软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(3):669–688 [doi: 10.13328/j.cnki.jos.006174]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

面向企业数据孤岛的联邦排序学习*

史鼎元[1,2,3], 王晏晟[1,2,3], 郑鹏飞[1,2,3], 童咏昕[1,2,3]

[1](软件开发环境国家重点实验室(北京航空航天大学),北京 100191)

[2](大数据科学与脑机智能高精尖创新中心(北京航空航天大学),北京 100191)

[3](北京航空航天大学 计算机学院,北京 100191)

通讯作者: 童咏昕, E-mail: yxtong@buaa.edu.cn

摘 要: 排序学习(learning-to-rank,简称 LTR)模型在信息检索领域取得了显著成果,而该模型的传统训练方法需要收集大规模文本数据.然而,随着数据隐私保护日渐受到人们重视,从多个数据拥有者(如企业)手中收集数据训练排序学习模型的方式变得不可行.各企业之间数据被迫独立存储,形成了数据孤岛.由于排序模型训练需要使用查询记录、文档等诸多隐私信息,数据孤岛难以融合打通,这制约了排序学习模型的训练.联邦学习能够让多数据拥有方在隐私保护的前提下联合训练模型,是一种打通数据孤岛的新方法.在其启发下,提出了一种新的框架,即面向企业数据孤岛的联邦排序学习,它同时解决了联邦学习场景下排序学习所面临的两大挑战,即交叉特征生成与缺失标签处理.为了应对多方交叉特征的生成问题,使用了一种基于略图(sketch)数据结构与差分隐私的方法,其相比于传统加密方法具有更高的效率,同时还具有隐私性与结果精度的理论保证.为了应对缺失标签问题,提出了一种新的联邦半监督学习方法.最终,通过在公开数据集上的大量实验,验证了所提方法的有效性.

关键词: 排序学习;企业数据孤岛;联邦学习;略图;差分隐私

中图法分类号: TP181

中文引用格式: 史鼎元,王晏晟,郑鹏飞,童咏昕.面向企业数据孤岛的联邦排序学习.软件学报,2021,32(3):669–688. http://www.jos.org.cn/1000-9825/6174.htm

英文引用格式: Shi DY, Wang YS, Zheng PF, Tong YX. Cross-silo federated learning-to-rank. Ruan Jian Xue Bao/Journal of Software, 2021,32(3):669–688 (in Chinese). http://www.jos.org.cn/1000-9825/6174.htm

Cross-Silo Federated Learning-to-Rank

SHI Ding-Yuan[1,2,3], WANG Yan-Sheng[1,2,3], ZHENG Peng-Fei[1,2,3], TONG Yong-Xin[1,2,3]

[1](State Key Laboratory of Software Development Enviroment (Beihang University), Beijing 100191, China)

[2](Beijing Advanced Innovation Center for Big Data and Brain Computing (Beihang University), Beijing 100191, China)

[3](School of Computer Science and Engineering, Beihang University, Beijing 100191, China)

Abstract: Learning-to-rank (LTR) model has made a remarkable achievement. However, traditional training scheme for LTR model requires large amount of text data. Considering the increasing concerns about privacy protection, it is becoming infeasible to collect text data from multiple data owners as before, and thus data is forced to save separately. The separation turns data owners into data silos, among which the data can hardly exchange, causing LTR training severely compromised. Inspired by the recent progress in federated

* 基金项目: 国家重点研发计划(2018AAA0101100); 国家自然科学基金(61822201, U1811463); 软件开发环境国家重点实验室(北京航空航天大学)开放课题(SKLSDE-2020ZX-15)

Foundation item: National Key Research and Development Program of China (2018AAA0101100); National Natural Science Foundation of China (61822201, U1811463); State Key Laboratory of Software Development Environment (Beihang University) Open Program (SKLSDE-2020ZX-15)

本文由“支撑人工智能的数据管理与分析技术”专刊特约编辑陈雷教授、王宏志教授、童咏昕教授、高宏教授推荐.

收稿时间: 2020-07-19; 修改时间: 2020-09-03; 采用时间: 2020-11-06; jos 在线出版时间: 2021-01-21

learning, a novel framework is proposed named cross-silo federated learning-to-rank (CS-F-LTR), which addresses two unique challenges faced by LTR when applied it to federated scenario. In order to deal with the cross-party feature generation problem, CS-F-LTR utilizes a sketch and differential privacy based method, which is much more efficient than encryption-based protocols meanwhile the accuracy loss is still guaranteed. To tackle with the missing label problem, CS-F-LTR relies on a semi-supervised learning mechanism that facilitates fast labeling with mutual labelers. Extensive experiments conducted on public datasets verify the effectiveness of the proposed framework.

Key words: learning-to-rank; data silo; federated learning; sketch; differential privacy

近年来,排序学习(learning-to-rank,简称 LTR)在信息检索领域取得了巨大的成功[1−3],这种成功在诸如谷歌(Google)与必应(Bing)等具备大量用户数据资源的商业搜索引擎上尤为显著.排序学习算法的有效性离不开大量训练数据的支撑,搜索引擎巨头公司的数据往往由搜索引擎与其上千万甚至上亿用户在每日的频繁交互中产生.然而,众多普通规模的中小型企业也有搭建搜索引擎或搜索系统的需求,如企业搜索(enterprise search)或邮件搜索,但其并不具备单独收集大规模用户数据的实力.中小型企业与企业之间,乃至企业的部门之间存在的数据孤岛(data silo)问题导致了大量数据分散存储,出于企业机密保护以及行政手续复杂等原因难以被打通.与此同时,随着越来越多的数据隐私保护法规如欧盟的“通用数据保护条例(general data protection regulation,简称 GDPR)”等的出台,即使企业在主观上愿意,其也无法再像以往一样自由地交换或分享含有用户隐私的数据.种种不利因素,使得中小型企业很难利用强大的排序学习算法与大数据的支撑搭建其自己的有效搜索系统[4].因此,如何为各企业克服数据孤岛障碍,在满足隐私保护的约束下训练有效的排序学习,是一个亟待解决的问题.

为了解决上述问题,在本文中,我们首次提出了面向企业数据孤岛的联邦排序学习(cross-silo federated LTR,简称 CS-F-LTR),其致力于协调多方企业或组织在不交换各自原始数据的前提下,联合训练有效的排序学习模型.一方面,与分布式排序学习不同的是:在本文场景下,训练排序学习模型所需的排序对象(如文档、电子邮件等)与查询语句均被分散地存储在各方,原始数据被动地被分隔且无法自由交互,只能由各方借助隐私保护的手段来生成排序学习所需的训练样本;而在分布式场景下,训练样本是提前由原始数据计算所得再进行数据分割的.同时,无论是训练样本的生成或是后续的训练过程中,各分布式节点的数据交互都不受隐私约束[5].另一方面,本文的联邦学习场景与现有的其他联邦学习也有着很大的不同.现有的联邦学习中,数据一般是被横向分割(horizontally-partitioned)或纵向分割(vertically-partitioned)[6].在横向分割场景中,联邦中每一方都拥有若干互不重叠的训练样本,每个样本包含了完整的训练特征(features)与标签(labels);在纵向分割场景中,联邦每一方都拥有完整的全样本,但是各个特征以及标签为不同方所拥有.然而,本文的数据分割场景不同于上述任意一个,我们将其命名为交叉分割(cross-partitioned).在交叉分割场景下,每一条训练样本的特征是由联邦任意两方的排序对象与查询语句交叉计算所得.举例而言,甲方拥有一条样本,其中查询关键词为“电影”,排序对象为一篇影评,标签为“相关”;乙方拥有一条样本,其查询关键词为“美食”,排序对象为一篇食谱,标签为“相关”.此时,如果遵从现有的横向联邦学习场景,联邦中共有两条标签均为“相关”的样本.而本文场景下,则将新增两个特征对应了“电影-食谱”与“美食-影评”同时标签未知的交叉样本.上述 3 种分割场景如图 1 所示.其中,特殊的交叉分割场景给本文的面向企业数据孤岛联邦学习框架带来了如下挑战.

(1) 交叉特征生成:如何在隐私保护约束下安全地连接各方的查询语句与排序对象,从而生成有效的排序学习样本特征.在现有的排序学习研究工作中[7],许多具有表现力的特征都是通过查询语句与排序对象联合计算所得(如查询语句中单词在排序文档中出现的频率等).因此,需要一种各方的查询语句与其他方排序对象之间安全交互、从而计算特征的方法;

(2) 缺失标签处理:如何应对交叉样本中的标签缺失问题.通过上文例子可以发现,交叉样本往往只具有训练特征而无法直接获取其准确的标签.对某一方来说,要想获知其文档在其他方查询语句下的相关度而不泄露各自的文档与查询语句信息是一件极其困难的事,这也将我们面临的问题转变为一个联邦场景下的半监督学习问题.

本文致力于解决上述两大挑战,并提出高效的联邦排序学习算法.

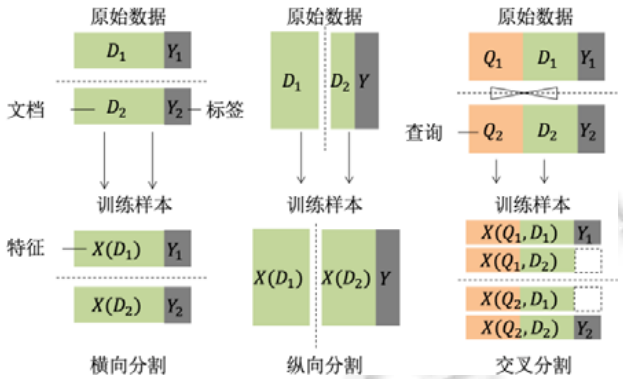


Fig.1 Three types of data partitions in federated learning

图 1 联邦学习中 3 种类型的数据分割

本文的主要贡献如下:

(1) 首次定义了企业数据孤岛场景下的联邦排序学习问题,并提出相应的解决框架,进一步指明了其两大研究挑战,即交叉特征生成与缺失标签处理.目前,尚未有任何面向企业数据孤岛的联邦排序学习相应研究成果;

(2) 为了应对交叉特征生成的挑战,本文提出了基于略图(sketch)数据结构的交叉特征生成算法,并在理论上证明了算法具有的隐私性与结果精度的保证;

(3) 为了应对缺失标签处理的挑战,本文提出了一种半监督联合训练算法,通过交互的标签生成器,来高效与准确地推断交叉样本对应的缺失标签;

(4) 本文通过公开数据集的大量实验,验证了所提算法的有效性.

