SECURE DATA ACCESS ON DISTRIBUTED DATABASE USING SKYLINE QUERIES

Sumayya C K[2], M. Natarajan[2]

[1]*M.Phil Research Scholar, Department of Computer Science, Thanthai Hans Roever College(Autonomous), Perambalur, India*
[2]*Assistant Professor, Department of Computer Science, Thanthai Hans Roever College(Autonomous), Perambalur, India*

Abstract - *The outsourced data in the distributed database server are quiet unsecure when compared with the current techniques and security measures. So that we propose a methodology based on skyline queries, along with that we include the user data as a illusion data and the data and the database are in an encrypted format, so that the distributed server has no knowledge about the data that has been saved by the user to the data owner. Data owner is the authorized person who transmits the data to the distributed database. The distributed server has no knowledge about the data that has been saved by the user. So that we use K nearest neighbour and CNN algorithm to obtain the better result for the proposed scheme. In case of intruder breach the intruder attains the illusion data and the intruder alert intimation will be given to the concern data owner and user.*

Key Words: Security, Information, Outsourcing, Encryption, Protection etc…

1. INTRODUCTION

As a developing computing model, cloud computing attracts increasing attention from both research and industry communities. Outsourcing data and computation to cloud server provides a cost efficient way to help large scale data storage and query processing. But, due to safety and privacy concerns, delicate data need to be endangered from the cloud server as well as other illegal users.

Modern progression in communication technologies has ended with the extensively and rise in use of the cloud resources by different users. The numerous resources used in the cloud comprise software, servers, network, storage etc. payable on demand according to their usage . The major drawback of cloud computing is the vulnerability of user data to malicious attack or intruders. The most new approach to stop the security encounters in the cloud used by researchers is cryptography and steganography. The both mentioned techniques are used to protect data but in different fashions. Cryptography distresses itself with the hiding of the content of a secret message whereas steganography deals with the concealment or hiding of a secret message from an unauthorized person. Since, users pay for their services according to the resource consumed, the need to evaluate the performance of various security techniques used in the cloud against the resources they consumed becomes imperative. The major aim of this work is to do performance study of digital text and image steganography. RSA cryptosystem is engaged for secret information privacy and verification. Steganography is a way of hiding unrevealed data in a cover object while communication happen between sender and receiver. The data types used for the inspection include text, image, audio and video whereas the system resources considered are encryption and decryption time, memory consumption, processing power usage and bandwidth utilization.

2. EXISTING SYSTEM

In this paper, we focus on the problem of secure skyline queries on encrypted data, another type of similarity search important for multi-criteria decision making. The skyline or Pareto of a complex dataset given a query point contains the data points that are not dominated by other points. A data point controls another if it is closer to the query point in minimum of one dimension and at least as close to the query point in every other dimension. The skyline query is particularly useful for selecting similar (or best) records when a single aggregated distance metric with all dimensions is hard to define. The hypothesis of kNN queries is that the relative weights of the features are known in advance, so that a single similarity metric can be computed between a pair of records accumulating the similarity between all attribute pairs. But, this assumption does not always hold in practical applications. In many situations, it is necessary to retrieve similar records considering all possible relative weights of the attributes (e.g., considering only one attribute, or an arbitrary combination of attributes), which is importantly the skyline or the "pareto-similar" records.

Our goal is for the cloud server to compute the skyline query given q on the encrypted data without revealing the data, the query q, the final result as well as any intermediate result to the cloud. We note that skyline computation (with query point at the origin) is a special item of skyline queries. In addition we propose a new technique that in case any intruder tries to access the data that has been stored by the client, the intruder receives an illusion effect of data through shoulder surfing algorithm i.e.: the data that has been stored by the client is in an freaky format, so that the intruder gets no knowledge about the data that has been stored in the cloud server and at the same time the client receives the intrusion alert for his desired data, from which the data owner or the client can improve their privacy.

2.1 Survey

A. FINDING K-DOMINANT SKYLINE CUBE BASED ON SHARING-STRATEGY

K-dominant skyline query has been proposed as an important operator for multi-criteria decision making, data mining and so on, this technology can decrease the large result sets of skyline query in high dimensional space. In this paper, a new idea was primarily proposed: k-dominant Skyline cube, which contains all the k-dominant skylines. Although existing algorithms can compute every k-dominant skyline, they lead to much repeat work because of no sharing result. We built two computation sharing strategies-ASCEND sharing strategy and DESCEND sharing strategy. Based on these two sharing strategies, two novel algorithms-BUA (Bottom-Up Algorithm) and UBA (Up-Bottom Algorithm) are planned to compute k-dominant skyline cube. Furthermore, complete hypothetical analyses and extensive experiments demonstrate that our algorithms are both efficient and effective.

