山西大学学报(自然科学版)47(6):1211—1220,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023117

基于隐私信息检索的大规模用电增信查询方法

李辉[1],黄祖源[1],田园[1],毛正雄[1],赵鹏[2],任雪斌[2],李亚男[3*]
(1.云南电网有限责任公司 信息中心,云南 昆明 650011;
2.西安交通大学 计算机科学与技术学院,陕西 西安 710049;
3.河南理工大学 软件学院,河南 焦作 454003)

摘 要:用电信用报告已经成为企业增信的重要凭证,但现有电力金融服务平台在提供用电信用报告查询时存在未保护查询方的查询偏好隐私信息,并且难以支持大规模数据库查询两方面问题。针对上述两个问题,提出一种基于隐私信息检索和不经意多项式计算的安全高效检索方法Effi-Retrieval。具体地,使用Paillier同态加密和不经意多项式计算实现查询方的偏好隐私和电力金融平台的数据库安全。此外,基于k-匿名方法在实现查询方的个性化隐私需求同时,结合哈希映射设计了最优分桶策略,用以降低查询方和电力金融服务平台间的通信开销。综上两方面策略,Effi-Retrieval将传统隐私信息检索的复杂度由数据库规模的指数函数降低为匿名参数k。最后,给出了Effi-Retrieval的安全性分析和用户隐私需求和承担费用对通信开销影响的数值实验。本文使用python的paillier同态加密库编写实验代码并在单台主机上模拟客户端与服务器间的交互,在FATE开源联邦平台提供的Credit Card数据集上进行了实验。结果表明,在数据规模300及以上的数据库上使用8 192密钥位数加密时,与不使用分桶策略的传统方法相比,Effi-Retrieval的密态多项式生成时间可降低50%以上,与不使用k-匿名方法的查询算法相比,Effi-Retrieval的检索时间可降低30%以上。

关键词:电力增信查询;隐私信息检索;不经意多项式计算;k-匿名;通信开销
中图分类号:TP309.7 **文献标志码**:A **文章编号**:0253-2395(2024)06-1211-10

Privacy Information Retrieval Based Credit Inquiry for Large-scale Electricity Users

LI Hui[1], HUANG Zuyuan[1], TIAN Yuan[1], MAO Zhengxiong[1], ZHAO Peng[2], REN Xuebin[2], LI Ya'nan[3*]
(1. *Network Information Center, Yunnan State Grid, Kunming* 650011*, China;*
2. *School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an* 710049*, China;*
3. *School of Software, Henan Polytechnic University, Jiaozuo* 454003*, China*)

Abstract: Electricity usage-based credit report is an important way for enhancing the query accuracy of enterprises' credit conditions. However, there are two main problems in existing credit inquiry service. The first problem is that existing methods do not protect the inquiry preference, which is the private information. The second is that existing private information retrieval method cannot be extended to larger-scale database due to the poor efficiency. To address the above problems, a novel inquiry method named Effi-Retrieval was proposed to simultaneously ensure the inquiry security and promote its efficiency. Particularly, Paillier homomorphic encryption and oblivious polynomial calculation were used to ensure the security of inquiry. Optimal binning strategy with k-

收稿日期:2023-04-22;**接受日期**:2023-06-21
基金项目:云南电网科技项目(YNKJXM20210141)
作者简介:李辉(1991-),男,云南昆明人,硕士研究生,中级工程师,研究方向为数据的挖掘与分析。E-mail:1518888868@qq.com
* **通信作者**:李亚男(LI Ya′nan),E-mail:liyn@hpu.edu.cn
引文格式:李辉,黄祖源,田园,等.基于隐私信息检索的大规模用电增信查询方法[J].山西大学学报(自然科学版),2024,47(6):1211-1220. DOI:10.13451/j.sxu.ns.2023117

anonymity method and hash mapping were used to achieve the personalized privacy requirements of the query party in practice. The complexity of privacy information retrieval was reduced from the exponential function of the database size to the anonymity parameter k. Finally, the security analysis of Effi-Retrieval was presented and numerical experiments were conducted to show the impacts of user privacy requirements and affordable fee on communication overhead. We used the Paillier homomorphic encryption library of Python to simulate the interaction between the client and the server on the same host, and conducted experiments using the Credit Card dataset provided by the FATE open-source federated learning platform. Given database consists of at least 300 items encrpyted by 8 192 key bits, experimental results showed that compared with the traditional method without binning strategy, the encryption polynomial construction time of Effi-Retrieval can be reduced by more than 50%, and compared with the query algorithm without k-anonymity method, the retrieval time can be reduced by at least 30%.

Key words: power credit enhancement inquiry; privacy information retrieval; oblivious polynomial computation; k-anonymity; communication cost

0 引言

电力数据与社会各行业间存在着天然的强耦合关系,电力数据是国民经济发展和企业生产规模的晴雨表,基于电力数据的分析可赋能多种行业领域。例如,在政务方面,对家庭用电监测和企业用电模式分析可帮助政府识别当地“空心村”(空心村是指随着我国城市化和工业化进程,大量的农村青壮年劳动力流向城市,导致农村常住人口减少,村庄内出现了许多闲置房屋和宅基地的情况)或企业复工复产情况,为决策提供数据支撑。在城建规划方面,结合人口地理数据和区域用电量分布,可为商业中心选址提供科学支撑。在金融领域,电力公司出具的企业用电和缴费报告则可作为在贷款审核中企业增信的重要凭证。

相对于使用历史贷款及违约信息作为主要指标评价企业信用的方法,其存在时效性差、覆盖面窄等缺点。而作为反映企业生产经营状况的时序变量,企业用电数据具有时效性强、及时性高、覆盖面广等显著特点,因此,基于用电数据的信用报告已经成为企业增信的重要凭证并已经纳入国家优化和健全社会信息体系的重要组成部分。例如,国家电网发布的线上产业链金融平台和南方电网发布的南方融e互联网金融服务平台中,均已经将上下游企业的用电行为作为商业保理、信托融资、融资租赁的增信凭证。

