Computer Science and Application 计算机科学与应用, 2021, 11(4), 795-802
Published Online April 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.114081

Hans汉斯

基于自适应嵌入的高抗毁数据水印技术研究

邹云峰[1]，胡　俊[2]，单　超[1]，于鹏飞[3,4]

[1]国网江苏营销服务中心(计量中心)，江苏　南京
[2]安徽继远软件有限公司，安徽　合肥
[3]全球能源互联网研究院有限公司，江苏　南京
[4]信息网络安全国网重点实验室，江苏　南京
Email: yupengfei@geiri.sgcc.com.cn

收稿日期：2021年3月9日；录用日期：2021年4月6日；发布日期：2021年4月13日

摘　要

国家大数据战略下，数据需要各个环节间流转、重组、使用，同时数据的管理控制权也随之转移，数据一旦泄露，数据提供者非常希望能够精确溯源是否是数据使用环节的责任。数据水印技术通过高仿真实数据内容，并参入标识信息，数据使用者不易察觉，并且有很强的隐蔽性和抗毁性，可有效满足数据安全流转需求。文章提出了一种基于自适应嵌入的高抗毁数据水印方法，可灵活针对目标数据的特性，在保证目标数据可用性的前提下，尽可能的保证数据水印嵌入率，提高数据水印的抗毁性。

关键词

数据安全，数据水印，精准溯源

Research on High Invulnerability Data Watermarking Technology Based on Adaptive Embedding

Yunfeng Zou[1], Jun Hu[2], Chao Shan[1], Pengfei Yu[3,4]

[1]State Grid Jiangsu Marketing Service Center (Metrology Center), Nanjing Jiangsu
[2]Anhui Jiyuan Software Co., Ltd., Hefei Anhui
[3]Global Energy Interconnection Research Institute Co., Ltd., Nanjing Jiangsu
[4]State Grid Key Laboratory of Information & Network Security, Nanjing Jiangsu
Email: yupengfei@geiri.sgcc.com.cn

Received: Mar. 9th, 2021; accepted: Apr. 6th, 2021; published: Apr. 13th, 2021

文章引用: 邹云峰, 胡俊, 单超, 于鹏飞. 基于自适应嵌入的高抗毁数据水印技术研究[J]. 计算机科学与应用, 2021, 11(4): 795-802. DOI: 10.12677/csa.2021.114081

Abstract

Under the National big data strategy, data needs to be transferred, reorganized and used among every chain. At the same time, the management control of data is also transferred. Once the data is leaked, the data provider hopes to be able to accurately trace whether it is the responsibility of the data use link. Data watermarking technology can effectively meet the needs of data security circulation by highly imitating the real data content and adding identification information, which is not easy to be detected by data users, and so data watermarking technology has strong concealment and invulnerability. This paper proposes a high invulnerability data watermarking method based on adaptive embedding, which can flexibly face to the characteristics of the target data, and then ensure the embedding rate of the data watermark as much as possible and the availability of the target data, and also improve the invulnerability of the data watermark.

Keywords

Data Security, Data Watermarking, Accurate Traceability

1. 引言

随着国家大数据战略的发展，数据要素已经成为我国经济发展的重要驱动力。2020 年 4 月 9 日，《中共中央、国务院关于构建更加完善的要素市场化配置体制机制的意见》发布，要求加快培育数据要素市场，推进政府数据开放共享，提升社会数据资源价值，加强数据资源整合和安全保护[1]。我国关键基础设施运营企业，比如能源、金融、电信等行业拥有海量的客户数据，不仅反映了社会用能、经济运行等宏观情况，而且反映了居民生活、企业经营等具体情况，价值巨大，社会开放共享需求迫切。然而近年来数据安全事件频繁发生[2]，国家法律法规对于个人信息保护越来越严格[3] [4] [5]，因此我们必须在保障数据安全的前提下，输出数据价值。

