



遗传算法在关系型数据库水印技术中的应用研究

丁霞

(福州职业技术学院 福建, 福州 350003)

Research on the application of genetic algorithm in relational database watermarking technology

Dingxia

(Fuzhou Polytechnic, Fujian, Fuzhou 350003)

Absrtact: at present, the current watermarking algorithm in relational database has been unable to meet the needs of its basic data management. In the process of watermark embedding, the number of tuples processed by the traditional genetic algorithm is far lower than the number of tuples actually existing in the database, so it has not been widely used in practical applications. This paper will deeply analyze and research the existing watermarking technology, and improve the relational database watermarking technology based on genetic algorithm. This technology uses genetic algorithm to select the optimal tuples in the relational database for matching, which significantly shortens the time of relational database watermark embedding, improves the efficiency of watermark embedding, effectively optimizes the relational database watermark, and makes a positive contribution to the development of relational database.

Keywords: genetic algorithm; Watermark technology; Database technology; matching

摘要: 目前, 关系数据库中现行的水印算法已无法满足其基础数据管理的需求。传统遗传算法在水印嵌入过程中, 其处理的元组数量远低于数据库中实际存在的元组数量, 因此并未在实际应用中得到广泛应用。本文将对现有的水印技术进行深入分析与研究, 并在遗传算法的基础上对关系数据库水印技术进行改进。该技术利用遗传算法在关系数据库中筛选出最优元组进行匹配, 显著缩短了关系数据库水印嵌入的时间, 提高了水印嵌入的效率, 有效地优化了关系数据库水印, 为关系数据库的发展做出了积极贡献。

关键字: 遗传算法; 水印技术; 数据库技术; 匹配

1 引言

随着计算机的进步, 数据库技术有了迅猛的发展, 数据库管理系统也逐渐获得了更多的重视与应用。然而, 数据库的安全性是每个用户都必须面对的问题。当我们在 Internet 网络上开放特定用户远程访问和查询权限时, 越来越多的黑客能够远程操作或窃取我们的数据库^[1]。为了减少和避免此类问题, 数据库水印技术应运而生。如今,

数据库水印优化方案在图像领域已备受关注并取得了发展。但同时, 由于关系数据库数据的特殊性, 其水印技术研究也面临一定难题。从目前的研究情况看, 国内外技术部门对该研究仍处于探讨阶段。毕竟遗传算法在数据库的水印嵌入方面缺乏相关科研技术, 尤其是关系数据库中的元组数量极为庞大, 水印嵌入的元组个数难以满足需求, 且嵌入过程复杂, 需耗费大量时间^[2]。因此,



本文将在相关研究基础上，提出一种新型的基于遗传算法的关系数据库水印技术，通过仿真发现，该算法能有效缩短嵌入时间。

1 关系数据库水印技术

在关系数据库里占据的数据量较大，不过其冗余度数值很小。为保证水印信息嵌入能和数据融合，且满足算法的可行性，我们得改进现有的算法。在某些情形下，关系数据并非全部具备数据型属性，还会夹杂一些音频、图像或视频等非数值型属性的数据^[3]。这种情况下，为进一步开展关系数据库水印技术的研究，我们可假设关系数据表中的所有数据都是数值型，在数据的可容忍范围内，关系数据表的属性值表示为能融合水印嵌入并接受数值改动^[4]。

1.1 水印嵌入

由水印密钥、分组数目以及关系数据共同构成属性水印和元组水印，属性水印在图中对数据列进行标识，元组水印则对数据行进行标识，进而形成一个水印矩阵，为数据在计算过程中的定位修改提供数据证明。水印嵌入的工作原理如图 1 所示。

由图 1 我们能够了解到，水印密钥在水印技术研究里起着关键的参数作用，它和分组数目一同决定了元组划分结果，同时也关联着水印信息的生成状况^[5]。

1.2 水印的提取与检测

水印检测模块涵盖提取水印和生成水印。水印的数据必须借助水印检测算法生成，才可以实现自检模式，水印生成方式和嵌入方式大致相同。关系数据库水印的提取与检测原理如图 2 所示。