现有的电力增信系统的隐私保护侧重于对企业和个人征集信息的保护,比如得到客户或征集平台授权前不得泄露客户信息,平台对外发布客户信用数据时去除与增信无关的结构化隐私信息等。但是在查询过程中金融服务平台和查询方之间采用的明文查询请求和原始数据结果反馈会造成两个方面的隐私安全风险。首先,明文的查询请求会导致查询方的隐私信息泄露。例如,若电力金融服务平台为半诚信的,则其能掌握某企业是否向特定的金融机构申请贷款。其次,明文形式的查询请求和结果反馈会导致企业的信息泄露。例如,攻击者可以通过信道监听获取查询方的查询对象以及查询结果,从而导致查询方利益受损。

为解决上述问题,本文基于隐私信息检索(Private Information Retrieval, PIR)技术设计一种支持关键词的电力增信报告查询方法。PIR是一种两方安全查询协议,其保证查询方在不披露查询信息的前提下完成对特定信息的查询。其工作原理为:查询方将查询关键词加密,并将密文发送给数据库所有方,后者执行约定的查询协议,并将计算的密文结果发送给查询方,查询方解密后得到相应信息。隐私信息检索的优势在于保护了查询方的查询信息,让数据提供方和外部攻击者无法获知查询方的查询偏好信息。将隐私信息检索技术用于电力增信系统能够有效地保护金融服务平台中电网的数据安全和查询方的查询偏好隐私,增强现有电力金融平台的隐私安全性。

由于隐私信息检索的计算和通信复杂度通常与数据库规模呈线性关系,因此当数据库规模很大时如何提升查询效率是需要解决的一个关键问题[1]。本文基于不经意传输协议实现电力增信查询[2],其主要思想为在满足实际场景中用户个性化隐私需求前提下,对满足隐私需

求的数据库子集上执行查询,通过最小化通信和计算开销设计最优分桶数量,并在每个桶上进行不经意多项式计算,以显著降低通信开销和计算成本。其中的个性化隐私需求基于k-匿名技术实现[3]。本文设计的电力增信查询系统架构如图1所示,其中参与方包括用电侧的家庭和企业用户、供电侧的电力金融服务平台、查询侧的银行等第三方金融机构,其实施流程步骤如下。

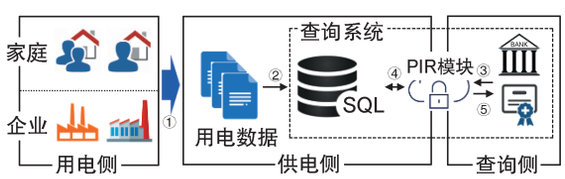

图1 基于隐私信息检索的电力增信查询架构图

Fig. 1 Framework of PIR-based power enhancement inquiry

步骤①:电网向家庭和企业用户提供电力服务,家庭用户产生历史用电量、缴费数据等用电数据;企业用户产生用电规模、分时段用电量、累计用电时间、缴费数据等用电数据。

步骤②:电网基于用户的用电数据,包括用电量、缴费时间、缴费金额、违约金、开户年限、缴费总次数、欠费总次数,生成用电信用报告数据库。

步骤③:银行等信用机构针对申请贷款的企业或个人向电力金融平台发起用电增信报告查询,查询支持关键词查询,并且目标关键词以密文形式发送。

步骤④:电网接收到的密文形式的目标关键词输入到PIR模块,并将输出的密文查询结果发送给查询方,在整个过程中电网不能获得查询方的查询信息。

步骤⑤:查询方将电网发送的计算后的密文结果进行解密和消息验证,若数据库中含有该查询对象则查询方得到正确查询结果,否则得到特定标识符表示无此关键词。

其中,步骤③—⑤是实现隐私信息检索的关键技术核心,需要供电侧和查询侧双方基于PIR协议进行交互,其详细技术细节见第4节。本文的主要贡献如下。

本文首次设计一种基于PIR的企业用电增信查询方法,该方法可用于电力金融平台用以保证电力金融平台的数据库安全和查询方的查询偏好隐私。

本文设计的PIR协议在满足k-匿名隐私安全前提下,可显著降低金融服务平台的计算开销,以及金融服务平台与查询方的通信开销,协议复杂度为$O(k+\sqrt{k})$。

本文证明了设计的PIR协议的安全性,并通过数值实验分析协议中的查询者的隐私安全需求和可承担费用对协议通信开销和用户计算开销的影响。

1 相关工作

1.1 PIR技术研究

PIR最早由Chor等[4]于1995年提出:对于给定的长度为n的二元数据库$\{x_1,x_2,\cdots,x_n\}$,其中$x_i\in\{0,1\}$,$i=1,2,\cdots,n$是数据的索引,用户向数据库发送查询第i个数据的请求,数据库服务器需要在不能获取信息i的前提下将x_i发送给用户。根据实现方法,PIR可分为信息论安全的隐私信息检索(Information-theoretic private information retrieval,IPIR)和计算安全的隐私信息检索(Computationally private information retrieval,CPIR)。考虑存储和通信复杂度,实际场景中通常采用CPIR。Chor等[5]指出仅需最小数量的数据库副本即可达到多项式的通信复杂度,并提出一种2-服务器端PIR协议,通信复杂度为$O(n^{\epsilon}),\epsilon>0$,但其要求的两个服务器不能共谋的要求在实际场景中不一定成立。Lipmaa等[6]使用长度可变的加法同态公钥密码系统,提出一种改进后的PIR协议,通信复杂度为$O\left[\log^2(n)\right]$,但计算复杂度至少为$O(n)$。随后,Kiayias等[7]提出了一种通信复杂度为常数级别的CPIR方案,设计了一种针对PIR的最优同态加密方案,其计算复杂度达到$O(\log m)$级别,其中m是同态加密多项式的次数,但协议主体仍依赖同态加密,查询的计算复杂度达到$O(n)$;Li等[8]研究了在多值检索下的PIR技术中降低通信复杂度的问题,将通信复杂度降低到了$O\left[\log(n)\right]$,但上述研究仅考虑如何降低通信复杂度,没有综合考虑时间复杂度和通信复杂度,有可能产生高额的计算开销。近年来,

