

基于区块链的数据溯源可信查询方法

张学旺^{1,2}, 冯家琦¹, 殷梓杰¹, 林金朝^{1,2}

1. 重庆邮电大学软件工程学院, 重庆 400065

2. 重庆大学微电子与通信工程学院, 重庆 400044

摘要: 针对区块链数据溯源系统中轻节点验证溯源信息时面临的存储问题, 引入一种优化 Merkle 树动态追加性能的数据结构 Merkle 山脉 (Merkle mountain range, MMR), 将区块链上完整的区块头存入 MMR; 提出一种溯源数据高效可信的验证方法, 降低了区块包含证明所需信息的大小。在此基础上, 设计了一种基于区块链的数据溯源系统的方案, 将数据溯源所需的通用模块封装起来, 通过接口开放给溯源应用调用。该方案只需轻节点存储一个最新区块的信息, 就可以有效地验证溯源信息是否存在于区块链上。

关键词: 区块链; 数据溯源; Merkle 山脉; Merkle 树

中图分类号: TP301

文章编号: 0255-8297(2021)01-0042-13

Trusted Query Method for Data Provenance Based on Blockchain

ZHANG Xuewang^{1,2}, FENG Jiaqi¹, YIN Zijie¹, LIN Jinzhao^{1,2}

1. School of Software Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

2. School of Microelectronics and Communication Engineering, Chongqing University, Chongqing 400044, China

Abstract: In order to reduce the storages needed in verifying provenance information of light clients in blockchain data provenance system, this paper firstly introduces a data structure titled by Merkle mountain range (MMR), which optimizes the dynamic append performance of Merkle trees and stores all block headers on blockchain in the MMR. Then we propose an efficient and reliable verification method for data provenance to reduce the size of the proof information required for the proof of the block contain. On this basis, a scheme of data provenance system based on block chain is designed, which encapsulates the common modules required for data provenance and opens them to the provenance application through interfaces. This scheme enables light clients to effectively verify whether the

收稿日期: 2020-11-12

基金项目: 国家重点研发计划基金 (No.2019YFC1511300); 工业和信息化部 2020 大数据产业发展试点示范项目基金 (工信厅发函〔2020〕47号); 重庆市技术创新与应用发展专项重点项目基金 (No.cstc2020jscx-fyzxX0026); 渝北区大数据智能化科技专项重点项目基金 (No.2020-02) 资助

通信作者: 张学旺, 副教授, 研究方向为数据安全与隐私保护、大数据与智能数据处理、区块链与物联网、网络通信软件。E-mail: zhangxw@cqupt.edu.cn

provenance information is contained in the block chain as long as they keep the information of the latest block in storage.

Keywords: blockchain, data provenance, Merkle mountain range (MMR), Merkle tree

随着信息技术的快速发展,数据的产生和流转快速增长。海量的数据通过网络在不同应用系统之间共享融合,使得数据的产生方式和流转方式呈现出多样性、复杂性的特点。对于一些对数据可信性和完整性要求较高的领域,如食品安全、司法取证、版权管理等,在查看和操作数据时都要验证数据的可信性和完整性。如果缺少对数据流过程的记录,就不易发现数据被非法篡改,也不能保障数据的可信性和完整性。因此,利用数据溯源技术记录和追踪数据的流过程变得尤为重要。

数据溯源可以理解作为一种用来记录数据流过程中变化的技术。Wang 等^[1]在研究多源异构数据库之间的数据流转时,首次提出了一种可以解决“数据来自何处”以及“在追溯数据的过程中使用了哪些中间数据”这两个问题的方案,从此开启了数据溯源的研究。Buneman 等^[2]正式定义并使用了数据溯源术语,丰富了数据溯源的含义,并从“why”和“where”两个角度阐述数据溯源。

传统数据溯源系统的存储和管理方式都是中心化的,尽管这种方式具有查询速度较快和使用较简单等优点,但是存在着单点问题和管理权限中心化问题,可能会破坏溯源信息的完整性和可信性。引入区块链技术后将数据溯源信息存储在区块链中,凭借区块链的去中心化、时序数据、集体维护、可编程和安全可信等特点,可以有效地保障数据溯源信息的可信性和完整性^[3]。文献[4]提出了一种基于区块链技术的云数据溯源模型框架 ProvChain,将溯源数据存储存储在区块链中,保证了数据的安全性。文献[5]设计了一个基于区块链的云服务提供商之间的数据共享模型,可以为云计算提供数据溯源和验证方法。文献[6]将开放溯源模型(open provenance model, OPM)与区块链结合起来,设计并实现了一个科学溯源管理模型—SmartProvenance。文献[7]利用区块链技术为物联网中的射频识别(radio frequency identification, RFID)数据设计了一种安全的数据溯源模型,提高了RFID数据的安全性。文献[8]提出了一种双链模型,也就是将实体多维授权信息和动态数据分别存储在两条链上,并在此基础上设计了一种物联网动态数据信息的分层溯源机制。文献[9]提出了一种基于区块链的去中心化数据溯源方法,同时设计了一套溯源数据管理的智能合约,将溯源数据存储到区块链上,可以确保用户获得的溯源数据真实可靠。

区块链技术的引入虽然解决了溯源信息的可信存储和管理问题,但是还不能有效解决如手机、物联网设备等资源受限的轻节点在不可信网络中验证溯源信息所需的证明数据过大的问题。在现有的区块链系统中,轻节点验证一笔交易的有效性需要借助简单支付验证(simplified payment verification, SPV),而SPV需要验证者(轻节点)拥有完整的区块头信息才可以完成验证。一条区块链中区块头的数量是随着区块的高度呈线性增长的,因此对于一个区块高度过高的区块链,维护其完整的区块头信息会占用轻节点较多的存储资源,显然SPV不太适用于溯源系统中轻节点的溯源信息验证。

