第37卷 第17期
Vol.37 No.17

计 算 机 工 程
Computer Engineering

2011年9月
September 2011

• 安全技术 •

文章编号：1000—3428(2011)17—0105—03 文献标识码：A 中图分类号：TP309

基于最佳桶划分的数据库加密存储与查询

林艳纯，王晓明

(暨南大学计算机科学系，广州 510632)

摘 要：现有支持快速查询的数据库加密方案未给出字符字段的划分值提取方法。针对该问题，提出一个基于最佳桶划分的数据库加密存储与查询方案。该方案利用最佳桶划分思想提取索引字段的划分值，通过提取代加密字段的索引值以及转换查询数据库的 SQL 语句实现加密字符型数据的存储和查询。实验结果表明，该方案具有较高的首次查询命中率和稳定的过滤效率，能提高加密系统的整体性能。

关键词：数据库加密；加密存储与查询；最佳桶划分；密文索引；过滤效率

Database Encryption Storage and Query Based on Optimal Bucket Partition

LIN Yan-chun, WANG Xiao-ming

(Department of Computer Science, Jinan University, Guangzhou 510632, China)

【Abstract】The database encryption scheme which supports the rapid inquires does not give the partition value extraction method of character and field, in order to solve the problem, this paper presents a database encryption storage and query based on optimal bucket partition. It chooses partition value based on optimal bucket partition, and the query and storage for encrypted string data is according to extract index value of agent encryption field and converse SQL statement of inquiry database. Experimental result shows that this scheme has higher first inquiry shooting and stable filtering efficiency, it can improve the encryption system performance.

【Key words】database encryption; encryption storage and query; optimal bucket partition; ciphertext index; filtering efficiency

DOI: 10.3969/j.issn.1000-3428.2011.17.034

1 概述

随着数据库技术的不断发展，数据库系统被越来越多地应用到生活中，数据库很容易成为黑客攻击的目标。然而，普通的访问控制等安全措施已不能满足一些重要部门和一些敏感数据的安全。采用加密算法对数据库中数据进行加密后，数据的安全性虽然得到提高，但加密后的数据会丢失其原有的有序性、相似性、可比性等，限制加密数据的查询效率。因此，密文数据库的检索问题已经成为目前数据库研究的热点之一。本文分析文献[1]的快速查询的数据库加密方案，指出加密方法中对划分值提取的方法存在缺陷，并针对该缺陷，提出一个基于最佳桶划分的数据库加密存储与查询方案。

2 相关研究

查询密文数据的性能低下是在查询存储密文数据的数据库时所面临的最大问题之一。关于数据库加密的研究主要集中在以数值型数据作为加密对象方面，而针对与字符型数据的加密研究相对较少。

文献[2]提出一种基于桶索引的数据库加密方案，使用过滤技术缩小解密范围的方法，从而提高密文的查询效率，但该方法建立密文索引后，密文索引信息的泄漏比较大。文献[3]对文献[2]的分桶进行改进，提高查询命中率，并证明桶划分建立密文索引的安全性与对多个敏感字段的划分次数和划分桶的数目有关，但没有给出提高密文索引安全性的具体方案。文献[4]提出利用基于最佳桶划分策略建立密文索引，从而获得最优的查询命中率，平衡密文索引的安全性和查询效率的问题，但该方法主要集中在对数值型数据的加密研究。关于字符型数据的加密研究，文献[5]利用特征函数将字符串映射到一个二进制位串，并将属性值的相似性转化到二进制位串的比较，有效地实现加密数据的模糊匹配如 LIKE 运算，但对于加密数据的范围查询如大于、小于则无能为力。文献[6]对加密数据连接多个关键词的检索提出了有效的方案。

