Open Journal of Legal Science 法学, 2023, 11(6), 5130-5135
Published Online November 2023 in Hans. https://www.hanspub.org/journal/ojls
https://doi.org/10.12677/ojls.2023.116734

Hans汉斯

文本与数据挖掘著作权合理使用分析及其认定

王思涵

江南大学法学院，江苏 无锡

收稿日期：2023年8月14日；录用日期：2023年8月23日；发布日期：2023年11月10日

摘　要

文本与数据挖掘技术的产生既是大数据时代资源利用的新机遇，又是我国著作权体系面临的新挑战。目前，允许并支持这种形式的研究以促进科技创新是世界的主流趋势，但我国《著作权法》并未明确其合法性，相关机构开展文本与数据挖掘研究的各阶段仍面临较大的侵权风险。结合文本与数据挖掘的技术特点，以及我国著作权体系下合理使用的认定标准，将其认定为合理使用具有理论依据及现实意义。由于构建开放式“合理使用”条款不符合我国司法实践，在我国刚完成第三次《著作权法》修订的背景下，应当考虑在《著作权法实施条例》中增设文本与数据挖掘例外，允许出于科学研究目的，对合法获取的作品进行文本与数据挖掘，以回应数字时代对作品利用的现实需求，更好地实现《著作权法》的立法目标。

关键词

数字资源，文本与数据挖掘，合理使用，著作权法

Copyright Fair Use Analysis and Determination of Text and Data Mining

Sihan Wang

School of Law, Jiangnan University, Wuxi Jiangsu

Received: Aug. 14th, 2023; accepted: Aug. 23rd, 2023; published: Nov. 10th, 2023

Abstract

The emergence of text and data mining technology is both a new opportunity for resource utilization in the era of big data, and a new challenge that our copyright system needs to face. At present, it is the mainstream trend in the world to allow and support this form of research to promote scientific and technological innovation, but China's Copyright Law does not clarify its legitimacy, and the relevant organizations are still facing a greater risk of infringement in the various stages

文章引用: 王思涵. 文本与数据挖掘著作权合理使用分析及其认定[J]. 法学, 2023, 11(6): 5130-5135.
DOI: 10.12677/ojls.2023.116734

of carrying out text and data mining research. Combined with the technical characteristics of text and data mining, as well as the criteria of fair use under China's copyright system, recognizing it as fair use has theoretical basis and practical significance. Since the construction of open-ended "fair use" provisions is not in line with China's judicial practice, in the context of China's third revision of the Copyright Law, consideration should be given to the addition of text and data mining exceptions to the Implementing Regulations of the Copyright Law, which allow text and data mining of legally obtained works for scientific research purposes, in order to respond to the realistic demand for the utilization of works in the digital era and better realize the legislative goal of the Copyright Law.

Keywords

Digital Resource, Text and Data Mining, Fair Use, Copyright Law

1. 问题的提出

在数字网络技术飞速发展的大背景下，实践中对数字化作品的利用与传统著作权保护模式产生了一系列不相适应的问题。文本与数据挖掘(Text and Data Mining, TDM)作为当前数字资源利用的重要技术性支撑，在文化传播以及科学研究等方面都发挥着重要作用。而文本与数据挖掘技术在适用著作权合理使用等制度时仍存在理论上、立法上以及司法上的争议。尤其新《著作权法》修订后，如何在现有法律框架下保障文本与数据挖掘技术的实施并对可能产生的著作权侵权问题进行防范，为科学技术的发展扫清障碍是本文研究的主要问题。

2. 文本与数据挖掘技术侵权风险分析

文本与数据挖掘是一个多步骤过程，具体可以分为四个阶段，即文本与数据收集阶段、数据预处理阶段、建立数据模型阶段以及结果分析阶段。这其中就可能涉及到一些侵权行为。由于建立数据模型阶段一般不会涉及著作权侵权问题，所以下文主要对其余三个侵权风险较大的阶段进行分析。

