A New Scheme for Range Queries over Encrypted Data

Shanyue Bu
Huaiyin Institute of Technology, School of Computer Engineering, Huaian, China
Email: bushanyue@126.com

Yue Zhang and Kun Yu
Huaiyin Institute of Technology, School of Computer Engineering, Huaian, China

Abstract—Cloud servers could provide secure services to data management for encrypted sensitive data, however, the difficulties of querying these data by data owners increase. To solve the problem, this paper proposes a new scheme for range queries over encrypted data. In particular, the indices and interval trapdoors of sensitive data are first created by using circle mapping. Then, these indices and interval trapdoors are encrypted. During range query processing, cloud servers can efficiently retrieve and return query results after completing an assertion. Moreover, the correctness, security and computation complexity are presented and analyzed in details. Compared with the existing approaches, this method is more efficient and secure.

Index Terms—Cloud Storage; Circle Mapping; Indexing; Interval Trapdoor

I. INTRODUCTION

Cloud storage and database outsourcing have become a new trend in database management. Data Owners (DO) outsource data management to Cloud Servers (CS), which are equipped with powerful hardware, software and plenty of network resources. The cloud servers can provide efficient data management, such as database storage and queries, to DO in order to reduce their costs on data management [1]. Due to concerns on reliability and security of the CS, data security has been investigated as an important research topic in cloud storage.

Agrawal et al. in IBM research center proposed an Order Preserving Encryption Scheme, or OPES, that preserves the order of sensitive data in indices for efficient queries over encrypted data [2, 3]. However, their method exposes the order information of sensitive data to attackers. If a portion of plaintext and ciphertext are accessible in advance, the attackers can perceive all the values of the sensitive data. Hacigümüş et al. developed an advanced approach that could partition sensitive data in buckets for efficient range queries over encrypted data [4-7]. Because each bucket contains sensitive data in such a way that their values are limited in a particular range, qualified ciphertext can be retrieved by deciding which range they are in during range query processing. Li et al. designed a prefix-preserving encryption scheme for index creation. They used prefix-matching methods to support range queries. However, their method discloses the distribution information of sensitive data through interval trapdoors [8]. Cheng et al. proposed an authorization-based access control approach, but parts of data encryption have to be completed by DO, which incurs extra overheads to the method during attribute and user privilege revocation [9-12]. Pervez et al. developed an Oblivious Term Matching (OTM) approach that allows users to customize their queries and reduces the costs of query processing [13]. In [14], an efficient single-assertion-based method for range queries over sensitive data was proposed. But more efforts are required in network communication improvement.

This paper proposes a new scheme for range queries over encrypted data. By using circle mapping, cloud servers are able to determine whether the sensitive data is in a particular query range after executing a single assertion. This method can reduce information leakage of data during query processing. Moreover, our method introduces extra encryption on indices and interval trapdoors, which increases security of query processing. Our scheme has the advantage of good security, small amount of calculation, and less resource consumption.

II. PRELIMINARIES

Definition 1: Given a semicircle of radius $D$ in a two-dimensional coordinate, and a point $A$ on the semicircle, let $\beta$ denote the angle between line $OA$ and x-axis, $d$ denote the length of the projection of line $OA$ on x-axis.

So, $\cos \beta = \frac{d}{D}$, $\beta = \arccos \frac{d}{D}$. As shown in Figure 1, $\beta$ is the value of mapping $d$ to a circle of radius $D$.

© 2014 ACADEMY PUBLISHER
doi:10.4304/jcp.9.11.2656-2660
There are two roles, Data Owners (DO) and Cloud Servers (CS), in this approach. We assume that all sensitive data on CS has been encrypted, so that the real values of these sensitive data are not accessible by CS. But indices and interval trapdoors of the sensitive data can be perceived by CS. Our approach is described as below:

(1) DO creates indices based on $d$, and encrypts them with a key $K$. Then, the encrypted indices $I$ and sensitive data are uploaded to CS.

(2) DO creates a interval trapdoor $T$ based on the range of $d$ of sensitive data, which is represented by the upper bound and lower bound of query ranges.

(3) CS evaluates range queries by using the indices $I$ and interval trapdoor $T$, and sends qualified encrypted data to DO.

