文章编号：2095-2163(2020)02-0236-04 中图分类号：TP309.2 文献标志码：A

大数据背景下在线学习用户的隐私保护研究

张建珍[1]，牛煜艳[2]，李 强[1]
（1 山西机电职业技术学院 信息工程系，山西 长治 046000；
2 中国农业银行股份有限公司长治市分行 办公室，山西 长治 046000）

摘 要：大数据时代，数据挖掘与数据分析技术为改善在线学习用户的学习体验，预测学习效果，推荐学习课程发挥了重要作用，但用户在线学习过程中的信息安全和隐私保护问题逐渐引起了信息技术人员和教育界人士的关注。通过对用户网络隐私界定，数据足迹形成，分析现有大数据技术背景下隐私保护策略，提出在线学习用户隐私保护模型，从平台及用户两方面提出建议，为大数据背景下在线学习用户的隐私保护提供一种思路。

关键词：大数据；在线学习；用户隐私；数据分析；信息挖掘；信息安全；隐私保护

Research on privacy protection of online learning users in the context of big data

ZHANG Jianzhen[1], NIU Yuyang[2], LI Qiang[1]
(1 Department of Information Engineering, Shanxi Institute of Mechanical & Electrical Engineering, Changzhi Shanxi 046000, China; 2 Agricultural Bank of China Changzhi Branch Office, Changzhi Shanxi 046000, China)

【Abstract】 In the era of big data, data mining and data analysis technology play an important role in improving the learning experience of online learning users, predicting the learning outcomes and recommending learning courses. But the problem of information security and privacy protection in the online learning process has gradually attracted the attention of information technology personnel and educators. By defining users' network privacy and forming data footprint, this paper analyzes the privacy protection strategy under the background of existing big data technology, proposes the privacy protection model of online learning users, and gives out suggestions in the views of both the platform and users, so as to provide an idea for the privacy protection of online learning users under the background of big data.

【Key words】 big data; online learning; user privacy; data analysis; information mining; information security; privacy protection

0 引 言

随着大数据技术发展，用户在互联网上留下的数据足迹中蕴含的巨大价值不断被发掘，如个人在购物网站上某商品页面停留时长及浏览商品种类可能成为购物平台下次推送特定种类商品广告的依据；个人在某新闻网站对某类新闻的点击及浏览时间也成为了新闻网站筛选并推送特定新闻的依据。信息技术大发展也促进了教育信息化，并为有学习意愿的人提供了更多的选择以提升职业能力，充实兴趣爱好。自 2012 年盛行的 MOOC 成为在线学习的主要模式。借助大数据技术，MOOC 平台为教育机构改善教学设计、支持教育决策，完善课程建设、开展教育科学研究等提供了依据，同时也为用户进行课程推荐、学习效果预测提供依据。

在现阶段，利用大数据技术[1]进行以提升产品功能、服务质量等为目的的研究中，不仅通过数据获得了个人真实的行为习惯信息，越来越多的个人敏感信息也被发现。

电商平台、社交平台、医疗及金融平台由于涉及商业利益，人们对用户隐私的警惕性普遍较高，因此，以上平台大数据的用户隐私保护问题即已成为平台本身、同时也是用户关注的焦点。但就在线学习而言，大多均以知识和技能提升为目的，关于用户隐私信息较少引起关注，进而在线学习用户在大数据背景下可能涉及的隐私泄漏及隐私保护还没有受到足够重视。本文研究以下 3 个问题：

Q1：在线学习用户网络隐私。

Q2：在线学习用户的网络足迹。

Q3：在线学习用户隐私保护策略。

1 研究现状

国内的研究主要立足于对 CNKI 数据库的挖掘分析。考虑到 2012 年 MOOC 在中国开始呈现的普及态势，因此，选取了 2013～2018 年的时间范围发表的文献，并在结果中检索同时包含“大数据”、“隐私保护”两个关键词，而据 2018 年 12 月 26 日中国知网检索结果可知，一共获得 353 篇有效文献。大

基金项目：山西机电职业技术学院院课题(JKY-17027)研究成果之一。

作者简介：张建珍(1980-)，女，硕士，副教授，主要研究方向：网络安全。

收稿日期：2019-09-18

多数文献都比较新，被引频次仍在逐步提高。如图 1 所示，除 2017 年出现小幅度下降外，整体呈现明显上升趋势。以“大数据”、“隐私保护”关键词基础上，添加“教育”关键词，仅检索到 2 篇文献，以“大数据”+“隐私保护”+“在线学习/网络学习”关键词则没有检索到任何有效文献。

