Advances in Social Sciences 社会科学前沿, 2020, 9(10), 1541-1547
Published Online October 2020 in Hans. http://www.hanspub.org/journal/ass
https://doi.org/10.12677/ass.2020.910214

Hans汉斯

大数据背景下隐私问题研究现状及热点分析

——基于84篇期刊论文的统计与内容分析

何　宁，陈永进，杨大容

重庆大学公共管理学院，重庆
Email: hening@cqu.edu.cn

收稿日期：2020年9月22日；录用日期：2020年10月8日；发布日期：2020年10月15日

摘　要

大数据不仅在信息技术行业备受瞩目，更成为变革科研、商业、政府运作方式乃至人类思维的一个热点，但同时也使得隐私保护受到更大挑战。本文基于CNKI平台，对2013~2018年间发表的关于我国大数据背景下隐私问题相关研究的84篇论文进行梳理，并从文献增长与分布特征、论文刊源、发表作者群体、研究主题进行统计与内容分析。结果发现：1) 我国大数据背景下隐私问题相关研究正处于萌芽阶段，总体数量偏少；2) 期刊分布较为分散，缺少核心区期刊和高质量研究成果；3) 研究群体范围在逐渐扩大，但是新老作者结构不均衡，没有形成稳定的研究队伍，核心影响力较弱；4) 大数据背景下隐私问题现状和对策建议是该研究的热点，具体集中在对不同来源信息隐私侵犯情况的分析讨论以及技术层面和体系建设两方面的应对措施。

关键词

大数据时代，隐私问题，内容分析

Research Status and Hot Issues of Privacy in the Context of Big Data in China

—Statistical and Content Analysis of 84 Papers

Ning He, Yongjin Chen, Darong Yang

School of Public Affairs, Chongqing University, Chongqing
Email: hening@cqu.edu.cn

Received: Sep. 22nd, 2020; accepted: Oct. 8th, 2020; published: Oct. 15th, 2020

Abstract

Big data have not only attracted much attention in the information technology industry, but also

文章引用: 何宁, 陈永进, 杨大容. 大数据背景下隐私问题研究现状及热点分析[J]. 社会科学前沿, 2020, 9(10): 1541-1547. DOI: 10.12677/ass.2020.910214

become a hot spot to change the way of scientific research, business, government operation and even human thinking. However, it also makes privacy more of a challenge. Based on the platform of CNKI, this paper sorts out 84 papers published in 2013~2018 on privacy issues in the context of big data in China; the statistical and content analysis is made from the characteristics of literature growth and distribution, the source of papers, the group of authors, and the research topics. The results show that: 1) the research on privacy issues in the background of big data in China is in the embryonic stage, and the total number is relatively small; 2) the distribution of journals is scattered, a number of core area journals have not been formed, and the research results are lacking; 3) the research community is expanding, however, the structure of new and old authors is not balanced, a stable research team has not been formed, and the core influence is weak; 4) the status quo of privacy issues in the context of big data and suggestions for countermeasures are the focus of this research, which is embodied in the analysis and discussion of information privacy harm from different sources, as well as the technical level and system construction.

Keywords

Big Data Era, Privacy Issues, Content Analysis

1. 引言

2013 年是被称为“大数据元年”的一年，至此大数据时代开启新的浪潮[1]。党的十九大明确指出要推动互联网、大数据、人工智能的发展，政府将加速建设国家数据治理平台、着力在大数据技术上实现“弯道超车”[2]。那么，大数据是什么？2011 年国际数据中心 IDC 在其报告中给出界定：大数据被设计于从大规模多样化的数据中通过高速捕获、发现和分析技术提取数据的价值[3]。陈威霖认为大数据的发展存在两个阶段，收集整理数据并对数据进行挖掘、预测分析研究的 1.0 时代，以及由数据 + 社会科学理论方式驱动的 2.0 时代[4]。在 2.0 时代，网络数据越来越广泛的应用和暴露出来的弊病倒逼人们去研究其与社会发展的相互关系。其中弊病之一便是信息泄露、隐私侵犯等行为越来越频繁，比如数据收集、挖掘将个人信息泄露从而另作他用，数据预测使个体的健康状况、信用偿还能力、犯罪概率等个体隐私数据掌握在拥有大数据的机构手中，而且，无论个体是否有所察觉，更加全面的监控都将使其隐私无所遁形[5]。如何在享受大数据带来的科技红利时保护自己的隐私，成为了生活在大数据时代的人群时刻关心的问题，学术界在该方面的研究也有了一定程度的进展。因此，本文借助 CNKI 平台，统计与大数据背景下隐私问题主题相关的期刊论文，通过对文献分布和主题内容的分析展示 2013~2018 年间该研究的现状及热点情况，以便为之后相关主题研究提供参考。