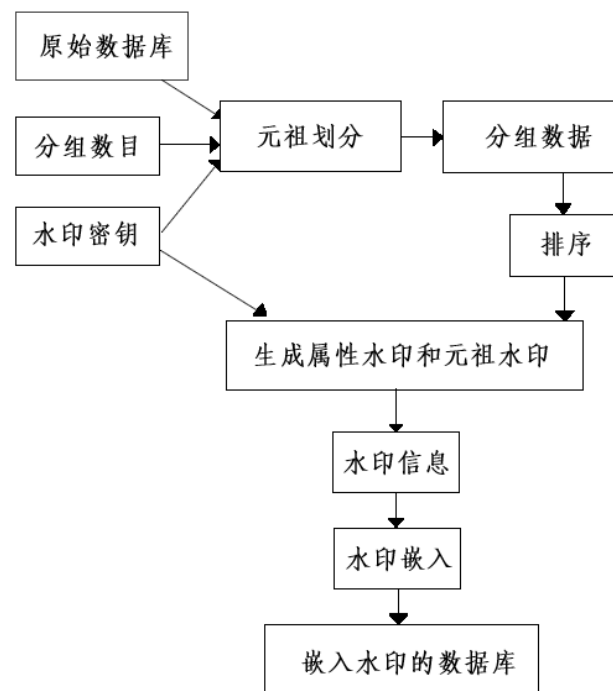


图 1 数据库水印嵌入

从图中能够看出，水印嵌入过程存在特定的数据顺序，不过这些数据在实际嵌入时通常处于离散状态^[6]。此时，若要保证水印信息的提取和嵌入顺序同步，就必须确保两者分组时的步调一致。所以，数据拥有者需在获取水印密钥和分组数目之后，才能开展水印信息处理工作。

2 基于遗传算法的水印技术

遗传算法也称作 GA (Genetic Algorithm)，它是一种不直接针对具体问题进行搜索的方式。该理论最早由 Holland 于 1975 年提出，是模拟达尔文遗传机制与自然生物进化过程，进行全程随机搜索并找到最终解决方案的一种优化算法。改进后的算法能出色地对数据库进行优化，大幅缩短了关系数据库的水印嵌入时间，提升了水印嵌入



效率,切实实现了关系数据库水印的优化^[7]。

2.1 算法描述

假定关系数据库里的数据有少量属性值发生改变,但这不会影响数据库的有用性。把数字水印信号和文本水印信号建立对应关系,接着将数字水印信号数值通过水印嵌入算法,用可忍受的误差值来评估,水印即为关系数据库的误差值。假设每个元组中仅有一个属性可用于水印信息嵌入,

计算共分四个步骤:(1)开展水印预处理,将水印转换为二进制比特流后进行分组设计,同时进行纠错编码整理;(2)利用遗传算法筛选最佳元组进行水印嵌入,并依据水印的信息容量选择嵌入元组数量,最后通过遗传算法挑选最优元组。(3)对元组进行标记并分组,借助密钥 K_s 和前一步骤筛选出的元组主属性,给元组分组排序。(4)通过元组排序分组,把水印信息嵌入元组分组之中。

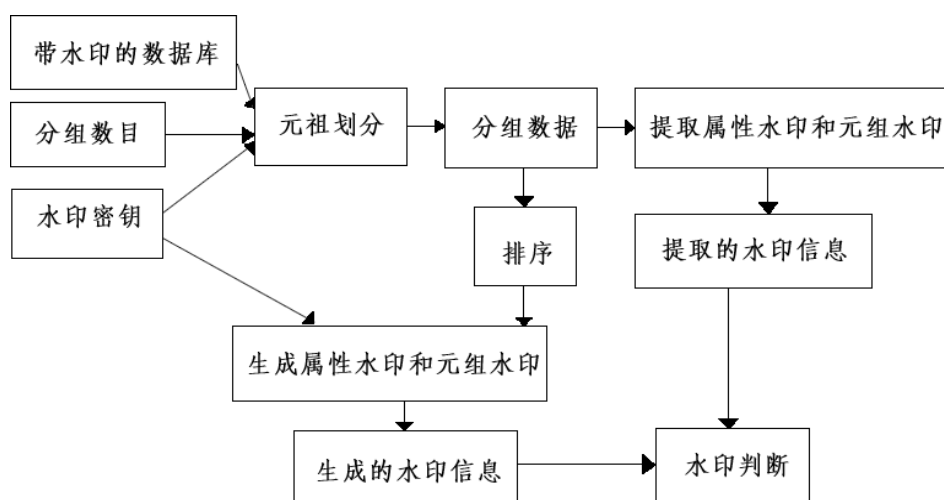


图2 数据库水印提取与检测

2.2 水印预处理

如何将文本水印转换成计算机能识别的二进制码非常重要,水印预处理就是要解决这一问题。首先用一字符 t 代表水印,对照 ASCII 码表将字符 t 转换成一串二进制代码,将这一串二进制码用 w' 表示, w' 长度为 $L=|w'|$ 。然后将一串二进制码分组,每三个二进制码为一组,最后剩余不够三个为一组的,进行补位操作,补上一个或两个 0 数字,并采用 3_bit_set 数据进行比特组显示^[8]。

