RESEARCH ARTICLE OPEN ACCESS

A Survey On Content Based Information Retrieval Using Privacy Preserving Framework

Prof. Priti Mithari, Komal Satam, Sayali Satpute, Pranita Umarjikar
Department of Computer Science
Pune University, DYPIET Ambi
India

ABSTRACT
Content based information system in which content is used for information retrieval. User sends query to Content Based Information Retrieval System (CBIR). If content is sensitive and user don't want to disclose this query to Content Based Information Retrieval system then there is need of privacy preservation. In previous system, that is content-based multimedia information retrieval contains browsing and also search paradigms, learning, semantic queries, new features and different media types, high performance indexing, and evaluation methods. The major challenge in the MIR system is semantic search with significance on the discovery of concepts in media with difficult backgrounds. The other system gives efficient algorithms for the exact nearest neighbour problem. Robust Sparse Hashing (RSH) is Nearest Neighbour (NN) retrieval. This approach is boosted by the victory of dictionary understanding for sparse coding. Vector Quantization (VQ) is magnificent quantization method from signal processing that permits the modeling of feasibility density function by the allocation of prototype vectors. Fast Library Appropriate Nearest Neighbour is a library for accomplishing fast appropriate Nearest Neighbour (NN) search in high dimensional spaces. A Discrete Cosine Transform (DCT) conveys a finite sequence of data points in terms of a sum of cosine function moving at different frequency. In current content based information retrieval system, one problem is that both client and server sides are loaded with the same tasks. To solve this problem, the common task is assigned to the third party in such a way that privacy should also be preserved. This framework is tested using large image database. It provides two layers of protection. First, robust hash value are used as queries to secure original content. Second, in future to decrease the ambiguity of the server the client can choose to omit some bits in hash values, two different robust hash algorithm are used. One is random projection and other is discrete wavelet transform. This improves the performance of retrieval.
Keywords: — Information Retrieval, Multimedia Database ,Indexing, Data Privacy.

I. INTRODUCTION

Nowadays, multimedia content is hugely produced and occured throughout an area. Content-based search methods have been developed to usefully locate content in a large scale database. They are used by Content Based Information Retrieval (CBIR) [1] to satisfy keyword based methods in applications like recommendation, near-duplicate detection etc. A user provides a set of queries to the system and in return it provides proper information from database as in response. Issues like query or database involves privacy-sensitive information. In environment, the roles of database owner, user and service provider can be worked with different parties, which do not believe each other. A privacy problem arises when an untrusted party needs to acess the private credential of another party. So necessary action should be taken to secure the information. The challenge is that the search has to be performed without opening the original query. This induces the need for PCBIR system.

Information retrieval handles the representation, storage and retrieval of unstructured information. Classical information retrieval deals mainly with text.The progress of multimedia databases and of the web have given new interest to IR. A Multimedia database is a collection of related information which contains one or more primary media data
types like text, images, animation sequences, audio and vid eo. Data privacy is the relation between collection and dissemination of information, technology, the public expectation of privacy, and the legal and political issues. Privacy concerns exist where personally identifiable data or sensitive information is collected, stored, used, and finally destroyed in digital form. Improper disclosure control can be the root cause for privacy issues. Data privacy problems can arise in response to information from a vast range of sources.

In previous system Multimedia information retrieval (MIR) system tells about the search for knowledge in all its forms. The fundamental issue has been how to improve multimedia retrieval using content-based techniques. Content-based methods are vital when text comments are incomplete [1]. The content-based technique can potentially improve retrieval precision even when text comments are present by providing additional insight into the media collections [2].

Privacy Content Based Information Retrieval (PCBIR) framework works for both public and private database. PCBIR provides design for large scale database. It offers multiple levels of privacy protection. It is easy to configure and generalize. Granularity of privacy protection is not considered in PCBIR solution. According to the application, cost of privacy protection is adjustable. This framework provide flexible trade-offs and can be used in a heterogeneous network. The proposed framework is an SRR approach. The key elements are robust hashing piece-wise inverted indexing. Any robust hash algorithm can be used as a module and any feature an be converted to hash value. The level of privacy protection is controlled by privacy policy. These elements work together according to a new Privacy Content Based Information Retrieval (PCBIR) protocol. Two different robust has algorithm are used to show the compatibility of the framework.

