一种高效安全的去中心化数据共享模型

董祥千[1),3)]，郭兵[1)]，沈艳[2)]，段旭良[1)]，申云成[1)]，张洪[1)]

[1)](四川大学计算机学院 成都 610065)

[2)](成都信息工程大学控制工程学院 成都 610054)

[3)](成都东软学院 成都 611844)

摘 要 数据开放共享是推动数据相关产业发展的源动力，然而，现有的数据共享模型，如数据市场，数据提供方将数据上传至数据存储中心，数据需求方下载数据以实现分析。这种模型存在如下缺陷：1）以关键字为基础的数据检索无法高效发现可连接数据集； 2）数据交易缺乏透明性，无法有效检测及防患交易参与方串谋等舞弊行为；3）数据所有者失去数据的控制权、所有权，数据安全无法保障。为此，本文借助区块链技术建立一种全新的去中心化数据共享模型。首先从共享数据集中提取多层面元数据信息，通过各共识节点建立域索引，以解决可连接数据集的高效发现问题；其次，从交易记录格式及共识机制入手，建立基于区块链的数据交易，以实现交易的透明性及防串谋等舞弊行为；最后，依据数据需求方的计算需求编写计算合约，借助安全多方计算及差分隐私技术保障数据所有者的计算和输出隐私。实验结果表明，本文提出的域索引机制在可接受的召回率范围内，连接数据集查准率平均提高 22%。而以时间及交易区块数相结合的共识机制则能兼顾低交易频率与高交易频率双重需求。同时，与加密方式相比，在保证数据安全前提下，本文提出的安全计算模型平均节省处理时间 6s。

关键词 数据共享；区块链；域索引； 安全多方计算； 差分隐私

中图法分类号 TP391

An Efficient and Secure Decentralizing Data Sharing Model

DONG Xiang-Qian[1),3)] GUO Bing[1)] SHEN Yan[2)] DUAN Xu-Liang[1)] SHEN Yun-Cheng[1)] ZHANG Hong[1)]

[1)](Department of Computer Science, Sichuan University, Chengdu 610065)

[2)](Department of Control Engineering, Chengdu University of Information Technology, Chengdu 610225)

[3)](Chengdu Neusoft University, Chengdu 611844)

Abstract Data opening and sharing is the source power for promoting the innovation of data related industries. However, the typical data sharing model existing at present, e.g., data market, in which data provider uploads their data to a centralized repository and data demander needs to download their required data to analysis, has the following flaws. 1) The key-based data retrieval method cannot efficiently find the linkable datasets. 2) As lack of transparency in the process of data transactions, it is not efficient to detect the transaction collusion or other frauds among the parties. 3) The data owners lose the power of controlling their own data, which causes no guarantee of data ownership and data security. In this paper, to solve the above problem, we proposed a new

本课题得到国家自然科学基金重点项目(61332001)；国家自然科学基金项目(61772352，61472050)；四川省科技计划项目（2014JY0257，2015GZ0103）；成都市科技惠民技术研发项目（2014-HM01-00326-SF）董祥千，男，1975年生，博士研究生，讲师，主要研究领域为个人大数据，E-mail: dongxiangqian@nsu.edu.cn. 郭兵（通信作者1），男，1970年生，博士，教授，博导，CCF高级会员，主要研究领域为绿色计算、个人大数据，E-mail: guobing@scu.edu.cn. E-mail: liq@scu.edu.cn. 沈艳（通信作者2），女，1973年生，博士，教授，主要研究领域为智能终端、智能仪器，E-mail: shenyan02@163.com. 段旭良，男，1982年生，博士研究生，讲师，主要研究领域为个人大数据，E-mail:5025968@qq.com. 申云成，男，1979年生，博士研究生，讲师，主要研究领域为个人大数据，E-mail:403953413@qq.com. 张洪，男，1980年生，博士研究生，讲师，CCF会员，主要研究领域为个人大数据，E-mail: 945389781@qq.com.

blockchain-based decentralization data sharing model. Firstly, we extracted the multi-sided metadata information from the shared data set for efficiently discovering linkable datasets and had the consensus nodes to set up domain index. Secondly, with the help of consensus mechanism, we implemented data transaction based on blockchain to achieve transparency and to prevent conspiracy. Finally, according to the computing needs of the data owners, we compiled the computation contract to ensure the computation and output privacy of the data providers by using the secure multi-party computation and differential privacy. The experimental results show that the domain index mechanism proposed in this paper increases the average precision by 22% without substantially reducing the recall rate. And the modified consensus mechanism, which combines time and transaction block number, takes both low trading frequency and high trading frequency into account. At the same time, on the premise of ensuring data security, compared with the encryption method, our method proposed saves the processing time of nearly 6s.

Keywords data sharing, blockchain, domain index, secure multi-party computation, different privacy

1 引言

当前，大数据产业面临“人人有数据，人人缺数据”的“数据孤岛”式困境，其有效解决途径是建立合理、高效的数据共享模型。数据开放、数据交易是当前常见的两种数据共享模型。

数据开放源自 21 世纪初期的数据开放运动，倡导数据的免费、自由使用。包含以下四个阶段①：1）数据集的选择；2）制定相关的开放协议或规则；3）使数据可获取；4）使数据易于发现。其中数据“可获取”指数据以机器可读的格式发布及提供下载；数据“可发现”则一般通过建立中心化的数据目录来实现。数据开放的外在表现形式是由政府部门支撑的数据开放平台②。

数据交易源于 19 世纪中叶的股票市场及报刊行业，是以数据价值为基础的一种数据商品交易行为。数据交易一般通过数据交易平台实现，即数据提供方将数据以一定契约（协议）提交到交易平台，数据需求方在支付一定的费用后从数据平台获取数据，数据交易平台在完成交易后支付数据提供方一定比例费用。数据交易的表现形式包括数据市场③④、数据银行[1]等。

数据开放、数据交易的特点如表 1 所示，本文称这类数据共享模型为传统数据共享模型。Sharemind⑤[2,3,4]是传统数据共享模型的典型代表

表 1 两种常见的传统数据共享模型特点

共享模型	例子	特点
数据开放	美国国家数据开放平台	1）免费开放，自由获取、使用以及再发布，但数据供给方的数据规模与服务能力有限。 2）在个人隐私、企业商业秘密和国家安全等信息脱敏或保护前提下，几乎不受数据产权、管理权等权益的限制（特例除外）。
数据交易	国云数据；Quandl	1）解决的核心问题：数据确权与溯源保护、数据计量与计价和数据收益模型。 2）交易标的物是数据拷贝，即原始数据或经过加工后的数据，数据的分析和使用往往脱离数据供给方的控制，无法有效解决数据的产权、管理权及数据安全等问题，权益难保障。

大学期间工作与不能按时毕业之间的关系？--因为大学期间工作而不能按时毕业，甚至退学？确定 ICT 类学生和非 ICT 类学生的工作习惯？

图 1 查询用例

(说明：(1) 如文献[2]所述，解决该用例需要联合教育数据集和税收数据集。(2)该用例既用于本节模型说明，也用于第 5 节实验评估)

，其解决图 1 所示用例的过程可简要描述为：

1）数据导入及预处理（即 ETL）阶段，具体过程如图 2 所示。

2）查询分析阶段

为获取最终结果，针对不同的数据集，Sharemind 执行了 6 大类，共 49 个具体的查询。其中针对教育类数据集的查询有 9 个，其它都是针对分析表的查询。

① http://opendatahandbook.org/
② https://www.data.gov/
③ http://www.moojnn.com
④ https://www.quandl.com/
⑤ https://sharemind.cyber.ee/

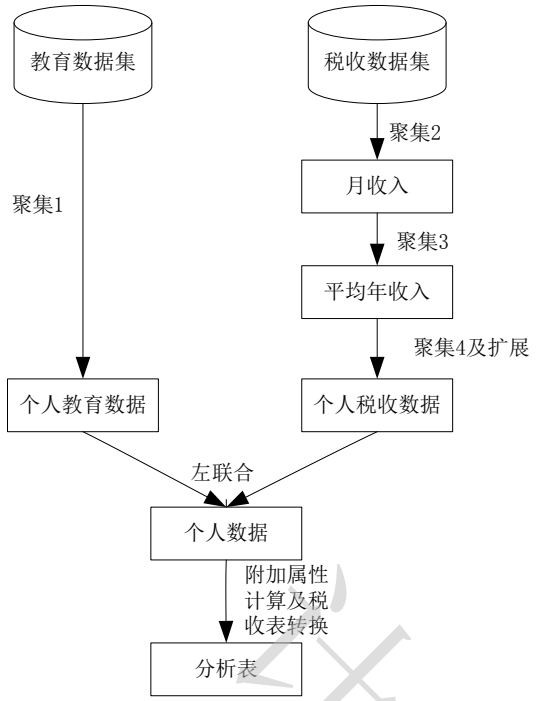


图 2 Sharemind 解决用例 1 的过程

(说明：各聚集步骤的含义如下，聚集 1：按 person ID 与 curriculum ID 组合；聚集 2：统计在一个月内某人薪水总和；聚集 3，统计在一年内某人薪水总和；聚集 4：在计算涉及的年限内，将所有薪水信息按 person ID 进行统计，形成扩展表)