2. 研究数据和方法

2.1. 数据来源

本文用作分析的文献来源是中国知网(CNKI)数据库。首先，本文设置检索限制为“期刊”，再通过关键词 = “大数据”并含“隐私”准确检索，得到中文文献 85 篇，论文发表时间在 2013~2018 年间[1]。其次，通过文献标题与摘要的阅读，删除访谈、会议记录等非学术论文后得到分析样本 84 篇。最后，使用 Excel 对上述检索的论文进行资料整理和统计分析。

[1]根据检索方式，最早一篇相关文献发表于 2013 年，检索时间截止 2018 年 12 月 14 日。

2.2. 研究方法

本文通过对检索得到的 84 篇论文进行统计和内容分析，探讨我国近六年来“大数据背景下隐私问题”相关研究，具体包括文献数量、发表期刊和作者群体特征，以及大数据背景下隐私问题相关研究主题热点及变化趋势。

3. 文献分布现状

3.1. 数量分析

某一学科或研究领域在一定时期内的发展水平和速度可以用该阶段发表论文的数量和变化轨迹来反映[6]。根据普赖斯文献增长规律，某一学科研究自诞生起会经历以下几个阶段：第一阶段的萌芽时期在文献数量上表现为不稳定增长；第二阶段的大发展时期表现为指数性增长；第三阶段的成熟期呈现增长减缓特点；第四阶段的完善时期文献数量会日趋减少[6]。因此，本文选取“文献累积数”和“文献累积率”[2]两个指标来判断大数据背景下隐私问题相关研究的发展现状。结果如表 1 所示。

Table 1. Annual distribution of privacy issues in the context of big data
表 1. 大数据背景下隐私问题研究文献年度分布

刊文年度	刊文数量	文献累积数	文献累积率
2013	1	1	——
2014	10	11	1000%
2015	21	32	190.90%
2016	23	55	71.88%
2017	15	70	27.27%
2018	14	84	20.00%

从表 1 的“刊文数量”可以看出，2013~2016 年间关于大数据背景下隐私问题的相关研究在 CNKI 中呈现快速增长状态，2016 年显示为峰值 23 篇文献，2017 年以后发文数量开始缓慢下降；通过表 1“文献累积数”、“文献累积率”分析，关于大数据背景下隐私问题的相关研究还处于研究领域的第一阶段，即萌芽阶段。2013~2014 年，文献迅猛增长，这与“大数据元年”开启有极大的关系，其后，文献累积率在 20%~1000%之间波动，显然波动幅度较大，表明我国大数据背景下隐私问题的相关研究在不稳定状态。因此，本文认为大数据背景下隐私问题相关研究还处于萌芽阶段，总体上发文数量远远不够，研究力度亟待提升。

3.2. 期刊分析

文献期刊统计可以分析一项研究的文献空间分布特点，找到该研究领域的核心期刊，为后来研究者提供深入研究依据[6]。本文检索到的 84 篇论文一共涉及 70 种刊物，包括 CSSCI、CSCD 在内的核心及以上期刊 23 种，约占总刊物的 32.86%，总发文数量为 25 篇，约占发表论文总数的 29.76%，表明在检索的 84 篇期刊学术论文中，研究水平参差不齐，相对较少成果性研究。结果如表 2 所示。

Table 2. Distribution of publications of research papers on privacy issues in the context of big data
表 2. 大数据背景下隐私问题研究论文发表刊物层次分布

期刊层次	期刊数量(种)	百分比	发文数量(篇)	百分比
核心及以上期刊	23	32.86%	25	29.76%
其他公开期刊	47	67.14%	59	70.24%
总计	70	100%	84	100%

[2]“文献累积数”是指该年及以前文献的累加数量；“文献累积率”是指该年发文数量占上一年度文献累积数量的百分比。

根据布拉福德定律，发表某一研究领域论文的期刊可以分为核心区、相关区和边缘区三种，本文主要分析其核心区期刊。核心区期刊数量计算公式为 $r_0 = 2\ln(e^E*Y)$[3]，r_0 表示核心区期刊数量，Y 为发文量最多期刊所发表的文献数量[7]。带入本文数值计算得出，$r_0 = 2\ln(e^{0.5772}*4) \approx 4$，即刊文量排在前四位的期刊位于我国大数据背景下隐私问题研究的核心区，即法制与社会、新闻研究导刊、数字技术与应用、科学与社会。这 4 类期刊约占所有期刊数量的 5.71%，一共发文 13 篇，仅占发表论文总数的 15.48%，表明大数据背景下隐私问题研究的核心影响力相对较弱。表 3 列举出检索文献中发文数量 2 篇及以上的所有刊物，一共有 9 种，约占刊物总数的 12.86%，共发文 23 篇，约占发表论文总数的 27.38%，这表明大部分期刊只刊登过一次大数据背景下隐私问题的相关研究，分布较为分散。