3 基于最佳桶划分的数据库加密存储与查询

3.1 文献[1]方案分析

文献[1]提出的加密字符串的索引字段包括划分值和特征值。划分值是指所有属性值域的一个映射，用于支持字符串数据的精确匹配查询，而特征值采用基于对偶编码函数的映射方法，用于支持字符串数据的模糊查询。然而，文献[1]方案对于划分值的设置并没用提出详细方案，只是指出数值型字符域值的划分通常采用等宽或等深划分的方法，对于字符型字段的域值划分则尽量使每个划分包含的元素数目大致相等，即等深划分。无论是对数值型的等宽划分，还是对字符型的等元素划分，都不能证明使首次查询所返回的结果中包含伪记录最少，且其过滤效率不能保持在稳定水平。

3.2 方案实现

针对文献[1]方案中对划分值提取方法的不足，本文提出基于最佳桶划分算法[3]的思想，改进其划分函数以提取索引的划分值，并根据新的索引字段划分值的提取方法更改查询

基金项目：国家自然科学基金资助项目(61070164)；广东省自然科学基金资助项目(8151063201000022)；广东省科技计划基金资助项目(2010B010600025)

作者简介：林艳纯(1988－)，女，硕士研究生，主研方向：数据库技术，信息安全；王晓明，教授、博士

收稿日期：2011-03-02 **E-mail：**lamsumg@gmail.com

数据库的方法。方案主要包括加密数据库中加密字符型数据的存储以及加密字符型数据的查询 2 个部分。对于加密字符型数据的存储，主要工作在于提取代加密字段的索引值；对于加密字符型数据的查询，主要工作在于对两阶段查询数据库 SQL 语句的转换。

3.2.1 加密字符型数据的存储

设原始关系模型为 $R(X_1,X_2,\cdots,X_r,\cdots,X_n)$。若 X_r 为需要加密的属性，则加密后将以关系模型 $R(X_1,X_2,\cdots,X_r^E,\cdots,X_n,X_r^S)$ 存储，其中，X_r^E 是 X_r 经过加密后的密文字符串；X_r^S 是根据 X_r 提取的索引字段。索引字段由待加密数据的划分值以及特征值组成。

算法 1 加密字符型数据的存储

输入 所有待加密的字符串 $str_i(1\leqslant i\leqslant N)$，划分数量为 M

输出 各字符串的密文 str_i^E，各字符串的索引字段 str_i^S

算法流程如下：

(1)统计所有待加密字段的字符串值 v_i 及其频率 f_i，得到 $D=(V, F)$，其中，$V=\{v_1,v_2,\cdots,v_N\}$；$F=\{f_1,f_2,\cdots,f_N\}$，并按字母升序(或降序)排序。

(2)使用文献[3]中的最佳桶划分算法作为划分函数 $QOB(D, M)$，得到最优划分情况下的最小代价值 $H[N][M]$。

(3)通过记录最佳桶划分情况的数组 H，得到 M 个划分区间 $\{p_1,p_2,\cdots,p_M\}$，通过散列函数为每个划分分配对应的标识符 $Ident(p_i)$。

(4)为每个待加密字符串 str_i 提取索引字段及加密字段：

1)利用第(2)步中的划分方案，对待加密字符串 str_i 提取其所属划分值，$str_i^{S1}=Map(str_i)$。

2)使用文献[1]中对偶编码函数把待加密的字符串映射到二进制字符串，从而提取 str_i 的特征值 $str_i^{S2}=PC(str_i)$。

3)输出划分值和特征值组成索引字段，$str_i^S=str_i^{S1}+str_i^{S2}$。

4)调用加密算法，输出各字符串的密文 $str_i^E=E(str_i)$。

3.2.2 加密字符型数据的查询

根据加密数据的存储方案，采用两阶段查询的方法实现加密字符串的 SQL 查询。首次查询是粗糙查询，主要是通过字符串的索引字段进行过滤，但过滤结果中可能包含“伪记录”。二次查询是精确查询，将解密首次查询过滤后所得的结果进行对应的 SQL 查询。本方案采用随机映射函数，查询 $X_i=v$ 时，对应映射函数为 $Map_{R.Xi}(v)=Ident_{R.Xi}(p_j)$，其中，$p_j$ 是包含 v 的划分。对于查询 $X_i>v$ 时，对应的映射函数应为 $Map_{R.Xi}^{>}(v)$，其中，映射值应该包含比 v 更大的值所有的划分 $\{p_i,p_{i+1},\cdots,p_j\}$。同理对于查询 $X_i<v$ 时，对应的映射函数应为 $Map_{R.Xi}^{<}(v)$，其中，映射值应该包含比 v 更小的值所有的划分 $\{p_i,p_{i+1},\cdots,p_j\}$。根据划分值映射的不同，对查询语句的转换规则做相应的修改，具体转换规则如下：