在文献[2]中，尽管采用了安全多方计算确保计算隐私，但原始数据集被批量下载至 Sharemind 平台，仍然存在数据隐私泄露风险。

可见，虽然传统数据共享模型是突破数据使用困境的有效途径之一，但该模型仍然存在如下尚待解决的问题：

P1. 数据汇聚并存储在第三方平台，数据的分析或使用脱离数据提供方控制，无法有效解决数据使用过程中的透明性(Transparency)、可信性(Trust)、可控性(Control)、增值性(Value)等需求[5]，也是现有数据共享模型存在的主要问题。

P2. 对数据集的描述仅限于数据基本信息层面，如数据集名称、数据格式、数据集大小等，缺乏数据使用历史及数据分析工具的说明。同时，以关键词为基础的索引方式仅考虑关键词在文档中的出现频率（即 TF-IDF），不能反映文档之间的可连接性。

P3. 数据所有者及数据需求方无法查询特定数据集的使用情况，数据集扩展使用不方便。

P4. 数据更新不及时，大多是过时的数据，用户参与度低。

以上问题揭示传统数据共享模型在不同层面存在的问题：其中问题 1 属于数据安全计算范畴；问题 2 属于数据发现范畴，问题 3、4 则涉及数据集的可扩展性及更新问题。本文提出的去中心化数据共享模型正是针对上述问题设计的一种新型数据共享模型：以恢复数据提供方对数据的可控性为基础，以在数据提供方完成数据安全计算和分析为核心；通过域索引及接口机制获取计算数据集及计算任务，通过区块链技术控制用户行为及数据流通。从而解决数据共享过程中数据集高效发现、交易管理、数据安全分析与计算等方面的问题。

本文的主要贡献有：

1）将数据迁移转化为计算迁移[6]，建立去中心化的数据共享模型。原始数据储留在由数据提供者完全控制的数据空间[7]中。在受控的数据空间内完成数据的分析与计算，只有符合安全性要求（满足计算隐私和输出隐私）的结果数据才会脱离数据所有者的控制，发送到数据需求方。同时数据空间自动记录数据访问日志，通过日志分析实现数据监控、溯源等操作。该设计为恢复数据的透明性、可信性、可控性、增值性[5]等提供了条件，同时也能更好地实现对用户访问控制的精细化处理[8]。

2）采用分布式哈希表（DHT）[9]及局部敏感哈希（LSH）[10,11]技术，建立去中心化的数据索引机制。首先提取数据集多侧面元数据信息，如数据集所有者、生成时间、模式信息、使用情况（工具，使用频率）等基本信息以及数据集的内容信息，从而建立数据集域索引机制。其次，根据数据域的相似性（关联度），将相似数据域的索引存储在邻近的网络节点中。既能提高数据搜索速度，也有利于相似数据集检索。数据索引的去中心化存储，降低了中心化索引节点的单点故障率，有效提高数据查询效率。

3）建立基于区块链技术的透明化交易模型，以确保数据交易参与方的公平性、诚实性[12]。依据参与方遵守协议的程度，动态调整参与方的信誉度，只有信誉度较高的节点才有机会成为共识节点，从而激励各参与方遵守系统协议。区块链技术也用于数据访问策略控制，防串谋控制等。

4）将计算写入智能合约脚本，使网络具备计算处理能力。同时，组合安全多方计算与差分隐私技术保证计算隐私和输出隐私。

本文的结构安排：

第一部分在分析现有数据共享模型特点的基

础上，针对该模型存在的主要问题，提出去中心化数据共享模型的基本思想。第二部分概括介绍系统相关工作及本文涉及的相关技术性问题。第三部分和第四部分是本文的核心，其中，第三部分从模型计算范式出发，概括出系统模型的层次结构及各层功能；第四部分则详细阐述了模型各层所涉及的具体技术及相关协议的实现细节。第五部分展示系统实验过程、结果以及模型主要组件的性能指标。最后，第六部分是对本文工作的总结及对未来工作的展望。

2 相关工作

为应对数据共享中存在的问题，世界经济论坛组织提出以用户为中心的数据处理模型及其实现的关键原则：透明性(Transparency)、可信性(Trust)、可控性(Control)、增值性(Value) [5]。学术界提出的数据空间概念及其新的编程范式是实现这一转变的有效途径。

数据空间的概念由 Michael Franklin 等人在 2005 年的 SIGMOD 会议上首次提出[7]，指与组织或个人相关数据组成的数据集合。这些数据不以数据存储的地理位置划分，而以数据关联的主体划分。随后各种数据空间模型相继出现，如(1)iDM(iMeMex)数据模型[13]，将个人数据以资源视图或资源视图类的形式进行组织，并采用专用的查询语言 iQL 实现数据检索；(2)VDS(即虚拟数据空间模型 Virtual Dataspace Model)[14]是面向工业数据应用定制的一种数据空间模型，使用 4 元组(S_{UR}, D_S, DR_S, S_S)来表示。其中 S_{UR} 是用户需求的主体， D_S 表示数据源集， DR_S 是数据源之间的关系集， S_S 是服务集。该模型将不同用户的需求、数据及其服务抽取出来形成虚拟空间，系统内部实现虚拟空间到物理数据空间之间的映射。

MIT 人类动力学实验室教授 Pentland 等人[15]，在对个人隐私信息保护的研究中提出的“安全问答 (SafeAnswer) ”模型则是基于数据空间的一种新的编程范式。 SafeAnswer 是在用户个人数据空间中运行的软件，使用用户的敏感数据集进行分析，这些分析及其结果被数据提供方完全控制，只有“安全” 的数据才会发送给数据需求方。

去中心化数据共享模型使用逻辑上独立的数据空间(第 3 节中则将数据空间抽象成节点)来管理数据。本文假设与主体（个人或机构等）关联的数据都属于该主体，如某人的消费数据、统计数据、医疗数据等都属于该人，并由其数据空间统一管理。数据的所有权问题以及数据与数据空间之间的关联等问题详见文献[15]。抽象的数据空间构成了本文提出的去中心化数据共享模型的基石。

文献[8]采用离链存储数据（加密的数据，以 DHT 的方式存储），将数据的存储地址保存在公共账本中。该系统区为两种类型的交易 T_{access} 及 T_{data}，前者实现用户与某一服务(用户安装的应用)之间权限访问及更新；后者用于查询及存储数据。与该文相比，本文在以下方面进行了改进：

1）建立特定的域索引机制，以提高连接数据集的检索速度，保证结果数据集可连接性。

2）摒弃文献[8]采用的数据加密存储方式，以安全多方运算及差分隐私代替，从而提高数据处理效率。

3）利用智能合约，使系统具备计算处理能力，而不仅仅用于存储与检索信息。

本文涉及的技术问题主要包括：

1）区块链

区块链是分布式去中心化账本[16,17]，用无限状态机（五元组）描述，如式 2-1：

$$(\Sigma, S, s_0, \delta, F) \quad \text{式 2-1}$$

其中 Σ 是交易(Transaction)的集合； S 是区块(Block)的集合，以链式结构组织； s_0 是初始状态，即创世块； δ 是状态转换函数，即 $\delta : S \times \Sigma^* \rightarrow S$，由共识算法实现； F 是终止状态（为空）。

t 时刻区块 B_t 的结构如式 2-2 所示：

$$B_t \leftarrow Hash([tx_1, tx_2, \ldots tx_n] \,||\, B_{t-1} \,||\, t) \quad \text{式 2-2}$$

其中 tx_n 表示当前区块中的第 n 个交易。

区块链是利用密码学技术保证数据的不可篡改性和不可伪造性、利用分布式节点共识算法生成和更新数据、利用自动化脚本代码 (智能合约)编程和操作数据的一种全新的去中心化基础架构与分布式计算范式[18]。

2）安全多方计算

安全多方计算（Secure Mutiparty Computation, SMC）是解决两个或多个用户间敏感数据计算的有效方法。其基本特征是保证参与方在不泄漏敏感数据的前提下，协作完成某项计算任务。安全多方计算模型可表述为[19]：

在一个分布式网络中，有 n 个互不信任的参与方 $P_1, P_2, \cdots, P_n$，每个参与方 P_i 秘密输入 x_i，他们需要共同执行函数 F:$(x_1, x_2, \cdots, x_n) \rightarrow (y_1, y_2, \cdots y_n)$

其中 y_i 为参与者 P_i 获得的相应输出。在函数 F

的计算过程中，要求任意参与者 P_i 除 y_i 外，均不能获知其他参与者 $P_j(j\neq i)$ 的任何输入信息。

安全多方计算存在如下两方面的问题[12]：

1、如果不满足大部分参与方诚实可信，安全多方协议不能保证公平性。

2、安全多方协议仅负责计算的安全性，不能确保用户提供“真实”的输入且尊重输出结果。

针对以上问题，文献[12]提出基于区块链的“限时承诺”解决方案，限时承诺分为两个阶段，即承诺阶段和公开阶段。例如，设承诺方的原始值为x，在承诺阶段，承诺方随机选择 r，计算 $s:=x||r$ 后向每个接收方发送 $\tilde{h}=\vec{H}(s)$(其中 $\vec{H}$ 为哈希函数，如SHA-256)，同时，承诺方还需支付一定量的押金；在公开阶段，承诺方向每个接受方发送 s，接受方验证 $\tilde{h}$ 是否等于 $\vec{H}(s)$。如果相等则恢复 x。如果在限定时间内承诺方未公开承诺，则押金被转到接收方。