B. SECURE OUTSOURCED SKYLINE QUERY PROCESSING VIA UNTRUSTED CLOUD SERVICE PROVIDERS

Recent years have witnessed a growing number of location- based service providers (LBSPs) outsourcing their points of interest (POI) datasets to third-party cloud service providers (CSPs), which in turn answer various data queries from mobile users on their behalf. A key challenge in such systems is that the CSPs cannot be completely trusted, which may return fake query results for various bad motives, e.g., in favor of POIs willing to pay. As an important type of queries, location-based skyline queries (LBSQs) ask for the POIs that are not spatially dominated by any other POI with respect to some query position. In this paper, we offer three novel structures that permit effective confirmation of any LBSQ result repaid by an untrusted CSP by embedding and discovering a novel neighboring relationship among POIs. The effectiveness and efficiency of our schemes are thoroughly analyzed and evaluated.

C. FAST REVERSE SKYLINE PROCESSING WITHOUT PRE-COMPUTATION

Reverse skyline questions, regain a set of objects whose stimulating horizon comprises a given query point, are useful and valued for many applications such as business location and environmental monitoring applications. Though there are numerous methods for handling reverse skyline queries, they are based on pre- processing. Since they waste time and space to pre-compute necessary data and to manage the pre-computed data, they are not feasible for some applications. In this paper, we propose a robust algorithm to fast and efficiently compute reverse skyline queries without pre - computation. Since the proposed algorithm is based on a branch-and-bound approach, it can access correct nodes by minimizing excessive traverses. It can also decrease candidates by using two snipping methods introduced throughout this paper, and it is efficient over frequently changing datasets since it does not pr e-compute and maintain any data. To verify a performance of our algorithm, extensive experiences are conducted. The experimental result shows that the proposed algorithm is better to its competitors.

D. SECURE K NEAREST NEIGHBORS QUERY FOR HIGH- DIMENSIONAL VECTORS IN OUTSOURCED ENVIRONMENTS

Due to the volatile rise of data in both the aspects of dimensionality and volume, performing k nearest neighbors search over cloud environments has been gradually receiving more attention among researchers in the field of database cloud computing. But, the main experiment for swapping k nearest neighbors search from the local server (i.e., traditional way) to the third - party cloud is, that the database which always comprises series of delicate information has to be kept secret against the cloud. In this work, we propose a pair of solutions towards Secure k Nearest Neighbors(SkNN) query in outsourced environments. By skillfully utilizing coarse quantization and the cryptography techniques Advanced Encryption Standard(AES) and Paillier homomorphic encryption, we construct a secure Inverted File(IVF) and compute encrypted approximate distances directly to search for high-dimensional facts in the third- party cloud provider, and finally find the better tradeoff between the search quality and security. Experimental study over real datasets and practical environments validate our solutions' feasibility, completeness, and practicality. Compared to the state -of-the-art, the proposed solutions resolve the SkNN of high-dimensional data novelly, have very limited response time and provide high privacy protection on the side of both the User and the cloud provider.

E. A TWO-PHASED REFINEMENT ALGORITHM TO PROCESS REVERSE SKYLINES WITHOUT PRE-PROCESSING

Converse skyline questions are hard to process on account of the enormous amount of calculations for testing candidates because current algorithms for reverse skylines are generally based on pre-processing. Though pre-processing decreases the number of calculations on processing queries, it requires re-computations of pre-processed result every time data change. To overcome this restriction, we suggest an efficient algorithm to reduce the number of computation in processing reverse skyline queries with a two-phased refinement step. Before cleansing the final update from candidates, the strategic algorithm has an extra refinement stage for reducing the number of candidates, so that it can handle reverse skyline queries more effectively without any pre-processing. Since not based on pre-processing, our algorithm is more apt for frequently updated data. Trial results show that the performance of the proposed algorithm is better than those of the existing pre-processing-based ones.

3. PROPOSED SYSTEM

In the proposed system along with the encrypted skyline queries we implement illusion data, intruder breach and the encrypted database. The user add or upload his data to the distributed database along with the skyline queries and the additional features added in the process are the data are added in distributed server in an illusion format, so that we can prevent the data from the breach of intruders and along with that the database has been encrypted so that there will be quiet better security to the user data. Along with that intruder alert message will be given to the user from the data owner if any abnormal activities occur in the distributed server.