本文旨在解决在区块链溯源系统中轻节点客户端验证溯源信息需要占用过多存储资源的问题,并提出一种基于区块链的数据溯源系统的通用实现方案。

1 相关技术

1.1 区块链

区块链由化名“中本聪”的学者于2008年提出并在比特币中实现^[10],是一种分布式共享

数据账本技术,具有去中心化、不可篡改性、时序性等特点。

1) 去中心化

在区块链网络中的每个节点都共同管理维护着一份数据账本,账本上的每一笔记录都要经过全网节点的合法性校验;一笔交易只有得到足够多节点的背书认证才会写入区块链。

2) 不可篡改性

区块链以链式结构存储数据,除了创世区块外,每个区块都存储了上一个区块的 Hash 值。如果上游区块 A 中的数据被修改,其 Hash 值随之改变,那么下游区块 B 中的上一个区块 A 的 Hash 值也会发生改变,区块 B 的 Hash 值也就相应地发生改变。

3) 时序性

每个区块中的时间戳字段赋予区块链时序性,使得区块链中存储数据具有良好的溯源性。

区块链中的每个区块都分为区块头 (block head) 和区块体 (block body),其中区块头中包含前一区块的 Hash 值 (prev hash)、区块高度 (height)、时间戳 (timestamp)、随机数 (nonce) 以及 Merkle 树根 (Merkle root) 等字段,而完整的 Merkle 树则存放在区块体中,如图 1 所示。

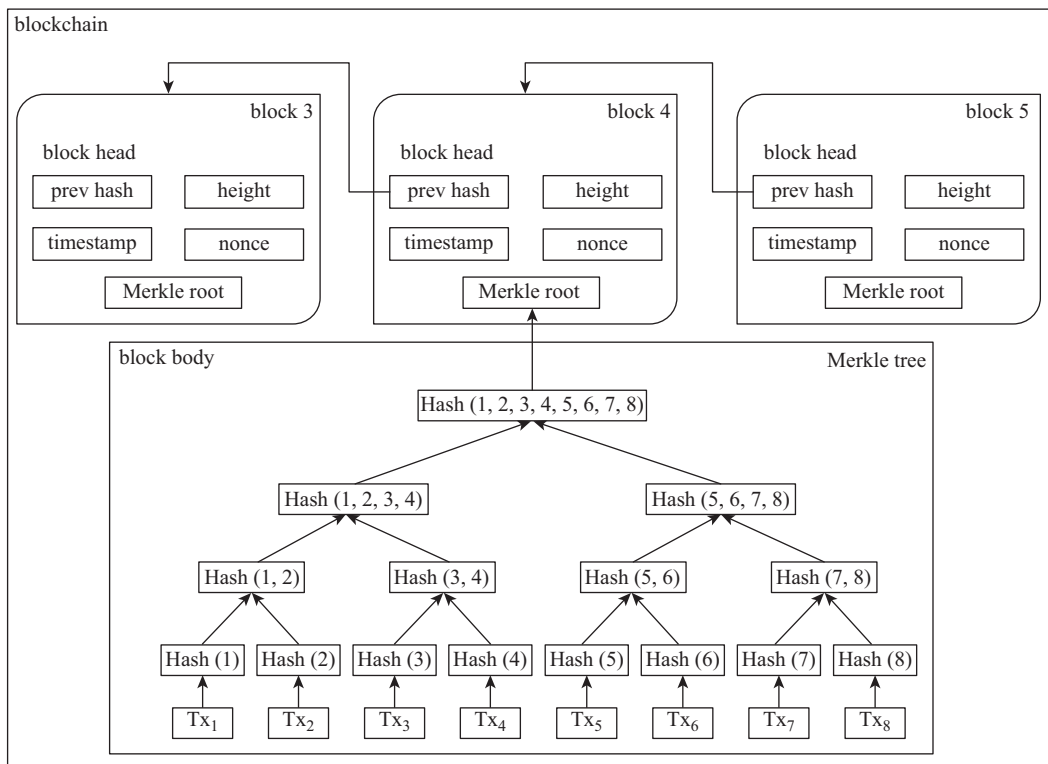


图 1 区块数据结构

Figure 1 Block data structure

1.2 Merkle 树

Merkle 树 (Merkle tree, MT) 是 Ralph Merkle^[11] 提出的一种二叉 Hash 树。Merkle 树的结构如图 2 所示,其叶子节点存储数据元素的 Hash 值,而除叶子节点之外的节点存放的都

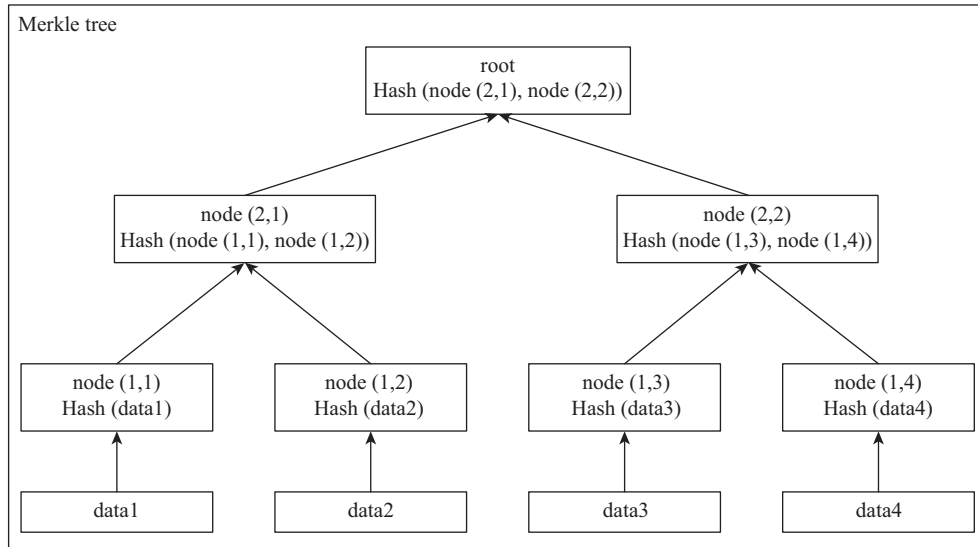


图2 Merkle 树

Figure 2 Merkle tree

是左右子节点的 Hash 值。节点中 Hash 值的计算公式为

$$\text{node}_{i,j} = \text{Hash}(\text{node}_{i-1,2j-1}, \text{node}_{i-1,2j}) \quad (1)$$

