

Cardinality-based Inference Control in OLAP Systems: An Information Theoretic Approach

Nan Zhang
Department of Computer
Science Texas A&M University
College Station, TX, 77843
nzhang@cs.tamu.edu

Wei Zhao
Department of Computer
Science Texas A&M University
College Station, TX, 77843
zhao@cs.tamu.edu

Jianer Chen
Department of Computer
Science Texas A&M University
College Station, TX, 77843
chen@cs.tamu.edu

ABSTRACT

We address the inference control problem in data cubes with some data known to users through external knowledge. The goal of inference controls is to prevent exact values of sensitive data from being inferred through answers to online analytical processing (OLAP) queries. We present an information theoretic approach for cardinality-based inference control, which simply counts the number of cells that all queries have covered thus far to determine whether a new query should be answered. Compared to previous approaches in sum-only data cubes, our new approach has a more general framework (applies to MIN, MAX and SUM) and is more effective.

Categories and Subject Descriptors

H.2.8 [Information Systems]: Database Applications—*Data Mining*; H.2.7 [Information Systems]: Database Administration—*Data warehouse and repository, Security, integrity, and protection*

General Terms

Algorithms, Security

Keywords

Data Mining, OLAP, Inference Control, Information Theory

1. INTRODUCTION

In this paper, we address issues related to the protection of sensitive data from being disclosed through answers to aggregate queries. We focus on restriction-based inference control problem. This problem has been extensively studied in statistical databases (SDB) literature. Many algorithms have been proposed and analyzed. A good survey can be found in [1]. However, few of them can meet the instant response time requirement in OLAP systems.

Several studies have been carried out on efficient inference control in OLAP systems [2–14]. A cardinality-based mechanism was proposed in [14]. However, most existing approaches have constraints either on the applications (e.g., classification) or on the ag-

gregation operations (e.g., SUM-only). We introduce an information-theoretic restriction-based approach to remove such constraints. Our new approach has the following important features to distinguish it from previous approaches.

- Our approach applies to MIN, MAX and SUM. We first unify the discussion of aggregate operations by addressing their information theoretic common properties.
- Our approach is more effective than previous approaches. The result in [14] is actually a special case of our general formulation when compromisability is equal to 2 and dimensionality of query is equal to 1 (which will be defined in Sect. 2). In other cases, our approach can accept more queries while keeping individual data confidential.
- Our approach is easy to implement. Our offline calculation algorithm can be readily integrated into the materialization of cuboids. The overhead of our online query restriction algorithm is substantially smaller than previous restriction-based algorithms in SDB.

The rest of this paper is organized as follows. In Sect. 2, we present our models, review previous approaches and introduce the basic idea of our new approach. The inference control algorithm and related components are discussed in Sect. 3. The information theoretic approach is provided in Sect. 4, followed by a final remark in Sect. 5.

2. APPROACHES

In this section, we will first introduce our models of data cubes, queries, inferences and compromisability. Based on these models, we review the cardinality-based approach introduced in [14]. We will point out the problems associated with this approach which motivates us to design a new inference control method, based on information theoretic techniques.

2.1 Models of Data Cube and Queries

Consider an n -dimensional $D_1 \times D_2 \times \dots \times D_n$ data cube X . Some cells of X are either vacant¹ or insensitive (thus can be known by users through external knowledge). An example of data cube X is shown in Tab. 1. The sales amount of a product in a particular month is considered to be sensitive. Since the bookstore began to sell *used books* and *CD* in February, the sales amounts of *used books* and *CD* in January are known to users as unavailable. Moreover, the bookstore used the sales amounts of *books* and

¹i.e., The users know that the cell is vacant. See the example below for details.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DOLAP'04, November 12–13, 2004, Washington, DC, USA.
Copyright 2004 ACM 1-58113-977-2/04/0011 ...\$5.00.