水印嵌入与水印提取是一组互为反向操作的过程,用信息接收代表水印提取,用信息发送代表水印嵌入。因此,在水印预处理方面,我们将采用通讯方式进行,在处理过程中提取纠错编码技术进行分组纠错,修复数据。稳定水印嵌入中的 Robust。在文中,我们采用的是偶校验编码技术对信息码元尽心错误修整^[9]。

$$S = a_{n-1} \oplus a_{n-2} \oplus \dots \oplus a_0$$

其中 S 为校正子,又被称之为伴随式。 $S=0$ 正解, $S=1$ 错误。假设编码中有信息码元 3 位,分别

为 $a_0a_1a_2$, 监督码元 3 位,分别为 $a_3a_4a_5$ 。根据多数判决原则与异或运算法则,我们可以列出偶校验的一致监督关系。

2.3 遗传算法优化选择水印嵌入元组

为减少数据的大规模修改,本文选择一定元组数量的水印信号来延长数据库的可用性。在选择多大的元组数量上,我们参照了文献资料,选择比数据库的元组更小的数量来进行试验,真正实验了数据的大规模修改,然后通过遗传算法选择最优水印嵌入元组数据^[10]。

(1) 确定适应度函数

$$fitness(d_i) = \frac{\delta_i \times \sum_{j=1}^M |d_j^2 - d_j^2|}{\varphi_i}$$

公式中为 ψ_i 元组 i 的权重系数; δ_i 为元组 i 在 d 范围内可容忍的最大误差; M 为数据库元组个数。

(2) 初始化数据,随机选用 $r \in R$ 且 $rl > \xi$ (rl 表示元组 r 有效位长度)来产生第一代种群染色体,



用 R' 代表了嵌入水印最终得到的最优元组集合。

(3) 采用轮盘选择操作, 先对初始群体进行轮盘选择, 保留质优个体。

(4) 进行交叉运算, 通过一定交叉概率来对个体采用单点交叉运算小生境遗传算法中交叉概率多为常数, 维持在 0.5~0.8 之间。

(5) 变异运算: 按照一定概率对单个染色体进行二进制的变异运算, 变异因子范围在 0~1 之间, 变异率维持在 < 0.5 。

(6) 小生境淘汰运算法: 将 M 个个体与 N 个个体进行何静, 得到 $M+N$ 个个体的新群体, 并通过 $M+N$ 个个体求出 X_i 和 X_j 之间距离 $\text{Dist}(i,j)$, $\text{Dist}(i,j) < L$ 时, 进行个体适应度大小比较, 对适应度低的个体进行罚函数处理。

$$F_{\min}(X_i, X_j) = \text{Penalty}$$

(7) 进行终止条件判断, 依照 $M+N$ 个个体适应度分别对个体进行降序排列, 记忆前的 N 个个体进行终止条件判断处理, 促进循环的终止条件有以下两种: 一是在种群中各异适应度的方差小于设定值则终止, 二是设定了最大终止代数并满足条件。条件满足的前提下, 进行结果计算输出, 若条件不满足时, 进行进化代数计数器, $t=t+1$, 并将 M 个个体作为下一代群体 $P(t)$ 选择^[11]。进行 $N \times \theta$ 个适应度函数值最大的染色体选择, 该染色体即为待嵌入水印最优元组集合 R' 。

2.4 元组分组

将元组按照三个为一组进行分组, 同时每个分组包括三个不同的属性。分组个数用公式表示为 $m=N \times \theta / 3s$, 在每个子集的属性中进行水印信号的嵌入, S 表示水印嵌入次数。

2.5 水印嵌入元组分组

$\text{Watermark}(\text{attribute}, Ks, \text{text}, r, p)$

$t \leftarrow \text{text_to_flow}(\text{text})$ // 将文本水印转换为二进制

$w \leftarrow \text{stream_W}(t, Ks)$ // 将文本水印转换为比特流

$W' \leftarrow \text{group_triple}(w)$ // 将初始比特流分组, 3 比特 / 组

$n \leftarrow \text{length}(w) / 3$ // n 为分组个数

$W' \leftarrow \text{Even}(W')$ // 对分组比特流 W' 进行偶校验编码

$R' \leftarrow \text{select}(R)$ // 用遗传算法求出最优嵌入元组

$V_i \leftarrow \text{Vector_attribute}(\text{attributes}, 3, Ks)$ // 属性组成向量, 每个向量都包含了 3 个属性