本文第 1 节将介绍相关工作.第 2 节将给出问题定义.第 3 节与第 4 节将分别讨论交叉特征生成与缺失标签处理相对应的解决方案.第 5 节将展示实验结果.最后,在第 6 节进行文章总结.

1 相关工作

1.1 隐私保护的信息检索

信息检索系统通常包含同时来自于服务器端与客户端的大量数据,在当下流行的基于排序学习的信息检索框架下,这些大数据经常被用来训练排序学习模型.随着排序模型的使用日渐广泛,用于训练排序模型的数据种类也逐渐增多,如客户端的浏览器的查询历史信息、移动端的时空轨迹信息[8]、甚至是医学数据[9]等.这些数据能够反映用户的网络浏览习惯、出行规律和健康状况,属于隐私敏感的信息.这些信息难以通过简单的脱敏手段来消除隐私泄露隐患,历史上就曾发生过著名的 AOL 数据隐私泄露事件[10].现如今,大多数基于隐私保护的信息检索算法都致力于保护客户端数据隐私免受来自恶意或半诚实的服务器的攻击.其中,一项早期研究成果提出了名为隐私信息检索(private information retrieval,简称 PIR)的模型[11].然而,该模型过于理想化,需要通过数据库的复制来实现隐私保护,因此缺乏实际应用价值.还有一些研究通过评估查询语句所访问数据的敏感程度,实现隐私泄漏风险控制[12],这种方法的查询结果严重受限于数据的私密程度.一些更加实用的方法通过混淆化的查询语句来保护用户的隐私,其混淆化的手段也是多样的,如重构查询语句[13,14]、插入掩蔽语句[15]、或是满足差分隐私(differential privacy,简称 DP)的噪声[16]等.另一大类模型被称为对称隐私信息检索(symmetrically-private information retrieval,简称 SPIR),其在防御恶意服务器攻击的同时,也考虑了恶意用户的攻击,同时保护了双方的数据隐私[17].为了实现对称化的隐私保护[18,19],许多基于安全关键词查询的方法被提出,其中较为著名的有可搜索对称加密(searchable symmetric encryption,简称 SSE)[20]以及保序加密(order preserving encryption,简称 OPE)[21]等.同时,也有研究工作考虑到了与本文相似的具有多数据拥有方的数据共享场景,并提出了基于保

序加密的解决框架[22].然而,这些工作都只是针对简单的关键词搜索问题,而排序学习问题与之相比则复杂得多.目前,有少数研究工作考虑到了排序学习中的隐私保护,但往往局限于特定的排序学习算法,如基于树集成模型的算法[23].不同于以往的基于隐私保护的信息检索相关工作,本文所提的联邦排序学习考虑了在联合学习过程中同时保护各个联邦参与方的隐私.由于排序对象与查询语句的隐私同样都属于隐私保护范畴,因此本文期望实现联邦中任意两两参与方的对称隐私要求,这是以往工作所不曾涉及的.同时,简单地套用加密算法来实现任意两方的对称隐私,也将在计算与通信效率上变得不可行.

1.2 联邦学习

联邦学习(federated learning,简称 FL)最早由谷歌提出[24],它致力于实现多方在数据隐私保护前提下共同完成机器学习任务.个人数据隐私是当今互联网最值得关注的问题之一,随着欧盟“通用数据保护条例(general data protection regulation,简称 GDPR)”的颁布,大型互联网公司无法再自由地收集用户数据用于机器学习.因此,谷歌于 2016 年提出了针对安卓手机用户的联邦学习方法,它可在保证用户数据不出本地设备的前提下训练一个联合模型[24].文献[4]对谷歌的联邦学习概念进行了扩展,给出了更加通用的定义,并针对数据的不同切分方法,把联邦学习分为横向联邦学习、纵向联邦学习与联邦迁移学习.无论在哪种场景中,都具有几大共同的挑战.首先是数据隐私保护方法的设计,现有的大部分工作都基于多方安全计算(secure multi-party computation,简称 SMC)的相关方法,设计基于同态加密、秘密共享等加密技术的联合机器学习算法[25],也有一部分工作基于随机扰动算法,如差分隐私(differential privacy,简称 DP),设计基于噪声扰动的安全机器学习框架与算法[26−28].另一大挑战是传输成本问题,由于受到原始数据不能出本地的限制,需要在多方之间交互计算机器学习的中间参数,如梯度,该交互过程轮数较多,因此,如何节约传输成本以及减少交互轮数成为了关键.文献[29]对此提出了提高通信效率的优化方法.同时,如何有效地激励用户参与到联邦学习的过程中来,也是目前研究的重点之一.文献[30]提出了一种基于夏普利值的联邦学习多方贡献评估机制,可以高效地计算每个参与方对联合模型的贡献,以此公平地分配利益.除此之外,一些现有工作还研究了联邦学习的一些其他问题,如多任务学习问题[31]、数据非独立同分布问题[32]、模型可解释性问题[33]等.文献[34]总结了大量现有工作并给出了更加详细的联邦学习综述,根据其不同应用场景,将之分为面向企业数据孤岛(cross-silo)的联邦学习与面向终端设备数据孤岛(cross-device)的联邦学习.在后者中,往往涉及成百上千甚至千万级别的用户,因此,如何缩减通信开销成了其面临的最大挑战.而前者,即本文对应的面向企业数据孤岛场景,涉及数据拥有方一般少于 100,因此通信成本变得不那么重要.然而,由于每一方拥有的数据量一般要大于个人用户的数据,因此,如何缩减计算成本成为了更为重要的挑战.本文研究与现有面向企业数据孤岛联邦学习的不同之处在于考虑了一种以往工作不曾涉及的数据分割场景,即交叉分割,使得研究问题更具挑战性.值得一提的是,另一项最新工作[35]也考虑到了联邦排序学习,然而其与本文的本质区别在于:其场景为面向终端设备的联邦学习,数据也是普通的横向分割,这使得现有联邦学习方法能够很容易地适用于其问题.相比之下,本文侧重于企业搜索等面向企业数据孤岛的应用场景,问题也更加具有挑战性.

1.3 半监督排序学习

在排序学习中,获取足够且高质量的带标签数据需要花费大量人力成本.因此,半监督学习(semi-supervised learning,简称 SSL)经常被用于只有部分数据带标签的排序学习中.文献[36]提出了一种归纳排序算法,其中,无标签的文档将根据与有标签文档的相似度而自动生成对应的标签.文献[37]则采用了一种转导推理方法,利用半监督学习来寻找更优的特征.文献[38]为直推式半监督学习引入了聚类的方法,降低了迭代中产生的噪声,提高了排序学习模型性能.事实上,如今处于研究前沿的几种半监督学习算法,如 Temporal Ensembling[39],Mean Teacher[40]等都可以被简单地应用于排序学习上,因为排序学习也可以被看作是一般的分类问题.然而,在本文所涉及的联邦学习场景下,无标签的数据被分散存储在了各方,如何有效地在开销有限的前提下为之打上标签成为挑战.

1.4 Sketch数据结构

为了在保护隐私的同时提升计算效率,本文用到了 Sketch 数据结构,它是一种用于近似推断大规模数据流上诸如元素出现频率等统计量的方法[41],其中比较常用的两种频率点估计方法分别为 Count Sketch[42]与 Count-Min (CM) Sketch[43].除了大规模数据流的压缩以外,它在许多领域都有广泛的应用,例如大规模机器学习中的梯度压缩[44]与重要项查询[45]等.在某些场景下,它也可以与差分隐私(differential privacy,简称 DP)相结合,同时兼顾数据隐私与计算开销[46,47].有文献证明:在特定假设条件下,Count Sketch 可以在不加入任何噪声的前提下直接满足差分隐私所需的条件[48],因此也被称为"免费的隐私(privacy for free)".在本文中,我们将证明这种免费的隐私无法适用于我们的问题场景.因此,我们也将提出一种新的基于 Sketch 的差分隐私定义.

2 问题描述

本节给出面向企业数据孤岛联邦场景下的排序学习问题的正式定义.

2.1 面向企业数据孤岛的联邦场景

设联邦 F 中有 n 个企业,$F=\{P_1,P_2,\ldots,P_n\}$.每个企业 P_i 持有排序对象(文档)数据集合 $D_i=\{d_{i,1},d_{i,2},...,d_{i,|D_i|}\}$ 和查询数据集合 $Q_i=\{q_{i,1},q_{i,2},...,q_{i,|Q_i|}\}$.其中,文档和查询都可以看作是由全体单词集合 Γ 中的某些单词组成的多重集.全体数据可以表示为 $D=\bigcup_i D_i$ 和 $Q=\bigcup_i Q_i$.每篇文档和每个查询之间存在一个相关度分数(即标签),表示为 $rel_i(d,q)$,其中,$d\in D_i,q\in Q_i$.由于每个企业只能够访问它自身所持有的文档数据和查询数据,所以它只能计算自己内部文档与查询的相关度.然而,它的查询可能也会和其他企业中的文档相关,或反之,其他企业中的查询和它的文档相关,但是这些相关度分数都因为数据不能相互访问而无法直接计算.

与其他机器学习问题类似,开展排序学习训练任务需要构造训练数据集(X,y),其中,X 表示从文档数据 D 和查询数据 Q 中提取的特征,标签 y 则是相关性分数.我们假设特征提取函数 $\Phi:D\times Q\to R^m$ 是已知的.该函数依据某篇文档和某个查询,生成交叉特征,如词频(term frequency,简称 TF)、BM25[49]和 LMIR[50]等.

在没有联邦学习的情况下,第 i 个企业依据自身所持有的数据 D_i 和 Q_i 构造训练数据集(X_i,y_i),作为待训练排序模型 $M_i(\theta,x)$的输入.训练过程为标准的经验风险最小化问题,即 $\theta_{opt}=\arg\min_\theta E_{(x,y)\in(X_i,y_i)}[L(M_i(\theta,x),y)]$.其中,$L$ 是损失函数.在面向企业数据孤岛的联邦场景中,每个企业还可以和其他企业开展合作,来生成更多的训练数据 X_i'.如前所述,这些训练数据没有标签,在这种情况下,各企业需要联合训练一个全局模型.