PIR技术的一些变体也相继被研究,包括具有极大距离可分码(Maximum distance separable code,MDS)编码存储的PIR[9]、多消息的PIR[10]、多轮次交互PIR[11]、服务器串通下的安全PIR[12]、基于考虑辅助信息的PIR[13]、对称安全的PIR[14-15]。Li等[9]利用MDS编码降低了对数据库不共谋的要求,但增加了存储成本;Banawan等[10]主要用于一次查询多条消息的场景,优化了查询多条消息的计算效率,但会增加查询的通信量;Wang等[11]基于上次交互结果降低存储成本,但需要更多的计算时间;Jia等[12]虽然允许服务器在一定程度上的串通,但仍不符合实际需求;Chen等[13]利用辅助信息简化查询降低查询成本,但如何获取好的辅助信息仍然比较困难;陈磊磊等[14]利用量子安全的密钥交换网络实现了对称安全的PIR,但带来了较大的计算成本;俞志斌等[15]探讨了对称PIR的总通信成本与条件秘密披露之间的关系,设计了特定数据数目下达到最小通信的方法,但仍然需要两个共谋的服务器。本文考虑单服务的实际场景,并综合通信和计算复杂度设计最优的分桶数量降低通信和计算开销。

1.2 PIR应用研究

PIR技术已经被应用于医疗系统中药物或疾病的查询[16],证券市场中股票代码的查询[17],互联网搜索引擎中用户的查询[18]以及身份认证系统中的身份查询[19]等,这些场景的共同特点是查询方发起的查询内容具有隐私性,查询方不期望服务器获得其具体查询内容。Badr等[20]将PIR技术与区块链相结合应用在智能停车场系统中,以帮助司机检索附近的优惠停车区域并匿名预订停车位。丁佳晨等[21]提出了一种基于比特币的PIR支付协议,通过控制交易兑现的条件,使得如果服务方相互串通,则服务方会遭受经济上的损失,由此降低共谋的可能性。Attia等[22]研究了PIR技术在非编码存储的受限数据库中的应用,进一步扩展了PIR技术的应用范围。目前,尚未有文献将PIR技术应用到电力增信应用中。

2 企业模型和背景知识

为叙述方便,服务器表示金融服务平台,代表服务提供方,用户表示银行等金融机构,代表查询方,数据库为企业用电信用报告。为方便表述,以企业用户为对象介绍基于PIR的用电增信报告查询方法。

2.1 系统模型

正文内容服务器$\mathcal{S}$有关于企业的用电信用报告数据库$\mathcal{X}=\{x_1,\cdots,x_i,\cdots,x_n\}$,每条数据$x_i=\{w_i,m_i\}$包括关键词$w_i$以及对应用电信用报告$m_i$,其中$n$为数据库规模。如图2所示,关键词$w_i$为企业名称或统一社会信用代码,具有唯一标识性;信用报告m_i包括当月电费、当月是否欠费、历史逾期交费金额、历史欠费金额。

关键词(w_i)		增信报告内容(m_i)
企业名称	统一信用代码	(当月电费,当月是否欠费,历史逾交金额,历史欠费金额)
企业a	Code a	12万元,否,0万元,0万元
企业b	Code b	21万元,是,30万元,3.7万元
…		…
企业n	Code n	8.6万元,否,10万元,5.2万元

图2 用电信用报告数据库结构示例

Fig. 2 Toy example of power enhancement database

基于PIR技术的用电增信查询模型见图3,包含八个步骤。步骤(1)—(3)中,用户和服务器首先基于k-匿名隐私需求确认支付费用μ,然后用户$\mathcal{U}$支付费用μ并将包含w_*在内的关键词列表上传至服务器。步骤(4)—(5)中,服务器在用户指定的数据库子集上执行分桶和多项式构建。步骤(6)中,用户上传关键词w_*的1到m次加密密文。步骤(7)中,服务器在密文上执行多项式计算。步骤(8)中,用户解密对应桶上的密文得到查询结果:如果存在i使得$w_i=w_*$,则用户$\mathcal{U}$获得m_i,否则获得标记符$\perp$表示数据库不包含目标关键词。

系统需求包含查询正确性、用户隐私安全和服务器数据安全三部分,系统的功能需求实现主体是PIR协议的设计。查询正确性指如果服务器和用户都是诚实地执行协议,则对于输入w_*,用户获得m_i使得$m_i=m_*$,或者$\perp$当$m_*\neq m_i,\forall i=1,\cdots,n$。用户隐私安全指对于任意的服务器$\mathcal{S}'$和输入$\boldsymbol{\mathcal{X}},w_*,w'_*$,服务器$\mathcal{S}$不能在多项式计算时间内区分输入$\boldsymbol{\mathcal{X}},w_*$和$\boldsymbol{\mathcal{X}},w'_*$。服务器数据安全指对于理想用户$\mathcal{U}'$和实际协

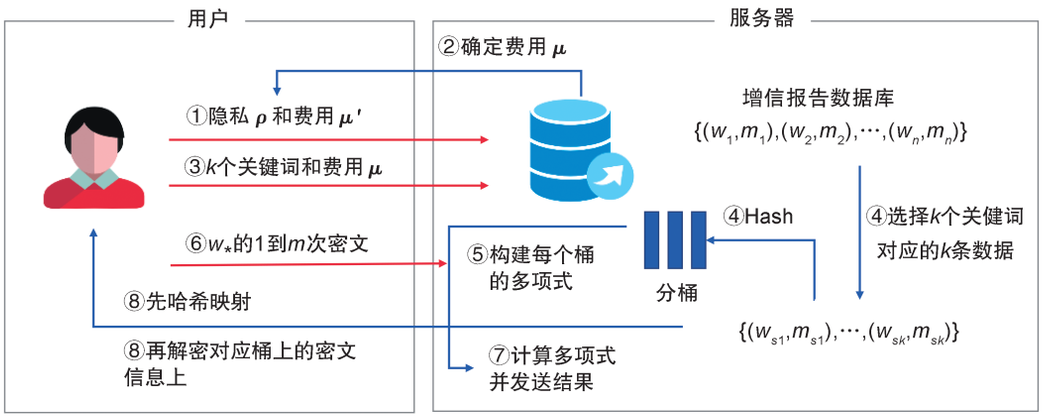

图3 基于隐私信息检索技术的用电信用报告查询流程

Fig. 3 Pipeline of PIR based inquiry

议中的用户$\mathcal{U}$,他们对于任意输入$(\mathcal{X},w_*)$得到的结果在多项式计算时间内不可区分,即用户不能从查询结果中得到额外的数据库信息。