Advantages

Enormously increases the authentication procedures of entire data that has been stored in the distributed servers.

User has a great influence on his data or the information that has been outsourced to the distributed or te cloud server.

Along with the data stored in the owner side the users who stores their data has an enormous features to provide perfect authentication to their information that has been decentralized.

Reduces the intruder breach up to the core which enhances the security features.

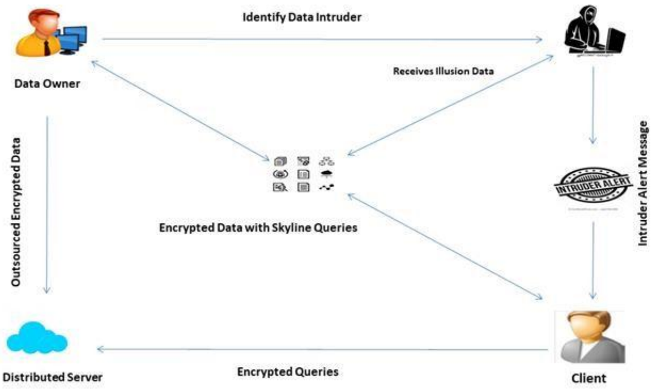

Fig -2: Architecture Diagram

development of drought monitoring and prediction tools. To further document and understand seasonal differences. This work presents a Soil Moisture Forecasting Ensemble Model (SMFEM) by joining the features of various machine learning approaches.

KNN algorithm

KNN algorithm is one of the uncertain classification algorithm and it is one of the widely used learning algorithms. ... KNN is a non-parametric, lazy learning algorithm. Its purpose is to use a database in which the data points are divided into several classes to predict the classification of a new sample point. K nearest neighbours is a simple algorithm that carries all available cases and categorizes new cases based on a similarity measure (e.g., distance functions). KNN has been used in numerical estimation and pattern recognition already in the beginning of 1970's as a non-parametric technique.

Overview

•Understand k nearest neighbour (KNN) – one of the most popular machine learning algorithms

•Learn the working of kNN in python

•Choose the correct value of k in simple terms

Introduction

In the four years of my data science career, I have made more than 80% classification models and just 15-20% regression models. These ratios can be more or less widespread throughout the industry. The reason behind this preference towards classification models is that most analytical problems involve making a decision. For instance, will a customer attrite or not, should we target customer X for digital campaigns, whether customer has a high potential or not etc. These analysis are more insightful and directly linked to an implementation roadmap. In this article, we will talk about another widely used machine learning classification technique called K-nearest neighbours (KNN) . Our focus will be mainly on how does the algorithm work and how does the input parameter affect the output/prediction.

When do we use KNN algorithm?

KNN can be used for both classification and reversion predictive problems. But, it is more broadly used in classification problems in the industry. To calculate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2. Calculation time

3. Predictive Power

Let us take a few instances to place KNN in the scale :

KNN algorithm fairs across all parameters of considerations. It is commonly used for its easy of interpretation and low calculation time.

Algorithm

An instance is categorized by a major vote of its neighbours, with the case being assigned to the class most common amongst its K nearest neighbours measured by a distance function. If K = 1, then the case is simply assigned to the class of its nearest neighbour.

It should also be noted that all three distance methods are only valid for continuous variables. In the case of definite variables the Hamming distance must be used. It also propose the problem of standardization of the numerical variables between 0 and 1 when there is a mixture of numerical and categorical variables in the dataset.

Choosing the optimal value for K is best done by first examining the data. In general, a large K value is more specific as it reduces the overall noise but there is no guarantee. Cross-validation is additional way to retroactively determine a good K value by using an independent dataset to validate the K value. Factually, the ideal K for most datasets has been between 3-10. That produces much better results than 1NN.

Example:

Consider the following data concerning credit default. Phase and Credit are two numerical variables (predictors) and Default is the target.

We can now use the training set to classify an unknown case (Phase=48 and Credit=$142,000) using Euclidean distance. If K=1 then the nearest neighbour is the last case in the training set with Default=Y.

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

With K=3, there are two Default=Y and one Default=N out of three closest neighbours. The forecast for the unknown case is again Default=Y.