2.1. 文本与数据收集阶段

第一阶段是文本数据的获取与输入，文本与数据收集阶段研究者为了进行计算研究往往需要复制期刊文章或者数据集。当研究人员将文本通过数字化的方式在有形物质载体上进行再现，且该文本被相对稳定和持久地“固定”在有形物质载体之上形成复制件时，该行为就受到著作权法中复制权的控制[1]。此外，如果未经权利人许可破坏或避开技术措施，则可能违反《著作权法》第 49 条第 2 款的规定。我国虽然有可以避开技术措施的例外规定，规定学校课堂教学或科学研究可以利用技术规避措施获取受版权保护的作品，但仅限于当作品无法通过正常途径获取时，才可适用该例外情形。

2.2. 数据预处理阶段

数据预处理阶段研究人员需要对文章进行重新格式化，可能将 PDF 格式转换为 XML 或其他结构化数据格式。此外，数据预处理还要进行数据标记，对复杂的文本信息标记类别标签。这一步需要从非结构化的自然语言文本中抽取事实信息，并形成结构化数据输出。简而言之，这就是将文本转化为数据的

过程。根据“思想–表达二分法”原理，只有在保留原作品基本表达的情况下通过改变原作品，创作出新作品的行为属于著作权法意义上的改编行为，也就是说如果仅仅根据原作品的思想创作出新作品并非受改编权控制的行为。有学者主张通过计算机技术对文本进行重新“转码”的行为侵犯权利人的翻译权或改编权[2]。但从计算机“转码”行为的技术特征来看，计算机对样本的识别与转码是一一对应的，所以计算机转码仍然是复制行为，受复制权的控制。

2.3. 结果输出阶段

最后一阶段就是结果输出阶段，该阶段通过之前设定的收集、处理、建模后得出最后的搜索结果或产生新的内容。该阶段是否侵权则主要取决于最后产生结果的表现形式。根据研究人员的研究目的，最后的结果分析可能包含收集样本中符合要求的部分表达或简短摘要，也可能仅仅包含不属于作品的一些数据信息。如果这一阶段产生的作品再现了第一阶段收集到的受保护作品的表达，可能侵犯原作品著作权人的复制权。而如果最后产生的内容为仅运用原作品思想创作的新作品或根本不构成作品的信息则不构成侵权。同时如果对构成侵权的分析结果进行传播则会侵犯传播权。

3. 文本与数据挖掘合理使用分析及认定

著作权合理使用制度是为了满足社会公众对知识和信息的需要，在一定条件下对著作权人的专有权利进行限制的一种平衡制度。合理使用通过防止著作权人对作品绝对垄断从而达到促进社会主义文化和科学事业繁荣发展的立法目标。我国加入的《伯尔尼公约》以及《TRIPS 协定》等国际条约均规定了对专业权利限制规定时应当遵循“三步检验标准”，由于“三步检验标准”已经转换为我国国内立法，新《著作权法》更是明确了三步检验法标准的适用。所以，在进行文本数据挖掘的合理使用分析时也应当结合三步检验标准以及其他著作权合理使用理论进行考量与认定[3]。

3.1. “非表达性使用”的认定

“非表达性使用”指的是任何复制行为，其目的不是使人类能够享受、欣赏或理解被复制的表达。公共政策要求著作权法保护的作品仅限于对思想的独创性表达，作品中抽象的思想是不受保护的。所以区分对原始文本的“表达性使用”与“非表达性使用”是认定合理使用的重要因素之一。“非表达性使用”行为本身并非以再现作品的独创性表达为目的，其与原作品也并非具有竞争性的替代关系，所以对于作品的“非表达性使用”不会影响原作品的使用，也不会不合理损害权利人的合法权益[4]。

在文本与数字挖掘过程中，计算机对文本或数据库的复制是为了生成一系列有研究价值的信息，而这些信息并不涉及作品中的表达性内容，也并非为了替代原有作品。与著作权侵权判断中区分作品思想与表达的原理相同，著作权人不能阻止读者提取和复制作品中包含的事实与思想。允许未经授权对版权作品进行文本挖掘等“非表达性使用”是符合版权法基本构造的，因为版权法只关注将作者的原始表达传达给大众，而文本数据挖掘并不会与版权所有者的专有权相冲突。所以，纯粹出于非表达性目的的复制作品并不侵犯作者的著作权，应当认定为合理使用。