式中, $\text{node}_{i,j}$ 表示第 i 层 (自下而上) 的第 j 个节点 (从左向右) 的值。

根据式 (1) 依次上溯便可以生成一棵 Merkle 树, 而 Merkle 树的根节点 (Merkle root) 便是整棵树的摘要。

Merkle 树主要有 2 个优点: 1) 对 Merkle 树中任意数据的篡改都会导致该节点的父节点及父节点以上的节点发生改变, 最终导致根节点发生变化。2) 在无需存储数据集的情况下, 可以验证指定数据元素是否存在于数据集中。

Merkle 树广泛应用于无法保证数据可信性的 P2P 网络, 例如在区块链平台中使用 Merkle 树来存储交易信息, 并利用 Merkle 证明来验证一个交易是否存在于区块链中。Merkle 证明的详细流程参见 2.1 节。

1.3 Merkle 山脉

Merkle 山脉 (Merkle mountain range, MMR) 是 Peter Todd^[12] 提出的一种 Merkle 树的变种数据结构。在 Merkle 山脉中, 除叶子节点外其余节点的值都是其左右子节点的 Hash 值。Merkle 山脉可以提供类似于 Merkle 证明的 Merkle 山脉证明, 能够证明一个叶子节点是否存在于 Merkle 山脉之中。与 Merkle 树的不同之处在于: Merkle 树是一棵完美二叉树, 而 Merkle 山脉允许出现不完美的情况, 或者说 Merkle 山脉本身是由多个完美二叉树组成的数据结构。将 Merkle 山脉设计成仅追加的数据机构, 在插入时无需重构 Merkle 山脉。这个特点非常适用于对区块链中区块头的承诺, 只要构建包含全部区块头的 Merkle 山脉, 就可以证明一个区块是否存在于区块链中。

如图 3 (a) 所示, 在一个 Merkle 山脉中可以有多个子 Merkle 树 (称为山峰), 这些 Merkle 树的根称为峰顶; 多个山峰合起来形成一片山脉, 这也是 Merkle 山脉的由来。在 B_7 后追加

一个新区块 B_8 ，可以发现 Merkle 山脉的 3 座山峰变成了整个山峰，这是因为在 Merkle 山脉中出现两个高度相同的山峰会触发合并操作。合并操作是对两个高度相同山峰的峰顶节点进行 Hash 运算后生成一个新节点，该节点的左右子节点是需要合并的两座山峰的峰顶节点。一旦追加 B_8 ， B_7 和 B_8 两个区块就会触发合并操作，合并成峰顶为 H_5 、高度为 2 的新山峰。因为已经存在一个高度为 2、峰顶为 H_4 的山峰，所以继续触发合并操作。依次执行合并操作直到没有两个高度相同的山峰为止，最终形成一座如图 3 (b) 所示的山峰。

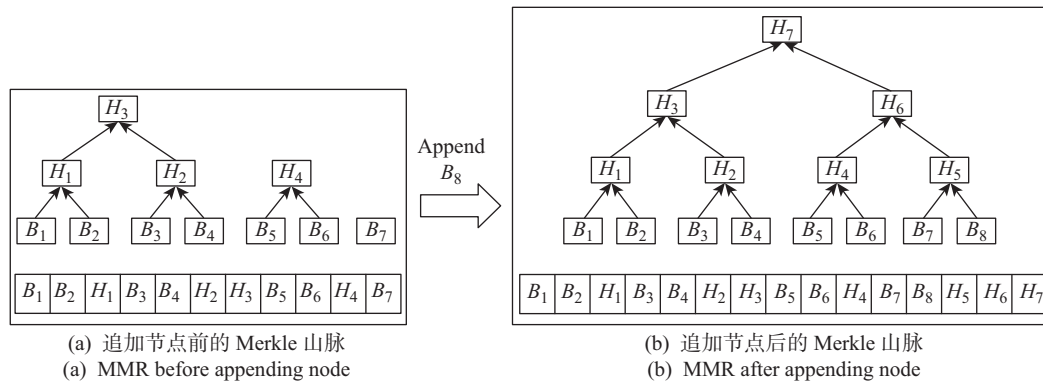


图 3 Merkle 山脉追加节点

Figure 3 MMR append node

当 Merkle 山脉中只有一座山峰时，峰顶节点是 Merkle 山脉的根 Hash。如图 3 (a) 所示，在 Merkle 山脉中有多个山峰的情况，此时需将多个峰顶节点的 Hash 值合并，得到一个新的 Hash 值，这个 Hash 值是 Merkle 山脉的根 Hash。例如计算图 3 (a) 中 Merkle 山脉的根 Hash，需依次计算 $\text{Hash}(H_3, \text{Hash}(B_7, H_4))$ ，获得 Merkle 山脉的根 Hash。

2 溯源数据的高效可信验证方法

在基于区块链的数据溯源系统中，全节点在本地维护了完整的区块链，因此可以独立地验证数据溯源信息的完整性和真实性。随着时代的发展，溯源数据验证的发起设备越来越倾向于智能手机、物联网设备等轻节点。这些设备的网络带宽、存储等硬件资源有限，不能像全节点一样实时维护完整的区块链信息。区块链系统如比特币、以太坊等都为轻节点验证交易信息提供了 SPV 方式，但 SPV 不太适合于溯源系统中轻节点的溯源信息验证。例如在以太坊中一个区块头的大小约为 508 字节，截止到 2020 年 9 月，以太坊区块的高度将近 1 080 万，全部区块头大概有 5G。维护如此庞大的区块头信息对于轻节点来说非常不友好。为解决此问题，需要分析以 SPV 方法验证一条溯源数据的真实性和可信性的过程，验证过程可以分为两个步骤：