Table 1: Example of a Bookstore

Sales ($\times \$100$)	Jan	Feb	Mar	Sum
Book	192	Known	220	S_B
Used Book	Known	20	Known	S_U
CD	Known	30	87	S_D
Video	168	Known	96	S_T
Sum	S_1	S_2	S_3	

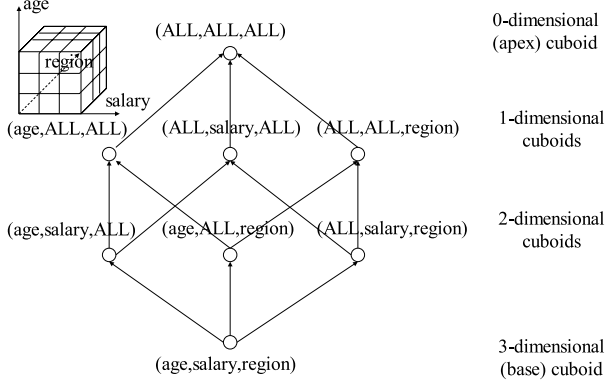


Figure 1: lattice of cuboids

videos in February as well as the sales amount of *used books* in March to make advertisement. Thus, these sales amounts are also known to users (as insensitive). We use a user co-occurrence matrix defined below to represent whether a cell is known by users through external knowledge.

Definition 1. (User Co-occurrence Matrix) For a given $D_1 \times D_2 \times \dots \times D_n$ data cube X , the user co-occurrence matrix of X is a $D_1 \times D_2 \times \dots \times D_n$ matrix U with elements $u_{i_1, i_2, \dots, i_n} \in \{0, 1\}$. An element in U is equal to 1 (i.e., $u_{i_1, i_2, \dots, i_n} = 1$) if its corresponding cell in X is known by users. Otherwise $u_{i_1, i_2, \dots, i_n} = 0$. The number of cells in X known by users through external knowledge (i.e., the number of 1 in U) is represented by $t(X)$.

For example, the user co-occurrence matrix of data cube X shown in Tab. 1 is

$$U = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

We have $t(X) = 5$.

In OLAP systems, the requirement of instant response time restricts queries on a data cube X to those that can be derived from predefined cuboids of X . Each cuboid shows a view of the data cube at different summarization (*group by*) level. The lattice of these cuboids forms the data cube. Figure 1 shows an example of the lattice of cuboids. In general, a n -dimensional data cube X has 2^n cuboids, including $\binom{n}{k}$ k -dimensional cuboids. An n -dimensional cuboid is called a base cuboid of X . We denote the set of k -dimensional cuboids by Γ_k and the set of all cuboids by Γ .

In this paper, we only consider queries on the cuboids of a data cube. We define query as below.

Definition 2. (Query) Given an n -dimensional data cube X , a query $Q\langle F, W \mid F \in \{\text{MIN}, \text{MAX}, \text{SUM}\}, W \subseteq \Gamma_n\}$ on X satisfies

1. the answer to Q is to calculate the aggregate function F on W . i.e., $AQ = F(W)$.
2. there exists a (unique) cuboid $S \in \Gamma$ such that a cell of S is the aggregation (*group by*) of W .

We call S the corresponding cuboid of Q and the above cell in S (denoted by s) the corresponding cuboid cell of Q . We call Q an $(n - k)$ -dimensional query if and only if S is a k -dimensional cuboid. Following previous notions in SDB, we denote the set of answers to all queries on X by AQ .

An intuitive explanation of the dimensionality of Q is the dimensionality of the dice of X that is aggregated by Q . For example, given the 2-dimensional data cube in Tab. 1, there are 8 legitimate queries: $S_1, S_2, S_3, S_B, S_U, S_D, S_T$ and the sum of all cells in the data cube. The first 7 queries are 1-dimensional queries while the last one is 2-dimensional.

2.2 Models of Inferences and Compromisability

An inference problem occurs when the exact value of sensitive data is disclosed to users through answers to queries. In OLAP systems, this means that the exact value of a cell of the data cube can be determined by users through answers to OLAP queries and external knowledge. For example, in the 2-dimensional data cube shown in Tab. 1, an obvious inference problem occurs on the answer to S_U . A user can simply infer that the sales amount of used books in February satisfies²

$$x_{U2} = S_U. \quad (1)$$

A straightforward solution is to block queries to S_U such that no more “direct” inference exists. However, there still exists another inference problem in the data cube.