III. RANGE QUERIES SCHEME

Our approach consists of 7 steps, in which step 1-3 are in the initialization phase, and step 4-7 are in range query processing phase.

(1) Data Owner (DO) constructs a 2-by-2 invertible matrix $K_1$, $K_2$ denotes the inverse of $K_1$, so, $K_1 \times K_2 = I$.

(2) DO selects a large value $D$ in such a way that, for any sensitive data $d, d \in [-D, D]$. Then, $\beta$ is calculated by mapping $d$ to a circle of radius $D$.

(3) DO calculates indices $I = K \begin{bmatrix} \sin \beta \\ \cos \beta \end{bmatrix}$, and uploads encrypted sensitive data $d$ and indices $I$ to CS.

(4) During range query processing, DO calculates $\beta_1$ and $\beta_2$ by mapping the lower bound $d_1$ and upper bound $d_2$ to a circle of radius $D$.

(5) DO calculates the trapdoor $T'_1 = [\cos \beta_1 - \sin \beta_1]K_2$, and $T'_2 = [-\cos \beta_2, \sin \beta_2]K_2$. Let the interval trapdoor $T = \{T_1, T_2\}$. If $T_1 = T'_1$, then $T_2 = T'_2$. If $T_1 = T'_2$, then $T_2 = T'_1$. In other words, $T_1$ and $T_2$ are randomly equal to either $T'_1$ or $T'_2$.

(6) DO uploads $T = \{T_1, T_2\}$ and sends a range query request to CS.

(7) After receiving the range query request, $(T_1 \times I)(T_2 \times I)$ is calculated by CS. If $(T_1 \times I)(T_2 \times I) \geq 0$, then the corresponding sensitive data $d$ of the indices $I$ is sent back to DO. These sensitive data is the result, the real values of which fall in the range $[d_1, d_2]$ specified by the range query.

IV. VERIFICATION OF AUTHENTICITY

Theorem 1: As shown in Figure 2, given a semicircle of radius $D$ ($D > 0$) in a two-dimensional coordinate and three points $A_1$, $A_2$, $A_3$ on the semicircle, $\beta$, $\beta_1$ and $\beta_2$ are the angles between line $OA_1$, $OA_2$, $OA_3$ and $x$-axis, respectively. $0 \leq \beta, \beta_1, \beta_2 \leq \pi$. If $-D \leq d, d_1, d_2 \leq D$, and only if line $OA$ locates between line $OA_1$ and $OA_2$, then $\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0, d_2 \leq d \leq d_1$.

Proof: Figure 2 shows that $\cos \beta_1 = \frac{d_1}{D}, \cos \beta = \frac{d}{D}, \cos \beta_2 = \frac{d_1}{-D}$.

(1) When line $OA_1$ locates between line $OA_1$ and $OA_2$, we observe that, $\beta_2 \geq \beta \geq \beta_1$, $\beta_2 - \beta \geq 0$, $\sin(\beta_2 - \beta) \geq 0$, $\beta - \beta_1 \geq 0$ and $\sin(\beta - \beta_1) \geq 0$.

Thus, $\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0$.

Since $\cos \alpha (0 \leq \alpha \leq \pi)$ is a decreasing monotonic function, the inequalities $\cos \beta_2 \leq \cos \beta \leq \cos \beta_1$ and $\frac{d_2}{-D} \leq \frac{d}{D} \leq \frac{d_1}{D}$ hold when $\beta_2 \geq \beta \geq \beta_1$. So, $d_2 \leq d \leq d_1, d \in [d_1, d_2]$.

(2) In Figure 3, when line $OA$ locates at the right hand side of line $OA_1$ and $OA_2$, we can get $\beta_2 \geq \beta_1 > \beta$, $\beta_2 - \beta \geq 0$, $\sin(\beta_2 - \beta) \geq 0$, $\beta - \beta_2 \leq \beta_1 - \beta$, $\sin(\beta - \beta_1) \leq 0$.

Thus, $\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \leq 0$.

Since $\cos \beta_2 \leq \cos \beta \leq \cos \beta_1$ and $\frac{d_2}{-D} \leq \frac{d}{D} \leq \frac{d_1}{D}$, which implies $d_2 \leq d \leq d_1, d \in [d_1, d_2]$.