我国关键基础设施运营数据不仅会在内部各个部门间流转，也会向工商、司法等政府部门，金融、能源、互联网公司等企事业单位提供。数据一旦对外提供，数据的管理控制权将随之转移，如果发生数据泄露，责任难以定位。《中华人民共和国网络安全法》规定网络运营者应当按照网络安全等级保护制度要求，防止网络数据泄露或者被窃取、篡改[6]。《数据安全管理办法(征求意见稿)》规定网络运营者有履行数据安全保护义务[7]。如果责任不能厘清，则极大限制了数据管理者的数据共享积极性，因此采取必要的技术手段，尽可能避免数据泄露，加强数据泄露之后的定责能力，十分必要。

文章提出了一种基于自适应嵌入的高抗毁数据水印方法，并采用当前主流的微服务技术架构，针对数据离线对外提供场景，提供高抗毁、高仿真的数据水印，一旦数据被恶意泄露，即可根据数据水印，定位责任环节，同时数据水印也可警示数据使用人员的安全意识，防止数据管理不当，非法泄露。

2. 相关知识

2.1. 概述

数据水印技术又被称为数字水印、数据库水印，能够将特定的标识信息嵌入到数字产品的内容当中，

例如数据库、文档、软件等，同时不影响原数字产品的使用价值，也不容易被第三方探知和破坏。当数字产品流转给用户使用时，用户只能在许可范围内使用该数字产品，不能再次转发给第三方使用，这既符合绝大部分商品销售的商业规则，也是数据广泛开发共享的基本保证。使用的数据水印技术，不仅是对数据权属的标识，而且可以准确的进行数据泄露责任追溯，保护数据提供者的利益。

2.2. 技术现状

2002 年 Agrawal 和 Kiernan 首次提出数据库水印的概念，以此来标记哪些是数值型的数据[8]。随着数据水印技术的发展，总的来说，大致可分为失真和无失真的两类水印[9]。数据水印的嵌入首先应该根据不同业务的差异化需求，尽量保证数据的可用性；其次数据水印应该具有完整唯一标识，保证数据泄露时溯源的准确性；最后数据水印应该具有很强的隐蔽性和抗毁性，保证数据在流转过程中，水印不会被攻击破坏。为了同时实现上述三个目标，国内外学者提出多种数据水印方法。一种思路是基于差异扩展的数据水印，同时考虑失真约束，控制数据水印的嵌入率，例如 Bhattacharya 和 Cortesi 提出的将元组划分置换后嵌入水印方法[10]。为了优化嵌入效果，有学者提出了遗传算法和直方图移位算法[11]，但是前期计算量非常大。

为了更好的保证数据水印嵌入率，同时保证数据水印的隐蔽性和抗毁性，文章针对使用范围较广的数值型数据，开展了基于自适应嵌入的高抗毁数据水印技术研究，包括数据水印自适应嵌入，以及数据水印精准提取，并得到了实验验证。

3. 算法设计

3.1. 自适应分组

为保证水印更加均匀地分散在目标数据中，我们设计了自适应分组算法，根据不同数据库的内容特性，计算最优的分组方式，具体算法如下：

$$A_g = \left\{ x_{i,j} \mid \text{if } hash\left(k_i\right) \bmod \left(\tau \cdot 2^m\right) = g \right\} \tag{3.1}$$

其中，$x_{i,j}$ 代表第 i 行第 j 列的属性值，k_i 代表序号为 i 的元组主键值，τ 代表编码后的水印码字长度，$g \in \left\{0,1,\cdots,\tau \cdot 2^m - 1\right\}$ 为分组序号，参数 m 的取值范围由数据水印位数和目标数据量之间的幂的逆运算值取整确定，一般为自然数。

通过上式，我们把第 j 列中所有属性项分为 $\tau \cdot 2^m$ 组，每一组用来嵌入一位水印信息，很明显，当 m 大于 0 的时候，每一位水印嵌入到了 2^m 分组上。为了评估 m 的取值对于最终抗毁性的影响，我们定义了计算分组的离散度。

我们首先对第 j 列属性按照主键值进行排序：

$$\left\{x_{1,j}, x_{2,j}, \cdots, x_{n,j}\right\} \to \left\{x_{p(1),j}, x_{p(2),j}, \cdots, x_{p(n),j}\right\} \tag{3.2}$$