(1)精确匹配查询：如 $X_i>v$，$X_i=v$，$X_i<v$ 等。

如果查询条件为 $X_i=v$，则可基于索引字段对加密数据进行双重过滤。匹配其所隶属的划分值，及其所对应的特征值：

$$Trans(X_i=v)\rightarrow X_i^S=(Map_{R.X_i}(v)+PC(v))$$

如果查询条件为 $X_i>v$，可基于索引字段对加密数据进行双重过滤。匹配其所隶属的划分值，及其所对应的特征值：

$$Trans(X_i>v)\rightarrow X_i^S\ \ LIKE\ (Map_{R.X_i}^{>}(v)+(n\ bit)\text{'?'})$$

其中，'?' 表示单个字符的通配符。

同理，当查询条件为 $X_i<v$ 时：

$$Trans(X_i<v)\rightarrow X_i^S\ \ LIKE\ (Map_{R.X_i}^{<}(v)+(n\ bit)\text{'?'})$$

其中，'?' 表示单个字符的通配符。

(2)模糊匹配查询：如 $X_i\ LIKE\ c_1c_2\cdots c_k$ 等。

如果查询条件为 $X_i\ LIKE\ c_1c_2\cdots c_k$，则可基于索引字段的特征值对加密数据进行过滤：

$$Trans(X_i\ LIKE\ c_1c_2\cdots c_k)\rightarrow X_i^S\ LIKE\ s$$

$$s[i]=\begin{cases}\text{'1'} & (\exists j)(h+Hash(c_i,c_j)),\ i\leqslant j\leqslant k-1\\ \text{'?'} & \text{其他},1\leqslant i\leqslant h+m\end{cases}$$

其中，'?' 表示单个字符的通配符。

(3)复合查询：由 and 和 or 把 2 个或多个简单查询语句组合。

通过查询条件的转换，可转换相关复合查询语句：

$$Trans((X_i\ op_1\ v_1)\ \text{and}\ (X_i\ op_2\ v_2))\rightarrow Trans(X_i\ op_1\ v_1)\ \text{and}\ Trans(X_i\ op_2\ v_2)$$

$$Trans((X_i\ op_1\ v_1)\ \text{or}\ (X_i\ op_2\ v_2))\rightarrow Trans(X_i\ op_1\ v_1)\ \text{or}\ Trans(X_i\ op_2\ v_2)$$

其中，op_1、op_2 为查询逻辑运算符。

算法 2 加密字符型数据的查询

输入 SQL 查询语句，元数据(各划分所分配的字符，特征函数等)

输出 返回查询数据集

算法流程如下：

(1)首次查询，即粗糙查询，具体步骤为：

1)利用元数据(各划分所分配的字符，特征函数等)的规则，将原 SQL 查询中对加密字段的条件语句转换为其对应的索引字段条件值；

2)执行已经转换的 SQL 语句，抛弃过滤字段。

(2)二次查询，即精确查询，具体步骤为：

1)解密首次查询返回的记录，存储到临时表中。

2)转换 SQL 语句，将原始 SQL 语句中的关系表用临时表替代。

3)执行转换后的 SQL 语句，得到查询结果。

3.3 改进算法分析

改进算法的分析如下：

(1)划分桶数量的选取。对于划分桶数量的选取，应该针对待加密字段域值进行设定。从查询效率来说，提高划分桶数量的值有利于提高查询效率，但并不是划分桶数量越高越好。一方面是因为当划分桶数量提高到一定的值后其过滤效率只会停留在一定程度而不再上升，每个划分桶中包含多个不同的元素。另一方面因为如果刻意提高划分桶数量会使得索引字段的安全性降低，如果要使过滤效率达到 100%，那么划分值的设置不仅没有意义，还破坏了原方案的安全性，攻击者凭划分值即可推测原字符串。本方案划分桶数量的选取主要从两方面考虑：1)对应加密字段域值的设定；2)在条件 1)的前提下考虑最高可行的划分桶数量。