(3) As shown in Figure 4, when line $OA$ locates at the left hand side of line $OA_1$, $OA_2$, then $\beta \geq \beta_1 > \beta_2$, $\beta_2 - \beta < 0$, $\sin(\beta_2 - \beta) < 0$, $\beta - \beta_2 \geq 0$ and $\sin(\beta - \beta_1) \geq 0$.
So, \( \sin(\beta_2 - \beta) \sin(\beta - \beta_1) < 0 \),
\[
\cos \beta_1 \geq \cos \beta_2 > \cos \beta \quad \text{and} \quad \frac{\beta_2 - \beta}{D} \geq \frac{\beta - \beta_1}{D},
\]
which implies \( d < d_2 \leq d_1, d \notin [d_1, d_2] \).

Therefore, if and only if line \( OA \) locates between line \( OA_1 \) and \( OA_2 \), then
\[
\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0,
\]
\[
d_2 \leq d \leq d_1, d \notin [d_1, d_2].
\]

Theorem 2: As shown in Figure 2, given a semicircle of radius \( D (D > 0) \) in a two-dimensional coordinate and three points \( A_1, A, A_2 \) on the semicircle. \( \beta_1, \beta \) and \( \beta_2 \) are the angles between line \( OA_1, OA, OA_2 \) and \( x \)-axis, respectively, \( 0 \leq \beta, \beta_1, \beta_2 \leq \pi \), \( -D \leq d, d_1, d_2 \leq D \), if and only if
\[
\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0,
\]
then line \( OA \) locates between line \( OA_1 \) and \( OA_2 \), \( d_2 \leq d \leq d_1 \).

Proof: By Theorem 1, line \( OA \) is outside of line \( OA_1 \) and \( OA_2 \), \( d \notin [d_1, d_2] \), if \( \sin(\beta_2 - \beta) \sin(\beta - \beta_1) < 0 \). Thus, line \( OA \) locates between line \( OA_1 \) and \( OA_2 \), \( d_2 \leq d \leq d_1 \), \( d \notin [d_1, d_2] \), if and only if
\[
\sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0.
\]

By Theorem 2, the corresponding values of \( [d_1, d_2] \) and \( d \) after the circle mapping are \( [\beta_1, \beta_2] \) and \( \beta \), respectively. If \( d_2 \leq d \leq d_1 \), \( \sin(\beta_2 - \beta) \sin(\beta - \beta_1) \geq 0 \), which implies that there still exists an assertion that can represent the range relation after the circle mapping.

Theorem 3: In step 7 of our approach, the assertion \( (T_2 \times I)(T_2 \times I) \geq 0 \) can be used to search sensitive data for particular range queries.

Proof:
\[
(T_2 \times I) = [\cos \beta_1 - \sin \beta_1] K_2 \times K_1 \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}
\]
\[
= [\cos \beta - \sin \beta] \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}
\]
\[
= \sin \beta \cos \beta_1 - \cos \beta \sin \beta_1
\]
\[
= \sin(\beta - \beta_1)
\]
\[
(T_2 \times I) = [-\cos \beta_2 \sin \beta_2] K_2 \times K_1 \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}
\]
\[
= [-\cos \beta_2 \sin \beta_2] \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}
\]
\[
= (\sin \beta_2 \cos \beta - \cos \beta_2 \sin \beta)
\]
\[
= \sin(\beta_2 - \beta)
\]
So, \( (T_2 \times I)(T_2 \times I) = \sin(\beta_2 - \beta) \sin(\beta - \beta_1) \), and our approach is proved.

V. SECURITY ANALYSIS

Our approach assumes that the sensitive data has been encrypted. Attackers cannot obtain the real values of the sensitive data, but they can possibly perceive the order of the sensitive data from the indices and interval trapdoors provided by Data Owners (DO). Moreover, the real values of sensitive data could be potentially predicted after its order is available. Thus, the security of indices and interval trapdoors in our approach is analyzed as follows, which indicates that our approach can satisfy the security requirement:

A. Security of Indices

To ensure security of indices, neither sensitive data nor the order of any two sensitive data can be derived from indices by attackers.