保证 $k_{p(k_1),j} \ge k_{p(k_2),j}$, if $p\left(k_1\right) > p\left(k_2\right)$，$p(i)$ 代表排序后的第 i 个元组。排序后按照下式计算分组的离散度：

$$C_m = \sum_{t=1}^{\tau} \frac{\sum_{x_{i,j} \in AA_t} \left(p(i) - \bar{P}\right)^2}{\left|AA_t\right|}{t}, \quad \text{其中} AA_t = \left\{ A_g \mid \left\lfloor \frac{g}{2^m} \right\rfloor = t \right\} \tag{3.3}$$

其中，$\bar{P}$ 表示所有元组排序号的均值，A_g 的定义如式(3.1)。

最后，在根据下式选择使 C_m 最大的 m 值，计算最终的分组个数 S_w 。

$$S_w = \text{数据水印位数} \times 2^m, \text{if } \max\{C_m, m = 0,1,2,\cdots\} \tag{3.4}$$

S_w 即为自适应分组算法的分组结果，该分组方式能够保证水印在数据库中的分布更加均匀和离散，能够有效抵抗删除、更新等攻击。

3.2. 自适应嵌入强度

为保证数据库水印的抗毁性，特别是对于删除、更新等攻击的抵抗能力，文章通过计算适宜的水印嵌入强度，在单个属性项中嵌入合适的多位水印，在提高水印嵌入率的同时又兼顾了数据可用性，保证数据失真较小。

我们根据每一列的属性项的两个统计特性来计算嵌入强度：

$$E_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{3.5}$$

E_j 表示第 j 列属性值的均值，n 表示第 j 列属性列的属性个数。显然，均值越大，我们的修改量也可以更大，且不引起明显失真。

$$D_j = \frac{\sum_{i=1}^{n}\left(x_{ij} - E_j\right)^2}{n} \tag{3.6}$$

D_j 表示第 j 列属性值的方差，该值是用来判断该列的数值抖动情况，如果 D_j 的值较小，说明该列数值变化平稳，为了避免明显失真，应使修改量较小；反之，可以嵌入较大容量的水印信息。

针对上述统计特性，我们设计了强度判断公式：

$$e_j = \begin{cases} 1, & \text{if } E_j \in [0, g_1) \\ 2, & \text{if } E_j \in [g_1, \infty) \end{cases}, \quad d_j = \begin{cases} 1, & \text{if } D_j \in [0, g_2) \\ 2, & \text{if } D_j \in [g_2, \infty) \end{cases} \tag{3.7}$$

上式中，e_j 为均值条件下的嵌入强度等级，d_j 为方差条件下的嵌入强度等级，g_1 和 g_2 为设定的阈值，可根据实际情况进行调整。

$$t_j = \min\{e_j, d_j\} \tag{3.8}$$

t_j 为最终确定的嵌入强度等级，为保证数据失真较小，我们取均值条件下和方差条件下较低的那一个等级作为最终的强度等级。

3.3. 数据水印提取

数据水印的精准提取，关键在于数据水印标识位的识别，标识位识别出来后，就可以根据既定规律，准确提取出数据水印。由于在提取端，无法得知属性列是否嵌入水印，提取标识位是完全盲提取，因此需要对属性列进行遍历查询，确定是否存在标识位。

对于第 j 列中的一个属性项 $\hat{x}_{i,j}$,首先通过下式计算其标识位信息：

$$l_{i,j} = \begin{cases} \text{error}, & \text{if } \operatorname{mod}\left(x_{i,j}, Q\right) \neq 0; \\ 0, & \text{if } \operatorname{mod}\left(x_{i,j}, Q\right) = 0 \text{ and } \operatorname{mod}\left(\dfrac{x_{i,j}}{Q}, 2\right) = 0; \\ 1, & \text{if } \operatorname{mod}\left(x_{i,j}, Q\right) = 0 \text{ and } \operatorname{mod}\left(\dfrac{x_{i,j}}{Q}, 2\right) = 1; \end{cases} \tag{3.9}$$