为了简化各企业之间的协作场景,假设训练过程中存在一个中心服务器,以协调各方信息交互.同时假设中心服务器和各个企业都是诚实但好奇的(honest-but-curious),即半诚实的(semi-honest).它们会遵守协议的要求,但是也会根据收到的数据推断敏感信息.面向企业数据孤岛联邦学习的目标是:在保护各企业数据隐私的情况下,训练一个有效的全局模型.

2.2 联邦场景下的排序学习问题

基于第 2.1 节所述的通用联邦场景,排序学习的联邦场景问题定义如定义 1.

定义 1(联邦场景下的排序学习问题). 给定有 n 个企业的联邦和一个特征提取函数 Φ,联邦场景下的排序学习旨在让各企业协同训练一个全局的排序模型 M.其中,每个企业 P_i 拥有一个同其他企业合作产生的无标签数据集 X_i' 和自身持有的带标签数据集(X_i,y_i),训练过程需要满足以下条件.

- 共享性:对任意的两个企业 $P_i,P_j(i\neq j)$、任意查询 $q\in Q_i$ 和任意文档 $d\in D_j$,存在一个 $x_i'=X_i'$,满足 $x_i'=\Phi(d,q)$;
- 隐私性:在训练过程中,任意两个企业之间、企业和中心服务器之间的信息交互造成数据 D_i 和 Q_i 的隐私泄漏需要可控;
- 有效性:协同训练的模型 M 要有比各企业独立训练的模型 M_i 有更好的性能.

上面的 3 个条件对于一般的面向企业数据孤岛的联邦学习场景同样适用,但是本问题由于数据的分割造成了交叉特征生成难度大、标签丢失等诸多新挑战,克服这些挑战所需的技术将在下面两节介绍.

3 隐私保护的交叉特征生成

本节介绍基于 Sketch 的企业间隐私保护的交叉特征生成解决方案(如图 2 所示).该方法旨在保持共享性和隐私性的条件下,解决联邦排序学习中的企业间交叉特征生成问题.

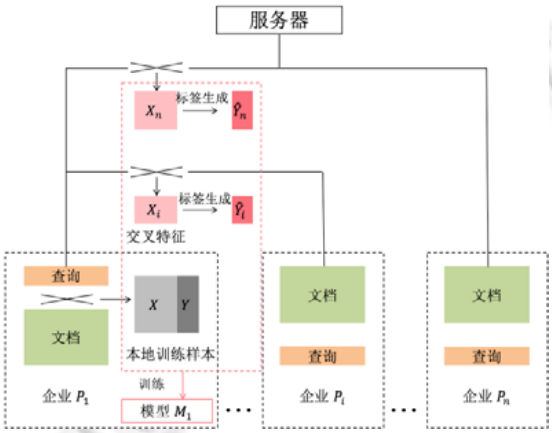


Fig.2 Mutual feature generation for P_1

图 2 P_1 的交叉特征生成

3.1 交叉特征介绍

为简化说明,我们使用以下两个交叉特征实例展示基于 Sketch 的交叉特征生成解决方案:词频(term frequency,简称 TF)和逆文本频率(inverse document frequency,简称 IDF).在面向企业数据孤岛的联邦场景下,两个特征定义分别如定义 2、定义 3 所示.

定义 2(跨企业数据孤岛词频). 假设 $P_i,P_j\in F,d\in D_j,q\in Q_i,t_1,t_2,\ldots,t_M$ 是查询 q 中的 M 个单词.t_k 在文档 d 中的的跨企业数据孤岛词频 $TF_{i,j}(t_k,d)=\dfrac{TC_{i,j}(t_k,d)}{l_d}$,其中,$TC$ 表示出现的次数,l_d 是文档的长度.假设文档长度不属于隐私信息可以直接传输,那么计算跨企业数据孤岛词频就等价于计算跨企业数据孤岛的单词出现次数.

定义 3(跨企业数据孤岛逆文本频率). 单词 t_k 的跨企业数据孤岛逆文本频率为

$$IDF_i(t_k)=\frac{\sum_j|D_j|}{\sum_j\sum_{d\in D_j}I(TF_{i,j}(t_k,d)>0)} \tag{1}$$

其中,I 是指示器函数,输入条件为真时值为 1,否则为 0.因为需要聚合各企业的词频,跨企业数据孤岛逆文本频率和传统计算方法略有不同,将在下文着重说明.

3.2 跨企业数据孤岛词频生成

本部分的目标是:在保护双方(提供文档的一方和提供查询的一方)数据隐私的情况下,计算跨企业数据孤岛词频.在这样一个多方场景下,任意两方需要相互多次查询对方的文档.传统的基于加密的方法在效率和灵活性上都很不足,因此,我们设计了基于 Sketch 的解决方案.该数据结构有如下优势:首先,它在构造后可重用,即有新企业加入不会导致其他已加入企业产生额外开销;其次,它响应查询速度很快,是常数时间复杂度;最后,由于使用了哈希函数,该数据结构很自然地隐藏了部分文本信息,在此基础上进行基于差分隐私的改进(详见第 3.2.2 节),可以实现更强的隐私保护.

首先给出 Sketch 数据结构和差分隐私的相关定义(分别为定义 4、定义 5).

定义 4(Sketch). 对于一篇文档 d,为其构造 Sketch $S=(Enc,f_S)$,其中:Enc 是一个编码函数,使得 $C_d=Enc(d)$ 是 d 的一个编码数据概要;而 f_S 是一个计数查询函数,满足 $f_S(d_t)$ 是单词 t 的计数查询结果.

以经典的 Count Sketch 为例(也可使用 Count-Min Sketch).Count sketch 的编码过程需要两个哈希函数集合 $H=\{h_1,h_2,\ldots,h_z\}$和 $G=\{g_1,g_2,\ldots,g_z\}$,它们是从 $h_i:\Gamma\to[1,w](w<<|\Gamma|)$和 $g_i:\Gamma\to\{-1,+1\}$中随机成对抽取形成的.每个单词 t 的编码方式为

$$C_d(a,h_a(t))\leftarrow C_d(a,h_a(t))+g_a(t),\forall 1\leqslant a\leqslant z \tag{2}$$

在把文档 d 编码后,将得到一个 z 行 w 列的表 $C_d(\cdot,\cdot)$.单词 t 的计数查询方式为

$$\overline{f}_t = f_C(d,t) = median_{a\in[1,z]}C_d(a,h_a(t)) \tag{3}$$

因为 Sketch 上的查询过程需要多个哈希函数,我们可以通过直接混淆这些哈希函数来实现查询方的隐私保护.而为了定量描述文档一方的隐私保护程度,需要定义计数查询的ε-差分隐私(ε-differential privacy,简称 ε-DP).因为依赖于 Sketch,ε-DP 的定义和常规定义略有不同.

定义 5(计数查询的ε-差分隐私). 一个随机算法 A 满足ε-DP,如果对任意相邻文档 d 和 d'(只相差一个单词)、任意单词 t 的计数查询 f_S 和所有 A 的可能输出结果 o,有:

$$\Pr[A(f_S(d,t'))=o]\leqslant e^{\varepsilon}\Pr[A(f_S(d,t))=o].$$

如果计数查询的结果满足ε-DP,那么文档方的隐私信息就无法推断.

接下来介绍跨企业数据孤岛词频特征生成方案的技术细节.假设第 i 个企业 P_i 有单词 t,它希望找到该单词在另一方 P_j 的文档 d 的词频.

(1) 构造 Sketch

首先,企业 P_j 构造文档 d 的 Sketch.构造 Sketch 是在各方开展通讯之前就完成的.各方使用相同的哈希函数构造 Sketch,所以在不同的企业数据中都可以查询.哈希函数可以通过企业间加密传输的方式(如 Diffie-Hellman 密钥交换协议)来生成索引,从而避免服务器推断哈希函数的有关信息.按照前述公式(2)可实现 Sketch 构造,其中,单词来自词汇表$|\Gamma|$,构造后将得到一个 z 行 w 列的表格.文档 d 有 l_d 个单词,如果哈希过程时间复杂度为 $O(1)$,则需要花费 $O(zl_d)$的时间来构造 Sketch.每个文档的 Sketch 都由各方在本地保存,只能由查询找到某个项的频率,构造过程如图 3 所示.

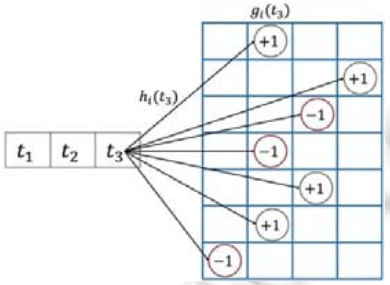


Fig.3 Construction of the Sketch from a document

图 3 一篇文档的 Sketch 构造过程

(2) 带混淆哈希

完成 Sketch 的构造后,为得到单词出现次数,P_i 需要使用 z 个 H 中的哈希函数将查询中的单词 t 进行哈希.由于隐私保护的要求,不能直接在 Sketch 上进行查询,而是随机从 H 中挑选 z_1 个哈希函数对单词 t 进行哈希.对其他的 $z-z_1$ 个哈希函数,他们的输入是随机从$|\Gamma|$中选取的,使得这些单词经过哈希后和 t 经过哈希的结果产生碰撞,从而实现混淆.经过了这样的哈希过程和混淆过程后,P_i 将得到一个长度为 z 的向量 $(h_a^{(i)}(t))_{(1\leqslant a\leqslant z)}$,其中有 z_1 个位置的哈希函数索引是由 P_i 所私有的,即:

$$h_a^{(i)}(t)=\begin{cases}h_a(t), & a\in PK_i\\ h_a(t'), & a\in B_{t'},cd(t,t')=b\end{cases},\forall 1\leqslant a\leqslant z \tag{4}$$

其中,PK_i 是 P_i 的私有哈希函数索引(1 到 z 随机排列后前 z_1 个值的集合),$cd(t,t')$是单词 t 和 t'发生碰撞的哈希函数个数.$B_{t'}$包含不属于 PK_i 的 z_1-b 个整数,且每个 $B_{t'}$彼此不重叠.因为单词 t 和 t'有 b 个哈希结果相同,我们也能

找到 z_1 个哈希函数结果和 t'一致.因此,任意其他的参与方无法通过暴力枚举 z_1 个哈希值区分原始的单词 t 和用于混淆的单词 t'.然而,我们也有 $E[|\{t'|cd(t,t')=b\}|]=\left(\frac{1}{w}\right)^b|\Gamma|$,所以如果 b 太大,就很难找到用来混淆的单词 t'.因此,存在一个隐私保护程度和数据精度的权衡.我们设置一个较小的 z_1 和 b 来实现相同的隐私保护程度,同时保持查询结果的可信度.哈希处理后,P_i 将把混淆的哈希向量发送给服务器,由服务器把该向量发送给 P_j 做进一步查询处理.