步骤 1 交易包含证明。采用 Merkle 证明就可以证明一条溯源数据是否存在于一棵 Merkle 树中，从而证明这条溯源数据是否存在于包含这棵 Merkle 树的区块中。

步骤 2 区块包含证明。将包含待验证溯源信息的区块头与本地维护的完整区块头中对应高度的区块头进行对比，从而判断包含待验证溯源信息的区块是否存在于区块链中。

步骤 1 是为了验证一条溯源信息是否存在于一个区块中，步骤 2 是为了验证一个区块是否属于区块链。在验证过程中，交易包含证明占用的资源并不多；区块包含证明部分的验证需要完整的区块头信息，这会占用过多的设备资源。文献 [13] 在 FlyClient 中采用 Merkle

山脉高效承诺机制, 将区块包含证明所需的完整区块头信息降低到对数级别。本文将借鉴 FlyClient 使用的 Merkle 山脉承诺机制完成区块包含证明, 降低了基于轻节点验证溯源信息所需要的资源, 从而提高了验证效率。

下面分别从交易包含证明和区块包含证明两方面介绍基于 Merkle 山脉承诺机制的 SPV 方法。

2.1 交易包含证明

交易包含证明的核心是基于 Merkle 树的 Merkle 证明, 其过程可以通过如下的例子加以说明。要证明图 4 中的 Tx_6 是否属于 Merkle 树, 需要提供图 4 中圈出的节点, 这些节点组成的集合就是 Merkle 证明集合。下面给出寻找 Merkle 证明集合的步骤。

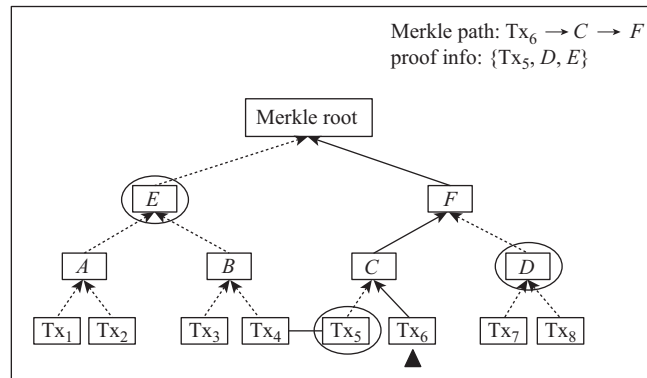


图 4 Merkle 证明

Figure 4 Merkle proof

步骤 1 从待验证的叶子节点依次向其父节点上溯, 直至到达 Merkle 树的根节点, 经过的路径称为 Merkle 路径 (Merkle path)。

步骤 2 将 Merkle 路径上所有节点的兄弟节点按照从下而上的顺序添加到 Merkle 证明集合中。

经计算可以得出 Tx_6 的 Merkle 路径为 $Tx_6 \rightarrow C \rightarrow F$, 如图 4 中实线部分所示。这条路径上所有节点的兄弟节点集合为 $\{Tx_5, D, E\}$, 此集合就是 Tx_6 的 Merkle 证明。将集合中的节点与 Tx_6 一起进行如下运算:

$$\text{Merkle root}' = \text{Hash}(E, \text{Hash}(\text{Hash}(Tx_5, Tx_6), D)) \quad (2)$$

最终可以获得 $\text{Merkle root}'$, 再将它与 Merkle 树真实的 Merkle root 进行比较, 如果相等则证明 Tx_6 存在于这棵 Merkle 树中且没有被篡改。

根据上述例子可以分析出 Merkle 证明的大小等于 $m - 1$, 其中 m 表示 Merkle 树的高度。因为 Merkle 树的叶子节点的数目为 $N = 2^{m-1}$, 所以可以推导出拥有 N 个叶子节点的 Merkle 树所包含证明信息的大小为 $\text{lb } N$ 。

2.2 区块包含证明

交易包含证明可以快速轻量地验证一笔交易是否存在于区块中是因为采用了 Merkle 树的承诺机制。在区块包含证明中, Merkle 树数据结构并不适合区块头的承诺, 这是因为随着区块链系统的运行, 区块的高度不停地呈线性增长, 而在 Merkle 树中追加数据需要重构

Merkle 树。若在区块包含证明中引入 Merkle 证明,就需要引入支持动态追加的 Merkle 山脉承诺机制对区块头进行承诺。每当有新的区块加入区块链,只要在 Merkle 山脉中动态追加新区块即可,而不必重构 Merkle 山脉。

引入 Merkle 山脉对区块链数据结构的改动很小,只要在区块头中增加一个 Merkle 山脉的根 Hash 即可。如图 5 所示,将从创世区块 $block_1$ 开始到 $block_{(n-1)}$ 为止的全部区块头一起组成 Merkle 山脉,并将 Merkle 山脉根 Hash 写入 $block_n$ 的 MMR root 字段。只要 $block_n$ 之前任意区块头的内容发生改变都会引起 $block_n$ 中 MMR root 值的改变。利用 Merkle 山脉承诺机制对区块头承诺后,就可以根据 Merkle 山脉证明完成区块包含证明。