3）差分隐私

不同数据集的组合及分析通常导致两种主要的隐私问题：计算型隐私及输出型隐私。前者通过安全多方计算实现，后者则涉及差分隐私的内容。(ε,δ)差分隐私的定义如式 2-3 所示[20]：

$$\Pr[M(x)\in\zeta]\leq e\varepsilon\Pr[M(y)\in\zeta]+\delta \qquad \text{式 2-3}$$

式中 M 是定义在特定数据集上的算法。x, y 是该数据集的任意子集，且 $||x, y||_1\leq 1$ 即 x, y 最多相差一条记录。ζ 是算法 M 的值域空间。

实现差分隐私的常用方法是向输出结果加入适量的噪声。例如常用的 Laplace 机制，是向输出结果加入服从 $Laplace(s/\varepsilon)$ 分布的噪声(s 是算法 M 的敏感度) 。

差分隐私满足组合定理,即如果第 i 次查询满足 ε_i 差分隐私，则这些查询的组合满足 $\Sigma\varepsilon_i$ 差分隐私。因而对于特定实体数据集，可以预先定义总的隐私预算 B_u。每次查询消费部分预算 ε_i，只有当 $\Sigma\varepsilon_i\leqslant B_u$ 的查询才是被允许的查询，从而保证数据集满足差分隐私要求，即满足输出隐私。

4）局部敏感哈希

数据集可连接性的实质是相似项搜索，其计算复杂度与项对数目成平方关系。这样，即使项间相似性的计算复杂度较低，但在项对较多时其计算复杂度也非常高。为此，学术界提出了近似搜索，局部敏感哈希(Locality Sensitive Hashing, LSH)是最常用的近似搜索技术之一，它将搜索范围集中在可能的相似项对上。

定义 2-1 局部敏感哈希(LSH)

若函数簇 H (从项到实数域的函数集)中的每一个函数 h 对任何项 p, q 都满足如下两个条件，则 H 是 (r_1, r_2, p_1, p_2)敏感函数簇，h 是 (r_1, r_2, p_1, p_2)敏感函数[21]：

(1) 如果 $sim(p, q)\geq r_1$，则 $\Pr_H[h(q)=h(p)]\geq p_1$

(2) 如果 $sim(p, q)\leq r_2$，则 $\Pr_H[h(q)=h(p)]\leq p_2$

式 2-4

其中 $sim(p, q)$ 称为项 p, q 之间的相似度，一般有 $p_1\geqslant p_2$。

定义 2-2. 最小哈希(minHash)

设随机置换函数 $\pi:\Omega\to\Omega$。则最小哈希函数[22]定义为：

$$h_\pi(x)=min(\pi(x)) \quad x\in\Omega \qquad \text{式 2-5}$$

符号 Ω 表示项的全体取值集合。

文献[21]证明了最小哈希函数在 Jaccard 相似度[23]下是局部敏感哈希函数。

在非结构化数据集中相似项的搜索是一项挑战性的工作。常采用“自顶向下”及“自底向上”的方法解决相似项搜索及数据发现问题。前者通过建立数据集的全局模式，如数据仓库中的表模式；后者是从数据集的分析出发，在抽取出其中的实体及其关系的基础上完成相似项搜索。

文献[24]利用自底向上的方式设计大规模数据集发现原型系统，该系统由三个组件构成，其中数据签名组件收集数据源中的实体；数据路径搜索组件建立数据源间的相似度；数据的披露及发现等则通过数据发现原语组件来完成。

GOODS [25]也属于自底向上的方法，该方法首先从数据集中抽取出各种基本元数据信息，如数据所有者，时间，模式信息等，以及数据集间的关系元数据信息，如相似度和溯源信息，之后将这些元数据信息以服务的方式披露给数据的使用者。

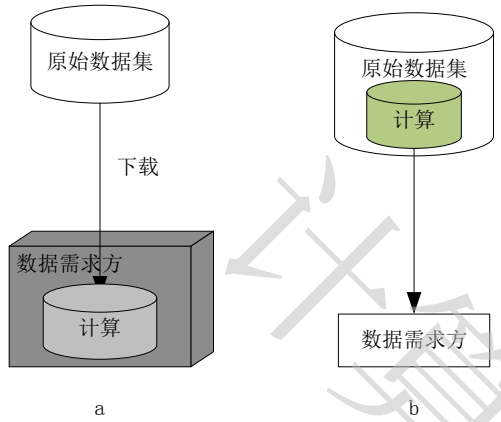
本文仅讨论了非结构化数据集的相关理论问题，实验部分是针对结构化数据展开的。

3 去中心化数据共享模型结构

本节从模型计算范式出发，概括模型层次结构及各层功能，在第 4 部分将重点阐述各层的实现细节。

3.1 模型计算范式

传统数据共享模型需下载原始数据集才能完成数据分析任务，如图 3-a 所示。数据所有者失去数据的控制权，是数据隐私泄露、数据误用（滥用）的根源。本文提出利用计算迁移[1]，将原本在数据需求方（或平台）完成的计算，迁移到数据提供方，如图 3-b 所示。其基本出发点是恢复数据的透明性、可信性、可控性。两种共享模型计算范式的根本区别如图 3 所示。



a：传统数据共享模型； b：去中心化数据共享模型

图 3 两种数据共享模型计算范式对比

去中心化数据共享模型的工作流程概括为：

1）数据提供方提取数据集元数据信息，并以数据域的形式发布（称为数据发布交易）。系统的共识节点负责为其建立索引，索引分布存储在系统的各个节点（具备索引存储权益）中。

2）数据需求方根据各自需求查询数据索引（称为数据查询交易），获取相关的数据集信息，并编写计算合约。

3）数据的分析与处理以智能合约的方式在各数据提供方安全地执行，并向数据需求方提供精确的、满足隐私需求的计算结果。

在以上流程中，区块链将各个部分有机串联在一起，承担交易执行及权限审计任务。

3.2 模型层次结构

如图 4 所示，系统模型由接口层，索引层，交易层以及数据层组成。各层中的节点只是逻辑功能上的分离，物理上可以是同一实体。

1）接口层

接口层是针对数据需求方的，其功能类似区块链技术中客户端钱包。系统通过钱包公钥识别网络身份，公钥在一定程度上保护了用户的身份信息。接口层主要完成两方面的功能：（1）根据数据需求，通过索引层，获取计算数据集；（2）根据计算数据集特性及数据需求编写计算合约。以上两种操作都被称为交易（系统中的所有事务都称为交易），交易被发布在区块链中，由各共识节点以“挖矿”的形式执行。

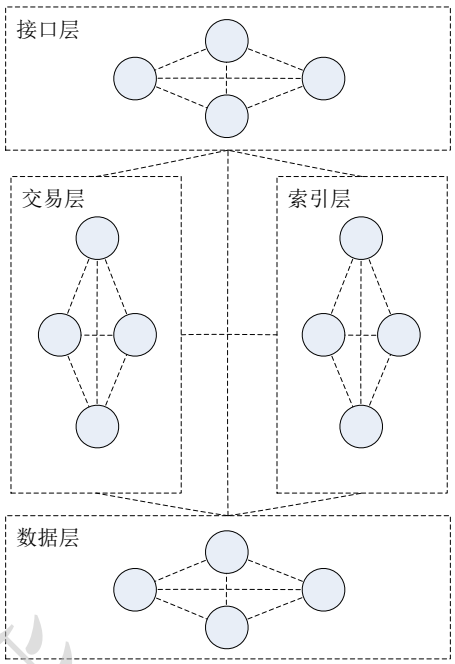


图 4 系统模型图

(说明:图中各个圆点代表节点实体 N_E)

2）索引层

数据索引是相似(可连接)数据集的搜索接口。如 3.1 节所述，为确保数据所有者的控制权，原始数据集是不公开的，公开的是数据的基本信息，包括数据集元数据信息及部分特征信息。数据集索引类似传统搜索引擎的索引，但具有显著的不同点：（1）索引层的外在表现形式是资源目录①，用户可以按资源分类标准浏览平台数据集及数据集的特征内容。（2）关联数据集搜索采用的是域索引机制（详见 4.2 节），即对由数据集的元数据属性信息及其特征值组成的域按照局部敏感哈希建立的索引。对索引层的检索获得计算相关的数据集，并将数据

① 资源目录的编排可参照《政务信息资源目录编制指南》：https://www.gzdata.com.cn/c69/20170714/i1963.html

集元数据信息返回给数据需求方。在去中心化数据共享模型中，索引文件也采用去中心化方式存储（即分布式哈希表技术(DHT)）。

3）交易层

交易指数据的处理逻辑，包括以下几种交易类型：索引及数据目录的生成，计算查询集的产生，数据需求方的计算需求，需求双方信誉评价。系统包括链上数据及链下数据，交易信息(不存储数据分析的结果信息及中间信息)存储在链上，索引信息及原始数据集信息则在链下存储。