(3) 查询结果扰动

本步骤中,P_j 将收到服务器发来的 z 维向量,然后使用 $h_a^{(i)}(t)$ 在 Count Sketch 上进行查询,得到 $C_{d,H}(a,h_a^{(i)}(t))$, $1\leqslant a\leqslant z$.把查询的结果直接发布将产生隐私泄露风险,因为恶意的攻击者可能会查询一些敏感单词来推断其他参与方的文档词频分布.因此,我们设计了一个 ε-DP 机制来扰动查询的结果,以保护文档隐私.我们的扰动机制采用拉普拉斯机制,扰动方式如下:

$$C'_d(a,h_a^{(i)}(t))=C_d(a,h_a^{(i)}(t))+N \tag{5}$$

其中, $N\sim Lap\left(\frac{1}{\varepsilon}\right)$,$\varepsilon$为隐私预算.对所有的哈希结果,只采样一个随机数,后续我们将证明,该机制满足ε-DP.随后,P_j 将把扰动后的 Sketch 查询结果通过服务器发送给 P_i.收到结果后,只需要通过 PK_i 恢复结果.最后的单词出现次数为

$$f_t=median_{a\in PK_i}C'_d(a,h_a^{(i)}(t)) \tag{6}$$

相关过程如图 4 所示.算法 1 和算法 2 分别描述了查询方和文档拥有者的操作过程.

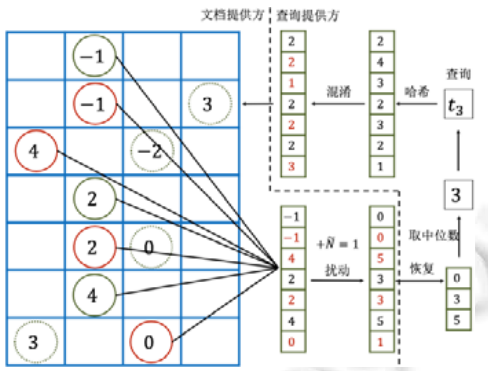


Fig.4 Thecount query of term from the Sketch

图 4 单词的词频查询过程

算法 1. 跨企业数据孤岛词频计算:查询方.

Input:待查询单词 t,哈希函数 H;

Output:估计的单词频率 f_t.

1: Q←空向量

2: PK←随机生成的哈希函数编号

3: **for** $1\leqslant a\leqslant z$ **do**

4: 依据公式(4)生成哈希值

5: 添加到 Q 结尾

6: 把 Q 发送给服务器

7: 从服务器收到查询结果

8: 去掉混淆项

9: 按照公式(6)估计 f_t

10: **return** f_t

算法 2. 跨企业数据孤岛词频计算:文档持有方.

Input:文档 d,单词的哈希值 Q,哈希函数 H;

Output:无.

1: 按照公式(2)构造 C

2: 从服务器接收向量 Q

3: F_Q←空向量

4: 采样 N

5: **for** $1 \leqslant a \leqslant z$ **do**

6: 注入噪声

7: 追加到 F_Q

8: 把 F_Q 发送给服务器

9: **return** None

3.3 跨企业数据孤岛逆文本频率生成

上文阐述了面向企业数据孤岛联邦场景下的词频特征生成过程,对于逆文本频率,过程将略有不同.跨企业数据孤岛逆文本频率的定义如第 3.1 节公式(1)所示.公式中分子部分表示企业所拥有的文档总数,这并非隐私信息可以直接公开.计算的主要挑战来自对分母部分的计算.一个朴素的解决方案是,在每篇文档上执行跨企业数据孤岛词频特征生成.然而,这样的解决方法十分低效,仅针对一个单词就需要进行 $O\left(\sum|D_j|\right)$ 次查询.造成这种低效的主要原因是:逆文本频率的计算需要语料库的全局信息,而不仅仅是单一的文档信息.为了提高效率,我们设计了一种新的方法如下.

首先,从 P_j 的所有个文档中生成一个新的文档.它的单词是所有在 D_j 中出现过的单词,而单词出现次数为所有出现过该单词的文档数量.为这个新文档构造一个 Sketch,这样,查询中的每个单词在此 Sketch 上查询的结果就是包含该单词的文档数量.后续步骤和跨企业数据孤岛词频生成相同,即:使用混淆单词隐藏查询的真实单词,为查询结果添加拉普拉斯噪声满足差分隐私.最终的时间复杂度可以被削减为 $O(n)$.

事实上,我们还可以进一步削减查询的时间.可以考虑通过安全聚合的方法把各方新文档的 Sketch 聚合为一个,存储在服务器上.这样,未来的查找可以在 $O(1)$时间内完成.

3.4 其他跨企业数据孤岛特征

上文我们主要介绍了词频和逆文本频率的跨企业数据孤岛隐私保护交叉特征生成方案,该方法可以很容易拓展到其他排序学习所使用的统计特征中.举例来说,几乎所有 LETOR 4.0[51]所使用的特征(网页链接数等文档重要度特征除外)都可以通过这两个基本操作来生成.例如,依据前面所介绍的跨企业数据孤岛词频和逆文本频率,可以定义跨企业数据孤岛 BM25[49]特征:

$$\mathrm{BM25}_{(i,j)}(d,q)=\sum_{k=1}^{l}\frac{IDF_i(t_k)TF_{i,j}(t_k,d)(k_1+1)}{TF_{i,j}(t_k,d)+k_1\left(1-b+\dfrac{bl_d}{avdl}\right)},$$

其中,$avdl$ 是文档的平均长度,l 是查询的长度,k_1 是超参数.不难发现:如果词频和逆文本频率是事先算好的,那么 BM25 可以在 $O(l)$的时间内计算出来.

同理可以定义 LMIR[50]的诸多特征:

$$LMIR.JM=(1-\lambda)\frac{TC_{i,j}(t,d)}{\sum_k TC_{i,j}(t_k,d)}+\lambda p(t\mid D_i),$$

$$LMIR.DIR=\frac{TC_{i,j}(t,d)+\mu p(t\mid D)}{\sum_k TC_{i,j}(t_k,d)+\mu},$$

其中,λ和μ是超参数,$p(t|D_i)$是单词在文档中的词频.

最后,对每个企业可以生成一个新的数据集 X'.该数据包含了自身查询语句和其他企业的文档产生的交叉特征,满足了共享性.

3.5 隐私保护与精度损失的理论分析

本部分分析我们的跨企业数据孤岛隐私保护交叉特征生成方案的隐私保护情况和精度损失情况.

3.5.1 隐私保护上界

差分隐私是联邦学习场景中定量衡量隐私保护程度的事实标准.从定义 5 可以看出,其核心思想是用概率之比衡量不可区分性.下面我们将证明我们的方法满足ε-差分隐私.

定理 1. 对任意计数查询,公式(5)的扰动方法满足ε-差分隐私.

证明:设文档 d'的长度为 $l_{d'}=l_d+1$.文档 d 和 d'是相邻的文档,即前 l_d 个单词相同,而 d'比 d 多一个单词 t'.由每对哈希函数彼此独立,可得:

$$\Pr[h_a(t)=h_a(t')]\leqslant\frac{1}{range(h_a)}=\frac{1}{w},\forall t\neq t'.$$

对任意一个确定的企业 i,我们分析相邻文档 d 和 d'中单词经过扰动后的不可区分程度.具体来说,分为两文档单词不同的情况($t\neq t'$)和单词相同的情况($t=t'$).

对$\forall t\neq t'$的情况,可得:

$$C_{d'}(a,h_a(t))=C_d(a,h_a(t))+RX_a\cdot g_a(t),\forall a\in PK_i.$$

其中,RX_a 是从伯努利分布 $\Pr[RX_a=1]\leqslant\frac{1}{w}$ 采样的结果.于是可得:

$$\begin{aligned}f_C(d',t)&=median_{a\in PK_i}C_{d'}(a,h_a(t))\\&=median_{a\in PK_i}(C_d(a,h_a(t))+RX_a\cdot g_a(t))\\&=median_{a\in PK_i}(C_d(a,h_a(t))+RY)\\&=f_C(d,t)+RY,\end{aligned}$$

其中,$RY\in\{+1,0,-1\}$.

令 $a_0=\arg median_{a\in PK_i}C_d(a,h_a(t))$,可得:

$$\Pr[RY=1\vee RY=-1]=\Pr[f_C(d',t)=h_{a_0}(t)+1\vee f_C(d',t)=h_{a_0}(t)-1]\leqslant\Pr[RX_{a_0}=1]\leqslant\frac{1}{w}.$$

注入拉普拉斯噪声 $N\sim Lap\left(\frac{1}{\varepsilon}\right)$,即在查询结果上增加随机数 N 后,可得:

$$\begin{aligned}\frac{\Pr[A(f_C(d',t))=o]}{\Pr[A(f_C(d',t))=o]}&=\frac{\Pr[N=o-f_C(d,t)-RY]}{\Pr[N=o-f_C(d,t)]}\\&=\Pr[RY=0]+\Pr[RY=1]\frac{\Pr[N=o-f_C(d,t)-1]}{\Pr[N=o-f_C(d,t)]}+\Pr[RY=-1]\frac{\Pr[N=o-f_C(d,t)+1]}{\Pr[N=o-f_C(d,t)]}\\&\leqslant 1-\frac{1}{w}+\frac{1}{w}e^{\varepsilon}\\&\leqslant e^{\varepsilon}.\end{aligned}$$

对于 $t=t'$的情况,我们有:

$$C_{d'}(a,h_a(t))=C_d(a,h_a(t))+1,\forall a\in PK_i.$$

因此,$f_C(d',t)=f_C(d,t)+1$,于是可得 $\frac{\Pr[A(f_C(d',t))=o]}{\Pr[A(f_C(d,t))=o]}\leqslant e^{\varepsilon}$.定理得证. □

从上面的证明过程可以看出:虽然我们的方法中带混淆哈希和拉普拉斯噪声都提供了单词的不可区分性,但是拉普拉斯噪声的方差直接反比于隐私预算ε,仅拉普拉斯噪声提供的不可区分性就足以提供满足ε-差分隐私的保护了.为了进一步充分利用带混淆哈希所产生的不可区分性,可以适当降低拉普拉斯噪声的注入量.