(2)正确性分析。在本方案的两阶段查询中，符合查询条件的记录不会在首次查询被过滤掉，而二次查询过滤掉的记录必定是不符合查询条件的记录，因此，查询处理结果一定包含了所需的正确集合。

(3)安全性分析。索引字段采用最佳桶划分的划分方法提取划分值后，每个划分中所含元素数量分布不一，每个划分中所包含元素所对应的频率大致相等，能有效抵抗统计攻击获得划分值和划分的对应情况，因此，一般情况下很难通过已知明文攻击获得划分值所对应的属性。

(4)查询效率分析。对比等宽或等深的划分方式，最佳桶划分查询代价是最低的，查询效率比其他 2 种方法高。从理论角度看，文献[3]提出以错检元组总数作为衡量桶划分策略

的参数，并已证明当划分桶区间包含了 N 个属性值时，错检元组总数为：

$$(N-1)\sum_{i=1}^{N} f_i = (N-1)\times F \approx N\times F$$

即在桶划分自区间数目被限定的情况下，使错检元组总数 $N\times F$ 最小，则查询的命中率最高。

通过计算划分的错检元组总数作为查询代价可比较基于最佳桶划分的过滤效率和等深划分的过滤效率。例如，设 D={(A,4), (B,4), (C,4), (D,10), (E,10), (F,4), (G,6), (H,2), (I,4), (J,2)}，其中，(A,4)表示输入中包含以 A 开头的字符串共有 4 个。设划分子区的上限数目为 N=4 个。采用最佳桶划分的划分方式得到分布为：{A,B,C}，{D,E}，{F,G}，{H,I,J}，其错检元组总数是 120。采用等深的划分方式时得到划分区间的分布为：{A,B,C}，{D}，{E,F}，{G,H,I,J}，其错检元组总数是 130。可见，采用最佳桶划分方式在查询时出现更少的错误元，具有更高的查询效率。图 1 是在随机 500 条字符串中，对照 2 种方式划分的查询代价。

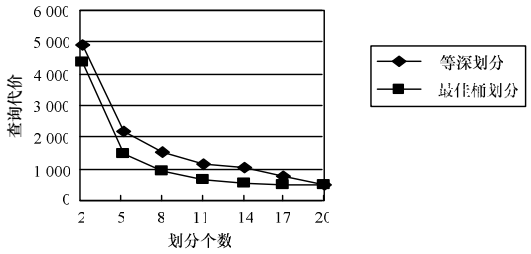


图 1　2 种划分方法的查询代价比较

4　仿真实验

仿真实验主要是测试采用不同划分方法建立索引的划分值，在查询时间代价与过滤效率等方面的性能，从而检验采用基于最佳桶划分提取划分值是否较传统的划分方法在查询效率上更优越。

实验仿真环境采用 Windows 7 操作系统，SQL Server 2008 数据库服务器。选用 SQL Server 中带有的示例数据库 AdventureWorks 中的数据表 Person.Address 前 1 000 条记录作为本次实验的数据源，并将 AddressLine1 字段作为敏感字段进行加密处理。对敏感字段的加密采用 AES 加密算法，密钥长度为 128 bit。

根据测试需求，实验分别采用等深划分、基于最佳桶划分以及先解密后查询的方法进行测试。对于等深划分、基于最佳桶划分的方法的测试，在特征值选取不变的情况下，分别用对应方法进行不同划分数量的加密存储与查询。每个实验数据都做 10 次，取其平均值。