2.2 预备知识

本文设计的基于PIR技术的用电增信查询方法中用到不经意多项式计算和Paillier同态加密知识。Paillier加密在域$M=Z_n$上满足加法同态性,即$E_{pk}(m_1)E_{pk}(m_2)=E_{pk}(m_1+m_2)$,其中$m_1,m_2$是两条给定的消息,$pk$是加密算法$E$的公钥。其加解密流程如下:(1)生成两个随机大素数p,q,令$n=pq$,$\lambda=lcm(p-1,q-1)$,$\mu=\lambda^{-1}\bmod n$;(2)则公钥为$pk=n$,私钥为$sk=(\lambda,\mu)$;(3)加密消息m,其密文为$c=(n+1)^m r^n \bmod n^2$,其中随机数$r\in Z_n$;(4)解密密文c,利用私钥$sk=(\lambda,\mu)$计算消息$m=L(c^\lambda \bmod n^2)\mu \bmod n$,其中$L(x)=(x-1)/(pq)$。

3 本文方法

3.1 基本思想

PIR技术中存在着用户隐私安全需求、服务器计算开销、服务器和用户间的通信开销三方因素的均衡问题。传统的做法是保证数据库级别的用户隐私安全,但由此产生的通信开销与数据库规模成正比关系。特别地,全国注册企业规模约6 000万个,南方电网五省(广东、文本、云南、贵州、海南)将近1 000万个,每次查询都消耗与数据库体量相当的通信和计算,开销将是巨大的,进而影响服务质量。传输数据库规模相当体量的密文将是巨大的开销,也会影响检索时间和服务质量。

为解决此问题,本文的设计思想为在满足用户实际隐私需求前提下,通过在数据库的子集上执行隐私信息检索,并结合分桶方法降低计算成本和通信开销。满足用户隐私需要的可行性基于多数实际查询场景中,用户并不需要数据库级别的隐私安全性。例如,用户查询某制造企业a的用电信用时,其隐私需求为确保服务器不能从制造业中判断查询目标“企业a”,而无须确保不能从全国企业中明确查询目标“企业a”。

因此,基于用电信用增信查询中的实际需求,本文的基本思想是实现用户查询目标的k-匿名不可区分,在此基础上设计并实现计算开销与通信开销的最优均衡,而最优均衡的实现则通过调节对k个条目的分桶数量。

3.2 Effi-Retrieval算法描述

本节中参量k,μ分别表示用户隐私安全程度和用户的费用开销,其中k是匿名隐私保护中的匿名参量,μ是用户支付的查询费用,其具体确定方法见3.3节。对于大小为n的数据库,可实现的最大匿名程度为$k=n$。详细协议流程如下。

1)用户设置隐私参数k以及支付费用μ(以元为计量单位),并将参量上传给服务器。

2)服务器基于公式(1):

$$f(x)=\begin{cases}\sqrt{k}, & \mu\geqslant\alpha k+\sqrt{k},\\ \mu-\alpha k, & \alpha k+1\leqslant\mu\leqslant\alpha k+\sqrt{k}。\end{cases}\tag{1}$$

计算最优分桶数L以及用户需支付的实际

费用μ,并将参量μ发送给用户。

3)用户支付服务器确认后的费用μ,并将包含目标关键词w_*在内的k个关键词上传给服务器,记为$\{w_1,\cdots,w_*,\cdots,w_k\}$。

4)服务器从数据库$\mathcal{X}$中选择与关键词$\{w_1,\cdots,w_*,\cdots,w_k\}$对应的$k$条记录$\{(w_i,m_i)\}_{i=1}^{k}$,并使用公开的哈希函数$H$将$k$个关键词映射到$L$个桶中,即$(w_i,m_i)\rightarrow H(w_i)$。令$m$表示哈希映射之后每个桶中包含记录数量的上界。

5)服务器构造多项式,对第j个桶,$j\in\{1,2,\cdots,L\}$,定义多项式

$$\begin{cases}P_j=\prod_{s=0}^{d_j}(w-w_s),\\ Q_j=\sum_{s=1}^{d_j}(m_s|0^l)(w-w_s+1)\cdot\prod_{t=1,t\neq s}^{d_j}\dfrac{w-w_t}{w_s-w_t}。\end{cases}\tag{2}$$

使得被映射到第j个桶中的元素(w_i,m_i),满足$P_j(w_i)=0$且$Q_j(w_i)=(m_i|0^l)$,其中d_j是第j个桶中对应的记录数量,l是统计安全参数。然后,选择整数型随机数r_j并定义多项式:

$$Z_j(w)=r_j*P_j(w)+Q_j(w)。\tag{3}$$

将$Z_j(w)$展开成关于w的标准多项式形式$Z_j(w)=\sum_{i=0}^{d_j}a_iw^i$。

6)用户首先基于Paillier同态加密生成公钥和私钥,然后用公钥对目标关键词的1到m次幂$w_*,w_*^2,\cdots,w_*^m$进行加密,并将密文$Enc(w_*),\cdots,Enc(w_*^m)$上传给服务器。