推论 1. 若注入的拉普拉斯噪声 $N\sim Lap\left(\frac{1}{\varepsilon'}\right)$,其中,$\varepsilon'=\ln\left[w\left(e^{\varepsilon}-1+\frac{1}{w}\right)\right]$,则公式(5)的扰动方法依然满足$\varepsilon$-差分隐私.

证明:根据定理 1 的证明,我们有:

$$\begin{aligned}\frac{\Pr[A(f_C(d',t))=o]}{\Pr[A(f_C(d,t))=o]}&=\frac{\Pr[N=o-f_C(d,t)-RY]}{\Pr[N=o-f_C(d,t)]}\\&=\Pr[RY=0]+\Pr[RY=1]\frac{\Pr[N=o-f_C(d,t)-1]}{\Pr[N=o-f_C(d,t)]}+\Pr[RY=-1]\frac{\Pr[N=o-f_C(d,t)+1]}{\Pr[N=o-f_C(d,t)]}\\&\leqslant 1-\frac{1}{w}+\frac{1}{w}e^{\varepsilon}\\&\leqslant e^{\varepsilon}.\end{aligned}$$

因为 $\varepsilon'=\ln\left[w\left(e^{\varepsilon}-1+\frac{1}{w}\right)\right]$,代入得:

$$\frac{\Pr[A(f_C(d',t))=o]}{\Pr[A(f_C(d',t))=o]}\leqslant 1-\frac{1}{w}+\frac{1}{w}e^{\varepsilon'}=1-\frac{1}{w}+\frac{1}{w}\left[w\left(e^{\varepsilon}-1+\frac{1}{w}\right)\right]=e^{\varepsilon}.$$

推论得证. □

3.5.2 精度损失上界

Count Sketch 的频率估计是一个无偏估计,其方差为 $\frac{F_2}{w}$,其中,$F_2=\sum_{1\leqslant k\leqslant l_d}f_{t_k}^2$.为了进一步降低精度的损失,可以考虑利用数据的偏度实现对方差的进一步控制.文档中,单词的频率通常满足齐夫定律,按照文献[52],F_2可以进行如下放大:

$$F_2^R=\sum_{1\leqslant k\leqslant l_d}f_{t_k}^2\leqslant\frac{c_z^2(r-1)^{1-2\varsigma}}{2\varsigma-1},$$

其中,$f_i=\frac{c_z}{i^{\varsigma}}$ 是齐夫定律中第 i 个最常出现的单词的频率.可以通过该上界实现对错误上界的控制.

定理 2. 对于一个单词 t,如果 z_1 被设定为 $O\left(\log\left(\frac{1}{\delta}\right)\right)$的,那么至少有 $1-\delta$的概率,可以保证词频估计有如下上界:

$$|f_t'-f_t|\leqslant\sqrt{\frac{16}{\varepsilon^2}+\frac{64}{w}\cdot F_2^R}.$$

证明:依据公式(6),我们有:

$$C_a(t)=f_t+\sum_{t':g_a(t')}g_a(t)g_a(t')f_{t'}+Lap\left(\frac{1}{\varepsilon}\right).$$

由 Count Sketch 估计的期望和方差以及拉普拉斯噪声的方差,可得:

$$E[C_a(t)]=f_t,$$

$$Var[C_a(t)]=\sum_{t':g_a(t')}g_a(t)g_a(t')f_{t'}^2+\frac{2}{\varepsilon^2}.$$

按照齐夫定律关于词频分布规律的假设,不论如何选取哈希函数,总有 7/8 的概率,使得 $r=\frac{w}{8}$ 的最频繁项和

给定的单词 k 在任意给定行都不会发生碰撞.因此, $E[\sum_{k>r} f_{t_k}^2]=\frac{F_2^R}{w}$.

由马尔可夫不等式 $\Pr\left[\sum_{k>r} f_{t_k}^2 \leqslant \frac{8F_2^R}{w}\right] \geqslant \frac{7}{8}$,于是我们有:

$$\Pr\left[Var[C_a(t)] \leqslant \frac{8F_2^R}{w}+\frac{2}{\varepsilon^2}\right] \geqslant \frac{7}{8} \tag{7}$$

由切比雪夫不等式:

$$\Pr\left[|C_a(t)-f_t| \geqslant \sqrt{\frac{64F_2^R}{w}+\frac{16}{\varepsilon^2}}\right] \leqslant \frac{1}{8}\cdot\frac{Var[C_a(t)]}{\frac{8F_2^R}{w}+\frac{2}{\varepsilon^2}} \tag{8}$$

由公式(7)和公式(8),可得:

$$\Pr\left[|C_a(t)-f_t| \geqslant \sqrt{\frac{64F_2^R}{w}+\frac{16}{\varepsilon^2}}\right] \geqslant 1-\frac{1}{8}-\frac{1}{8}-\frac{1}{8}=\frac{5}{8}.$$

因为哈希函数是独立的,由切诺夫界,最终可得:

$$\Pr\left[|f_t'-f_t| \geqslant \sqrt{\frac{16}{\varepsilon^2}+\frac{64}{w}\cdot F_2^R}\right] \leqslant e^{-O(z_1)}.$$

定理得证. □

对于长度为 l 的查询 q,我们使用 $f_q'=median_{a\in PK_i}\sum_{1\leqslant k\leqslant l} C_d'(a,h_a(t_k))$ 进行频数估计,则错误上界满足定理 3.

定理 3. 对于长度为 l 的查询 q,如果设置 $z_1=O\left(\log\left(\frac{1}{\delta}\right)\right)$,那么至少有 $1-\delta$ 的概率,可以保证出现次数的估计有如下上界:

$$|f_q'-f_q| \leqslant \sqrt{\frac{16l}{\varepsilon^2}+\frac{64l}{w}\cdot F_2^R}.$$

证明:首先证明任意两个单词 t_1 和 t_2 以及任意的 $a\in PK$,$C_a(t_1)$ 和 $C_a(t_2)$ 是独立的.这是因为我们有:

$$C_a(t_1)=f_{t_1}+\sum_{t':g_a(t')} g_a(t_1)g_a(t')f_{t_1}'+Lap\left(\frac{1}{\varepsilon}\right),$$

$$C_a(t_2)=f_{t_2}+\sum_{t':g_a(t')} g_a(t_2)g_a(t')f_{t_2}'+Lap\left(\frac{1}{\varepsilon}\right).$$

所以有:

$$\begin{aligned}E[(C_a(t_1)-f_{t_1})(C_a(t_2)-f_{t_2})] &= E[\sum_{t':g_a(t')} g_a(t_1)g_a(t')f_{t_1}'\cdot\sum_{t':g_a(t')} g_a(t_2)g_a(t')f_{t_2}'] \\ &= E[\sum_{t',t''} g_a(t_1)g_a(t')g_a(t_2)g_a(t'')f_{t'}^2] \\ &= 0.\end{aligned}$$

因此有:

$$E[\sum_{1\leqslant k\leqslant l} C_a(t_k)]=\sum_{1\leqslant k\leqslant l} E[C_a(t_k)]=\sum_{1\leqslant k\leqslant l} f_{t_k},$$

$$Var[\sum_{1\leqslant k\leqslant l} C_a(t_k)]=\sum_{t',k:g_a(t)=g_a(t_k)} f_{t'}^2+\frac{2p}{\varepsilon^2}.$$

按照定理 2 的证明,可得:

$$\Pr\left[|f_q'-f_q| \geqslant \sqrt{\frac{16l}{\varepsilon^2}+\frac{64l}{w}\cdot F_2^R}\right] \leqslant e^{-O(z_1)}.$$

定理得证. □

4 半监督协同排序学习

本节介绍半监督协同排序学习方法,来克服标签缺失的挑战.在一个企业所持有的数据之内,可以很容易地评估查询和文档生成的交叉特征样本的相关性,形成带标签的样本数据.但是跨企业数据孤岛所生成的交叉特征却无法直接评估相关性,因此不存在标签.现有的半监督学习方法并不能直接应用到本问题中,主要有两个原因:一是带标签和不带标签的数据都是分属各方的,如图 5 所示;二是半监督学习过程也需要考虑原始数据的隐私.为了解决上述问题,我们首先提出一个简单的半监督学习基准算法,然后再设计一个协同学习的方法.

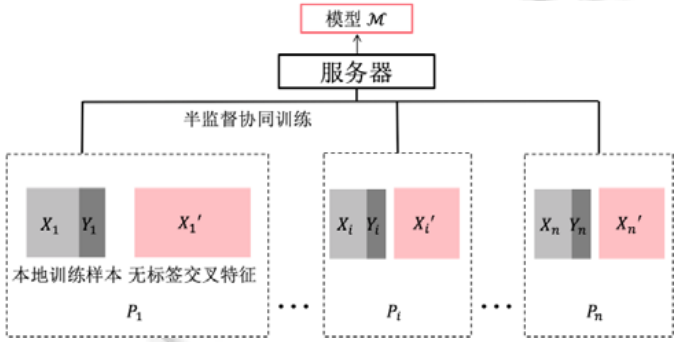


Fig.5 Framework of semi-supervised collaborative learning

图 5 半监督协同训练框架

4.1 基准算法:本地训练

每个企业拥有自身的带标签数据(X_i,y_i)和无标签数据 X',它可以在本地利用这些数据以半监督训练的方式得到一个模型.本地训练可以自然符合数据不离开本地的特点,所以简单地满足了隐私限制.本地训练中,可以使用一个通用的伪标签生成器来实现半监督学习.学习过程就是求解如下最优化问题:

$$\theta_{opt}=\text{argmin}_{\theta}[E_{(x,y)\in(X,Y)}[L(M_i(\theta,x),y)]+\beta E_{x'\in X'}[L(M_i(\theta,x'),y')]],$$

其中,y'是 x'由伪标签生成器生成的的伪标签,β是无标签数据的权重.伪标签生成器的设计有多种方法,我们简单地直接使用上一轮的模型给无数据打标签.最后,各方都会持有一个由自己训练的私有模型M_i作为最终结果.在这种方案中,生成的无标签数据并没有被充分使用,所以我们提出了一个改进的协同训练方案.