Table 3. Publication of two or more journals on privacy issues in the context of big data
表 3. 大数据背景下隐私问题研究发文 2 篇及以上期刊

期刊名称	发文数量(篇)	期刊名称	发文数量(篇)
法制与社会	4	网络安全技术与应用	2
新闻研究导刊	4	数字通信世界	2
数字技术与应用	3	今传媒	2
科学与社会	2	伦理学研究	2
现代工业经济和信息化	2		

3.3. 作者分析

作者群体分析可以反映某一研究领域的研究群体的成熟度和稳定性。本文从作者发文数量、作者增复数量两个指标对我国大数据背景下隐私问题研究的作者群体进行分析。根据洛特卡定律对作者群体的判断，如果发表 n 篇论文的作者数量约为发表 1 篇的作者数量的 $1/n^2$，同时发表 1 篇的作者数量约占所有作者数量的 60%，则该研究领域处于一个成熟时期[6]。按照该定律要求，本文仅统计检索的 84 篇文献的第一作者，从发表 n 篇论文的作者数量占作者总数的比例和发表 n 篇论文的作者数量占发表 1 篇文献的作者数量比例两个方面来描述作者发文数量指标。数据如表 4 所示。

Table 4. Statistical analysis of the number of articles published by authors
表 4. 作者发文数量统计分析

发文数量(篇)	作者人数(人)	占作者总数的比例	占发表1篇文献作者数量比例
3	1	1.23%	1.27%
2	1	1.23%	1.27%
1	79	97.53%	100%
总计	81	100%	

从表 4 可以得到：同一人发表大数据背景下隐私问题相关研究的论文最多的是 3 篇，有且仅有一位作者，约占作者总数的 1.23%；其次有一位作者发表了 2 篇相关的论文，同样约占作者总数的 1.23%；其余作者皆只发表了 1 篇大数据背景下隐私问题相关研究论文，约占作者总数的 97.53%。该数据初步表明在我国大数据背景下隐私问题研究领域，研究群体较为分散，几乎没有核心团队进行持续研究。同时，从表 4 中也可得到：发表 3 篇论文的作者数量和发表 2 篇论文的作者数量都只占发表 1 篇文献的作者数量的 1.27%。按照洛特卡定律，理论上发表 2、3 篇论文的作者应占发表 1 篇论文的作者数量的 1/4 和 1/9，

[3]E 为欧拉系数 0.5772。

但本文得到的数据明显低于该理论值，且只发表 1 篇大数据背景下隐私问题相关研究论文的作者数量约占作者总数的 97.53%，远远高于理论值 60%。进一步说明大数据背景下隐私问题相关研究还没有形成成熟的作者群体，不具备群体稳定性，需要培养可以持续钻研该领域的研究人员。

作者复量和增量指标通过一定时期内该研究领域发表论文作者重复、增加两方面进行描述。在一段时间内，期刊论文作者一定会有第一次在该领域发表论文的新作者，也会有多次在该领域发表论文的老作者。本文对 2013~2018 年在大数据背景下隐私问题研究领域发表期刊论文的作者总量记为 N，老作者记为 A，新作者记为 B，作者复量使用 A/N 表示，是作者多次在该领域发表论文的体现；作者增量使用 B/N 表示，是该研究领域中出现新作者的状况。本文统计的数据如表 5 所示。

Table 5. Statistical analysis of the authors' number of repetitions and additions
表 5. 作者复量、增量统计分析

年份	N	A	B	A/N	B/N
2013	1	0	1	0	1
2014	9	0	9	0	1
2015	21	1	20	0.05	0.95
2016	23	0	23	0	1
2017	15	0	15	0	1
2018	14	1	13	0.07	0.93

从表 5 数据可知：2013~2018 年间大数据背景下隐私问题研究领域的作者复量 A/N 的均值为 0.06，而该领域内的作者增量 B/N 在 0.93~1.00 之间，尤其是 2013、2014、2016 和 2017 年，作者增量 B/N 均为 1。按照洛特卡定律，A/N 的值越趋近于 1，表示该研究领域的作者群体比较稳定，核心作者相对较为集中，同时存在作者构成僵化、缺少后续人才的问题；而 B/N 的值越趋近于 1，表示该研究领域拥有许多新生力量，更新换代频繁，带来的问题是作者群体结构不稳定，缺乏核心人物。因此，本文认为该领域的作者群体比例结构不均衡，尽管该领域获得广泛关注，大量新生作者尝试进入该领域，研究前景光明，但是极小的作者复量也表明此时的研究队伍是极不稳定的，几乎一年更换一批研究学者，从而缺乏对大数据背景下隐私问题进行持续且深入的研究。