其中 Q 为数值型数据水印的量化公约数，原始属性项 $x_{i,j}$ 通过 Q 的映射，得到数据水印标识。如果 $x_{i,j}$ 与 Q 取模不等于 0，说明该属性项不是标识为，即为 error；如果 $x_{i,j}$ 与 Q 取模等于 0，可进一步进行取模运算，判断该位标识是 0 还 1。

为对 j 列中所有属性项通过上式(3.9)计算其标识位信息，然后进行标识位统计：

$$ST_\lambda = \frac{\sum_{\mathrm{mod}(Hash(k_i),S_l)=\lambda} l_{i,j}}{\left|\mathrm{mod}\left(Hash\left(k_i\right),S_l\right)=\lambda\right|} \tag{3.10}$$

其中||运算符表示返回集合中有效数值的个数，$\lambda \in [0, S_l - 1]$用于标记提取出来的 S_l 位标识位，而统计完成之后，第 λ 位标识位通过下式计算获得：

$$L_\lambda = \begin{cases} 0, & \text{if } ST_\lambda < \text{标识位为0的判断阈值} \\ 1, & \text{if } ST_\lambda > \text{标识位为1的判断阈值} \end{cases} \tag{3.11}$$

如果是自然采集得来的数据，ST_λ 值是接近近于 0.5 的。当进过标识位映射之后，ST_λ 值要么接近于 0，要么接近于 1；当数据水印受到局部破坏时，ST_λ 值会向 0.5 靠拢。

通过遍历匹配，可以完整的提取标识位，识别当前属性列是否含有标识位信息。

4. 软件设计及应用结果分析

4.1. 微服务能力设计

在互联网+新业态下，服务商为了追求客户粘性，业务升级和迭代迅速，因此相关能力微服务化设计成为业界趋势。文章提出的自适应嵌入的高抗毁数据水印溯源算法，也设计了微服务能力，供前端服务或者其他后台能力调用，设计架构如下图 1。

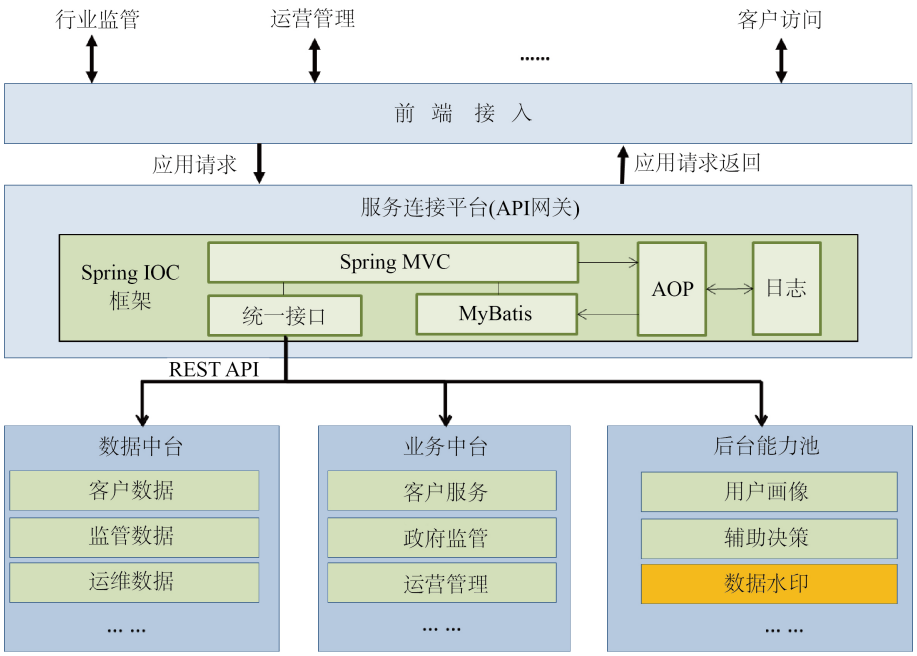

Figure 1. Design of micro-services for data watermark traceability
图 1. 水印溯源能力微服务设计