数据交易格式及共识机制是交易层设计的重点。

4）数据层

数据层主要针对数据提供方。数据提供方是一种抽象的独立实体（指与特定数据关联的具体的人或组织），它负责原始数据集的汇集、维护以及元数据的抽取、发布等。数据层的主要计算任务是根据数据需求方的请求以安全、私密的方式获取计算结果。与数据需求方类似，数据提供方也通过客户端钱包来管理或查询相关事务。钱包除完成上述功能外，还具备数据仪表盘功能，即监控数据、查看数据流向以及手动添加注释等。

从模型的分析可以看出，依据功能的侧重点不同，可以将网络中节点分为三类，即数据需求节点、交易节点以及数据提供节点。需要注意的是节点功能并不是固定的，即各类节点是平等的，如数据需求节点既可以参与数据交易的执行与验证，也可以作为数据的提供方。因此，用式 3-1 统一表示网络中的节点实体 N_E 。

$$N_E = \{A, D, P, I, C\} \quad \text{式 3-1}$$

其中，A 表示该节点的地址，即钱包的公钥，由区块链自身生成及维护。D 表示该节点维护的数据集，对于单纯的交易节点及数据需求节点，D 可以为空。P 表示节点的权属信息（如隐私偏好信息等）。I 表示该节点保存的索引集信息。C 是对节点算力（共识权益）的抽象。

4 各层相关技术

4.1 接口层

接口层主要实现计算数据集的生成及根据计算需求编写智能合约。

1）计算数据集的生成

计算数据集的生成采用类似自动问答系统[26~28]中相关文档生成技术，包括实体识别及扩展、候选数据集以及计算数据集生成三部分，如图 5 所示。其中实体识别及扩展首先结合数据集字典及自然语言处理技术，从问题中提取出名词、动词或形容词等实体。之后数据需求方根据需求对实体进行选择或修改，并填充一定量的实体属性信息，形成查询域。利用查询域检索域索引（域及域索引的概念见 4.2 节）得到候选数据集。用户通过浏览候选数据集属性（元数据信息），选择满足要求的目标数据集（即计算数据集）。其中，实体识别及扩展以及候选数据集的生成是在数据需求方的干预下，多次逐步求精的过程，如图 5 中虚线所指部分。

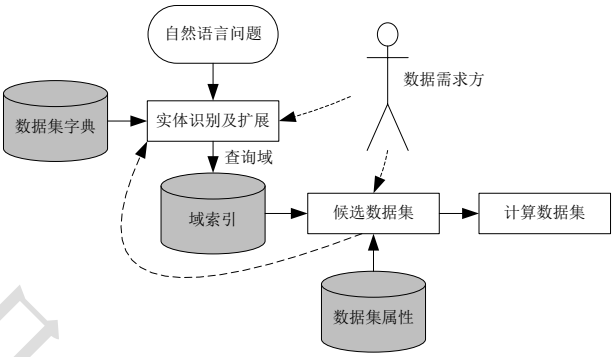


图 5 计算数据集生成过程

例如，对图 1 中的用例，可以提取如下关键词：

（大学期间） 工作 （不能按时） 毕业 退学 （ICT 类，非 ICT 类） 学生 工作习惯

通过自然语言处理技术及人工识别，易推断出核心关键词(即实体)包括：

学生、工作

关键词“学生”进行如下的扩展，即学生属性可能包括：学生 ID、性别、出生年月、课程 ID、学历层次、入学时间、学习状态、毕业（退学）日期等。

关键词“工作”的扩展如下：个人 ID，年，月，工作类别、薪资等①。

对实体“学生”、“工作”的各个属性都可以建立域，但考虑数据的可连接性，往往只对特定的属性（如主键属性）建立域，如学生 ID 域、课程 ID 域、个人 ID 域等，这些域称为查询域。

①数据集生成属于自然语言处理范畴，实验中假设这些域信息是已知的，并只处理英文字符集。

域检索指根据预设的阈值，在索引库中找出与域匹配度较高的项（类似 Top-N 近邻搜索）。需要特别说明与传统的搜索引擎衡量关键词指标的 TF-IDF 技术不同，这里采用了域关联度技术[29，30]。前者衡量查询词在文档中的出现频率；后者注重查询域与索引域的包含程度（关联程度）。

2）智能合约

以太坊（ethereum）的智能合约使网络在功能上类似计算平台，系统借助以太坊建立全新区块链网络。合约一旦被共识节点执行，则以太坊中的节点都会遵循相应的指令。利用这一技术，可以使网络中各节点协作完成相关计算任务。

接口层的智能合约分为以下两类：1）根据数据需求方的需求编写的计算合约；2）有关数据集使用策略的使用策略合约。例如，为验证合约账户的合法性，合约中需发送数字签名及对应的公钥[31~33]。数据集使用策略合约结构如表 2 所示：

表 2 数据集使用策略合约

```
contract policy {
    bytes signature;
    bytes pubkey;
    struct acList
    {
        //策略列表
    }
    function update()
    function get()
    function kill()   //kill this contact
}
```

数据需求方计算需求合约与此类似，这里不再赘述。

4.2 索引层

为处理海量数据集可连接搜索问题，模型采用域索引技术。为此，使用域重新定义数据集。

定义 4-1 域

数据集的特定属性或属性的组合称为域，如关系数据库表的主键可以构成一个域。使用域的取值集合来表示域，用符号D 表示。若符号Ω表示域取值的全集，则域的形式化定义如式 4-1 所示：

$$D=\{x|x\in\Omega\} \qquad \text{式 4-1}$$

例如对于表 3 所示的教育类数据集(部分截取信息)，采用 Hyperloglog[34]，可识别该数据集潜在键，包括：studentID，lessonID。因而可以建立 studentID 域及 lessonID 域，其中 studentID 域表示为：

$$D(studentID)=\{10001781,10000113,\cdots\}$$

表 3 教育类数据集的部分截取信息

studentID	lessonID	sex	birthday	educationLevel
10001781	12000532	1	1982/7/9	0
10000113	12000642	0	1982/3/21	2
10001318	12000968	1	1983/5/27	2
10000928	12000375	0	1983/1/29	1

使用域重新定义的数据集S 可以表示为：

$$S\equiv\{D_1,D_2,\cdots,D_n\} \qquad \text{式 4-2}$$

定义 4-2 域关联度

对于域Q,I，域关联度的定义如式 4-3 所示：

$$t(Q,I)=|Q\cap I|/|Q| \qquad \text{式 4-3}$$

式中$|\cdot|$表示集合的基（对有限集而言，即集合中元素的个数）。一般地，Q 表示查询域，I 表示索引域。域关联度$t(Q,I)\in[0,1]$，值越大，域之间的连接性越好。同样，若两个数据集中存在连接度较高的域对，则数据集的关联性（可连接性）就越高。

定义 4-3 域搜索

对于给定的域Q、域集I及关联度门限值$t^*\in[0,1]$，从域集I中搜索关联度大于t^*域的过程称为域搜索，形式化表示如式 4-4 所示：

$$\{X:t(Q,X)\geq t^*,X\in I\} \qquad \text{式 4-4}$$

域搜索目的是发现可连接性高的数据集。

定义 4-4 域索引

索引域I的 hash 值构成的索引称为域索引。

LSH 的一般做法是将签名向量（多次最小哈希得到的向量）分成 b 个分区(band)，每个分区内包含 r 行。则成为候选域的概率与 Jaccard 相似度s之间的关系如式 4-5 所示[21,30]

$$P(s\mid b,r)=1-(1-s^r)^b \qquad \text{式 4-5}$$

针对域关联度的非对称性，Shrivastava 及 Li[22]采用了填充新值的方法;而 Erkang Zhu 等人[30]则提出域关联度与 Jaccard 相似度s相互转换的解决思路，其关系如式 4-6 及式 4-7 所示：

$$\hat{s}_{x,q}(t)=t/(x/q+1-t) \qquad \text{式 4-6}$$

$$\hat{t}_{x,q}(s)=(x/q+1)s/(s+1) \qquad \text{式 4-7}$$

其中$x=|X|,q=|Q|,X,Q\subset D$，分别表示域X及域Q

的基。

域关联度与 Jaccard 相似度互换会引入伪正例及伪负例，为此结合[30]，本文提出如下的解决思路及处理步骤：

（1）预处理阶段

该阶段主要包括域及域值确定、限制与扩充，是针对数据提供方而言的。指其（a）按照特定的规则提取数据集中的域（元数据或属性）；（b）确定域值（例如，对于表示范围的数据，可以发布中值、最大（小）值（需考虑隐私方面的需求）等）；（c）根据域大小限制或扩充域值范围。

对结构化数据，既可以将不同属性（垂直划分）作为域，也可以视记录（水平划分）为域，各属性值作为域值；对于非结构化数据，通常将数据集（水平划分）作为一个域，并将数据集元数据作为域值。例如，在实验中，我们抽取了如表 4 所示的元数据域。