4. 内容分析

4.1. 关键词分析

关键词是一篇论文的核心所在，把握关键词的内涵能在很大程度上解读文章从而了解该领域的研究热点及现状，尤其是高频关键词几乎预示着该领域的研究热点。本文通过对能够表达论文核心内容的关键词在大数据背景下隐私问题研究领域的文献中出现的频次高低，来确定该领域的研究热点和发展动向[8]。对论文关键词通过人工甄别进行规范化处理：去除揭示论文内容专指性不强、含义过于宽泛的词，如“应用”、“分析”、“反思”；合并同义、近义关键词，如将“网络”与“互联网”合并为“网络”，“大数据”与“大数据时代”、“网络大数据”合并为“大数据”[9]。在检索到的 84 篇大数据背景下隐私问题相关研究的文献中，共计出现 134 个关键词，累积频次 344 次，平均出现频次约为 2.57 次，选取频次 ≥2 的关键词 16 个。结果见表 6。

Table 6. Keywords with frequency ≥ 2 in research papers on privacy issues in big data context
表 6. 大数据背景下隐私问题研究论文频次 ≥ 2 的关键词

序号	关键词	频次	序号	关键词	频次
1	大数据	84	9	隐私伦理	2
2	隐私	84	10	微信	2
3	安全	17	11	交易	2
4	保护	9	12	数据开放	2
5	伦理	5	13	政府信息	2
6	个人数据	5	14	数据挖掘	2
7	网络	5	15	被遗忘权	2
8	个人信息	3	16	认证	2

表 6 显示，在所检索的文献中，与“大数据”、“隐私”紧密相关的是对各种类型的信息、安全保护和伦理问题的讨论。“安全”在文献中出现了 17 次、“保护”出现了 9 次以及“伦理”出现了 5 次，在一定程度上代表着大数据背景下隐私问题的研究热点，表明我国学者重点关注在享受大数据带来的便利时存在的信息安全隐患、隐私伦理问题，以及如何对此进行安全保护。

4.2. 研究主题分析

研究主题是指某一学科领域研究中所涉及的现象或问题领域[10]。本文将检索的 84 篇大数据背景下隐私问题相关研究文献的有效关键词进行整理，并根据每篇论文对大数据背景下隐私研究所涉及的不同方面对论文进行分类，一篇论文可能存在多个主题分类中。一共划分为大数据背景下隐私问题现状和对策建议两类研究主题，并从关键词中提取具体的二级主题。

Table 7. Statistics of research topics related to privacy issues in the context of big data
表 7. 大数据背景下隐私问题相关研究主题统计

研究主题	二级主题	具体内容
A 问题现状	A1 应用	数据开放(1)、数据挖掘(1)、数据管理(1)、数据交易(1)
	A2 信息	个人数据(5)、个人信息(3)、微信(2)、交易(2)、政府信息(1)、脸书(1)、医疗(1)、社交网络(1)、假新闻(1)、传媒业(1)、“被直播”(1)、基因(1)、社会生活(1)、图书馆(1)、教育(1)
	A3 安全	网络安全(1)、大数据安全(1)、信息安全(1)、计算机安全(1)
	A4 伦理	隐私伦理(2)、科技伦理(1)、网络伦理(1)、道德伦理(1)
B 对策建议	B1 保护	数据保护(1)、法律保护(1)、隐私权保护(2)、个人数据保护(1)、个人信息保护(1)
	B2 治理	伦理治理(1)、信息治理(1)、数据治理(1)、社会治理(1)
	B3 对策	认证(2)、限制(1)、密码技术(1)、传媒技术(1)、风险访问控制(1)、加密(1) 法律途径(2)、信誉机制(1)、征信(1)、规范体系(1)、刑法体系(1)

从表 7 中可以看出，目前学术界对大数据背景下隐私问题相关研究的重点放在隐私问题现状研究和对策建议研究。

在隐私问题现状研究主题下，具体从大数据应用、信息、安全、伦理四方面进行展开描述。大数据时代主要是进行数据开放、数据挖掘、数据管理和数据交易，信息主要来源于个人社交网络、社会生活、医疗、教育、传媒以及政府信息，也是通过这些方面个人隐私权受到侵犯，如微信、脸书、基因、图书馆信息等易被盗用、窃取。因此，大数据时代更要考虑安全和伦理两方面。安全主要是指网络安全、大数据安全、信息