foreach vector of $V1$

foreach attribute attri of $V1$

foreach 3_bit_set of W' {

$i = 3\text{bit_to_int}(\text{eger}(3_bit_set))$

$\text{LSD}(\text{attri}) = (\text{MSD}(\text{attri}) + i) \bmod lo$

}

3 算法仿真

该实验需在处理器为 Intel2.4G 赫兹、电脑运行内存为 256 兆、硬盘总容量为 80GB, 且以 Windows XP 作为主要操作系统的环境下开展^[12]。RAD (快速应用开发) 环境处于 Java Eclipse Platform Version 3.0 系统里, 借助 ODBC (开放数据库互连) 连接 SQL Server 2003, 嵌入水印密钥是 sanxilifen, 分 8 个小组, 将可嵌入水印最小重要位数设为数字 2^[13]。以 SQL Server 2003 环境为依托构建数据库关系表, 表内所有数据皆为数值型, 还涵盖了主码等 6 种属性。实验期间持续通过数据变动对属性值修改、元组添加或删除以及属性列删除实施模拟攻击, 实验结果显示, 基于改进的遗传算法, 针对实验中的四类攻击, 数据均出现定位反应, 具备描述能力^[14]。ga 函数的迭代次数为 20 代, 海明距离设定的最大值为 1, 仿真算法时运用遗传算法的经验值来选取个体变异概率与交叉概率值, 值 $(Pc, Pm) = (0.65, 0.05)$, 罚函数值取经验值 $\text{Penalty} = 1/60$ 。通过实验可得知, 遗传算法对于各类常用数据库攻击实践有较强的抵抗性^[15]。

参考文献:

- [1] 张明远, 李成刚. 基于改进遗传算法的关系数据库数字水印优化方法 [J]. 计算机研究与发展, 2023, 60(5): 1123-1134.
- [2] 王雪峰, 刘洋. 遗传算法在关系型数据库水印嵌入中的参数优化研究 [J]. 计算机应用研究, 2022, 39(8): 2456-2462.
- [3] 陈志强, 赵芳芳. 基于多目标遗传算法的关系数据库鲁棒水印技术 [J]. 计算机科学, 2021, 48(10): 328-335.
- [4] 孙丽娜, 吴晓明. 遗传算法优化的关系数据库水印抗攻击性研究 [J]. 通信学报, 2020, 41(12): 156-165.



[5] 杨雪峰, 周涛. 基于遗传算法和离散余弦变换的关系数据库水印算法[J]. 计算机工程与应用, 2022, 58(15): 102–110.

[6] 刘佳明, 胡斌. 遗传算法在大型关系数据库水印容量优化中的应用[J]. 电子学报, 2021, 49(7): 1421–1429.

[7] 黄敏华, 林峰. 基于改进遗传算法的分布式数据库水印同步技术[J]. 计算机集成制造系统, 2020, 26(5): 1265–1274.

[8] 徐阳光, 郑洁. 遗传算法优化的关系数据库动态水印更新策略[J]. 计算机应用, 2023, 43(4): 1145–1153.

[9] 李娜娜, 王磊. 基于遗传算法和混沌加密的数据库水印系统设计[J]. 计算机工程, 2021, 47(9): 154–162.

[10] 赵鑫鑫, 马红梅. 遗传算法在时空数据库水印中的优化应用[J]. 计算机应用与软件, 2020, 37(6): 315–322.

[11] 吴迪迪, 周明明. 基于遗传算法和脆弱水印的关系数据库篡改检测[J]. 计算机科学, 2022, 49(5): 298–306.

[12] 刘芳芳, 陈刚. 遗传算法优化的多级关系数据库水印嵌入策略[J]. 计算机工程与设计, 2021, 42(8): 2214–2220.

[13] 孙伟伟, 张丽丽. 基于遗传算法和数字指纹的数据库版权保护方法[J]. 计算机应用研究, 2019, 36(12): 3675–3682.

[14] 胡静静, 杨帆. 遗传算法在云数据库水印分发优化中的应用[J]. 计算机工程与科学, 2020, 42(10): 1842–1850.

[15] 郑阳光, 刘伟. 基于自适应遗传算法的关系数据库可逆水印技术[J]. 计算机研究与发展, 2023, 60(3): 678–688.

作者简介: 丁霞 (1993 —), 女, 汉族, 福建福州人, 硕士, 福州职业技术学院讲师, 研究方向: 算法。