The rest of work is organized as follows: Section II is brief literature review. In Section III we conclude the work.

II. RELATED WORK

In recent survey by Muja and Lowe [3], describes that to search nearest neighbor matches to high dimensional data, it has two useful algorithms: the randomized k-d forest and the priority search k-means tree. To scale large data sets that would not fit single machine memory, they proposed a distributed nearest neighbour matching framework. All this research has been reveal as an open source library such as Fast Library For Approximate Nearest Neighbours (FLANN). It has collection of algorithm and has sound to work best for Nearest Neighbour search. A system for automatically picking. It gives accuracy up to 99 percent as a result. It has advantages like extensible for large High Dimensional data and high accuracy.

Mehran Kafai [4] describes Discrete Cosine Transform(DCT) hashing method for creating index structures for face descriptors. A hash index is formed, and further queried to search the images more similar to the query image. DCT are great significance and engineering from compression of audio and image. The use of cosine than sine function is difficult for compression because it turns out that small no of cosine function are needed. To appropriate a typical signal, for differential equation the cosine conveys a specific size of boundary condition. Discrete Cosine Transform (DCT) hashing algorithm has best retrieval accuracy and more efficient compared to other popular state of- the-art hash algorithms. It provides 88 percent retrieval precision as a result. It has advantages like fast and computationally cheap and outperforms than LSH, E2LSH and KLSH for nearest neighbour recall. The disadvantage is cost of computing the hash.

Anoop Cherian [5] proposes a new Nearest Neighbor (NN) framework: Robust Sparse Hashing (RSH). The key innovation of RSH is to use learned sparse code as hash code for speeding up sparse coding undergoes from a big drawback:- When data are noisy, for a query point, an appropriate match of hash code hardly happens, breaking the Nearest Neighbour (NN) retrieval. The algorithm is applied to NN retrieval for Scale Invariant Feature Transform (SIFT). For precise and fast NN retrieval, the ideas is to sparse code the data by using learned dictionary, and then produce hash codes out of these sparse codes. Results tell that Robust Sparse Hashing provides different accuracy with different dataset that is, 92 percent - MNIST dataset, 100 percent - SIFT dataset. It has advantages like fast Hash creation and best precise performance on SIFT and MNIST.

Benchang Wei [6] describes Projected Residual Vector Quantization (PRVQ) algorithm. The effectiveness of PRVQ algorithm is consolidated on two kinds of high dimensional vectors: GIST and vector of locally aggregated descriptors (VLAD). Vector Quantization (VQ) was firstly used for information compression. It works by partitioning a huge set of points information bunch having accurately the same number of points nearest to them. Every bunch is presented by its centroid point as in k-means and some another clustering algorithm. Projected Residual Vector Quantization (VQ) outperforms existing methods, for example product quantization (PQ), transform coding (TC), and Residual Vector Quantization (RVQ). It provides 30 ms per vector as a result of search time/ speed up parameter. It has advantages like high Precision and disadvantage is no cooperative framework.

III. CONCLUSION

In this paper, a wide survey of different approaches for privacy preserving data mining and analyses the algorithm available for each method. For large collection of data, it is important to maintain the privacy of sensitive data. There are various technologies such as Multimedia Information Retrieval (MIR), Fast Library Appropriate Nearest Neighbour (FLANN), Discrete Cosine Transform (DCT), Robust Sparse Hashing (RSH), Projected Residual Vector Quantization (PRVQ). Every method has its own advantages and disadvantages. It will overcome the disadvantage in existing system is problem of performance and computation overhead by using hash generation process to the third party server with privacy preservation. The algorithm which is used a robust sparse hashing algorithm for robust hash generation. It makes content based retrieval process accurate without revealing any information of interest.

IV. ACKNOWLEDGMENT

Authors would like to thank our Prof. Priti Mithari, Prof. Mangesh Manke, Prof. Sharmila Chopade for giving her valuable suggestion and providing several useful information.

REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Anoop Cherian, Suvrit Sra, Vassilios Morellas, Nikolaos Papanikolopoulos, "Efficient Nearest Neighbors via Robust Sparse Hashing", IEEE Transactions on vol. 23 , Issue: 8, 2014

[6] 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.

[7] 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.