Standardized Distance

One main difficulty in influencing distance measures straightly from the exercise set is in the situation where variables have unalike measurement scales or there is a mixture of numerical and categorical variables. For instance, if one variable is built on annual income in dollars, and the other is built on age in years then income will have a much higher effect on the distance calculated. One answer is to regulate the training set as shown below.

Using the even distance on the same training set, the unknown case returned a different neighbour which is not a good sign of robustness.

In design acceptance, the k-nearest neighbours algorithm (k-NN) is a statistic method used for classification and regression.[1] In both cases, the input consists of the kclosest training examples in the feature space. The output is based on whether k-NN is used for classification or reversion:

- In k-NN classification, the output is a class association. An object is classified by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply allocated to the class of that single nearest neighbour.

- In k-NN reversion, the output is the property value for the object. This value is the average of the values of k nearest neighbours.

k-NN is a kind of instance-based learning, or lethargic learning, where the function is only approximated locally and all computation is deferred until classification.

Both for classification and reversion, a beneficial technique can be to assign weights to the contributions of the neighbours, so that the nearer neighbours contribute more to the average than the more distant ones. For instance, a mutual weighting scheme contains in giving each neighbour a weight of 1/d, where d is the distance to the neighbour. The neighbours are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be believed of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data.

Parameter Selection

The finest optimal of k depends upon the data; generally, higher values of k decreases effect of the noise on the classification,[5] but make boundaries between classes less distinct. A good k can be chosen by various heuristic techniques (see hyperparameter optimization). The special case where the class is predicted to be the class of the nearby training sample (i.e. when k = 1) is called the nearest neighbor algorithm. The exactness of the k-NN algorithm can be severely degraded by the presence of noisy or irrelevant features, or if the feature scales are not reliable with their importance. Much research effort has been put into choosing or ascending features to improve classification. A chiefly popular[citation needed] method is the use of evolutionary algorithms to optimize feature scaling. Another prevalent method is to scale features by the shared information of the training data with the training classes.[citation needed] In binary (two class) classification problems, it is helpful to select k to be an odd number as this avoids tied votes. One popular way of choosing the empirically optimal k in this setting is via bootstrap method. The ability of two CNN models to classify soil echoes with different VWC is similar. However, the execution time is four times that of DNNR, and the different SNRs do notice influence the running time.

As of the previous results the DNNR specifies the perfect values by deep learning techniques and provides the exact scenario of each and every level of the data that has been inbuilt and uploaded by the users involved in the system. Due to deep learning technique the entire data and datasets are completely analyzed and classified for extraction to get an enhanced result that has been provided by the DNNR.

4. CONCLUSIONS

In this paper, we proposed a fully secure skyline protocol on encrypted data using two non-colluding cloud servers under the semi-honest model. It guarantees semantic safety in that the cloud servers knows nothing about the data including indirect data patterns, query, as well as the query result. In addition, the client and data owner do not need to participate in the computation.

We also offered a secure dominance protocol which can be used by skyline queries as well as other queries. Furthermore, we demonstrated two optimizations, data partitioning and lazy merging, to further reduce the computation load. Finally, we presented our implementation of the protocol and demonstrated the feasibility and efficiency of the solution. Along with this we introduce more new techniques like intruder breach, illusion data occurrences and the encrypted data as well as the encrypted data and database. So that the data that has been saved in the server are with quiet better privacy and security.

REFERENCES

[1] F. Baldimtsi and O. Ohrimenko. Sorting and searching behind the curtain. In FC 2015, pages 127–146, 2015.

[2] A. Beimel. Secret-sharing schemes: a survey. In International Conference on Coding and Cryptology, pages 11–46. Springer, 2011.

[3] J. L. Bentley. Multidimensional divide-and-conquer. Commun. ACM, 23(4):214–229, 1980.

[4] J. L. Bentley, H. T. Kung, M. Schkolnick, and C. D. Thompson. On the average number of maxima in a set of vectors and applications. J. ACM, 25(4):536–543, 1978.

[5] S. B¨orzs¨onyi, D. Kossmann, and K. Stocker. The skyline operator. In ICDE 2001.

[6] S. Bothe, A. Cuzzocrea, P. Karras, and A. Vlachou. Skyline query processing over encrypted data: An attribute-order- preserving-free approach. In PSBD@CIKM, pages 37–43, 2014.

[7] S. Bothe, P. Karras, and A. Vlachou. eskyline: Processing skyline queries over encrypted data. PVLDB, 6(12):1338–1341, 2013.