表 4 元数据域

类型	基本内容	用途及获取方法
基础元数据	数据集名称、内容、大小、格式、访问权限、创建（修改）时间以及用途等	确定数据集的基本用途，一般由数据提供者主动提供
溯源元数据	数据集操控历史、分析工具、下游（上游）数据集等	跟踪数据集的流通信息，通过对日志文件分析获取
基于内容的元数据	模式，数字水印，关键词域，高频词项，相似数据集等	用于关联搜索，通过对样例文件的扫描获取
语义元数据	语义知识图谱，如 RDF	用于语义搜索，通过文件扫描及本体匹配获取
自定义元数据	根据实际情况扩展的元数据信息	用于某些带指向性的用途，通过用户的使用反馈获取

由于域索引是公开的，域值的确定涉及数据集安全与隐私，可采用静态数据发布中的隐私保护策略（数据值模糊化、k-匿名[35]、l-多样性[36]、t-近邻等）确定域值。这些被发布的域值称为域的特征值，特征值的数量视需求而确定，需满足最小值(分区数与区内行数之积)要求。

由式 4-7 可知，当 $x/q \approx s$ 时，$\hat{t} \approx s$。因而可根据 x/q 的值确定不同的 b，r 值，进而由式 4-5 可实现候选集精度的动态调整。索引的建立如算法 1 所示。

（2）索引优化阶段

系统索引文件采用分布式哈希表(Chord)存储结构，即将索引文件存储在网络中的各个节点中，

算法 1 索引建立算法

输入：待索引域 I
辅助输入:按照幂律分布[30]预设的索引集及其对应的*(b,r)*值
输出：索引集
1. 根据域 I 的大小 $|I|$ 计算对应的索引集，及其对应的*(b,r)*值
2. 计算域 I 的 minhash 值，得到签名向量。
3. 根据*(b,r)*值划分签名向量，并将其 hash 到对应的桶中。

避免单点故障及过载等中心化索引所带来的问题。其具体实现过程如下：

抽取的域按表 5 的方式添加限定符前缀[25]（限定前缀实质是对域的预划分，即对不同领域共享数据集的划分）。在逻辑上将域分为多组，依据域的 LSH 值及各节点的 hash 值的近邻关系，域索引被存储在与其 hash 值相邻的节点中[37]。

各节点优先存储与该节点已存储索引关联度较高的索引信息，使得同一节点保存的索引信息的相似度较高。hash 索引的方式及高相似度索引文件的集中存储提高了数据检索的效率。

表 5 域 D 的限定符及其含义

限定符	含义
owned_by: D	数据集的所有者
proto: D	使用的协议或工具
read_by: D written_by: D	对数据操作（读/写）的名称
downstream_of: D upstream_of: D	上游/下游数据集的节点信息
kind: D	数据集的类型或模式
content: D	数据集内容信息。

4.3 交易层

数据需求方需要支付一定的费用以获取网络中(数据提供方)的数据；这样，数据提供方因提供数据服务而获得相应的报酬；另外，参与索引创建、存储以及计算数据集生成、信誉评级的共识节点也应获取一定额度的交易费用。信誉评级是对交易双方在交易过程中守信度的评估，信誉度低的节点参与交易的代价越高（对数据需求方而言，需要支付的费用越高；而对数据提供方而言能获取的费用则越低）。且信誉度越低，其成为共识节点的机会越低。数据提供方的激励主要体现在以一定比例获取数据需求方支付的交易费用。

1）交易类型及结构

系统中将一切事务都称为交易，系统交易包括：数据请求方发出的交易（分为计算数据集生成及计算合约的创建）；数据提供方发出的交易（分为数据使用策略，索引文件创建）以及针对数据供需双方的信誉评估机制。基于此，设计了特定的交

易结构，其主要数据结构如表 6 所示：

表 6 交易的主要数据结构

字段	说明
Type	交易类型。
Attributes	该交易所具备的额外特性。
Inputs/ Outputs	交易的输入/输出。
Scripts	交易的验证脚本。
ContributorSig	交易发起者签名。
Hash	该交易的散列值。

2）共识机制

系统交易只有确认后才能成为区块进而形成区块链，共识机制是在分布式环境下完成交易确认的算法。模型采用改进的委托拜占庭容错算法①（dBFT）[38]，原算法基于 POS(Proof of Stake，股权证明)，算法的改进主要体现在两个方面：

1) 引入信誉评级机制，更好地鼓励遵守协议的节点。

2）将原算法中区块的生成间隔以时间为单位的划分标准改为以时间及交易数量结合的划分标准。

改进后的共识算法引入了候选共识节点依据信誉度排序的机制。只有遵守协议（包括交易协议以及安全多方计算协议(4.4 节中描述)）的节点才能获得高排名，进而成为候选共识节点。可见，算法能够好激励遵守协议的节点。同时，改进后的算法并不会改变协议的 CAP 性质，仍然满足 $f = \lfloor (n-1)/3 \rfloor$ 容错能力，详细的证明过程参考文献[38]。

本文所采用的改进委托拜占庭容错算法如下所述：

设网络中节点数为 N，对每个参与的节点从 $0 \sim N-1$ 依次编号，并按可信度 $trust$ 降序排列，取前 n 个节点作为共识节点；设当前共识区块的高度为 h；并将一次共识从开始到结束所使用的交易数据的集合称为视图，记为 v；令节点 $p = (h - v) \bmod n$。

(1) 任意节点向全网广播交易数据，以及发送者的签名信息；

(2) 所有共识节点均独立监听全网的交易数据，并记录在内存；

(3) 节点 p 在经过时间 t 或系统中总的交易数量达到上限 u 后，发送提案：

$\langle \text{Request}, h, v, p, \text{block}, \langle \text{block} \rangle_{\sigma p} \rangle$

其中 $\langle \text{block} \rangle_{\sigma p}$ 表示节点 p 对消息散列 block 的签名。Request 是请求提案标志。

(4) 其它节点在收到提案后，对交易的合法性进行验证。如果不包含非法交易，则发送响应信息：

$\langle \text{Response}, h, v, i, \langle \text{block} \rangle_{\sigma i} \rangle$

否则开始下一轮共识。Response 是响应标志。

(5) 任意节点在收到至少 $n - f$ 个 $\langle \text{block} \rangle_{\sigma i}$ 后，共识达成并发布完整的区块。

(6) 任意节点在收到完整区块后，将包含的交易从内存中删除，并开始下一轮共识。

信誉度 $trust$ 的计算依据交易双方的正面评价数与负面评价数（分别用 P，F 来表示），α 为系数。节点 n 在第 i 次交易中的信誉度如式 4-8 所示：

$$trust_n^i = \frac{1}{1 + e^{-\alpha(P-F)}} \qquad \text{式 4-8}$$

4.4 数据层

本节主要讨论数据层的计算任务，即如何实现计算隐私（安全多方计算），以及输出隐私（差分隐私）。系统以关系运算为例，详细介绍了选择、连接、排序算法的实现，其它算法可基于该三种算法实现。

首先，注意到在选择（σ）、连接（join）、排序（sort）三种运算中，选择运算只涉及单个关系的运算，后两种运算满足交换律及结合律[39]，即若 R_1, R_2, R_3 表示关系，则连接和排序运算具备如下性质：

交换律:

$$R_1 \cdot R_2 \equiv R_2 \cdot R_1 \qquad \text{式 4-9}$$

结合律:

$$(R_1 \cdot R_2) \cdot R_3 \equiv R_1 \cdot (R_2 \cdot R_3) \qquad \text{式 4-10}$$

可见安全多方计算问题可以转化成安全两方计算（需要侧重考虑共谋攻击问题，本文通过区块链技术解决共谋攻击）。

不失一般性，本节采用 $|||$ 表示秘密共享数据。

1）选择(σ)

数据选择是指从关系 R 中获取满足特定条件的统计属性值（如统计计数、最大（小）值、算术平均值等）。参考[40,41,20,42]，本文设计的保护计算及输出隐私的安全多方选择算法如算法 2 所示。

算法 2 在半诚实模型下满足安全多方计算的安全性需求，同时也满足了差分隐私需求，确保了计

① https://github.com/AntShares/AntShares

算隐私以及输出隐私。

算法 2 保护输出隐私的安全多方选择运算σ

输入：数据查询请求方 P_1 秘密共享的查询向量 $\|x_i\|$，请求隐私预算 ε_j
辅助输入:查询处理方 P_2 的数据集 y 及其总的隐私预算 B_0，
输出：满足隐私需求的统计结果
1. P_2 随机扰乱数据集 y 的顺序（扰乱后的 y 用 $\tilde{y}$ 表示）,令 $i = 0$；
2. 若 $\|x_i\| = \|\tilde{y}_i\|$，且 $\varepsilon_j \le B_0$，则 $B_0 \leftarrow B_0 - \varepsilon_j$，并输出 $\sigma(\|\tilde{y}_i\|) + Laplace(s/\varepsilon_j)$,其中 s 是函数 σ 的敏感度；若 $\varepsilon_j > B_0$ 则提示超出隐私预算，并终止程序。
3. 若 $\|x_i\|$ 不等于 $\|\tilde{y}_i\|$ 则 $i=i+1$ 转步骤 2。
4. 若遍历完所有的 $\|\tilde{y}_i\|$ 记录，则终止程序。