实验 1　查询时间性能测试

实验 1 为了测试各种检索方法对应的查询时间性能，分别对基于最佳桶划分的检索方法、基于等深划分的检索方法以及传统先解密后查询的方法进行测试。为了简单起见，测试过程中只考虑查询条件 X_i=v 的情况，测试结果如图 2 所示。

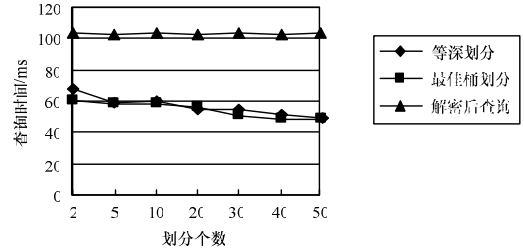


图 2　3 种划分方法的查询时间比较

由于先解密后查询的方法不存在划分的问题，因此无论划分值是多少对其时间性能上是一致的。先解密后查询的方法需要大量解密数据库中所有加密字段的数据，不仅系统开销大，花费的时间也很长。随着划分桶数量的增加，以上 2 种算法的查询用时明显下降，可见划分桶数量的增加明显提高了查询的时间性能。实验结果表明，基于最佳桶划分的检索方法和基于等深划分的检索方法相对查询花费的时间较少，且两者查询用时较接近。

实验 2　过滤效率测试

实验 2 通过测试 3 种不同的方法首次查询时对加密数据的过滤能力。

测试原理是：设数据表中共包含记录数 N 条，在首次查询时过滤后共 n_1 条记录，二次查询后得到共 n_2 条记录，则过滤效率=$(N-n_1)/(N-n_2)$。首次查询过滤记录数目为 $N-n_1$ 条，二次查询后过滤记录 $N-n_2$ 条。

实验分别对基于最佳桶划分和基于等深划分的存储方法进行提取索引值的加密数据进行查询，记录首次查询所得记录数量，再记录二次查询所得记录数量，从而计算出查询的过滤效率。为了简单起见，测试过程中只考虑查询条件 X_i>v 和 X_i<v 的情况，经过对比得到图 3。

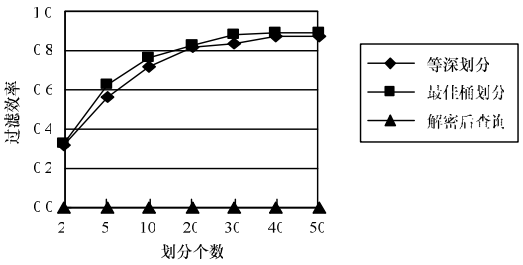


图 3　3 种划分方法的过滤效率比较

由于先解密后查询的方法是无所谓桶的划分，该方法是将所有数据先解密后再根据查询语句筛选记录，因此该方法不具备过滤功能，过滤效率为 0。由图 3 可以看出，基于最佳桶划分的检索方法以及基于等深划分的检索方法均具备对加密数据的过滤能力。随着划分数量的增加，两者的过滤效率有所提高。采用最佳桶划分比等深划分的过滤效率高，这说明在二次查询时，采用基于最佳桶划分的检索方法较采用基于等深划分的检索方法需要解密的伪记录少，从而减少解密开销。最佳桶划分的方法相比等深划分的方法具有更稳定的过滤效率。

以上实验分别对基于最佳桶划分的检索方法、基于等深划分的检索方法以及传统先解密后查询的方法进行查询时间性能以及过滤效率的测试，用实例验证了基于最佳桶划分的检索方法较传统的先解密后查询方法以及基于等深划分的检索方法更具优越性。体现在：基于最佳桶划分的检索方法较传统的解密查询方法具有更优的查询时间性能；在查询加密数据时，首次查询中表现出更高的过滤效率。实验结果验证了采用最佳桶划分的方法提取索引划分值，能提高首次查询时的命中率，减少首次查询中伪记录的出现，从而提高加密系统的性能。

5　结束语

本文通过采用基于最佳桶划分的方法改进划分值的提取方式，实验结果表明采用最佳桶划分比等深划分的过滤效率更高效稳定。今后将研究适用于动态数据的划分，并进一步

(下转第 110 页)