4.2 无标签数据的协同训练

一个直接的协同训练方法是:只使用各方的带标签数据,问题就被规约为标准的横向联邦学习.但是在这种情况下,针对查询的数据依然十分短缺,即每条查询所对应的特征空间不够大.此外,隐私保护也不应被忽视,因为梯度可能泄漏敏感信息.

为了能更好地使用这些丰富但无标签的、同时不包含 P_i 隐私信息的数据,我们设计了如下方法:首先,在协同训练之前,各方先在本地使用自身带标签数据训练一个分类模型 M_i;之后,各方把这个本地模型发送给服务器,服务器聚合这些模型作为一个标签生成器,即 $\bar{M}=Agg(M_1,M_2,...,M_n)$;$Agg(\cdot)$对各个模型的参数求平均,得到全局模型.$\bar{M}$ 由各方共享,且它也是标签生成器和排序模型的初始值.

各方使用 $\bar{M}$ 在本地对数据打标签,然后通过协同训练优化:

$$\theta_{opt}=\text{argmin}_{\theta}[E_{(x,y)\in(X,Y)}[L(M(\theta,x),y)]+\beta E_{x'\in X'}[L(M(\theta,x'),y')]] \tag{9}$$

可以以串行或并行的方式运行梯度下降算法.在每一轮迭代中,各方在本地使用一批带标签和无标签数据运行梯度下降算法,并提交梯度变化$\Delta\theta$给服务器.协同训练过程中使用的训练数据就是前文介绍的给类交叉特征.交叉特征的生成过程需要使用企业的原始文本数据,所以我们设计了带混淆哈希、拉普拉斯噪声注入等方法保护隐私.而协同训练的过程只是在使用这些特征,并不涉及企业原始数据,并且无标签数据的来源只有提交查询的 P_i 知道,即使攻击者有能力依据梯度更新反推出用于训练的交叉特征,仍然不会泄漏企业的原始文本数

据,所以我们不再对$\Delta\theta$使用额外的隐私保护手段.最后,随着θ的收敛,各方将获得一个全局模型 M.其他细节详见算法 3.

算法 3. 半监督协同训练算法

Input:带标签和无标签数据,其他参数;

Output:全局模型.

1: **for** 企业 P_i **do**

2: P_i 训练一个本地模型 M_i

3: $\bar{M} = Agg(M_1, M_2, ..., M_n)$

4: $M \leftarrow \bar{M}$

5: 使用 M 给无标签数据打标签

6: **for** t=1,2,… **do**

7: **for** 企业 P_i **do**

8: 收集带标签和无标签数据批数据

9: 根据公式(9)更新权重θ

10: **return** M

5 实验评估

本节介绍实验结果来证明所提方法的有效性.

5.1 数据集和实验设置

以 LETOR4.0[51]为代表的大部分标准数据集都只包含了已经提取好的特征,而没有原始的文档和查询数据,因此无法应用在我们的实验中.本文中,我们选择 MS MARCO 排序数据集(http://www.msmarco.org/dataset.aspx),从中采样部分数据用于实验.假设企业的数量为 4,每一方只有有限的带标签的查询结果.从 MS MARCO 中采样 4 个子集,每个包含约 200 条查询和 40 000 篇文档,每篇文档有大约 1 000 个单词.我们所提出的技术是为数据规模较小的企业提供帮助,因此设置单个企业数据孤岛的数据规模控制在 10^4 数量级.使用每条查询最相关的 100 篇文档的相关度分数作为标签.最相关的 10 篇文档被标记为非常相关(相关性分数为 2),排在 11 名~100 名的文档是相关(相关性分数为 1),100 名以外是不相关(相关性分数为 0).对每一方,生成约 2.4 万条带标签数据,并取其中 7 000 条样本数据构造全局测试集(共 2.8 万条).

在半监督学习模型的设置方面,我们采用点排序模型.模型结构为线性分类模型.损失函数设置为交叉熵,并且引入了 L2 正则项.本地训练和全局训练因为数据规模不同,分别设置迭代 1 000,2 000 轮次.更新过程采用批处理随机梯度下降,全局模型的更新在每一批数据处理完成后进行.而所提算法则取本地模型平均用来打标签,以开展线性回归训练.

交叉特征选取的主要依据是 LETOR4.0 所使用的公认排序学习特征.查询分别和文档的标题与正文交叉计算词频、逆文本频率、TF-IDF、BM25、LMIR.ABS、LMIR.DIR、LMIR.JM 属性,结合文档标题长度与正文长度,形成一条拥有 16 个特征的样本数据.与 LETOR4.0 相比,我们的排序对象是纯文本文档,不包含链接数量、网址深度等网页文档的特有特征.此外,除了我们所采用的查询语句和正文与标题分别计算的交叉特征外,LETOR4.0 还进一步把正文与标题看作整体,再与查询语句进行交叉特征计算.但是我们的排序对象正文长度大,把标题和正文结合和仅考虑正文相比差异不大,加之过多特征还会显著拖慢训练速度,所以我们并未引入这些特征.我们为每一方生成了 5.97 万条跨企业数据孤岛数据,用于半监督协同学习.可以看到:同单个企业数据孤岛所能产生的 2.4 万条数据相比,交叉特征生成的过程显著增加了训练样本数量.增加的数据对于模型训练的帮助将在后面的实验中具体说明.文档的预测顺序将按照模型给出的预测分数从高到低排序.评估排序好坏的指标有:

- 期望倒数排名(expected reciprocal rank,简称 ERR).评估相关度高的文档其排序位置是否靠前,计算方式为

$$ERR=\sum_{r=1}^{|D_i|}\prod_{i=1}^{r-1}\left(1-\frac{2^{rel_i}-1}{2^{rel_{\max}}}\right)\frac{2^{rel_r}-1}{2^{rel_{\max}}},$$

其中,rel_i表示排序在第 i 个位置文档的相关性分数(即 0,1,2),$rel_{\max}$为最大的相关性分数(即为 2);

- 平均准确率(mean average precision,简称 MAP).评估排序学习模型给出的顺序与按照实际相关度排序的差别,计算方式为

$$MAP=\frac{1}{|D_i|}\sum_{d\in D_i}\frac{rank(d)}{position(d)},$$

其中,$rank(d)$表示文档 d 按照相关度分数排序时的位置,$position(d)$为排序学习模型给出的位置;

- 归一化折损累计增益(normalized discounted cumulative gain,简称 *n*DCG).综合考虑相关性和排序位置的评估指标,计算方式为

$$nDCG=\frac{1}{IDCG}\sum_{i=1}^{|D_i|}\frac{2^{rel_i}-1}{\log(i+1)},$$

其中,IDCG 为归一化系数,其值为文档按照相关度分数降序排列时得到的 $\sum_{i=1}^{|D_i|}\frac{2^{rel_i}-1}{\log(i+1)}$;

- 前 10 名归一化折损累计增益(normalized discounted cumulative gain at 10,简称 *nDCG*@10).即排在前 10 文档的 *nDCG*.计算方式为

$$nDCG@10=\frac{1}{IDCG}\sum_{i=1}^{|10|}\frac{2^{rel_i}-1}{\log(i+1)}.$$

我们的方法是为全新的交叉分割联邦学习场景设计的,目前尚未有其他方法能够适用于这样的场景.这让我们很难找到其他的横向对比方法.因此,我们实验的重点在于检验使用交叉特征生成进行增广的数据集是否有助于排序模型的性能提升,于是选取以下 3 种方法同所提算法 CS-F-LTR 进行比较.

- Local:各方仅使用自己的带标签数据训练一个模型;
- Local+:各方使用自己的带标签和无标签数据训练一个半监督模型;
- Global:各方使用带标签数据协同训练一个模型,和横向联邦学习场景类似.

5.2 实验结果

5.2.1 主要结果

在表 4 中,我们记录了 Local,Local+,Global 和 CS-F-LTR 的 4 项评估指标.

Table 4 Comparison results

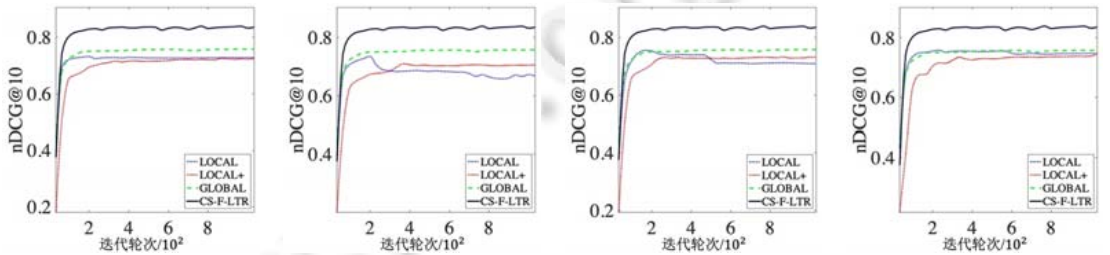
表 4 比较结果

		ERR	*nDCG*@10	*n*DCG	MAP
Local	企业 *A*	0.567 2	0.727 4	0.804 9	0.480 9
	企业 *B*	0.535 7	0.667 6	0.777 9	0.457 3
	企业 *C*	0.563 7	0.709 4	0.795 2	0.478 8
	企业 *D*	0.588 7	0.746 6	0.817 3	0.505 5
Local+	企业 A	0.562 8	0.725 7	0.802 6	0.477 6
	企业 *B*	0.544 8	0.706 4	0.793 2	0.477 9
	企业 *C*	0.576 6	0.731 4	0.809 1	0.493 6
	企业 *D*	0.598 8	0.747 5	0.824 7	0.503 9
Global		0.587 7	0.757 7	0.824 7	0.504 0
CS-F-LTR		**0.667 9**	**0.835 8**	**0.876 0**	**0.552 5**

可以发现:各方的 *nDCG*@10 指标在 Local 方法下结果为 0.7 左右,该方法只采用了本地数据来训练.而 Local+的方法使用了无标签数据,因此在一定程度上可以提高本地模型的表现.企业 *B*,*C* 的 *nDCG*@10 分别增长

了 5.8%,3.1%,企业 *D* 略微变好,企业 *A* 略微更差.这是因为各方的数据并不一定是独立同分布的.所以各方依据自身数据所训练的伪标签生成器在给交叉产生的互特征生成标签时,如果双方数据分布差异过大,打标签就可能出错.在 CS-F-LTR 中,通过模型平均和使用无标签数据,*nDCG*@10 有一个明显的提高,从 0.65~0.75 的范围升高到 0.83.同时,把 CS-F-LTR 和 Global 所训练的模型比较,后者虽然也聚合了数据但是没有使用半监督学习的方法,通过比较发现,*nDCG*@10 分数提高了.而对于其他的评价指标,分数也有一定的提高.