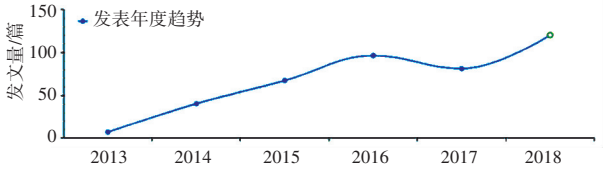

图 1　2013~2018 年中国知网收录相关论文情况

Fig. 1　CNKI included related papers during 2013~2018

通过文献研究发现，大数据技术以“获得知识与推测趋势”、“分析掌握个性化特征”、“辨识真相”为主要目标。社交、医疗、工作、学习、娱乐行业活动生成的大数据，经过机器学习、数据挖掘、回归分析等技术处理可以形成个人特征信息，刘雅辉等人[2]提出对个人信息采用分类分级保护技术，将个人信息分为 4 类，可以直接识别出特定个人的个人身份信息、与个人生活紧密相关的准标识符信息、通过某些信息可以关联得出的敏感信息、泄露可能导致风险的日志信息，同时还提出应划分企业和个人在隐私保护中的职责。

数据收集、分析、发布过程均存在隐私泄漏的安全隐患[1]。如根据用户签到行为或社交网络上下文推演用户兴趣点[3-4]，冯登国等人[5]提出通过修正数据精度，数据匿名发布、数据水印人工加扰等方法来保护隐私数据；最佳的隐私保护方案是加密所有数据。曹珍富等人[6]提出通过减少公钥加密使用次数来设计高效隐私保护外包的密文计算方法，并提出支持大属性集合的短密文高效可追踪可撤销属性基加密方案以控制密文访问。在数据分析和数据发布过程中涉及的隐私保护问题上，袁健等人[7]提出减少数据关联的冗余信息，通过自适应加噪技术为差分隐私保护生成合适数量噪声的方法。

教育大数据中用户隐私的保护尚未得到广泛的研究，就目前检索到的 2 篇文献来看，主要集中在用户知情权方面研究，如赵慧琼等人[1]基于教育信息化的数据收集过程中用户知情同意原则、数据分析过程中匿名原则、数据解释的公正原则提出大数据学习分析的安全与保护框架。周孟等人[8]以学生为对象，从隐私安全基本需求、隐私等级设置、隐私安全风险等 3 个维度研究教育隐私保护，讨论了教育数据采集中学生的知情权、所有权、选择权和控制权。

国外与本文研究主题较为相关的是 Jones 等人[9]在《Users or Students? Privacy in University MOOCS》进行了大学 MOOCs 中用户隐私的研究。研究中依据美国家庭教育权利和法案（The Family Educational Rights and Privacy Act，FERPA）对 MOOCs 平台可能涉及学习用户隐私信息进行分析，并对比了 Coursera、Blackboard CourseSites、EDX 三大平台用户数据收集政策，指出随着越来越多大学参与并陆续建设了自己的 MOOC，但没有具体指出在线学习用户隐私保护具体措施及可执行方案。

综上所述，大数据背景下对隐私保护的研究，主要以从技术角度对数据加密算法研究和从法律角度对隐私保护立法角度研究为主，对在线学习过程中可能存在的用户隐私泄漏及保护技术尚未得到技术及学者们的重视。

2　在线学习用户的网络隐私

2.1　隐私及网络隐私

隐私内涵根据社会、文化、技术背景不同而不同[10]，利益诉求也是影响隐私判别的重要因素。

随着网络的普及，出现了网络隐私，即个人隐私在网络中的延伸，自然人在网上的私人信息、私人空间和私人活动应当受到保护，不得随意搜集、复制、转载、下载、传播所知晓的他人隐私。

欧盟 1995 年 10 月通过《个人数据保护指令》，要求欧盟各国根据该指令调整制定本国的个人数据保护法。2013 年 11 月 26 日，联合国通过由巴西、德国发起的保护网络隐私权决议。中国有学者基于网络隐私提出了数据权的概念，数据权包括数据管理权、数据控制权等[11]，另外郭兵等人[12]基于银行个人货币资产管理模式及架构提出个人大数据资产管理。

2.2　在线学习用户的网络隐私

在线学习用户网络隐私包括广义网络隐私和狭义网络隐私。注册在线学习账户时使用的个人登录身份、邮箱地址、教育背景等属于广义网络隐私；而由在线学习用户的学习行为，经大数据技术分析生成的学习兴趣、常用登录地点、固定学习时间、发言讨论特点等属于狭义网络隐私。

3　在线学习用户的网络足迹

3.1　在线学习平台对用户数据收集情况

本文以国内较为流行的在线学习平台，如学堂在线（http://www.xuetangx.com/）、中国大学 MOOC（https://www.icourse163.org/）、华文慕课（http://

www.chinesemooc.org/),国外较早起步的三大平台,如 Coursera(https://www.coursera.org/)、Futurelearn(https://www.futurelearn.com/)、Edx(https://www.edx.org/),分析在线学习平台可能涉及对用户隐私信息收集的统计见表 1。