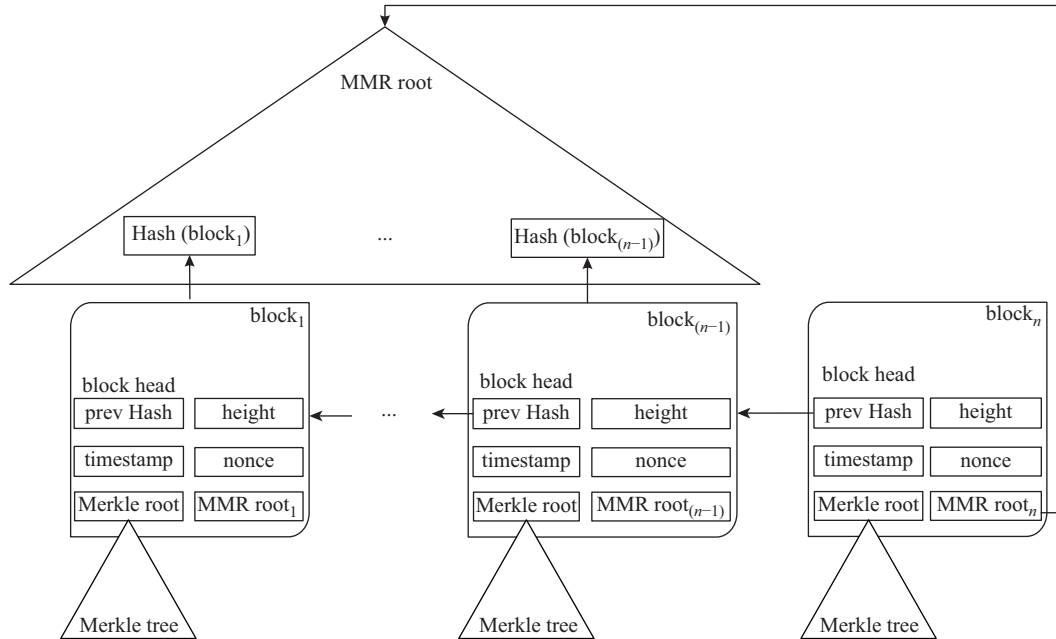


图 5 在区块链中添加 Merkle 山脉

Figure 5 Add MMR to blockchain

下面的例子可以说明如何利用 Merkle 山脉证明来完成区块包含证明的过程。如图 6 所示,若证明区块头 B_9 存在于区块链中,就需要提供图 6 中圈出节点的 Hash 值。这些圈出的节点集合称为 Merkle 山脉证明集合。

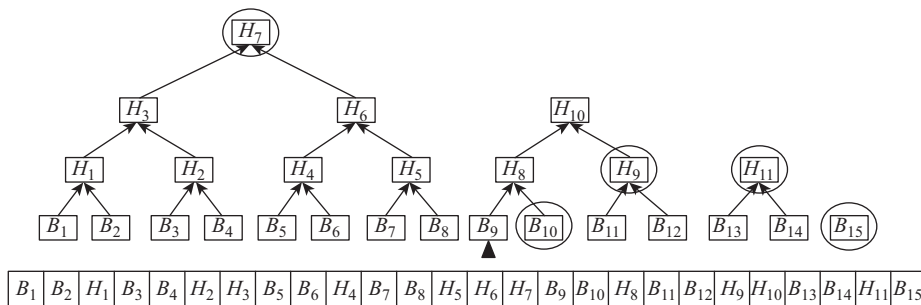


图 6 Merkle 山脉证明

Figure 6 MMR proof

下面给出寻找 Merkle 山脉证明集合的步骤。

步骤 1 从待验证的节点出发,自下而上地向其所在山峰的峰顶上溯,则经过的路径称为 Merkle 山脉路径 (MMR path)。

步骤 2 将 Merkle 山脉路径上所有节点的兄弟节点按照自下而上的顺序添加到 Merkle 山脉证明集合中。

步骤 3 将待验证节点所在的右侧山峰的峰顶加入 Merkle 山脉证明集合。

步骤 4 将待验证节点所在的左侧山峰的峰顶加入 Merkle 山脉证明集合。

对于寻找图 6 中 B_9 的 Merkle 山脉证明集合来说,步骤 1 可以得到 Merkle 山脉路径为 $B_9 \rightarrow H_8$,步骤 2 得到 Merkle 山脉证明集合为 $\{B_{10}, H_9\}$,步骤 3 得到 Merkle 山脉证明集合为 $\{B_{10}, H_9, H_{11}, B_{15}\}$,步骤 4 可以获得最终的 Merkle 山脉证明集合为 $\{B_{10}, H_9, H_{11}, B_{15}, H_7\}$ 。将 Merkle 山脉证明集合与待验证的区块头 B_9 一起进行如下运算:

$$\text{MMR root}' = H(H_7, H(H(B_9, B_{10}), H_9), H(H_{11}, B_{15})) \quad (3)$$

式中, $H()$ 表示 Hash 函数。最终可以根据 Merkle 山脉证明集合生成 Merkle 山脉的根 Hash 为 $\text{MMR root}'$,将它与真实的 MMR root 进行比较,如果相等则证明 B_9 存在于区块链中且没有被篡改。

本文借鉴了 FlyClient^[13] 中的轻节点认证方法,将 SPV 中的区块存在证明所需要的区块头数量 N 降低到 $\ln N$ 。轻节点只需在区块链网络中同步一个最新的区块头,再根据区块头中 MMR root 的值并结合区块包含证明和交易包含证明就可以有效地验证区块链上是否包含一笔交易。

3 基于区块链的数据溯源系统设计

3.1 系统架构

在第 2 节的 Merkle 树和 Merkle 山脉承诺机制的基础上设计一个基于区块链的通用数据溯源系统,系统架构如图 7 所示,自底向上依次分为区块链层、溯源模型层、通用溯源业务层和应用层。

1) 区块链层

将区块链的相关业务封装在该层,包括用于加解密算法和 Hash 算法的密码算法模块、用于保持网络中各个节点账本一致性的共识机制模块、用于区块链网络中各节点间通信的 P2P 网络模块以及用于账本持久化到本地存储(区块链文件系统或数据库系统)的数据持久化模块等区块链相关模块。

