosine Transform Locality-Sensitive Hashes for Face Retrieval, IEEE Transactions on multimedia vol. 16, No. 4, June 2014.

[3] Benchang Wei, Tao Guan, and Junqing Yu Huazhong University of Science & Technology, "Projected residual vector quantization for approximate nearest neighbor (ANN) search", Published by the IEEE Computer Society, 2014.

[4] Anoop Cherian, Suvrit Sra, Vassilios Morellas, Nikolaos Papanikolopoulos, "Efficient Nearest Neighbors via Robust Sparse Hashing", IEEE, 2014.

[5] Li Weng, Member, IEEE, Laurent Amsaleg, April Morton, and Stéphane Marchand- Maillet, "A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval ", IEEE Transactions On Information Forensics And Security, Vol. 10, No. 1, January 2015.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," ACM Trans. Multimedia Comput., Commun., Appl., vol. 2, no. 1, pp. 1–19, Feb. 2006.

[7] http://en.wikipedia.org/wiki/Contentbased_image retrieval#Content_comparison_using_image_distance_measure

[8] http://en.wikipedia.org/wiki/Contentbased_image___ retrieval#Content_comparison_using_image_distance_measure

[9] Z. Erkin et al., "Protection and retrieval of encrypted multimedia content: When cryptography meets signal processing," EURASIP J. Inf. Secur., vol. 2007, p. 20, Dec. 2007.

[10] 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.

[11] Peter, A. et al., "Privacy-Preserving Architecture for Forensic Image Recognition", IEEE Conference 2012.

[12] S. Rane and P. T. Boufounos, "Privacy-preserving nearest neighbor methods: Comparing signals without revealing them," IEEE Signal Process. Mag., vol. 30, no. 2, pp. 18–28, Mar. 2013.

Author Profile

Bhagyashree Khapli was born in Akola, Maharashtra in November 1987. She received the B.E. degree in Computer Science and Engineering from Babasaheb Naik College of Engineering, Pusad, SGBA Univ. in 2011 and pursuing M.E. degree in Computer Engineering from G. H. Raisoni College of Engineering and Management, Pune, in 2014-2015. Her research interests are in data mining and Computer Networks. She has worked as a Lecturer for 2 years. She has published 1 journal paper.

Manjushri A. Mahajan was born in Ahmadnagar, Maharashtra. She received the B.E. degree in Computer Engineering from College of Engineering, Ambajogai in 2004 and M.E. degree in Computer Engineering from Sinhgad College of Engineering, Pune, in 2013. Her research interests are in data mining and Computer networks. She has worked as a Lecturer for 11years. She has published 2 journal papers.