Consider a user who has obtained the answers to S_B, S_T, S_1 and S_3 . The user can derive the sales of CD in March by calculating

$$x_{D3} = S_1 + S_3 - S_B - S_T. \quad (2)$$

In this case, there is no more “straightforward” solution for inference control.

In order to separate “direct” inference problems in (1) from “indirect” problems in (2), we define compromisability of query, cuboid and data cube as below.

Definition 3. (Compromisability) Given an n -dimensional data cube X , the compromisability of a query $Q\langle F, W \rangle$ on X is inductively defined below.

1. If Q is a 1-dimensional query, the compromisability of Q is the number of cells covered by Q that is *NOT* known by users through external knowledge.
2. If Q is a $(k + 1)$ -dimensional ($k \geq 1$) query, there exists a set of k -dimensional queries \mathcal{Q}_0 such that the Q is the aggregation of all queries in \mathcal{Q}_0 . The compromisability of Q is the minimum compromisability of queries in \mathcal{Q}_0 . In other words, assume that the compromisability of Q is $L(Q)$, we have

$$L(Q) = \min_{Q_0 \in \mathcal{Q}_0} L(Q_0). \quad (3)$$

²Hereafter, we assume that the values of known cells are subtracted from answers to queries

The compromisability of the corresponding cuboid cell of Q is defined to be equal to the compromisability of Q (i.e., $L(s) = L(Q)$). The compromisability of X is equal to the compromisability of the cell in the 0-dimensional (apex) cuboid of X (note that there is only one cell in the 0-dimensional cuboid).

Assume there is no query with compromisability of 0. Due to Def. 3, a “direct” inference problem occurs if and only if a 1-dimensional query with compromisability of 1 is answered. Note that “direct” inference problem can be easily solved by blocking such queries. In order to simplify our discussion, hereafter we assume that there does not exist any legitimate query with compromisability less than 2.

2.3 Previous Approaches

To prevent the disclosure of sensitive data due to inference problems, countermeasures must be implemented in OLAP systems. A cardinality-based approach in SUM-only data cubes has been proposed in [14]. We briefly review this method below.

The inference control system in [14] is based on a three-tier model. Between the tiers of user queries and data cube, there is a tier named “predefined aggregations”, which contains pre-calculated aggregations that is free of inference problems. The system only answers queries that can be derived from these “safe” aggregations.

In order to calculate these “safe” aggregations, the data cube is first partitioned into disjoint chunks with the same dimensionality as the original data cube. After that, a cardinality-based criteria is tested on each chunk (i.e., sub-data cube [14]) to determine whether it is inference-free. For an inference-free chunk, all possible aggregations on the chunk are added to the safe list (i.e., “predefined aggregations”). Otherwise, none of them is added to the predefined aggregations.

The key observation here is the cardinality-based criteria. It is proved in [14] that a $D_1 \times D_2 \times \dots \times D_n$ sub-data cube X is inference free if $t(X) < 2D_l + 2D_m - 9$. Here D_l and D_m are the two smallest among D_1, D_2, \dots, D_n .

Clearly, the system should be measured by its capability in terms of answering more queries, while preventing inference problems from happening.

There are several problems with this approach. For example, the cardinality-based criteria on an n -dimensional data cube X is actually determined by the “worst” 2-dimensional $D_l \times D_m$ dice of X . Actually, even when $n = 2$, the cardinality bound is far from tight (as we will show in Sect. 3). Furthermore, if a chunk is not inference-free, NONE of the aggregations can be added to the safe list. Clearly there will be a lot of “safe” queries being rejected because of this.

These problems motivate us to develop a new approach that allows more queries to be answered. We describe the new approach in the next subsection.