2）连接(join)

数据连接（data join）指将不同数据集中具有相同关键字的记录合并在一起的过程。系统通过子查询的方式实现数据的连接，即秘密共享的数据在各方之间传递以实现最终的连接操作。结合文献[43,44]，本文设计的具备安全性要求的连接操作如算法 3 所示。

其中 $\|\pi_s\|$ 为伪随机置换簇，密钥 $\|s\|$ 唯一确定某一特定置换，且在参与多方计算各方之间秘密共享。

算法 3 数据连接的安全实现

输入：各方秘密共享的数据集 T_i，k_i 表示数据集的主键列
输出：各输入方秘密共享的等值连接数据集 T^*
1. 各计算方茫然的打乱各自的数据集 T_i，并用 $\|T_i^*\|$ 表示扰乱后的数据集，$\|k_i^*\|$ 为扰乱后的主键列。
2. 利用各方之间秘密共享的密钥 $\|s\|$ 选择随机置换函数 $\|\pi_s\|$
3. 各方依次利用置换函数 $\|\pi_s\|$ 评估查询主键列 $\|k_i^*\|$，并将值 $\|\pi_s(k_i^*)\|$ 依次传递给后续计算方。 后续各计算方依次与上一计算方发过来的结果连接，最后生成结果表 T^*

对于两个计算方，设各自维护的数据集中记录数分别为 m, n，则算法的时间复杂度为 $m\log n$。

3）排序(sort)

排序的实质是茫然排序，设 n 个计算方共享秘密向量 $x_1, x_2, \ldots, x_n$，则共享向量表示为 $\|x_1\|, \|x_2\|, \cdots, \|x_n\|$。排序的目的是按照特定的比较原则，确定向量的顺序。结合文献[45,46]，本文设计如算法 4 所示的茫然排序。

算法 4 茫然排序

输入：各计算方秘密共享的向量 $\|x_i\|$（形式上输入向量 $\|x\| = \|x_1\|, \|x_2\|, \cdots, \|x_n\|$）
输出：排序向量 $\|x'\|$
1. 茫然选择各计算方的输入向量（即扰乱 $\|x_i\|$ 之间的顺序，使计算不知道当前正在处理哪个具体的向量）。
2. 当 $1 \le i \le j \le n$ 时，并行秘密计算 $\|g_{i,j}\| = \|x_i\| \le \|x_j\|$
3. 根据解密的 $\|g_{i,j}\|$ 对向量 $\|x\|$ 排序，即获得排序向量 $\|x'\|$

该算法的最坏时间复杂度为 $o(n^2)$。

如第 2 节所述，多方计算实用化需解决“公平性”及“真实性”问题，本文利用区块链在多方计算中的审计功能来实现，即区块链担任与数据计算无关的“理想”审计者角色。

参考文献[47]，本文定义如下的“审计正确”。

设电路 $C = (c_1, c_2, \cdots, c_n)$，其中 c_i 是第 i 个计算参与方的具体计算电路，各计算方秘密共享的数据为 $\|x_i\|$，则 $\|y_i\| = \|c_i(x_i)\|$，即 $\|y_i\|$ 为各计算的潜在输出，τ_i 为评估 c_i 的协议誊本。则审计正确定义为：

$\|y_i\|$ 关于协议誊本 τ_i 是大概率事件。

5 实验评估

本节展示系统核心组件的性能，主要包括：索引性能，区块链网络性能及计算性能。

实验部署在 5 台物理服务器上。每台服务器配备两张独立网卡，每张独立网卡都设置有 200 个虚拟 IP 地址。将这些虚拟 IP 地址的哈希值作为 DHT 空间节点的 ID 值（共 2000 个节点），该值与 LSH 桶对应，同时以 IP 地址作为索引文件存储名。服务器的配置如下：处理器采用 Intel (R) Xeon (R) 2.0G，双 CPU；内存为 2GB；操作采用 Ubuntu 14.04.5 LTS server，内核版本号为 4.4.0-31-generic。5 台服务器分别命名 node1，node2，…，node5，其中 node4，node5 保存有原始数据集，既是数据提供方，又是安全计算的参与方。node1～3 中的任意一个节点都可作为数据需求方。所有节点都作为区块共识的候选节点。

分别采用构造数据集及真实数据集进行测试，数据集的基本信息如表 7 所示。

因无法获取文献[2]的原始数据集，系统采用 python 生成教育数据集(education)和税收数据集(salary)，并与文献[2]数据集属性一致。其中包括 6 年的教育数据集和 10 年的税收数据集。教育数据集的属性包括：个人 ID、课程 ID、性别、出生年月、学历层次（学士、硕士、博士、高等职业教育）、课程规定时长、就读学校，入学日期、学习状态（进行中、退学、毕业）、毕业日期/退学日期等 10 个属性，分别对其中的个人 ID，课程 ID 建立域索引。税收数据集的属性包括个人 ID、年、月、缴纳的社会保险费用、股息收入、雇主是否来自 ICT（或 ITL）等 6 个属性，只对个人 ID 属性建立域。类似地,分别对 student, CT slice 中的 student ID 集及 patent ID 建立域。

表 7 评估数据集

name	n	d	b	r
education	283470	10	10/5	[1-4]
salary	785	6	5	[1-4]
student[48]	1045	30	5	[1-4]
CT slice[49]	53500	384	10	[1-4]

(说明：education, salary 是根据文献[2]生成的数据集；student, CT slice 是真实数据集，数据集的描述参见相关文献。为增加数据集间的关联性，在 student 数据集中增加了 student ID 字段，其值来之 CT slice 中的 patent ID。表中 n 表示记录数，d 表示属性数，b 代表域划分的条数(bands)，r 是每条内的行数(rows))

软件实现依赖开源软件，其中索引部分依据 minHash LSH Ensemble① 构建；区块链采用 ethereum②技术，并修改了块结构以及共识机制；多方计算方面借鉴 obliv-c③[50]技术及编程语言[51]。实验评估结果如下：

5.1 索引性能

查全率（也称为召回率，recall）、查准率（也称为精度，precision）是搜索引擎的两个重要指标（另外一个重要指标，F 值依赖于查全率和查准率），也是衡量本系统的重要指标。本文根据不同的关联度阀值 t 与基准方法（即传统的关键字搜索）及 LSH ensemble[30]算法就查准率及召回率进行了比较，其结果如图 6、图 7 所示。

关键词搜索的关联度阈值定义为关键词向量与查询域向量的余弦距离。从图 6 中可以看出，在相似度阈值低于 0.7 时，随着相似度阈值的增加，精度都在一定程度降低，这是因为随着相似度的增大，被正确检索的项数有所减少；而当相似度阈值高于 0.7 时，因为伪正数量的迅速减少，各种方法的精度都呈现快速增加的趋势。总体来看本文提出方法的查询精度平均优于基准方法 22%，同时由于对 LSH 算法的改进，使得关联度阈值在[0.4,0.7]之间时，平均精度提高了 1.4%。图 7 显示，本文提出的索引及搜索方法仅导致召回率略微降低。

5.2 区块链网络性能

出块速度是大多数区块链网络的硬性限制，如比特币的出块速度大约为每 10 分钟一块。在块大小(如比特币网络是 1MB)一定的情况下，出块速度

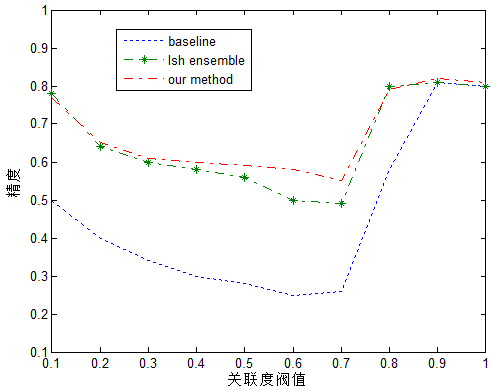


图 6 关联度与查询精度

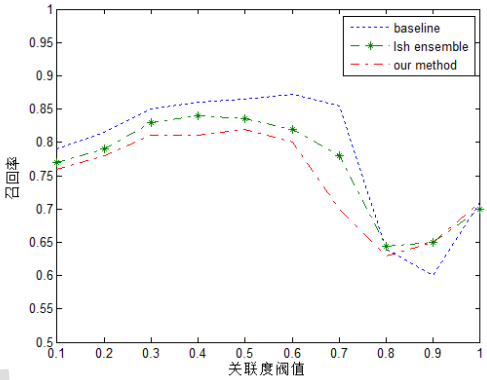


图 7 关联度与召回率

决定了单位时间内能够处理的交易数，可见，出块速度是影响系统的实时性及网络增长率的重要因素[44]。在 dBFT 共识算法中，时间间隔 t，上限交易数 u 以及网络节点数都与出块速度相关。

由于在不同时段的交易次数存在随机性，实验时按照一天中 internet 访问量统计规律④模拟交易量提交速率。并设定上限交易数 u =50, 共识节点数 n=5，分析区块大小在一天中的变化规律，结果如图 8 所示。