Theorem 4: Accessing the real values of sensitive data through indices \( I \) provided by our approach is difficult.

Proof: By definition 1 and step 3 in our approach, the indices can be calculated as
\[ I = K_1 \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}, \quad K_1 \quad \text{is a key represented by a 2-by-2 matrix, and}
\]
\[ \beta \quad \text{is the corresponding value of sensitive data} \quad d \quad \text{after the circle mapping. Apparently, if} \quad D \quad \text{and numbers in} \quad K_1 \quad \text{are large and encrypted, deriving} \quad d \quad \text{from the indices is as difficult as the factoring challenge for large integers.}
\]

Therefore, it is difficult for attackers to derive real information of sensitive data from the indices \( I \).

Theorem 5: Attackers cannot derive the order of sensitive data from indices in our approach.

Proof: Let \( P = \begin{bmatrix}
\sin \beta \\
\cos \beta
\end{bmatrix}, \quad K = \begin{bmatrix}
a & b \\
c & d
\end{bmatrix}, \quad K \quad \text{is an encrypted invertible matrix. Given a matrix} \quad C, \quad \text{there must exist multiple} \quad P \quad \text{'s} \quad \text{or} \quad \beta \quad \text{'s}, \quad \text{which satisfy} \quad C = K \times P. \quad \text{For example, given} \quad C_1 \quad \text{and} \quad C_2, \quad \text{there must exist multiple} \quad \beta_1 \quad \text{'s} \quad \text{and} \quad \beta_2 \quad \text{'s}, \quad \text{which satisfy} \quad C = K \times P.
\]
\[
\text{Give} \quad \beta_1, \quad \beta_2, \quad K = \begin{bmatrix}
a' & b' \\
c' & d'
\end{bmatrix}, \quad P_1 = \begin{bmatrix}
\sin \beta_1 \\
\cos \beta_1
\end{bmatrix}, \quad P_2 = \begin{bmatrix}
\sin \beta_2 \\
\cos \beta_2
\end{bmatrix}, \quad \text{thus}
\]
\[
C_1 = K' \times P_1' = \begin{bmatrix}
a' & b' \\
c' & d'
\end{bmatrix} \begin{bmatrix}
\sin \beta_1 \\
\cos \beta_1
\end{bmatrix} \quad (1)
\]
\[
C_2 = K' \times P_2' = \begin{bmatrix}
a' & b' \\
c' & d'
\end{bmatrix} \begin{bmatrix}
\sin \beta_2 \\
\cos \beta_2
\end{bmatrix} \quad (2)
\]
Let
\[
C_1 = C_1, \quad C_2 = C_2, \quad \beta_1 = \pi - \beta_1, \quad \beta_2 = \pi - \beta_2,
\]
\[
K' = \begin{bmatrix}
a' & -b' \\
c' & -d'
\end{bmatrix}, \quad \text{then}
\]
\[
P_1' = \begin{bmatrix}
\sin \beta_1 \\
\cos \beta_1
\end{bmatrix} = \begin{bmatrix}
\sin(\pi - \beta_1) \\
\cos(\pi - \beta_1)
\end{bmatrix} = \begin{bmatrix}
\sin \beta_1 \\
-\cos \beta_1
\end{bmatrix}, \quad C_1 = C_2 \quad (3)
\]
\[ P_2 = \begin{bmatrix} \sin \beta_1 \\ \cos \beta_1 \end{bmatrix} = \begin{bmatrix} \sin(\pi - \beta_1) \\ \cos(\pi - \beta_1) \end{bmatrix} = \begin{bmatrix} \sin \beta_2 \\ -\cos \beta_2 \end{bmatrix} \]

\[ C_2' = K^* \times P_2 = \begin{bmatrix} a' & -b' \\ c' & -d' \end{bmatrix} \begin{bmatrix} \sin \beta_2 \\ -\cos \beta_2 \end{bmatrix} = C_2 = C_2 \quad (4) \]

From equations (1),(2),(3),(4), given \( \beta_1 \) and \( \beta_2 \) that satisfy any \( C_1 \) and \( C_2 \), there must exist \( \beta_1' \) and \( \beta_2' \) that satisfy the same \( C_1 \) and \( C_2 \). If \( \beta_1 > \beta_2 \), there must exist \( \beta_1' \) and \( \beta_2' \) such that \( \beta_1' < \beta_2' \). Under the condition that \( \beta_1 \) and \( \beta_2 \) are unknown, there are two possible cases, either \( \beta_1 = \beta_1' \) and \( \beta_2 = \beta_2' \) or \( \beta_1 = \beta_1' \) and \( \beta_2 = \beta_2' \), and it cannot determine which case is true. Thus, the order of \( \beta_1 \) and \( \beta_2 \) is indeterminable.