2.4 Our New Approach

We do not need a third tier between user queries and data cube. Instead, we introduce some offline calculation onto the materialization of cuboids. During the materialization of any given k -dimensional cuboid C , we calculate a “cardinality bound” t for every cell of C (the detail of t will be discussed in Sect. 3). Besides, we have an online query restriction algorithm that determines whether a query should be answered. When a new k -dimensional ($k \leq n - 1$) query Q comes, our system first finds the corresponding cuboid cell c_0 of the query (which is in a $(n - k)$ -dimensional cuboid), then finds the set \mathcal{C} of cells in $(n - k - 1)$ -dimensional cuboids that contain c_0 (e.g., as in Fig. 1, $\langle \text{age} = 20, \text{all}, \text{all} \rangle$ contains $\langle \text{age} = 20, \text{salary} = 50,000, \text{all} \rangle$). Then for every cell in \mathcal{C} ,

we compare the “cardinality bound” t with the number of known cells covered by Q . Thus, we can determine whether a new query can be answered.

The key here is to properly design the “cardinality bound” t so that more queries can be answered free of inference problems. Given an n -dimensional data cube X , we prove that a k -dimensional query $Q(F, W)$ can be safely answered if every $(k+1)$ -dimensional dice X' in X that

1. contains W as a subset,
2. can be queried as a cell of a $(n - k - 1)$ -dimensional cuboid,

satisfies

$$t(X') < L(X')^k \sum_{i=1}^{k+1} d_i - (k+1)L(X')^{k+1} - 1. \quad (4)$$

Here d_1, d_2, \dots, d_{k+1} are the dimension domains of X' (i.e., X' is a $d_1 \times d_2 \times \dots \times d_{k+1}$ dice). $L(X')$ is the compromisability of the cell corresponding to X' . As we can see, the result in [14] is actually a special case of our result when $L(X') = 2$ and the dimensionality of query is equal to 1.

3. INFERENCE CONTROL ALGORITHM

Algorithm 1 Offline pre-calculation algorithm

Input: minimum compromisability l_0 .
 Given an n -dimensional data cube X ,
for $k \leftarrow n - 2$ **downto** 0 **do**
 for every k -dimensional cuboid C **do**
 for every element c in C **do**
 Assume c is the aggregation of an $(n - k)$ -dimensional
 $d_1 \times d_2 \times \dots \times d_{n-k}$ dice of X ,
 $l = \max(\text{compromisability of } c, l_0)$,
 the cardinality bound of c_j : $t = l^{n-k-1} \sum_{i=1}^{n-k} d_i - (n - k)l^{n-k} - 1$.
 end for
end for
end for

Algorithm 2 Online query restriction algorithm

Input: minimum compromisability l_0 .
 Given a new k -dimensional query Q , assume the compromisability of Q is $L(Q)$ and the number of known cells covered by Q is t_Q ,
 Find the corresponding cuboid cell c of Q . Without loss of generality, assume c is $\{x_1, x_2, \dots, x_{n-k}, \text{ALL}, \text{ALL}, \dots, \text{ALL}\}$,
for $j \leftarrow 1$ to $n - k$ **do**
 Find the cell $c_j = \{x_1, x_2, \dots, x_{j-1}, \text{ALL}, x_{j+1}, \dots, x_{n-k}, \text{ALL}, \text{ALL}, \dots, \text{ALL}\}$
 Find the cardinality bound t of c_j ,
 $t = t - t_Q$.
 if $t \leq 0$ **or** $L(Q) < l_0$ **then**
 Reject the query,
 else
 Answer the query.
end if
end for

The offline pre-calculation algorithm and the online query restriction algorithm are presented in Alg. 1 and Alg. 2, respectively.

The minimum compromisability l_0 is a parameter of the system. A larger l_0 can make the cardinality bound t larger. However, it may also reject many queries with cardinality less than l_0 . Due to space limit, we refer readers to [15] for the detailed proof of the correctness of our algorithm. We now discuss the information theoretic approach used to derive the cardinality bound on inference-free queries.

4. INFORMATION THEORETIC APPROACH

Readers not familiar with information theory are referred to the literature (e.g., [16]). Here we first model the inference problem in data cubes in an information theoretic way.

Lemma 1. Given an n -dimensional data cube X and its user co-occurrence matrix U , the inference problem occurs in X if and only if there exists a cell x_{i_1, \dots, i_n} in X such that $u_{i_1, \dots, i_n} = 0$ (i.e., the cell is not known by users through external knowledge) and

$$H(x_{i_1, \dots, i_n} | \mathcal{A}Q) = 0$$

Recall that $\mathcal{A}Q$ is the set of answers to queries. Here $H(x_{i_1, \dots, i_n} | \mathcal{A}Q)$ is the conditional entropy of x_{i_1, \dots, i_n} based on $\mathcal{A}Q$.