2) 溯源模型层

将 ProVOC 溯源模型融入区块链溯源,借助 ProVOC 模型中各种构件来描述数据溯源记录,并将溯源信息序列化 JSON 格式存储在区块链中。

3) 通用溯源业务层

为区块链溯源系统提供通用的业务服务,如对执行实体和查询用户进行权限认证的身份管理模块、将溯源数据提交到区块链中的溯源数据提交模块、查询区块链中溯源数据的查询模块以及验证溯源数据真实性和可靠性的验证模块等。在通用溯源业务层将这些模块封装起来,并以 API 接口的形式开放给溯源应用调用。

4) 应用层

本层与具体的溯源领域对接,根据溯源领域具体的需求和业务场景编写应用程序。在应用中调用通用溯源业务层提供的 API 接口以实现溯源数据上链、查询、验证等操作。

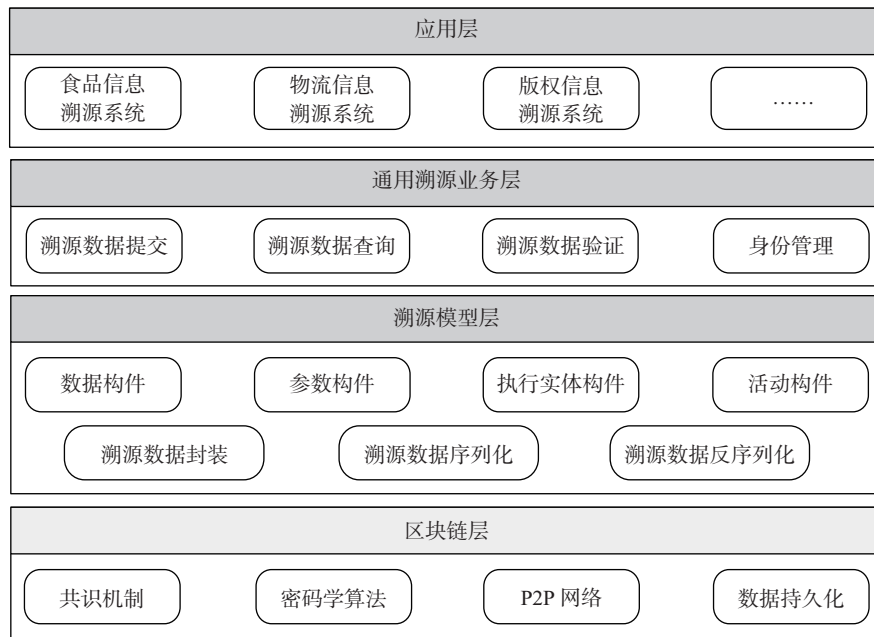


图7 数据溯源系统架构

Figure 7 Data provenance system framework

3.2 ProVOC 数据溯源描述模型

数据溯源的描述模型旨在建立一套溯源信息的通用描述标准和规范。第1个通用溯源模型是首届国际溯源和标注会议 (international provenance and annotation workshop, IPAW) 发布的 OPM^[14]。万维网联盟 (world wide Web consortium, W3C) 对 OPM 进行重大修改后发布了 PROV 溯源标准^[15]。中国于 2017 年 11 月 1 日正式发布《数据溯源描述模型》^[16] 标准, 并于 2018 年 5 月 1 日起正式实施该标准。国家标准中的数据溯源描述 (provenance vocabulary, ProVOC) 模型相对于其他溯源模型来说, 是一个较为轻量的模型, 可以根据具体的溯源领域灵活扩展。

选取较为轻量的 ProVOC 模型作为本文系统的溯源模型。ProVOC 模型由数据、活动和执行实体 3 个一级类构件组成。数据构件是对需要溯源的事物的描述, 包括参数和数据集 2 个二级类构件, 其中数据集可以根据具体的溯源领域进行定制。活动构件是由执行实体发起或控制的一个或多个连续的动作, 主要用来描述数据是如何从上一个状态转换到下一个状态的。执行实体则是指活动的发起者, 包括人类执行实体和非人类执行实体。各个一级类构件之间的关系如图 8 所示。

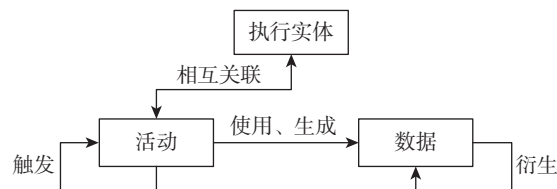


图8 ProVOC 模型结构

Figure 8 ProVOC model structure

3.3 溯源数据的可信存储

为了更好地保存并追踪数据的溯源记录,溯源信息需要依照 ProVOC 溯源模型中的定义转化为 JSON 格式的格式化溯源数据,然后保存到区块链中。下列 JSON 文件记录了一条数据溯源信息,其内容为人类执行体“group1.user1”使用非人类执行体“group1.computer1”,在时间为“2020-09-10T20:39:51”、坐标为“29.532326, 106.607960”的地点,执行了“合并”活动,操作“Hello”和“World”两条数据,最终将“Hello”和“World”合并为“Hello world”。