7)服务器根据每个桶的大小将对应的密文分别代入多项式$Z_j(w)$得到密文结果$\prod_{i=0}^{d_j}Enc^{a_i}(w_*^i),j\in\{1,2,\cdots,L\}$,并将结果发送给用户。基于Paillier同态加密性质,得到

$$\prod_{i=0}^{d_j}Enc^{a_i}(w_*^i)=\prod_{i=0}^{d_j+1}Enc(a_iw_*^i)=Enc\left(\sum_{i=0}^{d_j+1}a_iw_*^i\right)=Enc(Z_j(w_*)),\tag{4}$$

所以,查询目标w_*对应的信息m_*在服务器发送给用户的L个密文中。

8)用户利用公开的哈希函数H得到$H(w_*)$,则查询内容m_*在第$H(w_*)$个桶内。用户利用Paillier同态加密的私钥解密第$H(w_*)$桶的密文$\mathrm{Enc}\left(Z_{H(w_*)}(w_*)\right)$,得到$Z_{H(w_*)}(w_*)=m_*|0^l$。验证最后$l$位是否为0,若是,则$m_*$是查询结果;否则,输出$\perp$表示数据库的关键词不包含查询目标$w_*$。

说明 构造和计算不经意多项式是一种成本比较高的密码学方法,可采用Window策略[23]进一步降低计算开销。Window技术针对设置的哈希桶的数量估计出桶中最大的多项式次数,然后发送方可以选择一组基,所有的多项式中需要的幂可以根据这组基计算出来,这样可以减少通信和计算成本。

为直观理解算法流程,图4展示了某次查询的具体计算流程。在图4中,用户欲查询w_*="企业1"的用电信用报告,令公式(1)中的计算成本权重$\alpha=1$。为防止服务器获得用户的查询隐私,用户首先设置隐私参量$k=14$和承担费用$\mu'=20$并上传至服务器(第1步);服务器基于公式(1)计算出最优分桶数量$L=4$和确认费用$\mu=18$并将$\mu=18$发送给用户(第2步);用户完成费用支付后将包含"企业1"在内的14个关键词上传给服务器(第3步);服务器从数据库$\mathcal{X}$中选取关键词对应的14条记录,并利用哈希函数将其映射到4个桶中,每个桶包含的记录数量分别为4,4,3,3并且"企业1"被映射到第2个桶(第4步);服务器基于公式(2)构造多项式$P_j(w),Q_j(w)$,基于公式(3)构造多项式$Z_j(w)$(第5步);用户基于最高次幂4计算w_*对应地从1次幂到4次幂对应的Paillier密文并将其发送给服务器(第6步,需要先将字符型转化为整数型再进行次幂加密);服务器基于密文计算多项式$Enc\left(Z_j(w_*)\right),j=1,2,3,4$并将结果发送给用户(第7步);用户由哈希映射$H$知关键词"企业1"被映射到第2个桶,并解密第2桶对应消息(将整数型转化为字符型),得到"企业1"对应的用电信用报告"当月电费=12万元,当月是否欠费=否,历史逾交金额=0万元,历史欠费金额=0万元"。

3.3 Effi-Retrieval算法参数设定

为简便起见,用户和服务器之间的通信成本和计算开销是根据数据项数量计算的,因为在不同的运算中可能会涉及不同的量纲常数,

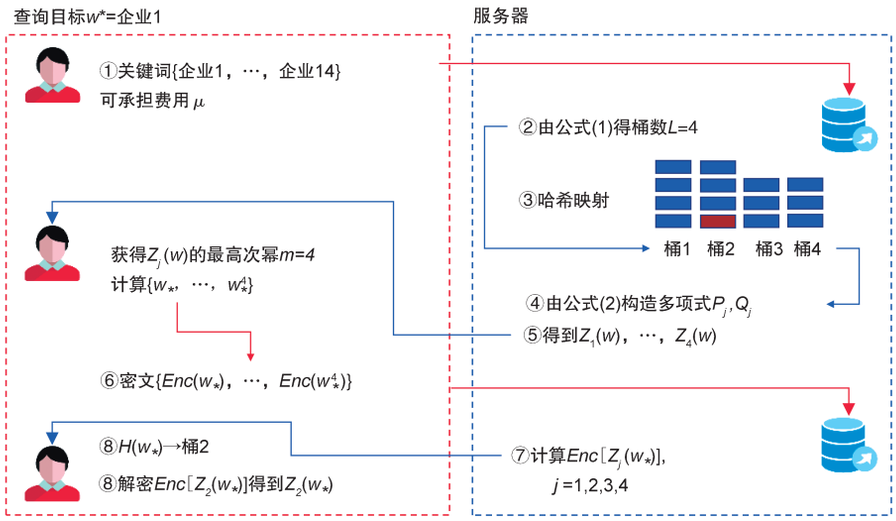

图4 查询关键词为“企业1”对应的计算流程示例

Fig. 4 Demonstration of inquiry with keyword "enterprise 1"

准确的通信成本应该考虑传输字节大小和密钥长度。

1)成本计算:根据前面系统模式和算法描述内部,假设数据库的大小为n,分桶数量为L。根据算法描述,通信开销包括三部分:(1)用户将k个关键词上传至服务器;(2)用户将w_*的1到m次密文发送给服务器;(3)服务器将L个桶上的多项式密文结果发送给用户。因此,通信开销可粗略地估计为$C_{\text{comm}}=k+m+L$。根据算法描述,计算成本包括用户的Paillier加解密计算和服务的多项式计算,考虑到承担费用为用户向服务器支付并且用户的计算成本远小于服务器的计算成本,所以这里仅考虑服务器的计算成本。由于服务器需要计算L个桶上的多项式密文结果$Z_j(w_*)$,而每个$Z_j(w_*)$的最高次幂为d_j,所以服务器的计算成本为$C_{\text{comp}}=\sum_{j=1}^{L}d_j=k$。