As we will see, the introduction of information theory brings us great convenience on discussion. Firstly, we transform the conditional entropy of x_{i_1, \dots, i_n} to the conditional entropy of answers to queries, which is much easier to analyze. For a given cell x in X , we have

$$H(x | \mathcal{A}Q) = 0 \Leftrightarrow H(\mathcal{A}Q) - H(\mathcal{A}Q | x) = H(x). \quad (5)$$

Consider an n -dimensional data cube X with user co-occurrence matrix U , we have

Lemma 2. The inference problem occurs in X if and only if there exists a cell x_{i_1, \dots, i_n} such that $u_{i_1, \dots, i_n} = 0$ and

$$H(\mathcal{A}Q) - H(\mathcal{A}Q | x_{i_1, \dots, i_n}) = H(x_{i_1, \dots, i_n}). \quad (6)$$

For the convenience of further discussion, hereafter unless indicated otherwise, all data cubes mentioned are n -dimensional, $D_1 \times D_2 \times \dots \times D_n$ data cubes. We define two functions as below. For all possible data cubes X with $t(X) = t_0$, $f_{\max}(t_0)$ and $f_{\min}(t_0)$ are the maximum and minimum entropies of answers to all queries on X , respectively.

$$f_{\max}(t_0) = \max_X \{H(\mathcal{A}Q) | t(X) = t_0\} \quad (7)$$

$$f_{\min}(t_0) = \min_X \{H(\mathcal{A}Q) | t(X) = t_0\} \quad (8)$$

Theorem 1. Given an n -dimensional data cube X with $t(X) = t_0$, no inference problem exists in X if $\forall x_{i_1, \dots, i_n}$ such that $u_{i_1, \dots, i_n} = 0$, we have

$$f_{\max}(t_0) - f_{\min}(t_0 + 1) < H(x_{i_1, \dots, i_n}). \quad (9)$$

PROOF. Due to Lemma 2, no inference problem exists if $\forall x_{i_1, \dots, i_n}$ such that $u_{i_1, \dots, i_n} = 0$, we have

$$H(\mathcal{A}Q) - H(\mathcal{A}Q | x_{i_1, \dots, i_n}) < H(x_{i_1, \dots, i_n}). \quad (10)$$

Given x_{i_1, \dots, i_n} , we can construct a new data cube X' that is the same as X except that x_{i_1, \dots, i_n} is known by users through external

knowledge. Denote the set of answers to queries on X' by $\mathcal{A}Q'$. Due to the definition of conditional entropy, we have

$$H(\mathcal{A}Q | x_{i_1, \dots, i_n}) = H(\mathcal{A}Q'). \quad (11)$$

Due to the definition of f_{\max} and f_{\min} , we have

$$H(\mathcal{A}Q) \leq f_{\max}(t_0), \quad (12)$$

$$H(\mathcal{A}Q | x_{i_1, \dots, i_n}) = H(\mathcal{A}Q') \geq f_{\min}(t_0 + 1). \quad (13)$$

In other words, no inference problem exists in X if $\forall x_{i_1, \dots, i_n}$ such that $u_{i_1, \dots, i_n} = 0$, we have

$$f_{\max}(t_0) - f_{\min}(t_0 + 1) < H(x_{i_1, \dots, i_n}). \quad (14)$$

□

In order to simplify our discussion, hereafter we use $H(x)$ to denote the minimum entropy of x_{i_1, \dots, i_n} satisfies $u_{i_1, \dots, i_n} = 0$. In other words,

$$H(x) = \min_{x_{i_1, \dots, i_n} | u_{i_1, \dots, i_n} = 0} H(x_{i_1, \dots, i_n}). \quad (15)$$

Due to the definition of information entropy, for any given query $Q(F, W)$ such that $F \in \{\text{MIN}, \text{MAX}, \text{SUM}\}$, the entropy of the answer to Q satisfies $H(x) \leq H(\mathcal{A}Q) \leq H(x) + O(\log |Q|)$. Here $|Q|$ is the number of cells covered by Q that is not known to users through external knowledge. Since $H(x)$ is the entropy of a cell that is not known to users, we can safely assume that $H(x) \gg \log |Q|$. In other words, we assume that

$$H(Q) = H(x). \quad (16)$$

Now we will derive an upper bound on f_{\max} . We use an example to illustrate the derivation.