JSON 溯源信息文件示例

```
{
  “执行实体”: [
    {
      “类型”: “人类执行体”,
      “公钥”: “A18O1/GFa0j1fACqUwfZ7THWwwRrJWBnbwNc6fNRyLsI”,
      “数字签名”: “N31XrMhASBMPOE/I/pdf8Yzp8uQwfmlu3xx4bXTW2QFqM
        mbDLQBqtnMMGwzr6F0SLUU/jpiPrZzZhPNsAmH8IA==”,
      “标识”: “group1.user1”,
      “执行”: “group1.computer1”
    },
    {
      “类型”: “非人类执行体”,
      “公钥”: “A8shcTCw8uTQthYXiliEdzQNy8GRkRvk1+JxeO7nP3Nc”,
      “数字签名”: “0DVat+YcMfR7RO8XsOF/iAkgx+Ay2AwTzdmAWZ6e0ytlpYl
        roZ4FHMZCYhvQRBG2BcTU0HVgR75QvgFz6nZaJg==”,
      “标识”: “group1.computer1”
    }
  ],
  “数据”: {
    “输入数据”: {
      “数据摘要”: “ff27ca250b4be7dde8504c4abcd1292f9c1f13fdcf864a89f29a95008e
        7b31b”,
      “数据集”: {
        “Data1”: “Hello”,
        “Data2”: “World”
      },
      “参数”: {
        “时间参数”: “2020-09-09 13:29:26”,
        “空间参数”: “29.532326, 106.607960”,
        “条件参数”: null
      }
    },
    “输出数据”: {
```

```

“数据摘要”：“64ec88ca00b268e5ba1a35678a1b5316d212f4f366b2477232534a8a
          eca37f3c”,
“数据集”：{
  “Data1”：“Hello World”
},
“参数”：{
  “时间参数”：“2020-09-10 20:39:51”,
  “空间参数”：“29.532326, 106.607960”,
  “条件参数”：null
}
}
},
“活动”：{
  “活动类型”：“合并”,
  “参数”：{
    “时间参数”：“2020-09-10 20:39:51”,
    “空间参数”：“29.532326, 106.607960”,
    “条件参数”：null
  }
}
}
}

```

每个执行实体都需要在溯源系统中注册并使用非对称加密算法生成一个公私钥对，其中私钥用于数字签名，公钥用于验证数字签名和标识执行实体的身份。执行实体需要使用自己的私钥对其操作数据的 Hash 值进行加密后生成数字签名。数字签名可用于验证该溯源信息是否为合法的执行实体发起的操作，同时还可以验证溯源信息有没有被篡改。该验证成立的前提是这条溯源数据必须存在于区块链中；若这条溯源数据没有通过可信性和真实性验证，其包含的证明信息是没有用的。

3.4 溯源数据的可信查询和验证

要实现第2节中描述的可信查询验证方法，就要在区块头中增加一个用于存储 Merkle 山脉根 Hash 值的字段，同时在上链验证的过程中增加对 Merkle 山脉根 Hash 字段合法性的验证过程。

图9描述了溯源数据查询流程。当一个轻节点需要查询溯源数据的真实性和可信性时，只需向区块链网络中发起查询请求即可；区块链网络中的全节点收到查询请求后在本地维护的区块链中查询指定数据的溯源信息，并生成交易包含证明和区块包含证明。将溯源信息、区块包含证明和交易包含证明打包成溯源数据包，然后返回给发起查询的轻节点用户。轻节点在发起查询的同时需要在区块链网络同步最新的区块头，以获得最新的 Merkle 山脉的根 Hash 进行 MMR 验证。轻节点收到溯源数据包后，利用溯源数据包中的 Merkle 证明集合和 Merkle 山脉证明集合就可以在本地完成溯源信息的可信性和真实性的验证。

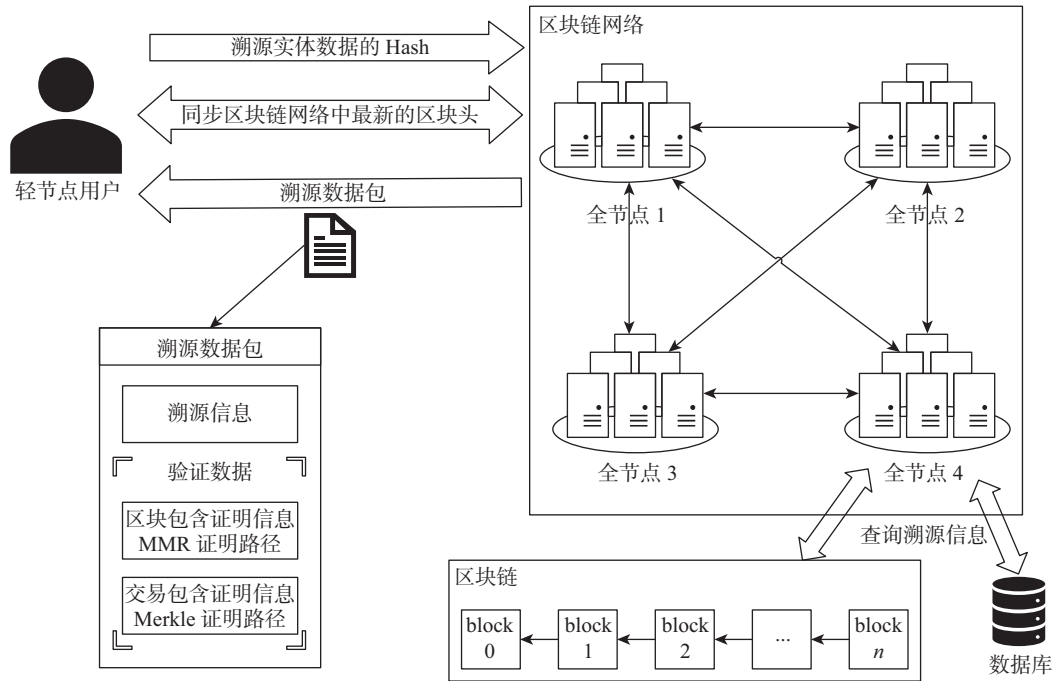


图 9 溯源数据查询流程

Figure 9 Data provenance query process

4 结 语

本文针对区块链数据溯源系统中轻节点查询验证溯源信息占用设备资源过多的问题, 引入可动态追加数据的变体 Merkle 树数据结构 MMR。然后, 基于 MMR 提出一种高效的溯源数据可信验证查询方法, 将 SPV 中需要的 N 个区块头信息减少到 $\lg N$, 大大降低了查询验证溯源信息所需要的设备资源。根据高效的溯源数据可信查询方法设计了一种基于区块链的数据溯源通用系统, 该系统基于 ProVOC 溯源模型将溯源信息封装成 JSON 文件, 利用非对称加密和 Hash 算法对执行实体的身份进行认证, 保证了溯源数据的可信存储。通过 API 接口将溯源服务提供给不同领域的溯源应用。