表 1 著名平台对用户信息收集情况统计

Tab. 1 Statistics of user information colleced by famous platforms

信息	学堂在线	中国大学 MOOC	华文慕课	Coursera	Futurelearn	Edx
电子邮箱	√	√	√	√	√	√
出生日期	√	√		√	√	√
所在地区	√	√	√	√	√	
教育背景	√	√	√	√	√	√
职业或职业目标	√	√		√	√	
兴趣爱好	√			√	√	

由表 1 可知,六大平台均支持邮箱注册,可以得出在线学习平台对用户电子邮箱的收集是必然的。作为个人信息的完善,大部分平台要求或希望用户完善年龄、教育背景、所在地区,职业目标或个人简介。随着在线学习平台商业化运营的走行态势,对注册用户信息的收集日渐精细,如中国大学 MOOC 平台,不仅收集用户邮箱、手机号,还收集用户身份证号。华文慕课是公益性开放共享慕课平台,以运用网络信息技术促进华文高等教育为使命,以为有学习和提升愿望的在校生、社会生、大学教师、大专院校提供学习机会,并不注重用户信息收集。

3.2 对比在线学习平台对个人数据的管理权限

在线学习用户在六大平台上对个人数据的管理权限见表 2。由表 2 可知,六大平台均支持用户修改个人信息和对数据管理的问题联系。中国大学 MOOC、Coursera、Futurelearn 和 Edx 四个平台提供删除个人信息的服务,但需要专门与平台联系,用户不可以自主删除自己的注册信息或学习记录。学堂在线、中国大学 MOOC、Coursera 和 Edx 四个平台为用户提供查看自身学习记录的监视数据功能。

表 2 在线学习用户在著名平台上对个人数据的管理权限对比

Tab. 2 Comparison of online learning users´ management rights of personal data on famous platforms

信息	学堂在线	中国大学 MOOC	华文慕课	Coursera	Futurelearn	Edx
修改个人信息	√	√	√	√	√	√
删除个人信息		√		√	√	√
问题联系	√	√	√	√	√	√
监视数据	√	√		√	√	
隐私政策	√	√		√	√	√

3.3 隐私政策

在线学习平台用户注册时,一般会有隐私政策告知用户平台收集哪些信息以及如何使用这些信息,如个人信息、学习表现、学习模式、上网 IP、以及使用、披露、分享用户信息的目的,但是,对于可能存在的用户隐私泄露并没有做出明确责任划分。

以上六个平台除华文慕课纯公益性质、未提供隐私政策外,其余五个平台均提供了隐私政策或服务条款,如可能收集的信息、如何收集和使用信息、可能分享、转让和披露的信息、如何保留、储存和保护信息、如何管理用户的信息。

4 在线学习用户隐私保护策略

用户搜索和浏览习惯、学习行为特点、年龄及教育背景、位置信息是为用户提供个性化服务以及进行营销推广的大数据分析的基础,也是进行隐私保护、防止第三方信息窃取的关键,如何平衡二者,是一个博弈[13]过程。

4.1 在线学习用户隐私保护模型

依据赵慧琼等人[1]的研究将大数据学习过程划分为数据收集、数据分析、数据解释等 3 个阶段,研究提出在线学习用户隐私保护应贯穿大数据学习的三个阶段,并且分别从用户和平台两方面加强隐私保护,保护模型如图 2 所示。

4.2 平台方面

在线学习用户一旦注册成为某平台用户后,接下来在该平台的一切学习行为均成为分析依据。平台通过特定分析模型进行用户行为大数据分析,探究学习者的学习过程与情境,总结其学习规律,进而

根据学习者特征及平台商业利益为学习者提供个性化自适应学习意见[14]。因此，如果平台基于用户数据与第三方开展合作研究时，存在用户数据漏洞问题。付玉香等人[15]提出基于迁移学习的多源数据隐私保护方法研究，不失为平台用户数据保护的优秀方法，就是先在本地使用 PATE-T 模型对隐私数据训练分类器，接着集合多方分类器，建立一个准确具有差分隐私的全局分类器，达到在不共享本地私有数据的情况下共享双方数据开展合作研究。