从图 8 可以看出，在交易量降低的时候，以时间间隔作为候选区块标准，当交易量较高的时候，则以区块数作为候选区块标准。可见在共识算法中附加交易区块数的限制，可以提高高峰期交易频率。

① https://ekzhu.github.io/datasketch/lshensemble.html
② https://github.com/ethereum/go-ethereum
③ https://github.com/samee/obliv-c
④ http://shcci.eastday.com/c/20160202/u1a9207794.html

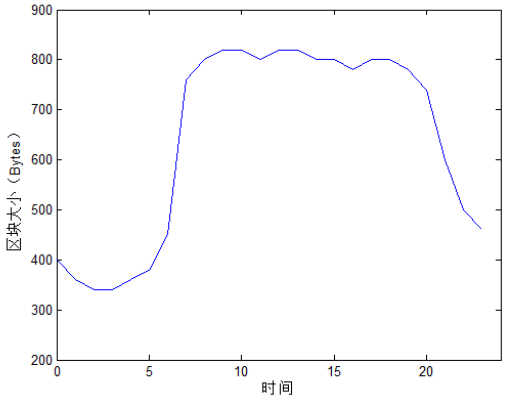


图 8 一天中区块大小模拟变化规律

5.3 计算性能

主要测试在同等隐私保护环境下，算法处理数据集的性能，即运行时间与数据集大小的关系。为此将测试数据集划分成不同大小的数据集，数据集大小（记录数）分别为：300，1000，10000，100000，200000。并与文献[8,52]的加密方案进行对比，结果如图 9 所示：

从图 9 可以看出，相对于数据加密存储，本文提出的方法显著降低了系统运行时间。

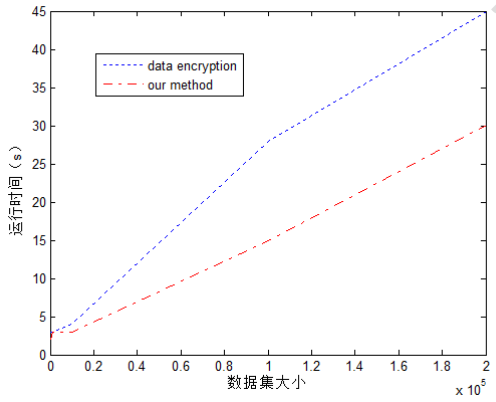


图 9 数据计算耗时对比

6 结论与展望

6.1 结论

大数据时代，数据作为一种数字资产是社会发展的原动力，去中心化数据共享模型作为一种全新的数据资产管理平台，有效地解决了数据所有权管理，隐私保护以及数据发现、数据安全计算、数据审计等。区块链技术在本文中的应用表现在以下几个方面：（1）将系统的所有事务处理都作为交易，统一了事务处理逻辑；（2）记录历史交易信息，并以密码学的方式保证信息的不可篡改和不可伪造性；（3）利用改进的 dBFT 共识算法来生成和更新数据，既解决了系统资源浪费也提高了系统事务处理能力；（4）利用智能合约来操纵原始数据，有效解决数据隐私及所有权问题；（5）解决了安全多方计算中的两个固有缺陷：在不满足大部分参与方诚实可信条件的情况下，不能提供公平性及协议只负责计算安全而不确保用户向协议提供真实的输入并且遵守输出结果；（6）解决数据集的使用策略问题等。

6.2 未来展望

为测试效率和准确性，我们开发了概念原型系统，创新性解决了实际需求中的问题。开发中也暴露出一些缺陷，总结下一步值得研究的一些问题主要包括：

1）计算扩展性

系统目前仅实现三种基本运算，即选择、连接运算和排序。虽然大多计算问题可以归结为这三种基本运算问题，如加法电路及乘法电路可扩展成通用电路[53,54]，但针对特定运算，其效率仍相对较低。系统实用化面临较多的计算类型，应根据不同的应用场景及数据类型，研究高效的计算挖掘算法，如针对图的计算、针对高维数据的计算等。另一方面将计算迁移到数据提供方，势必增加其所属托管服务器的负载。但这一思想正与现有大多数系统一致，如数据开放平台提供了数据分析、可视化能力。其中，增长最快速的当属基于 DaaS（数据即服务）理念的 API 服务。

2）数据定价

数据作为一种资产，价值属性是其固有属性。但是数据也是一种特殊的资产，如何对数据定价是业界广泛讨论的问题[21,22,55,56]。本系统作为共享交易系统，离不开数据的定价，同时数据交易价格也是系统激励机制考虑因素之一。

制约数据价格的主要因素有：数据价值、数据质量、成本价格（数据收集、传输、存储等产生的成本）等。

API 定价则根据服务的性质来确定，可以采用调用 API 次数确定相应的价格，如淘宝开放平台[①]采用的定价策略。

本系统的定价策略属于 API 定价，每笔交易有

① https://open.taobao.com

确定的价格。这种定价方式未考虑数据本身属性，因而很难反应数据固有价值。

3）系统效率

系统效率受以下因素制约：（1）检索性能；（2）交易频率；（3）多方计算性能。

对网络数据及服务而言，数据检索时间是重要因素。实验仅针对小规模数据，对大体量、异构数据集还有待定量分析及检验。

提高交易频率主要考虑增加区块大小及提高确认速度。但这两者都带来性能上的损失，前者形成超大区块体积，限制了在资源受限型设备上的应用；后者在一定程度上带来安全性能的损失[56,57]。根据系统交易特性设计更好的共识算法，是下一步提高系统交易速度应重点考虑的问题。

4）计算区块链

在区块链 1.0 时代，区块链作为一种分布式账本；区块链 2.0 时代加入了智能合约，使区块链具有了一定的智能；那么在区块链 3.0 时代，其智能程度应该进一步提高，能够“自适应”地处理各种需求，本系统正是基于这一思想，其适用性有待进一步检验。

参 考 文 献

[1] Guo Bing, et al. Personal Data Bank: A New Mode of Personal Big Data Asset Management and Value-added Services Based on Bank Architecture. Chinese Journal of computers, 2017 (1): 126-143(In Chinese) (郭兵等. "个人数据银行——一种基于银行架构的个人大数据资产管理与增值服务的新模式." 计算机学报 2017(1):126-143.)

[2] Bogdanov D, Kamm L, Kubo B, et al. Students and Taxes: a Privacy-Preserving Study Using Secure Computation. Proceedings on Privacy Enhancing Technologies, 2016, 2016(3):117-135.

[3] Dan Bogdanov. Sharemind: programmable secure computations with practical applications. PhD thesis, University of Tartu, 2013.

[4] R Talviste. Applying Secure Multi-party Computation in Practice PhD thesis, University of Tartu, 2016.

[5] Schwab K, Marcus A, Oyola J O, et al. Personal data: The emergence of a new asset class. Geneva: Forum World Economic, Technical Report:2011

[6] Zhang Wen-Li. Et al Computation Offloading on Intelligent Mobile Terminal. Chinese Journal of computers, 2016, 39(5):1021-1038 (In Chinese) (张文丽, 郭兵, 沈艳,等. 智能移动终端计算迁移研究. 计算机学报, 2016, 39(5):1021-1038.)

[7] Franklin, Michael, A. Halevy, and D. Maier. "From databases to dataspaces." Acm Sigmod Record 34.4(2005):27-33.

[8] Zyskind, Guy, O. Nathan, and A. '. Pentland. Decentralizing Privacy: Using Blockchain to Protect Personal Data. Security and Privacy Workshops IEEE, 2015:180-184.

[9] Zhang H, Wen Y, Xie H, et al. DHT Theory. Springerbriefs in Computer Science, 2013:5-22.

[10] Rao B C, Zhu E. Searching Web Data using MinHash LSH-International Conference on Management of Data. ACM, 2016:2257-2258.

[11] Zamora J, Mendoza M, Allende H. Hashing-based clustering in high dimensional data[J]. Expert Systems with Applications, 2016, 62:202-211.

[12] Andrychowicz, Marcin, et al. Secure multiparty computations on Bitcoin. Security and Privacy IEEE, 2014:443-458.

[13] Blunschi L, Dittrich J P, Girard O R, et al. A Dataspace Odyssey: The iMeMex Personal Dataspace Management System (Demo) CIDR 2007, Third Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 7-10, 2007, Online Proceedings. DBLP, 2007:114-119.

[14] Li, Yang, and C. Hu. Process Materials Scientific Data for Intelligent Service Using a Dataspace Model. Data Science Journal 15.3(2016).

[15] Montjoye, Yves Alexandre De, et al. openPDS: Protecting the Privacy of Metadata through SafeAnswers. Plos One 9.7(2014)

[16] Nakamoto, Satoshi. "Bitcoin: A peer-to-peer electronic cash system." Consulted (2009).

[17] Dennis R, Owen G. Rep on the block: A next generation reputation system based on the blockchain. Internet Technology and Secured Transactions. IEEE, 2015:131-138.

[18] Yong, etc Blockchain: The State of the Art and Future Trends. acta automatica sinica. 2016, 42(4):481-494. (In Chinese) (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.)

[19] Hazay, Carmit, and Y. Lindell. Efficient Secure Two-Party Protocols: Techniques and Constructions. Springer Berlin Heidelberg, 2010.