各企业在使用不同方法训练过程中,*nDCG*@10 指标随迭代过程变化如图 6 所示.可以看出:本地训练的两个方法(Local 和 Local+)因为使用的数据有限,性能显著低于全局方法训练的模型;而在两个全局方法(CS-F-LTR 和 Global)中,我们所提出的 CS-F-LTR 方法因为能够进一步使用无标签数据采用半监督学习的策略,因此性能更胜一筹.此外,对任一企业,采用全局方法进行训练得到的模型和仅在本地开展训练得到模型相比,在性能上有显著提升,这也从客观上推动了各企业破除“数据孤岛”开展合作.



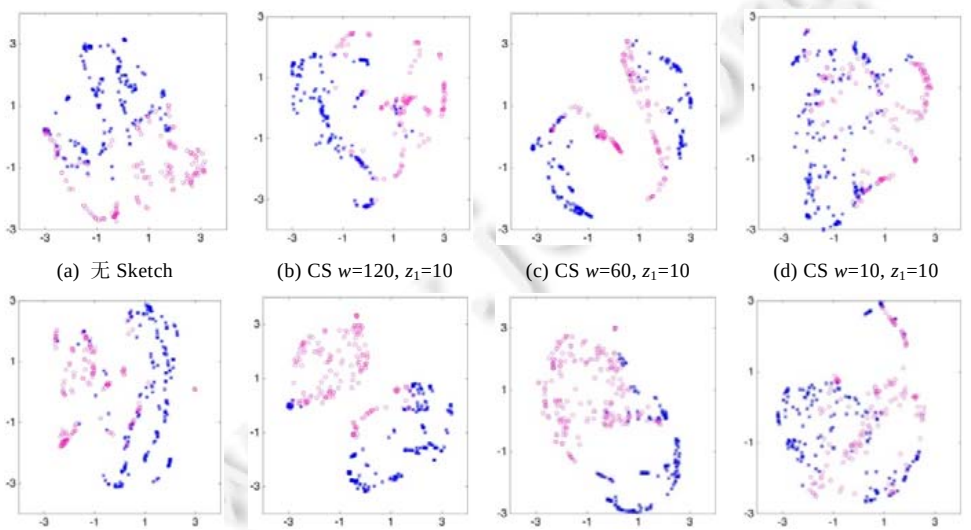
(a) 企业 *A* (b) 企业 *B* (c) 企业 *C* (d) 企业 *D*

Fig.6 Comparison among different solutions on each data silo

图 6 各企业数据孤岛中不同方法的性能比较

5.2.2 Sketch 的影响

我们评估了不同的 Sketch 策略:使用 CountSketch(CS)和 Count-MinSketch(CMS)两种 Sketch 类型,使用不同的哈希空间 w 和使用不同数量的哈希函数 z_1,实验结果如图 7 所示.



(a) 无 Sketch (b) CS w=120, z_1=10 (c) CS w=60, z_1=10 (d) CS w=10, z_1=10

(e) CMS w=10, z_1=120 (f) CS w=120, z_1=5 (g) CS w=120, z_1=3 (h) CS w=120, z_1=1

Fig.7 Evaluation of different sketch strategies

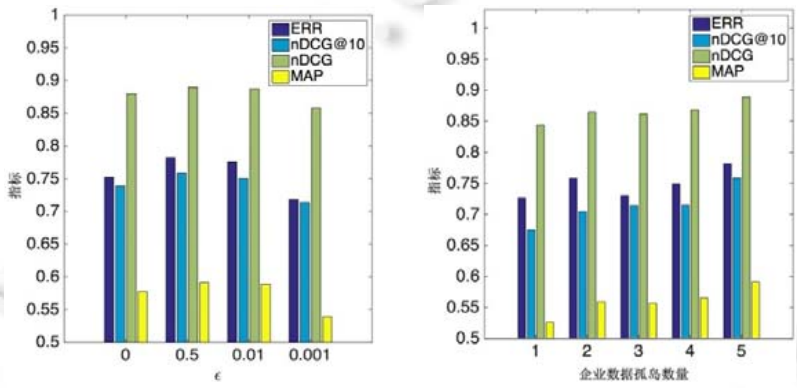
图 7 不同 Sketch 构造策略评估

从全局随机抽取 400 个正负样本(相关度分数大于 0 为正样本,否则为负样本),构造不同的 Sketch 来评估影

响.为了能更清楚地展示实验结果,通过 TSNE 把样本点嵌入到二维平面中.图 7(a)展现了没有 Sketch 的情况,图 7(b)对应 CS-F-LTR 中的 Sketch 构造方法.可以看到:在使用 Sketch 以后,不同类别之间的分界线依然是可以辨认的.我们也尝试使用相同的参数构造了 CM Sketch,结果如图 7(e).随着哈希空间 w 的减小,噪声在增加(图 7(c)和图 7(d)),这说明准确度关于哈希空间 w 十分敏感.哈希空间越小,越多的单词会彼此碰撞,词频和其他特征的计算结果就会越不准确.然而,当哈希函数的数量 z_1 减少时,结果会更加鲁棒.即使当 z_1 为 5(如图 7(f)所示)或者 3(如图 7(g)所示)时,依然有一个明显的分界.当 z 减小到 1 时,分界才开始模糊(如图 7(h)所示).这说明:当 z 确定时,一个较小的 z_1 依然可以保证较为准确的特征,同时保持较好的隐私保护.

5.2.3 隐私预算的影响

隐私预算的影响如图 8(a)所示,我们惊讶地发现:将少量噪声(ε为 0.5)注入到样本数据中,会带来更好的效果.可能的原因是噪声值增加到了无标签数据上,并且他们的伪标签可能不正确,这种情况下,增加噪声可以避免模型过拟合,从而提高泛化性能.随着噪声进一步增加,模型表现开始变差,但依然是可控的.半监督训练过程就像是一个微调过程,而带标签数据则为微调方向提供了一个很好的指引.



(a) 隐私预算影响 (b) 企业数据孤岛数量的影响

Fig.8 Impact of privacy budgetand number of datasiloson CS-F-LTR

图 8 隐私预算和企业数据孤岛数量对 CS-F-LTR 算法的影响

5.2.4 企业数据孤岛数量的影响

企业数据孤岛数量的影响如图 8(b)所示.可以看到:几乎所有的指标都在随着企业数量的增加而增长.*nDCG*@10 和 MAP 随着企业数量从 1 增长到 5 有着 8%的增长,这说明了 CS-F-LTR 的有效性.而 *n*DCG 的增长并不显著,这是因为每一条查询在测试集中有太多不相关的文档.*ERR* 的数值先降低后增加,最终增加了 5%.可能的原因是:当企业数量小的时候,各企业中数据分布差异大,非独立同分布特点明显,从而影响了某些指标.

5.2.5 实验总结

从以上结果可以看出,CS-F-LTR 方法和本地训练相比是更加有效的.更多的参与方可以提高模型的性能.同时,该方法关于隐私相关的参数也是鲁棒的,这说明它能够更好地平衡隐私保护的程度和模型的性能.然而,CS-F-LTR 的有效性也取决于各企业的数据分布情况.如果数据分布差异过大,增加更多的企业可能并不总是对提高性能有效,这也是符合常理的.

6 总结与展望

本文研究了在面向企业数据孤岛联邦场景下的排序学习问题.首先.针对这一特定问题,总结出了一种不同于现有横向或纵向联邦场景的新型面向企业数据孤岛的联邦场景——交叉分割联邦场景,并提出了联邦框架 CS-F-LTR,该框架可以帮助那些本地数据不足的企业间合作构建专有的文档检索系统.我们主要解决了该问题中的两大挑战:一是跨企业数据孤岛的交叉特征生成问题,即隐私保护限制造成了大量交叉特征数据无法被生

成和利用.为此,我们提出了基于 Sketch 与差分隐私的解决方法.可以在实现隐私保护满足差分隐私要求的前提下,实现误差可控的词频、逆文档频率等交叉特征的计算.二是生成数据标签缺失的问题,即用于交叉特征生成的查询和文档因分属不同企业无法直接评估相关性.我们使用半监督协同学习的方法解决了该问题.最终,在公开数据集上的实验结果验证了我们提出方法的有效性.

与此同时,联邦排序学习的未来研究仍存在诸多挑战.首先是数据非独立同分布的问题.各企业数据分布可能存在不均衡性,对模型性能会产生影响.不同企业所持有的数据不一定满足独立同分布的假设,这可能导致不同企业本地所训练的模型泛化性能降低,从而影响全局模型性能.其次是排序模型的拓展性问题.本文中,排序模型使用了经典的点排序模型;而对于基于文档对和文档列表的排序方法,因其标签和特征更加复杂,该框架中的特征生成和半监督学习方法需要进一步设计.最后是传输可靠性问题.虽然本文所提框架主要是针对企业间的联邦场景,但依然可能存在网络时延、掉线等因素破坏模型的训练进程,提高鲁棒性也是值得研究的问题.

目前,数据非独立同分布问题和传输可靠性问题是联邦学习领域的热点问题,已经有诸多研究可以参考.而针对排序模型的拓展性问题,则应该结合训练所需的样本结构设计新的协作方式和传输规划.未来,克服上述挑战将有助于彻底打通企业数据孤岛,并促成联邦排序学习的真实应用落地.

References:

[1] Burges CJC, Shaked T, Renshaw E, *et al*. Learning to rank using gradient descent. In: Proc. of the 22th Int'l Conf. on Machine Learning. 2005. 89−96.

[2] Chapelle O, Chang Y. Yahoo! Learning to rank challenge overview. In: Proc. of the 28th Int'l Conf. on Machine Learning. 2011. 1−24.

[3] Liu TY. Learning to Rank for Information Retrieval. Springer-Verlag, 2011.

[4] Yang Q, Liu Y, Chen TJ, *et al*. Federated machine learning: Concept and applications. ACM Trans. on Intelligent Systems and Technology, 2019,10(2):12:1−12:19.

[5] Chang EY, Zhu KH, Wang H, *et al*. Parallelizing support vector machines on distributed computers. In: Proc. of the 21st Neural Information Processing Systems. 2007. 257−264.