Example 1. Consider the 2-dimensional data cube in Table 1. Once we know any 6 elements in $\{S_B, S_U, S_D, S_T, S_1, S_2, S_3\}$, we may derive the seventh element (for example, S_1) by

$$S_1 = S_B + S_U + S_D + S_T - S_2 - S_3. \quad (17)$$

Thus $H(\mathcal{A}Q) = (D_1 + D_2 - 2 + 1)H(x) = 6H(x)$. Note that there is no constraint on the aggregate function to be SUM. For MIN and MAX, we can also find such an element. For example, assume $S_2, S_3 < S_1$. Once a user knows S_2, S_3, S_B, S_U, S_D and S_T , it may derive S_1 by

$$S_1 = \max\{S_B, S_U, S_D, S_T\}. \quad (18)$$

Follow this scheme, we can prove that $\forall t$, if only $(n-1)$ -dimensional queries and n -dimensional queries are allowed, we have

$$f_{\max}(t) \leq \left(\sum_{j=1}^n D_j - n + 1 \right) H(x). \quad (19)$$

In order to derive a lower bound on f_{\min} , we introduce maximum compromisable data cubes.

Definition 4. (Maximum compromisable Data Cube)

An n -dimensional data cube X is maximum compromisable if and only if its user co-occurrence matrix U satisfies

$$\exists l \in [2, \lfloor \frac{\min\{D_1, \dots, D_n\}}{2} \rfloor], s.t.$$

1. $\forall i_1, \dots, i_n \in [1, l], u_{i_1, \dots, i_n} = 0$,
2. $\forall i_1, \dots, i_n \geq l + 1, u_{i_1, \dots, i_n} = 0$,
3. All other elements in U are 1.

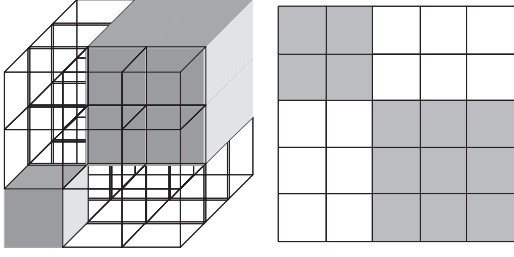


Figure 2: maximum compromisable data cubes

We can see that l is the compromisability of X .

Figure 2 shows two maximum compromisable data cubes. For an n -dimensional data cube that is maximum compromisable, we may divide the cells that is not known by users through external knowledge into two parts: one part is a $k \times k \times \dots \times k$ data cube and the other is a $(D_1 - k) \times (D_2 - k) \times \dots \times (D_n - k)$ data cube. Both of them have user co-occurrence matrices with all 0.

Theorem 2. (Main Theorem) Assume an n -dimensional data cube X_0 is maximum compromisable. The compromisability of X_0 is l_0 . If a data cube X satisfies

1. the compromisability of X : $l = l_0 > 1$,
2. $t(X) = t(X_0) - 1$,

then we have

$$H(\mathcal{A}\mathcal{Q}) \geq \left(\sum_{j=1}^n D_j - 1 \right) H(x). \quad (20)$$

PROOF. In order to simplify our discussion, we prove this theorem on 2-dimensional data cubes. Readers may easily extend this proof to n -dimensional cases.

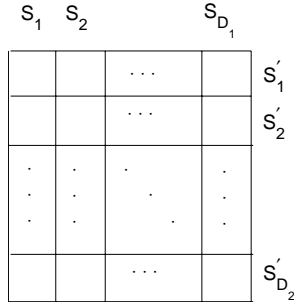


Figure 3: $D_1 \times D_2$ data cube X .