安全以及计算机安全；伦理问题体现在大数据使得隐私伦理、科技伦理、网络伦理、道德伦理受到挑战。

在隐私问题对策建议主题下，学者们主要从保护、治理和对策三方面进行研究。保护是指对个人数据、个人信息等进行数据保护，同时提供法律保护保障，达到隐私权保护的目的。从“治理”角度出发，学者们提出了伦理治理、信息治理、数据治理、社会治理研究方向。最后，对于大数据时代信息易被泄露、隐私易被曝光的问题，提出两方面对策：其一是技术方面，包括身份认证、密码技术、风险访问控制、信息加密、传媒技术等，其二是体系制度方面，建立征信系统、信誉机制、法律途径，形成刑法体系和规范体系。总体上来看，大数据背景下隐私问题相关研究论文在主题方面较为单一，重复性研究相对较多，研究成果较为浅薄，少有针对一个主题深入探索。

5. 结论

本文基于 2013~2018 年间发表在中国学术期刊网络出版总库(CNKI)上关于我国大数据背景下隐私问题相关研究的 84 篇论文统计与内容分析，描述了我国大数据背景下隐私问题相关研究的现状及热点，可以得出以下几方面的结论，为该领域未来发展提供一定程度的参考。

首先，从发文数量与发展阶段来看，我国大数据背景下隐私问题相关研究总体呈增长态势，处于萌芽阶段，但发文数量远远不够，研究力度亟待提升，未来研究应该加强对该领域的关注和重视。

其次，从发表期刊分布来看，大数据背景下隐私问题相关研究发表在核心及以上期刊的文献数量较少，体现出该领域的研究水平相对薄弱，研究成果较少，同时，大数据背景下隐私问题研究还未形成一批处于该领域核心区的期刊，论文分布较为分散。从期刊出版社角度，可以担负起社会责任，进行一次或多次相关主题的组稿或专栏约稿，提高影响力。

再次，从研究群体来看，新老作者比例结构不均衡，尽管有大量新生作者尝试进入该领域，但没有形成成熟的研究群体，不具备群体稳定性，缺乏核心影响力。这就要鼓励学者们积极投身该领域，不仅要在横向上扩展主题范围，多挖“坑”，更要在纵向上对某一点进行深入探讨，挖“深坑”。

最后，从关键词统计和内容分析来看，隐私问题的现状和对策建议是我国大数据背景下隐私问题的研究重点，具体集中在对不同来源信息泄露情况的分析讨论以及技术层面、体系建设的应对措施。一方面，以后研究在主题内容上应尽量多样化，寻找不同切入点探索大数据背景下的隐私问题；另一方面，尽管现有研究已经涉及多个学科，但是大多都还停留在理论层面，以后研究需要在结合多学科知识的基础上，将提出的理论技术付诸实践，检验其对策建议的可行性。

参考文献

[1] 姜盼盼. 大数据时代个人信息保护研究综述[J]. 图书情报工作, 2019, 63(15): 140-148.

[2] 柳亦博. 人工智能阴影下:政府大数据治理中的伦理困境[J]. 行政论坛, 2018, 25(3): 97-103.

[3] 吕耀怀, 曹志. 大数据时代的基因信息隐私问题及其伦理方面[J]. 伦理学研究, 2018(2): 86-91.

[4] 陈威霖. 脸书数据门所折射出的道德危机[J]. 新媒体研究, 2018, 4(7): 63-65.

[5] 李航. 大数据时代:网络隐私伦理问题探究[J]. 现代商业, 2018(29): 165-166.

[6] 陈新忠, 张亮. 我国研究生教育质量研究的轨迹、现状及热点——基于 1986-2016 年 CNKI 期刊的文献计量与内容分析[J]. 现代教育管理, 2018(6): 118-123.

[7] 江向东, 傅文奇. 十年来我国数字图书馆知识产权研究论文的统计分析[J]. 情报科学, 2008, 26(4): 580-585.

[8] 马费成, 张勤. 国内外知识管理研究热点——基于词频的统计分析[J]. 情报学报, 2006, 25(2): 163-171.

[9] 姜鑫. 国际图书情报领域“科学数据”研究进展述评——基于 SCI/SSCI 期刊论文的内容分析[J]. 现代情报, 2018, 38(12): 144-150.

[10] 彭长桂, 高俊山. 国内组织制度理论研究进展与评价——基于 CSSCI 数据的元分析和引证分析(2002~2008) [J]. 情报杂志, 2010, 29(7): 105-109.