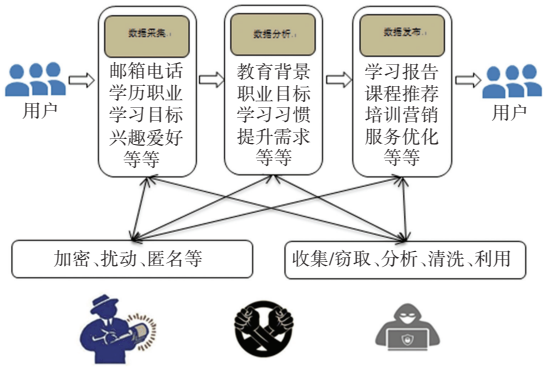

图 2　在线学习用户隐私保护模型

Fig. 2　Protection model of online learning users´ privacy

4.3　用户方面

任何隐私数据均来源于用户自身有意无意的提供。要想控制在线学习用户的隐私漏洞，最有效的方法在于用户对自身隐私的高度敏感。一方面，向平台提供信息时，能保持谨慎态度，非必要情形不予提供，或有选择性提供个人信息；另一方面，对强制要求提供的信息或带有倾向性推荐意见，能保持警惕，防止落入网络陷阱。

5　结束语

本文通过研究文献，提出在线学习用户的网络隐私范围；通过分析国内外较著名的六大在线学习平台，研究了在线学习用户的网络足迹，提出在线学习用户隐私保护应贯穿大数据学习的 3 个阶段，并且分别从用户和平台两方面加强隐私保护。大数据分析技术是把双刃剑，在为平台提供决策依据，为用户提供私人定制同时，也可以被用来挖掘个人隐私，从而导致在线学习用户的隐私泄露。因此，要想将在线学习这一现代学习模式效益最大化，平台必须重视注册用户信息保护，个人也必须谨慎对待一切要求提交的个人信息。

参考文献

[1] 赵慧琼，姜强，赵蔚. 大数据学习分析的安全与隐私保护研究[J]. 现代教育技术，2016，26(3)：5.
[2] 刘雅辉，张铁赢，靳小龙，等. 大数据时代的个人隐私保护[J]. 计算机研究与发展，2015，52(1)：229.
[3] 任星怡，宋美娜，宋俊德. 基于用户签到行为的兴趣点推荐[J]. 计算机学报，2017，40(1)：28.
[4] 任星怡，宋美娜，宋俊德. 基于位置社交网络的上下文感知的兴趣点推荐[J]. 计算机学报，2017,40(4)：824.
[5] 冯登国，张敏，李昊. 大数据安全与隐私保护[J]. 计算机学报，2014，37(1)：246.
[6] 曹珍富，董晓蕾，周俊，等. 大数据安全与隐私保护研究进展[J]. 计算机研究与发展，2016，53(10)：2137.
[7] 袁健，王迪，申泽宇. 大数据环境中交互式查询差分隐私保护模型[J]. 计算机应用研究，2019，36(6)：1782.
[8] 周孟，段智宸，上超望. 大数据时代教育隐私保护三重维度研究[J]. 广西广播电视大学学报，2016，27(3)：25.
[9] JONES M L，REGNER L. Users or students? Privacy in university MOOCS[J]. Science and Engineering Ethics，2016，22(8)：1473.
[10] 刘凌，罗戎. 大数据视角下政府数据开放与个人隐私保护研究[J]. 情报科学，2017，35(2)：112.
[11] 齐爱民，盘佳. 数据权、数据主权的确立与大数据保护的基本原则[J]. 苏州大学学报(哲学社会科学版)，2015，39(1)：64.
[12] 郭兵，李强，段旭良，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[J]. 计算机学报，2017，40(1)：126.
[13] 张伊璇，何泾沙，赵斌，等. 一个基于博弈理论的隐私保护模型[J]. 计算机学报，2016，39(3)：615.
[14] 姜强，赵蔚，王朋娇，等. 基于大数据的个性化自适应在线学习分析模型及实现[J]. 中国电化教育，2015，353(1)：85.
[15] 付玉香，秦永彬，申国伟. 基于迁移学习的多源数据隐私保护方法研究[J]. 计算机工程与科学，2019，41(4)：641.

(上接第 235 页)

[8] DONAHUE J，HENDRICKS L A，GUADARRAMA S，et al. Long-term recurrent convolutional networks for visual recognition and description[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston：IEEE，2015：2625.
[9] 朱煜，赵江坤，王逸宁，等. 基于深度学习的人体行为识别算法综述[J]. 自动化学报，2016，42(6)：848.
[10] 马钰锡，谭励，董旭，等. 面向智能监控的行为识别[J]. 中国图象图形学报，2019，24(2)：282.
[11] 卫星，乐越，韩江洪，等. 基于长短期记忆的车辆行为动态识别网络[J]. 计算机应用，2019，39(7)：1894.
[12] REDMON J，FARHADI A. Yolov3：An incremental improvement[J]. arXiv preprint arXiv：1804.02767，2018.
[13] WELCH G，BISHOP G. An introduction to the Kalman filter[M]. Chapel Hill：University of North Carolina，2001.