Suppose a $D_1 \times D_2$ data cube X does not satisfy (20). In other words,

$$H(\mathcal{A}\mathcal{Q}) < (D_1 + D_2 - 1)H(x). \quad (21)$$

Then there exists a proper subset of $\mathcal{A}\mathcal{Q}$, denoted by $\mathcal{A}\mathcal{Q}_0$, such that

1. There exists a query answer in $\mathcal{A}\mathcal{Q}_0$ that can be derived from other query answers in $\mathcal{A}\mathcal{Q}_0$. In other words, there exists $S \in \mathcal{A}\mathcal{Q}_0$ such that $H(S|\mathcal{A}\mathcal{Q}_0 \setminus S) = 0$.

2. No such query answer exists in any proper set of $\mathcal{A}\mathcal{Q}_0$. In other words, for any $\mathcal{A}\mathcal{Q}'_0 \subset \mathcal{A}\mathcal{Q}_0$, we have

$$\forall S \in \mathcal{A}\mathcal{Q}'_0, H(S|\mathcal{A}\mathcal{Q}'_0 \setminus S) > 0. \quad (22)$$

Without loss of generality, we assume that $\mathcal{A}\mathcal{Q}_0$ consists of S_1, \dots, S_{r_1} and S'_1, \dots, S'_{r_2} (this can be easily achieved by interchanging two rows or columns of X). A key point here is

Proposition 1.

$$\forall i_1 > r_1 \text{ and } j_1 < r_2, u_{i_1, j_1} = 1, \quad (23)$$

$$\forall i_2 < r_1 \text{ and } j_2 > r_2, u_{i_2, j_2} = 1. \quad (24)$$

PROOF. Suppose there exists $i_1 > r_1, j_1 < r_2$ such that $u_{i_1, j_1} = 0$. Then we have mutual information

$$I(S'_{r_2}; x_{i_1, j_1} | \mathcal{A}\mathcal{Q}_0 \setminus S'_{r_2}) > 0. \quad (25)$$

In other words, S'_{r_2} cannot be derived from $\mathcal{A}\mathcal{Q}_0 \setminus S'_{r_2}$. For any other S that satisfies $H(S|\mathcal{A}\mathcal{Q}_0 \setminus S) = 0$, we have

$$H(S|\mathcal{A}\mathcal{Q}_0 \setminus \{S, S'_{r_2}\}) = 0. \quad (26)$$

However, this contradicts the condition that no such query answers exists in any proper set of $\mathcal{A}\mathcal{Q}_0$. Thus, there does not exist such i_1 and j_1 that satisfy $u_{i_1, j_1} = 0$. Similarly, we can prove that $\forall i_2 < r_1$ and $j_2 > r_2, u_{i_2, j_2} = 1$. \square

Based on the definition of compromisability, we have $r_1, r_2 \geq l$. Due to Definition 4, if $l > \min(\lfloor \frac{D_1}{2} \rfloor, \lfloor \frac{D_2}{2} \rfloor)$, we have $t(X) > t(X_0)$. If $l \leq \min(\lfloor \frac{D_1}{2} \rfloor, \lfloor \frac{D_2}{2} \rfloor)$, we have

$$t(X) \geq r_1(D_2 - r_2) + r_2(D_1 - r_1) \quad (27)$$

$$\geq k(D_1 + D_2) - 2k^2 \quad (28)$$

$$= t(X_0). \quad (29)$$

However, this contradicts our assumption that $t(X) = t(X_0) - 1$. Thus, we have

$$H(\mathcal{A}\mathcal{Q}) \geq \left(\sum_{j=1}^n D_j - 1 \right) H(x). \quad (30)$$

Theorem 3. (Bound) Given an n -dimensional data cube X with compromisability $l > 1$, X is inference-free if

1. only $(n - 1)$ -dimensional and n -dimensional queries are allowed,
2. $t(X) < l^{n-1} \sum D_j - nl^n - 1$.