[6] Yang Q, Liu Y, Chen TJ, *et al*. Federated learning. Communications of the CCF, 2018,14(11):49−55 (in Chinese with English abstract).

[7] Yin DW, Hu YN, Tang JL, *et al*. Ranking relevance in Yahoo search. In: Proc. of the 22nd ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2016. 323−332.

[8] Li C, Shen DR, Kou Y, *et al*. Diversity-aware *k*NN query processing approaches for temporal spatial textual content. Pattern Recognition and Artificial Intelligence, 2017,30(1):64−72 (in Chinese with English abstract).

[9] Hou YX, Duan L, Li L, *et al*. Search of genes with similar phenotype based on disease information network. Ruan Jian Xue Bao/Journal of Software, 2018,29(3):721−733 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5445.htm [doi: 10.13328/j.cnki.jos.005445]

[10] Barbaro M, Zeller T. A face is exposed for AOLsearcherno. 4417749. New York Times, 2006.

[11] Chor B, Goldreich O, Kushilevitz E, *et al*. Private information retrieval. In: Proc. of the 36th IEEE Symp. on Foundations of Computer Science. 1995. 41−50.

[12] Liu YM, Zhou HF, Wang ZH, *et al*. Purpose fusion: The risk purpose basedprivacy-aware data access control. Chinese Journal of Computers, 2010,33(08):1339−1348 (in Chinese with English abstract).

[13] Pang HH, Ding XH, Xiao XK. Embellishing text search queries to protect user privacy. Proc. of the VLDB Endowmen, 2010,3(1): 598−607.

[14] Rebollo-Monedero D, Forné J. Optimized query forgery for private information retrieval. IEEE Tans. on Information Theory, 2010, 56(9):4631−4642.

[15] Murugesan M, Clifton C. Providing privacy through plausibly deniable search. In: Proc. of the SIAM Int'l Conf. on Data Mining. 2009. 768−779.

[16] Gaboardi M, Arias EJG, Hsu J, *et al*. Dual query: Practical private query release for high dimensional data. In: Proc. of the 31th Int'l Conf. on Machine Learning. 2011. 1170−1178.

[17] Gertner Y, Ishai Y, Kushilevitz E, *et al*. Protecting data privacy in private information retrieval schemes. Journal of Computer and System Sciences, 2000,60(3):592−629.

[18] Freedman MJ, Ishai Y, Pinkas B, *et al*. Keyword search and oblivious pseudorandom functions. In: Proc. of the Theory of Cryptography Conf. 2005. 303−324.

[19] Weng L, Amsaleg L, Morton A, *et al*. A privacy-preserving framework for large-scale content-based information retrieval. IEEE Trans. on Information Forensics and Security, 2015,10(1):152−167.

[20] Curtmola R, Garay JA, Kamara S, *et al*. Searchable symmetric encryption: Improved definitions and efficient constructions. Journal of Computer Security, 2011,19(5):895−934.

[21] Agrawal R, Kiernan J, Srikant R, *et al*. Order-Preserving encryption for numeric data. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 2004. 563−574.

[22] Zhang W, Lin YP, Xiao S, *et al*. Privacy preserving ranked multi-keyword search for multiple data owners in cloud computing. IEEE Trans. on Computers, 2016,65(5):1566−1577.

[23] Ji SY, Shao JJ, Agun D, *et al*. Privacy-Aware ranking with tree ensembles on the cloud. In: Proc. of ACM SIGIR Conf. on Research and Development in Information Retrieval. 2018. 315−324.

[24] Konecný J, McMahan HB, Yu FX, *et al*. Federated learning: Strategies for improving communication efficiency. CoRR, 2016, abs/1610.05492.

[25] Bonawitz K, Ivanov V, Kreuter B, *et al*. Practical secure aggregation for privacy-preserving machine learning. In: Proc. of ACM Conf. on Computer and Communications Security. 2017. 1175−1191.

[26] McMahan HB, Ramage D, Talwar K, *et al*. Learning differentially private recurrent language models. In: Proc. of Int'lConf. on Learning Representations. 2018.

[27] Jiang D, Song YF, Tong YX, *et al*. Federated topic modeling. In: Proc. of the 28th ACM Int'l Conf. on Information and Knowledge Management. 2019. 1071−1080.

[28] Wang YS, Tong YX, Shi DY. Federated latent dirichletallocation: A local differential privacy basedframework. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. 2020. 6283−6290.

[29] McMahan B, Moore E, Ramage D, *et al*. Communication-Efficient learning of deep networks from decentralized data. In: Proc. of the 20thInt'l Conf. on Artificial Intelligence and Statistics. 2017. 1273−1282.

[30] Song TS, Tong YX, Wei SY. Profit allocation for federated learning. In: Proc. of the IEEE Int'l Conf. on Big Data. 2019. 2577−2586.

[31] Smith V, Chiang CK, Sanjabi M, *et al*. Federated multi-task learning. In: Proc. of the 31st Neural Information Processing Systems. 2017. 4424−4434.

[32] Zhao Y, Li M, Lai L, *et al*. Federated learning with non-iiddata. arXiv preprint arXiv, 2018, 1806.00582.

[33] Pan RS, Han DM, Pan JC, *et al*. Visualization for federated learning: Chanllenges and framework. Journal of Computer-Aided Design & Computer Graphics, 2020,32(4):513−519 (in Chinese with English abstract).

[34] Kairouz P, McMahan HB, Avent B, *et al*. Advances and open problems in federated learning. CoRR, 2019, abs/1912.04977.

[35] Kharitonov E. Federated online learning to rank with evolutionstrategies. In: Proc of the Int'l Conf. on Web Search and Data Mining. 2019. 249−257.

[36] Amini MR, Truong TV, Goutte C. A boosting algorithm for learning bipartite ranking functions with partially labeled data. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval. 2008. 99−106.

[37] Duh K, Kirchhoff K. Learning to rank with partially-labeled data. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval. 2008. 251−258.

[38] Zhang X, He B, Luo TJ, *et al*. Performance analysis of clustering-based transductivelearning. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2865−2876 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4726.htm [doi: 10.13328/j.cnki. jos.004726]

[39] Laine S, Aila T. Temporal ensembling for semi-supervised learning. In: Proc. of the Int'l Conf. on Learning Representations. 2017.

[40] Tarvainen A, Valpola H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: Proc. of the 31st Neural Information Processing Systems. 2017. 1195−1204.

[41] Muthukrishnan S. Data streams: Algorithms and applications. Foundations and Trends in Theoretical Computer Science, 2005.

[42] Charikar M, Chen KC, Farach-Colton M. Findingfrequent items in data streams. Theoretical Computer Science, 2004,312(1):3−15.

[43] Cormode G, Muthukrishnan S. An improved data stream summary: The count-min sketch and its applications. Journal of Algorithms, 2005,55(1):58−75.

[44] Jiang JW, Fu FC, Yang T, *et al*. SketchML: Accelerating distributed machine learning with data sketches. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 2018. 1269−1284.

[45] Zhu WN, Kairouz P, Sun HC, *et al*. Federated heavy-hitters discovery with differential privacy. CoRR, 2019, abs/ 1902.08534

[46] Dwork C. Differential privacy. In: Proc. of the 33rd Int'l Colloquium on Automata, Languages and Programming. 2006. 1−12.

[47] Liang WJ, Chen H, Wu YC, *et al*. Differential privacy under continual observation. Ruan Jian Xue Bao/Journal of Software, 2020,31(6):1761−1785 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6042.htm [doi: 10.13328/j.cnki.jos.006042]

[48] Li T, Liu ZX, Sekar V, *et al*. Privacy for free: Communication-efficient learning with differential privacy using sketches. CoRR, 2019, abs/1911.00972.

[49] Robertson SE. Overview of the Okapi projects. Journal of Documentation, 1997,53(1):3−7.

[50] Ponte JM, Croft WB. A language modeling approach to information retrieval. In: Proc. of the ACM SIGIR Conf. on Research and Development in Information Retrieval. 1998. 275−281.

[51] Qin T, Liu TY. Introducing LETOR 4.0 datasets. CoRR, 2013, abs/1306.2597.

[52] Cormode G, Muthukrishnan S. Summarizing and mining skewed data streams. In: Proc. of the SIAM Int'l Conf. on Data Mining. 2005. 44−55.

附中文参考文献:

[6] 杨强,刘洋,陈天健,等.联邦学习.中国计算机学会通讯,2018,14(11):49−55.

[8] 李晨,申德荣,寇月,等.多样性感知的时空文本信息的 *K*NN 查询处理方法.模式识别与人工智能,2017,30(01):64−72.

[9] 侯泳旭,段磊,李岭,等.基于疾病信息网络的表型相似基因搜索.软件学报,2018,29(3):721−733. http://www.jos.org.cn/1000-9825/5445.htm [doi: 10.13328/j.cnki.jos.005445]

[12] 刘逸敏,周浩峰,王智慧,等.Purpose 融合:基于风险 purpose 的隐私查询访问控制.计算机学报,2010,33(8):1339−1348.

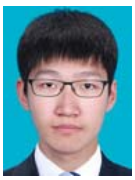
[33] 潘如晟,韩东明,潘嘉铖,等.联邦学习可视化:挑战与框架.计算机辅助设计与图形学学报,2020,32(4): 513−519.

[38] 张新,何苯,罗铁坚,等.基于聚类的直推式学习的性能分析.软件学报,2014,25(12):2865−2876. http://www.jos.org.cn/1000-9825/4726.htm [doi: 10.13328/j.cnki.jos.004726]

[47] 梁文娟,陈红,吴云乘,等.持续监控下差分隐私保护.软件学报,2020,31(6):1761−1785. http://www.jos.org.cn/1000-9825/6042.htm [doi: 10.13328/j.cnki.jos.006042]

史鼎元(1998－),男,本科,主要研究领域为联邦学习,时空大数据分析处理,众包计算,群体智能,隐私保护.

郑鹏飞(1996－),男,硕士,CCF 学生会员,主要研究领域为联邦学习,时空大数据分析处理,众包计算,群体智能,隐私保护.



王晏晟(1994－),男,博士,CCF 学生会员,主要研究领域为联邦学习,时空大数据分析处理,众包计算,群体智能,隐私保护.

童咏昕(1982－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为联邦学习,时空大数据分析处理,众包计算,群体智能,隐私保护.