参考文献:

- [1] WANG Y R, MADNICK S E. A polygen model for heterogeneous database systems: the source tagging perspective [C]//Proceedings of the 16th International Conference on Very Large Data Bases, San Francisco, California, 1990: 519-538.
- [2] BUNEMAN P, KHANNA S, WANG-CHIEW T. Why and where: a characterization of data provenance [C]//International Conference on Database Theory. Heidelberg, Berlin: Springer, 2001: 316-330.
- [3] 袁勇, 王飞跃. 区块链技术发展现状与展望 [J]. 自动化学报, 2016, 42(4): 481-494.
YUAN Y, WANG F Y. Blockchain: the state of the art and future trends [J]. Acta Automatica Sinica, 2016, 42(4): 481-494. (in Chinese)
- [4] LIANG X, SHETTY S, TOSH D, et al. Prochain: a blockchain-based data provenance architecture in cloud environment with enhanced privacy and availability [C]//Proceedings of the 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2017: 468-477.

- [5] XIA Q I, SIFAH E B, ASAMOAH K O, et al. MeDShare: trustless medical data sharing among cloud service providers via blockchain [J]. *IEEE Access*, 2017, 5: 14757-14767.
- [6] RAMACHANDRAN A, KANTARCIOGLU M. Smartprovenance: a distributed, blockchain based data provenance system [C]//*Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, 2018: 35-42.
- [7] 刘耀宗, 刘云恒. 基于区块链的 RFID 大数据安全溯源模型 [J]. *计算机科学*, 2018, 45(增刊 2): 367-368, 381.
LIU Y Z, LIU Y H. Security provenance model for RFID big data based on blockchain [J]. *Computer Science*, 2018, 45(Suppl.2): 367-368, 381. (in Chinese)
- [8] 乔蕊, 曹琰, 王清贤. 基于联盟链的物联网动态数据溯源机制 [J]. *软件学报*, 2019, 30(6): 1614-1631.
QIAO R, CAO Y, WANG Q X. Traceability mechanism of dynamic data in Internet of things based on consortium blockchain [J]. *Journal of Software*, 2019, 30(6): 1614-1631. (in Chinese)
- [9] 张国英, 毛燕琴. 一种基于区块链的去中心化数据溯源方法 [J]. *南京邮电大学学报(自然科学版)*, 2019, 39(2): 91-98.
ZHANG G Y, MAO Y Q. Blockchain-based decentralized data provenance method [J]. *Journal of Nanjing University of Posts and Telecommunications (Natural Science)*, 2019, 39(2): 91-98. (in Chinese)
- [10] NAKAMOTO S. Bitcoin: a peer-to-peer electronic cash system [EB/OL]. [2020-09-28]. <https://bitcoin.org/bitcoin.pdf>.
- [11] MERKLE R C. A digital signature based on a conventional encryption function [C]//*Conference on the Theory and Application of Cryptographic Techniques*. Heidelberg, Berlin: Springer, 1987: 369-378.
- [12] TODD P. Merkle mountain range [EB/OL]. [2020-09-28]. <https://github.com/opentimestamps/opentimestamps-server/blob/master/doc/merkle-mountain-range.md>.
- [13] BÜNZ B, KIFFER L, LUU L, et al. Flyclient: super-light clients for cryptocurrencies [C]//*2020 IEEE Symposium on Security and Privacy*, 2020: 928-946.
- [14] MOREAU L, CLIFFORD B, FREIRE J, et al. The open provenance model core specification [J]. *Future Generation Computer Systems*, 2011, 27(6): 743-756.
- [15] W3C. Prov-o: the prov ontology [R/OL]. [2020-09-28]. <https://www.w3.org/TR/prov-o/>.
- [16] 中华人民共和国国家质量监督检验检疫总局, 中国国家标准化管理委员会. 信息技术数据溯源描述模型: GB/T 34945-2017 [S]. 北京: 中国标准出版社, 2017.
- [17] 范贤丽, 范春晓, 吴岳辛. 基于区块链和 IPFS 技术实现粮食供应链隐私信息保护 [J]. *应用科学学报*, 2019, 37(2): 179-190.
FAN X L, FAN C X, WU Y X. Realization of privacy protection of food supply chain based on blockchain and IPFS [J]. *Journal of Applied Sciences*, 2019, 37(2): 179-190. (in Chinese)
- [18] 祝烈煌, 高峰, 沈蒙, 等. 区块链隐私保护研究综述 [J]. *计算机研究与发展*, 2017, 54(10): 2170-2186.
ZHU L H, GAO F, SHEN M, et al. Survey on privacy preserving techniques for blockchain technology [J]. *Journal of Computer Research and Development*, 2017, 54(10): 2170-2186. (in Chinese)
- [19] 王芳, 赵洪, 马嘉悦, 等. 数据科学视角下数据溯源研究与实践进展 [J]. *中国图书馆学报*, 2019(5): 79-100.
WANG F, ZHAO H, MA J Y, et al. Research and practice progress of data provenance from the perspective of data science [J]. *Journal of Library Science in China*, 2019(5): 79-100. (in Chinese)
- [20] 明华, 张勇, 符小辉. 数据溯源技术综述 [J]. *小型微型计算机系统*, 2012, 33(9): 1917-1923.
MING H, ZHANG Y, FU X H. Survey of data provenance [J]. *Journal of Chinese Computer Systems*, 2012, 33(9): 1917-1923. (in Chinese)

(编辑: 秦 巍)