自适应嵌入的高抗毁数据水印溯源算法模块在微服务软件框架下，可以以一个独立的模块部署在虚拟化容器中，并与数据流转的申请审批、脱敏处理、日志审计应用协同工作。当前端应用发起一个业务请求时，由业务中台中的应用服务按需调用数据中台数据和后台能力池中功能，如果是向外流转数据，在准备数据的同时，调用数据水印能力进行数据处理，然后流转数据。当安全人员接到数据泄露举报时，可以通过运营管理接口，调用数据水印溯源能力，精准定位是哪一位数据使用者泄露了数据。

4.2. 模拟验证

文章模拟了电力营销客户数据流转共享的场景，比如数据分析人员进行客户经营分析的业务，在数据外发前添加数据水印，如果向外流转的数据发生泄露，验证是否可以通过已添加的数据水印进行溯源。

(1) 验证环境

模拟验证的软硬件环境，包括 1 台关系数据库服务器，1 台应用服务器(部署数据水印溯源服务)，和 1 台导出文件服务器用于临时存储生成的数据文件(表 1)。

Table 1. Software and hardware verification environment
表 1. 模拟验证软硬件环境

序号	名称	设备类型	设备配置	操作系统	中间件
1	应用服务器	X86 服务器	2台 PC 服务器 8核/16 GB 内存/50 GB	CentOS7.5	——
2	导出文件服务器	X86 服务器	1台 PC 服务器 8核/16 GB 内存/500 GB	CentOS7.5	Tomcat 8.5.54
3	数据库服务器	X86 服务器	2台 PC 服务器 8核/32 GB 内存/500 GB	CentOS7.5	Tomcat 8.5.54

(2) 模拟样本数据

模拟样本数据主要采用个人基础信息和用电数据，包括姓名、电子邮箱、手机号码、用户名、密码，还包括客户当日用电量、当月累计用电量、本年度每月的用电量等等。样本数据的来源主要通过各属性的特点由代码随机自动生成，共计 20000 条。部分样例数据如下表 2。

Table 2. Some verification sample data
表 2. 部分模拟样本数据

No.	Name	Email	TEL	Today EC	This month EC	Jan EC	Feb EC	MarEC	Apr EC
1	吴刚	12700350@qq.com	18643950959	15.7	287	292	260	271	……
2	秦精鸿	327985541@qq.com	13851699236	9.6	216	380	363	356	……
3	章洪家	1009576282@qq.com	13853922283	3.2	19	306	283	272	……
4	于嘉禾	yjm981174@163.com	13513536249	23.1	448	481	457	441	……
5	胡晓南	huxnan@yahoo.com.cn	13776768541	11.2	215	297	289	291	……
6	程光	88780488@qq.com	15050806332	13.4	238	286	283	258	……
7	魏巍	13500357664@163.com	13500357664	18.9	315	317	313	298	……
8	王涛	540928109@qq.com	13813969224	2.3	16	185	190	152	……
9	憨志刚	HZG2006@163.com	13611506820	13.1	242	324	294	307	……
10	刘奕	592424917@qq.com	17023328401	15.2	264	356	327	318	……

(3) 水印添加及溯源结果

为了验证数据水印的隐蔽性，针对上述数据文件添加数据水印，并流转至数据运维人员、数据分析人员、业务运营人员、客户服务人员和安全管控人员，数据分析生成的相关统计曲线在容忍的精度范围之内，相关人员均未发现数据水印。

为了验证溯源的准确性及抗破坏性，针对上述数据文件的六种数据泄露情况进行溯源验证：

第一种情况：全量的 20000 条已经添加水印的数据文件发生泄漏；

第二种情况：已下载的数据文件随机 10000 条数据发生泄漏；

第三种情况：已下载的数据文件随机 1000 条数据发生泄漏；

第四种情况：已下载的数据文件随机 100 条数据发生泄漏；

第五种情况：已下载的数据文件随机 10 条数据发生泄漏；

第六种情况：已下载的数据文件随机 5 条数据发生泄漏。

每种数据泄露情况重复验证 100 次，局部泄露的数据范围随机选择，可以当数据全量泄露，或者少量泄露(大于 100 条)，溯源的准确率均达到 100%。而当泄露数据为 5 条时，数据水印遭到破坏，溯源的平均准确率约为 60% (表 3)。