2)参数决策:参数决策的目的是最小化通信开销$C_{\text{comm}}=n+\frac{n}{L}+L$和计算成本$C_{\text{comp}}=k$,同时满足约束$k\leqslant n$和$\mu\geqslant L+\alpha k$,其中$\alpha$表示计算成本转化为通信开销的权重,其值由用户和服务器协商。由于用户的查询目标w_*必须满足k-匿名约束,所以计算开销为固定的k。因此,该问题可表示为

$$\min C_{\text{comm}}=k+\frac{k}{L}+L,$$
$$s.t.\ k\leqslant n,$$
$$\mu\geqslant L+\alpha k。\tag{5}$$

显然,当$L=\sqrt{k}$时,通信成本最小。但是由于用户承担的费用μ需要满足约束$\mu\geqslant L+\alpha k$,所以$\min C_{\text{comm}}$的取值需要考虑理论最优分桶数量$L=\sqrt{k}$是否满足约束$\mu-\alpha k\geqslant\sqrt{k}$。如果有$\mu-\alpha k\geqslant\sqrt{k}$,则$C_{\text{comm}}$的最小值为$C_{\text{comm}}=k+2\sqrt{k}$;如果$\mu-\alpha k<\sqrt{k}$,则$C_{\text{comm}}$的最小值为$C_{\text{comm}}=k+\frac{k}{\mu-\alpha k}+\mu-\alpha k$,此时$L=\mu-\alpha k$。因为分桶数量$L$满足$L\geqslant 1$,所以有$\mu\geqslant\alpha k+1$。这说对于给定的$k$-匿名隐私需求时,用户需要承担的费用存在下界$\alpha k+1$。综上分析,当用户给定隐私参量$k$之后,最优的分桶数量$L$满足:

$$L=\begin{cases}\sqrt{k}, & \mu\geqslant\alpha k+\sqrt{k},\\ \mu-\alpha k, & \alpha k+1\leqslant\mu<\alpha k+\sqrt{k}。\end{cases}\tag{6}$$

4 安全性分析和实验评估

本节首先分析Effi-Retrieval协议的安全性,然后给出隐私参量k和μ对通信开销影响的数值分析结果。

4.1 Effi-Retrieval算法参数设定

从协议的正确性、服务器的安全性、用户的隐私性以及面向恶意用户的安全性四个方面对Effi-Retrieval协议的安全性进行分析。

1)**协议的正确性**:正确性指协议是否能够

正确执行,能否达到协议的目的,即用户和服务器按照协议的要求进行交互能否获得正确的查询结果。

在协议完整执行的情况下,协议的正确性是可以保证的。首先,用户和服务器商定协议参数即费用。服务器根据参数建立哈希映射并且构造多项式,用户发送不同次数的密文,服务器计算多项式的结果后以密文形式返回。由于用于映射的哈希函数是公开的,因此只要用户提供了正确的关键字的密文,就可以在相应的哈希槽中得到关键字对应的信息,还原出自己需要的结果。

2)**服务器的安全性:**服务器的安全性指在协议正常执行的情况下,用户只能得到其查询的关键词对应的信息,无法得到数据库中的任何其他信息。

在用户和服务器正常进行交互的情况下,用户得到的内容为L个Paillier密文,L为分桶的数量。这些密文是由用户发送的加密关键字代入L个多项式计算得到。由协议的描述可知,其中只有一个密文在解密之后能够得到对应的消息,其他密文得到的都是无意义的信息,由于多项式中添加了随机数,因此用户无法根据结果进行推测。

3)**用户的隐私性:**在PIR协议中,用户的隐私性指的是用户发送查询之后,服务器只知道用户进行了查询,不知道用户的查询内容。

本文所提协议使用了k匿名的思想,用户发送给服务器的信息为k个关键词以及其中的一个查询目标关键词w_*的1到m次Paillier密文$Enc(w_*),Enc(w_*^2),\cdots,Enc(w_*^m)$。此时服务器只知道用户的查询范围在$k$个关键词当中。服务器能否在多项式时间内得到用户查询内容取决于Paillier同态加密的安全性。

4)**针对恶意用户的安全性:**恶意用户指不按照协议的正常步骤进行交互的用户,这类用户会通过发送错误的信息来获取额外的服务器信息,协议需要保证这类用户无法获取任何相关数据。

由于用户接收到的服务器的信息为$Z_j(w_*)=r_jP_j(w_*)+Q_j(w_*)$,其中$r_j$为随机数,所以对于用户的每次查询,服务器均以随机数r_j生成$Z_j(w)$。因此,用户不能通过发起重复查询并根据反馈结果反向求解多项式的系数。此外,若用户恶意篡改密文$Enc(w_*^i),i=1,\cdots,m$,则用户非但不能反解出服务器的多项式构造,且在支付一定费用后不能获得正确的查询结果。

4.2 实验评估

根据公式(4),隐私参量k和费用参量μ会影响协议的通信开销,本节展示三者之间的关系以及与不使用k-匿名隐私检索的多项式生成和查询时间。采用的数据集是FATE开源联邦平台提供的Credit Card数据集,其行数为30 000,针对不同的数据库设置规模从中随机抽取特定子集。

首先,用户可指定其隐私需求k和承担费用μ。根据公式(1)的第一个式子,对于给定的参量k,即使用户愿意承担更多的查询费用,但当费用$\mu\geqslant\alpha k+\sqrt{k}$此时的最优分桶数量$L=\sqrt{k}$为固定值。即协议的通信开销存在下限,而不会随着用户愿意承担的费用而无限降低,这与实际情况相一致。根据公式(1)的第二个式子,当费用$\alpha k+1<\mu\leqslant\alpha k+\sqrt{k}$时,随着用户愿意承担的费用的增加,协议的通信开销会随之降低,即用户可得到更快速的查询响应。

图5展示了协议通信成本和用户愿意承担费用间的关系。对于给定的隐私参量k,随着用户愿意支付的费用μ的增大,协议的通信开销会逐渐降低。也就是说,用户可以通过支付更多的费用换取更快的服务响应。根据公式(1)的第2个式子,此时的分桶数量L随着μ的增加而增大,所以每个桶对应的记录数量将逐渐减少,所以用户需要上传的Paillier加密信息也越少,虽然此时服务器需要向用户发送更多的信息(与桶的数量相同),但通信开销的总合逐渐降低。同时,图5展示了对于给定的隐私参量k,协议通信开销存在下界。除此之外,图5展示了对于给定的用户支付费用μ,随着隐私参量k的减小,协议的通信开销也随之减小,即用户可以通过降低隐私需求换取更快的查询响应。