PROOF. Note that a maximum compromisable data cube X_0 with compromisability l satisfies $t(X_0) = l^{n-1} \sum D_j - nl^n$. Due to Theorem 2, $\forall t \leq t(X_0) - 1$, we have

$$f_{\min}(t) \geq \left(\sum_{j=1}^n D_j - n + 1 \right) H(x). \quad (31)$$

Due to (19),

$$f_{\max}(t) \leq \left(\sum_{j=1}^n D_j - n + 1 \right) H(x). \quad (32)$$

Thus for $t < l^{n-1} \sum D_j - nl^n - 1$, we have

$$f_{\max}(t) - f_{\min}(t + 1) < H(x). \quad (33)$$

Due to Theorem 1, X is inference-free. \square

Due to space limit, only a preliminary result is presented here. This result is effective enough for 2-dimensional data cubes. For data cubes with higher dimensionality, readers are referred to [15] for more complicated results to justify our algorithms in Sect. 3.

5. FINAL REMARKS

In this paper, we propose an information-theoretic inference control approach in OLAP systems. In comparison with previous approaches, we introduce an offline pre-calculation algorithm (which can be readily integrated into the materialization of cuboids) and an online query restriction algorithm to enhance the data availability while keeping sensitive data confidential. An upper bound on the cardinality of cells known by users through external knowledge is derived to make data cubes inference-free. Our work is preliminary and many extensions can be made. For example, we are currently investigating how to combine distortion-based and restriction-based inference control approaches. Distortion based approaches provide a higher data availability. However, it is restricted to specific applications (e.g., classification). Restriction-based approach, on the other hand, does not have such constraint. These two approaches are implemented on different layers in the OLAP system model. Thus, the combination of these two approaches is feasible and may lead to a more efficient inference control mechanism.

6. REFERENCES

- [1] N. R. Adam and J. C. Worthmann, "Security-control methods for statistical databases: a comparative study," *ACM Comput. Surv.*, vol. 21, no. 4, pp. 515–556, 1989.
- [2] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. of the ACM SIGMOD Conf. on Management of Data*. ACM Press, May 2000, pp. 439–450.
- [3] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in *Symposium on Principles of Database Systems*, May 2001, pp. 247–255.
- [4] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in *Proc. of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, June 2003, pp. 211–222.
- [5] J. M. Kleinberg, C. H. Papadimitriou, and P. Raghavan, "Auditing boolean attributes," in *Symposium on Principles of Database Systems*, 2000, pp. 86–91.
- [6] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in *Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, July 2002, pp. 217–228.
- [7] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in *Proceedings of 28th International Conference on Very Large Data Bases (VLDB 2002)*, August 2002, pp. 682–693.
- [8] J. Vaidya and C. Clifton, "Privacy preserving association rule mining in vertically partitioned data," in *Proc. of 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (KDD)*, July 2002, pp. 639–644.
- [9] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data," in *Proc. of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery (DMKD'02)*, June 2002, pp. 24–31.
- [10] W. Du and M. J. Atallah, "Privacy-preserving statistical analysis," in *Proc. of the 17th Annual Computer Security Applications Conf.*, New Orleans, Louisiana, USA, December 2001, pp. 102–110.
- [11] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *Newsletter of the ACM Special Interest Group on Knowledge Discovery and Data Mining*, vol. 4, no. 2, pp. 12–19, January 2003.
- [12] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in *Proc. of the ACM SIGMOD Conference on Management of Data*. ACM Press, May 2000, pp. 439–450.
- [13] R. Canetti, Y. Ishai, R. Kumar, M. K. Reiter, R. Rubinfeld, and R. N. Wright, "Selective private function evaluation with applications to private statistics," in *Proc. of 20th annual ACM symposium on Principles of distributed computing*. ACM Press, 2001, pp. 293–304.
- [14] L. Wang, D. Wijesekera, and S. Jajodia, "Cardinality-based inference control in sum-only data (extended version)," in *European Symposium on Research in Computer Security (ESORICS 2002)*, 2002, pp. 55–71.
- [15] N. Zhang, W. Zhao, and J. Chen, "A new scheme on cardinality-based inference control in OLAP systems," Texas A&M University, Tech. Rep. available at <http://www.cs.tamu.edu/people/nzhang/ANSCICOS.pdf>, 2004.
- [16] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, ser. Wiley Series in Telecommunications. New York, NY, USA: John Wiley & Sons, 1991.