Table 3. Data watermark tracing results
表 3. 数据水印溯源结果

序号	泄露数据量	平均准确率
1	泄露全量数据(20000条)	100%
2	泄露10000条数据	100%
3	泄露1000条数据	100%
4	泄露100条数据	100%
5	泄露10条数据	92%
6	泄露5条数据	60%

5. 总结

文章面向数据共享流转业务场景，开展了基于自适应嵌入的高抗毁数据水印方法，可灵活针对目标数据的特性，在保证目标数据可用性的前提下，尽可能的保证数据水印嵌入率，提高数据水印的抗毁性。经设计原型验证，本文提出的水印溯源技术具有较好的溯源准确性以及抗破坏性。文章研究成果针对数值型数据，虽然大部分数据分析人员的数据需求在于数值统计分析，然而随着社会数据安全意识的提高，以及国家个人信息保护法律法规的陆续出台，文本型数据和自然语言型数据也将是数据流转中安全防护的重点，开展文本型数据和自然语言型数据的水印溯源算法研究将是需要进一步研究的方向。

基金项目

文章研究成果由国网江苏省电力有限公司科技项目“电力营销敏感数据安全防护关键技术研究及应用”(项目编号 Grand No.J2020007)支持。

参考文献

[1] 中共中央国务院. 关于构建更加完善的要素市场化配置体制机制的意见[N]. 新华社北京, 2020-4-9(电).

[2] 张衡. 大数据安全风险与对策研究——近年来大数据安全典型事件分析[J]. 信息安全与通信保密, 2017(6):

102-107.

[3] 张晓娟, 王文强, 唐长乐. 中美政府数据开放和个人隐私保护的政策法规研究[J]. 情报理论与实践, 2016, 39(1): 38-43.

[4] 安小米, 穆勇, 王薇,等. 我国涉及隐私的个人信息保护与管理法律法规状况及要求分析[J]. 北京档案, 2011(5): 19-21.

[5] 王铮, 曾萨, 安金肖, 等. 欧盟《一般数据保护条例》指导下的数据保护官制度解析与启示[J]. 图书与情报, 2018(5): 119-125.

[6] 中华人民共和国网络安全法(2016 年 11 月 7 日第十二届全国人民代表大会常务委员会第二十四次会议通过)[EB/OL]. http://www.npc.gov.cn/wxzl/gongbao/2017-02/20/content_2007531.htm, 2017-02-20.

[7] 中华人民共和国司法部.《数据安全管理办法(征求意见稿)》[EB/OL] http://www.moj.gov.cn/news/c-ontent/2019-05/28/zlk_235861.html, 2019-08-28.

[8] Agrawa, R. and Kiernan, J. (2002) Chapter 15-Watermarking Relational Databases. *VLDB* '02: *Proceedings of the* 28*th International Conference on Very Large Databases*, Hong Kong, 20-23 August 2002, 155-166.

[9] Halder, R., Pal, S. and Cortesi, A. (2010) Watermarking Techniques for Relational Databases: Survey, Classification and Comparison. *Journal of Universal Computer Science*, **16**, 3164-3190.

[10] Bhattacharya, S. and Cortesi, A. (2013) A Distortion Free Watermark Framework for Relational Databases. *Proceedings of the International Conference on Software and Data Technologies*, **2**, 229-234.

[11] Hu, D.H., Zhao, D. and Zheng, S.L. (2019) A New Robust Approach for Reversible Database Watermarking With Distortion Control. *IEEE Transactions on Knowledge and Data Engineering*, **31**, 1024-1037. https://doi.org/10.1109/TKDE.2018.2851517