图6展示了用户上传的通信开销与用户愿意承担费用间的关系。对于给定的隐私参量k,随着用户愿意支付的费用μ的增大,其需要

上传服务器的Paillier加密的密文信息越少。原因在于随着分桶数量的增大,每个桶上多项式的最高次幂(桶中对应的记录数量)减小。但同时,由于桶的数量存在上界,所以对于给定的k,用户需要上传的密文数量存在一个下界。除此之外,图6展示了对于给定的支付费用μ,随着隐私参量k的减小,用户需要上传的密文数量也随之减小,即用户可以通过降低隐私需求换取上传通信开销。

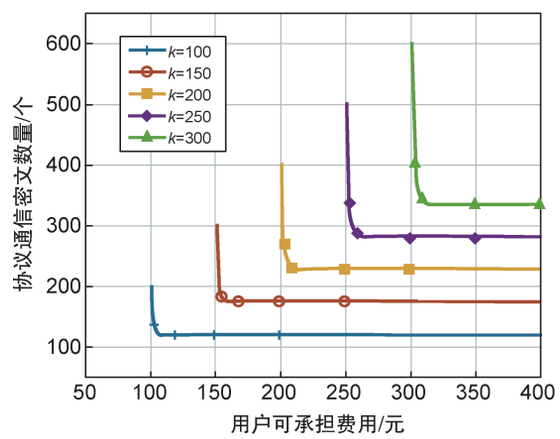

图5 用户可承担费用μ对协议通信开销的影响

Fig. 5 Impacts of μ on protocol communication

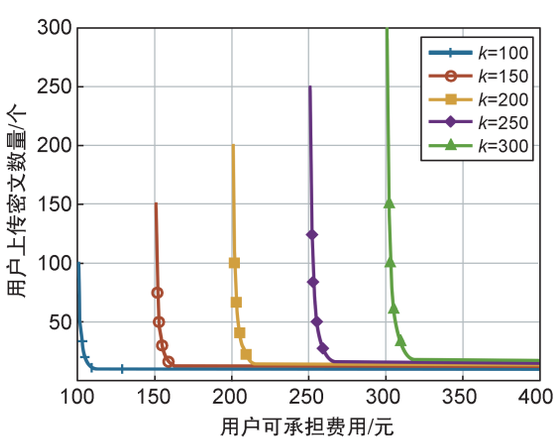

图6 用户可承担费用μ对上传通信开销的影响

Fig. 6 Impacts of μ on upload communication

表1展示了当密钥位数为8 192时,Effi-Retrieval与传统PIR方法(即不使用k-匿名)在不同规模数据集上的计算效率。由表1可知,当匿名参数$k=200$时,Effi-Retrieval的多项式生成时间和查询时间与数据库大小无关,而传统PIR方法的多项式生成和查询时间则与数据库大小呈指数增长,并且当数据规模为300及以上时,与不使用分桶策略的传统方法相比,Effi-Retrieval生成密态多项式的时间可降低50%以上,与不使用k-匿名方法的查询算法相比,Effi-Retrieval的检索时间可降低30%以上。当密钥位数设置为4 096时,两种对比算法的性能有类似的结论。

表1 密钥位数为8 192时Effi-Retrieval和传统数据库隐私信息检索计算效率对比

Table 1 Comparison of computation efficiency between effi-retrieval and traditional PIR (key length = 8 192)

数据库大小	生成多项式时间/s		查询时间/s	
	Effi-Retrieval (k=200)	传统检索方法	Effi-Retrieval (k=200)	传统检索方法
200	7.27	7.27	6.19	6.19
300	7.27	15.18	6.23	10.83
500	7.27	61.02	6.19	22.43
1 000	7.27	234.76	6.01	58.73
2 000	7.28	846.65	6.14	135.79
5 000	7.27	>10 min	6.13	>10 min

5 结语

本文针对现有用电增信报告业务中未考虑查询信息安全的问题,基于不经意多项式计算和Paillier同态加密技术提出一种隐私信息检索方法。该方法结合实际场景中用户的隐私需求,通过将用户查询时的数据库隐私安全替换为k-匿名隐私安全,以实现将隐私信息检索方法的$O(n^{\alpha})$,$\alpha>0$级的通信开销降低为$O\left(k+\sqrt{k}\right)$,其中n为数据库规模而k为隐私安全需求。本文提出的隐私安全检索方法可实现基于关键词的信息检索,并证明了无论服务器或用户是否诚信,该方法均可保证参与方的信息安全。最后的数值分析验证了隐私参量k和用户支付费用μ与协议通信开销的分析结论。

本文提出的隐私信息检索方法中使用了Paillier同态加密,该方法需要用户计算关键词的若干次幂的密文,从而增大的用户的计算和上传开销。下一步工作中,将使用层次同态加密技术,该技术支持服务器对密文进行有限次的加法和乘法运算。因而,用户只需要将关键词的密文,从而降低用户的计算和通信开销。

参考文献:

[1] ALMUTAIRI M M, ABI SEN A A, YAMIN M. Survey of PIR Approach and Its Techniques for Preserving Pri-

vacy in Iot[C]//Proceedings of the 2021 8th International Conference on Computing for Sustainable Global Development. New Delhi: India. 2021: 417−421. DOI: 10.1109/INDIACom51348.2021.00074.

[2] YU M K, YANG K C, WEI L B, *et al*. Practical Private Information Retrieval Supporting Keyword Search in the Cloud[C]//2014 Sixth International Conference on Wireless Communications and Signal Processing (WCSP). Hefei: IEEE. 2014: 1−6. DOI: 10.1109/WCSP.2014.6992210.