[20] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[J]. Foundations & Trends® in Theoretical Computer Science, 2014, 9(3):211-407.

[21] Rajaraman A, Ullman J D. Mining of massive datasets. Cambridge University Press, 2016.

[22] Shrivastava A, Li P. Asymmetric Minwise Hashing for Indexing Binary Inner Products and Set Containment. International Conference on World Wide Web. International World Wide Web Conferences Steering Committee, 2015:981-991.

[23] Kosub S. A note on the triangle inequality for the Jaccard distance. 2016.

[24] Fernandez R C, Abedjan Z, Madden S, et al. Towards large-scale data discovery: position paper. International Workshop on Exploratory Search in Databases and the Web. ACM, 2016:3-5.

[25] Alon Halevy, Flip Korn, Natalya F. Noy, etc. Goods: Organizing Google's Datasets. International Conference on Management of Data. ACM, 2016:795-806.

[26] Ferrucci D A, Brown E W, Chu-Carroll J, et al. Building Watson: An

Overview of the DeepQA Project. Ai Magazine, 2010, 31(3):59-79.

[27] Litvak M, Last M. Graph-based keyword extraction for single-document summarization Mmies 08 Workshop on Multi-source Multilingual Information Extraction & Summar. 2008:17--24.

[28] Mihalcea R, Tarau P. TextRank: Bringing order into texts. Association for Computational Linguistics, 2004.

[29] Stoica I, Morris R, Karger D, et al. Chord: A scalable peer-to-peer lookup service for internet applications Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications. ACM, 2001:149-160.

[30] Zhu E, Pu K Q, Pu K Q. LSH ensemble: internet-scale domain search. Proceedings of the Vldb Endowment, 2016, 9(12):1185-1196.

[31] Huh S, Cho S, Kim S. Managing IoT devices using blockchain platform International Conference on Advanced Communication Technology. IEEE, 2017.

[32] Christidis K, Devetsikiotis M. Blockchains and Smart Contracts for the Internet of Things. IEEE Access, 2016, 4:2292-2303.

[33] Mattila J, Seppälä T. Industrial Blockchain Platforms: An Exercise in Use Case Development in the Energy Industry Etla Working Papers, 2016.

[34] P. Flajolet, E. Fusy, G. O., and F. Meunier. Hyperloglog: The analysis of a near-optimal cardinality estimation algorithm. Analysis of Algorithms (AOFA), 2007.

[35] Latanya sweeney. k-anonymity: a model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 2002, 10(05):557-570.

[36] Machanavajjhala A, Gehrke J, Kifer D, et al. L-diversity: privacy beyond k-anonymity. International Conference on Data Engineering. IEEE, 2006:24-24.

[37] Liu Y, Cui J, Huang Z, et al. SK-LSH: an efficient index structure for approximate nearest neighbor search[J]. Proceedings of the Vldb Endowment, 2014, 7(9):745-756.

[38] Zhengwen Zhang A Byzantine Fault Tolerance Algorithm for Bl-ock Chains. http://docs.neo.org/zh-cn/node/consensus/whitepaper.html Retrieved 2017.8 (In Chinese) (张铮文 一种用于区块链的拜占庭容错算法. http://docs.neo.org/zh-cn/node/consensus/whitepaper.html)

[39] Ioannidis Y E. Query optimization[J]. Acm Computing Surveys, 1996, 28(1):121-123.

[40] Kamm, P., and L. Laud. Applications of Secure Multiparty Computation. IOS Press, 2015.

[41] Akter M, Hashem T. Computing Aggregates Over Numeric Data with Personalized Local Differential Privacy. Information Security and Privacy. 2017.

[42] Kifer D. Privacy and the Price of Data. ACM/IEEE Symposium on Logic in Computer Science. IEEE Computer Society, 2015:16-16..

[43] Laur S, Talviste R, Willemson J. From Oblivious AES to Efficient and Secure Database Join in the Multiparty Setting Applied Cryptography and Network Security. Springer Berlin Heidelberg, 2013:84-101.

[44] Laur S, Willemson J, Zhang B. Round-Efficient Oblivious Database Manipulation Information Security. Springer Berlin Heidelberg, 2011:262-277.

[45] Hamada K, Kikuchi R, Dai I, et al. Practically Efficient Multi-party Sorting Protocols from Comparison Sort Algorithms Information Security and Cryptology – ICISC 2012. Springer Berlin Heidelberg, 2012:202-216.

[46] Dan B, Laur S, Talviste R. A Practical Analysis of Oblivious Sorting Algorithms for Secure Multi-party Computation Secure IT Systems. Springer International Publishing, 2014:59-74.

[47] Baum C, Damgård I, Orlandi C. Publicly Auditable Secure Multi-Party Computation[M] Security and Cryptography for Networks. 2014:175-196.

[48] P. Cortez. Student performance data set. https://archive.ics.uci.edu/ml/datasets/Student+Performance, Retrieved 2017.8

[49] F. Graf, H.-P. Kriegel, M. Schubert, S. Poelsterl, and A. Cavallaro. Relative location of ct slices on axial axis data set. https://archive.ics.uci.edu/ml/datasets/Relative+location+of+CT+slices+on+axial+axis. Retrieved 2017.8

[50] Samee Zahur, David Evans. Obliv-C: A Language for Extensible Data-Oblivious Computation. http://pdfs.semanticscholar.org/2d25/81b990fd8b2df020cea5a6392b15f771bf0a.pdf Retrieved 2017.8

[51] Doerner J, Evans D, Shelat A. Secure Stable Matching at Scale. ACM Sigsac Conference on Computer and Communications Security. ACM, 2016:1602-1613.

[52] Zyskind G, Nathan O, Pentland A. Enigma: Decentralized Computation Platform with Guaranteed Privacy[J]. Computer Science, 2015.

[53] Xiong Ping, et al. A Survey on Differential Privacy and applications Chinese Journal of computers, 2014,37 (01): 101-122(In Chinese)(熊平等. 差分隐私保护及其应用[J]. 计算机学报, 2014, 37(01):101-122.)

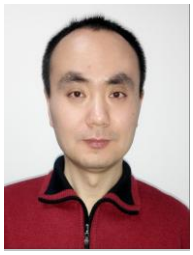
[54] Pettai M, Laud P. Combining Differential Privacy and Secure Multiparty Computation. The Computer Security Applications Conference. 2015:421-430.

[55] Nget R, Cao Y, Yoshikawa M. How to Balance Privacy and Money through Pricing Mechanism in Personal Data Market. 2017.

[56] Shen Y, Guo B, Shen Y, et al. A Pricing Model for Big Personal Data. Tsinghua Science & Technology, 2016, 21(5):482-490.

[56] Wang H, Chen K, Xu D. A maturity model for blockchain adoption. Financial Innovation, 2016, 2(1):12.

[57] Yuan Y, Wang F Y. Towards blockchain-based intelligent transportation systems. IEEE, International Conference on Intelligent Transportation Systems. IEEE, 2016:2663-2668.



Dong Xiang-Qian, born in 1975, Ph. D. candidate, lecturer. His current research interests include information security and personal privacy protection .

Guo Bing, born in 1970, Ph. D., professor. His current research interests include green computing and personal big data

Shen Yan, born in 1973, Ph. D., professor. Her research interests include smart terminal and instruments.

Duan Xu-Liang, born in 1982. Ph.D. candidate, lecture. His current research interests include personal big data, big data cleaning and data mining.

Shen Yun-Cheng, born in 1979, Ph. D. candidate, lecturer. His current research interests include personal big data and big data pricing.

Zhang Hong, born in 1980, Ph. D. candidate, lecturer. His current research interests include mobile internet data provenance and personal big data.

Background

More and more applications rely on sharing-data to obtain new insights. These datasets vary in their formats, change every day, and especially reside in different repositories. The power to select subsets of interest from these untamed datasets and protect the legal right of the data owners has become the major bottleneck. Data Lake, as one of the widely accepted conceptual model for data sharing management, is a set of centralized repositories containing vast amounts of raw data (either structured or unstructured), described by metadata, organized into identifiable data sets, and available on demand. But the problems, such as data discovery, rights management, data security, centralized storage etc., exposed in the process of data maintain impede it to develop further. BlockChain promises a new technology innovation to solve these flaws, as it guarantees transparency over how applications work and leave an irrefutable record of activities, provide strong incentives for honest behavior.

At present, many applications for above problems have been built. They address the problems, such as decentralized PKI service, identity management, incentive scheme, privacy enforcing computation, while ignore the issue of how to efficiently discovery linkable datasets, to provide the primitive operations of these datasets, and to design transaction for these specific applications. This paper tries to present the data sharing model architecture, which solves the problem of data discovery, data pricy and transaction management, which boosts by linkable data discovery through domain index, protects computing and output privacy through security primitive operator, meanwhile designs special transaction data format and protocol to control the process of data transaction. This work is partly supported by the State Key Program of National Natural Science Foundation of China under Grant No.61332001, the National Natural Science Foundation of China under Grant No. 61772352 and 61472050. Our group has been working on the personal big data and BlockChain. Many papers have been published in international conferences and journals, such as CBD, Tsinghua Science and Technology etc.