Since \( \beta_1 = \arccos \frac{d_1}{D} \), \( \beta_2 = \arccos \frac{d_2}{D} \), given any two indices \( I_1 \) and \( I_2 \), if we can find an order of a pair of sensitive data, such as \( d_1' > d_2' \), we can also find another pair of sensitive data in reverse order, \( d_1' < d_2' \). Thus, there are two possible cases, either \( d_1 = d_1' \) and \( d_2 = d_2' \) or \( d_1 = d_1' \) and \( d_2 = d_2' \), and we cannot determine which case is true. So, the order of \( d_1 \) and \( d_2 \) is indeterminable.

Therefore, attackers cannot perceive the order of any two sensitive data \( d_1 \) and \( d_2 \) through indices \( I \), which indicates that attackers cannot perceive the order information of all sensitive data through indices \( I \) in our approach.

(2) Security of interval trapdoors

The interval trapdoor in our approach is

\[ T = [\cos \beta_1 - \sin \beta_1, \sin \beta_2 - \cos \beta_2] \]

Like theorem 4, we can prove that deriving the values of \( D, d, d_1 \) and \( d_2 \) from \( [\cos \beta_1, -\sin \beta_2] \)

\[ [\cos \beta_1, -\sin \beta_2] \]

is difficult, if \( D \) and numbers in \( K \) are large and encrypted.

Like theorem 5, we can prove that attackers cannot derive the order of \( \beta_1 \) and \( \beta_2 \) from the interval trapdoor \( T \), which implies that the upper-bound and lower-bound of the range cannot be calculated from \( T \). Therefore, the query range and the order of sensitive data in the range are inaccessible to attackers.

So our interval trapdoors are secure.

VI. PERFORMANCE ANALYSIS

On the aspect of space complexity, like the method in [14], our approach maintains an invertible matrix \( K \) and a range \( D \) of values of sensitive data in clients. The inverse of \( K \) can be calculated. Let \( r \) be the size of each data record. The space complexity of our approach is \( O(4r^2+1) = O(1) \), which is independent to the number of data records. However, let \( n \) denote the number of data records. The size of each bucket is \( t \). Each bucket contains \( i \) data records. Thus, the space complexity of the method in [4] is \( O((n+i)\lambda) = O(n) \), which increases as the number of data records grows.

On the aspect of computation complexity, the interval trapdoors of the method in [14] has three components, but our approach consists of two components. Thus, the computation cost of our approach in clients is one third less than the one in [14] during query processing. Moreover, our approach is more secure, because it reduces network communication and discloses less information to CS. Let \( a, b \) and \( c \) be the computation complexity of multiplication, subtraction and comparison, respectively, the computation complexity of assertion processing in our approach is \( O(5am+2bn+cn) = O(n) \).

For example, if the number of data records in each bucket is identical, the computation complexity of the method in [4] is \( O(cn) = O(n) \) in the best case, in which there is only one bucket in each query range (we assume the method does not apply any optimization in query processing). In the worst case that query range includes all the buckets, the computation complexity becomes

\[ O((cn^2)/i) = O(n^2) \].

Therefore, our approach is more efficient in terms of both space and computation complexity.

VII. CONCLUSIONS

This paper proposes a new scheme of queries over encrypted data. The scheme utilizes trapdoors constructed by the upper bound and lower bound of query range, and achieves a transmission from range determination to single-assertion, which avoids potential risks that disclose the distribution information of sensitive data during query processing, and reduces the space requirement in clients. Moreover, our approach can avoid disclosing both sensitive data itself and its order and distribution information, which provides higher security and more efficient query processing.

REFERENCES