[3] SWEENEY L. *k*-anonymity: A Model for Protecting Privacy [J]. *Int J Uncertain Fuzziness Knowl Based Syst*, 2002, **10** (5): 557−570. DOI: 10.1142/S0218488502001648.

[4] CHOR B, GOLDREICH O, KUSHILEVITZ E, *et al*. Private Information Retrieval[C]//Proceedings of IEEE 36th Annual Foundations of Computer Science. Milwaukee: IEEE, 1995: 41−50. DOI: 10.1109/SFCS.1995.492461.

[5] CHOR B, GILBOA N. Computationally Private Information Retrieval (Extended Abstract) [C]//Proceedings of the Twenty−ninth annual ACM symposium on Theory of computing. New York: ACM, 1997: 304−313. DOI: 10.1145/258533.258609.

[6] LIPMAA H. An Oblivious Transfer Protocol with Log-squared Communication[C]//Proceedings of the 8th International Conference on Information Security. New York: ACM, 2005: 314−328. DOI: 10.1007/11556992_23.

[7] KIAYIAS A, LEONARDOS N, LIPMAA H, *et al*. Optimal Rate Private Information Retrieval from Homomorphic Encryption[J]. *Proc Priv Enhancing Technol*, 2015, **2015**(2): 222−243. DOI: 10.1515/popets-2015-0016.

[8] LI Y T, CHANG Y X, CHENG M Q, *et al*. Multi-value Private Information Retrieval with Colluding Databases *via* Trace Functions[J]. *Inf Sci*, 2021, **543**: 426−436. DOI: 10.1016/j.ins.2020.07.006.

[9] LI J, KARPUK D, HOLLANTI C. Towards Practical Private Information Retrieval from MDS Array Codes[J]. *IEEE Trans Commun*, 2020, **68**(6): 3415−3425. DOI: 10.1109/TCOMM.2020.2980833.

[10] BANAWAN K, ULUKUS S. Multi-message Private Information Retrieval: Capacity Results and Near-optimal Schemes[J]. *IEEE Trans Inf Theory*, 2018, **64**(10): 6842−6862. DOI: 10.1109/TIT.2018.2828310.

[11] SUN H, JAFAR S. Multi-round Private Information Retrieval: Capacity and Storage Overhead[J]. *IEEE Trans Inf Theory*, 2018, **64**(8): 5743−5754. DOI: 10.1109/TIT.2018.2789426.

[12] JIA Z Q, SUN H, ALI JAFAR S. Cross Subspace Alignment and the Asymptotic Capacity of X-secure T-private Information Retrieval[J]. *IEEE Trans Inf Theory*, 2019, **65** (9): 5783−5798. DOI: 10.1109/TIT.2019.2916079.

[13] CHEN Z, WANG Z Y, ALI JAFAR S. The Capacity of T-private Information Retrieval with Private Side Information[J]. *IEEE Trans Inf Theory*, 2020, **66**(8): 4761−4773. DOI: 10.1109/TIT.2020.2977919.

[14] WANG C, KON W Y, NG H J, *et al*. Experimental Symmetric Private Information Retrieval with Measurement-device-independent Quantum Network[J]. *Light Sci Appl*, 2022, **11**: 268. DOI: 10. 1038/s41377-022-00959-6.

[15] WANG Z S, ULUKUS S. Communication Cost of Two-database Symmetric Private Information Retrieval: a Conditional Disclosure of Multiple Secrets Perspective[C]//2022 IEEE International Symposium on Information Theory (ISIT). Espoo: IEEE. 2022: 402−407. DOI: 10. 1109/ISIT50566. 2022. 9834819.

[16] 陈磊磊, 陈兰香, 穆怡. 医疗辅助诊断系统中新型的双向隐私保护方法[J]. 密码学报, 2021, **8**(1): 167−182. DOI: 10.13868/j.cnki.jcr.000429.
CHEN L L, CHEN L X, MU Y. Novel Two-way Privacy Protection Method in Medically Assisted Diagnosis System[J]. *J Cryptologic Reseatch*, 2021, **8**(1): 167−182. DOI: 10.13868/j.cnki.jcr.000429.

[17] 俞志斌. 隐私保护的私有信息检索在云环境中的应用[D]. 重庆: 西南大学, 2015.
YU Z B. Application of Private Information Retrieval with Privacy Protection in Cloud Environment[D]. Chongqing: Southwest University, 2015.

[18] BENNY B. A Local and Intelligent Web Information Retrieval System[D]. University of Alberta, 2021.

[19] NAKAMURA T, INENAGA S, IKEDA D, *et al*. Anonymous Authentication Systems Based on Private Information Retrieval[C]//2009 First International Conference on Networked Digital Technologies. Ostrava: IEEE. 2009: 53−58. DOI: 10.1109/NDT.2009.5272083.

[20] BADR M M, AL AMIRI W, FOUDA M M, *et al*. Smart Parking System with Privacy Preservation and Reputation Management Using Blockchain[J]. *IEEE Access*, 2020, **8**: 150823−150843. DOI: 10.1109/ACCESS.2020.3016945.

[21] 丁佳晨, 俞能海, 林宪正, 等. 基于比特币的私有信息检索支付协议[J]. 信息安全学报, 2019, 4(6)1−9. DOI: 10.19363/J.cnki.cn10-1380/tn.2019.11.01.
DING J C, YU N H, LIN X Z, *et al*. Bitcoin-based Payment Protocol for Private Information Retrieval[J]. *J Cyber Secur*, 2019, 4(6)1−9. DOI: 10.19363/J. cnki. cn10-1380/tn.2019.11.01.

[22] ATTIA M A, KUMAR D, TANDON R. The Capacity of Private Information Retrieval from Uncoded Storage Constrained Databases[J]. *IEEE Trans Inf Theory*, 2020, **66**(11): 6617−6634. DOI: 10.1109/TIT.2020.3023016.

[23] CHEN H, HUANG Z C, LAINE K, *et al*. Labeled PSI from Fully Homomorphic Encryption with Malicious Security[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York: ACM, 2018: 1223−1237. DOI: 10.1145/3243734.3243836.