3.2. 转换性使用的认定

在著作权侵权判断中，对原作品的使用是否构成“转换性使用”以及其“转换性”程度也是衡量是否构成合理使用的重要因素之一。如果对于原作品的使用并非为了单纯地再现原作品本身的文学、艺术价值或者实现其内在功能或目的，而是通过增加新的美学内容、新的视角、新的理念或通过其他方式，使原作品在被使用过程中具有新的价值、功能或性质，那么对该作品的使用就构成“转换性使用”[5]。由于对原作品的“转换性使用”能够创造出新的价值或功能，其符合《著作权法》促进社会主义文化和

科学事业发展与繁荣的立法目的。

在文本与数字挖掘中，通过对文本的处理、建模后得出的搜索结果或产生的内容显然改变了原作品的功能与目的，其输出结果不仅增加了新的价值，还具有与原作品不同的特征与功能。大多数时候，通过文本挖掘所获得的信息是无法通过人类的阅读所获取的，其产生的新信息与新见解具有高度的“转换性”[6]。由于在文本数据挖掘中对文本进行复制的目的是挖掘出数据中潜在的模式或趋势，这种使用具有高度的“转换性”，且这种“转化性使用”不会与作品的正常利用相冲突也没有不合理地损害作者的利益，应为法律所允许。

3.3. 市场替代效果的认定

对原作品的使用是否会产生市场替代效果是合理使用判断中最为重要的因素之一，因为其直接与权利人的利益相关联。在进行判断时需要考虑复制品是否给市场带来了原件或其衍生物的竞争性替代品，从而剥夺了权利人的收入。市场替代效果的认定与前述“转换性使用”的认定是密不可分的，因为越是为了实现与原作品不同目的而进行的复制，其越不可能产生与原作品形成竞争关系的替代品。但对市场替代效果的判断也不是绝对的，在一些高度转换性的使用中，即便使用对原作品已有或潜在的市场有所损害，也不是必然被认定为是不合理的从而构成版权侵权，还需要在权利人损失与社会公众利益之间进行衡量[7]。

一般来说，出于文本数据挖掘目的而进行的复制由于具有高度转换性，其并不会产生市场替代效果。但有观点认为，由于市场是动态的，文本数据挖掘可能会限制权利人潜在市场的扩张，使权利人失去许可他人利用其作品进行文本挖掘的机会。但该观点在考虑市场替代效应时，将考虑范围扩展到了所有现有及潜在的市场中。对权利人应获得的市场利益进行判断时，应限制在可认知范围内的市场利益，如果将所有现有市场及可能的潜在市场均考虑在内会不合理的损害社会公众利益。在判断是否不合理地损害著作权人的合法利益时，也应借鉴比例原则，考察被告的行为所增进的公共利益与对原告所造成的损害是否成比例[8]。就文本数据挖掘而言，如果权利人开发或许可一个市场用于文本数据挖掘，这当然是被允许的，但权利人不能因此阻止他人进入合理使用市场。

4. 我国文本与数据挖掘合理使用路径选择

4.1. 构建开放式“合理使用”条款存在困难

我国著作权法的立法模式虽然在一定程度上借鉴了英美法系版权制度，但总体上沿袭了大陆法系著作权体系。因而更加注重对作者权利的保护，而不同于英美法系将社会公众利益放在首位。这就决定了合理使用等制度肩负的立法目标需要保障各方权利主体的利益平衡，避免因过度开放的例外条款而不合理损害著作权人的利益。封闭式的合理使用条款虽然在适应高速发展的科学技术带来的冲击上缺乏灵活性，但其确定性也保障了司法不会任意蚕食著作权人的合法权利，避免了著作权例外的过分扩张。

目前，我国著作权法采取的有条件的例外模式，有效平衡了著作权人与社会公众利益。而如果采取美国式的开放性著作权例外模式，将是否构成合理使用的认定完全交由司法机关，无疑会赋予法官过大的自由裁量权[9]。由于我国是成文法国家，法官并没有足够的司法判例及相关审判经验的支持，采用开放式合理使用条款使法官陷入无明确规范性条款适用的情境，不仅无法实现司法的可预期性，也难以消除由于新技术不断产生所带来的不确定性。所以，构建开放式“合理使用”条款即不符合我国著作权法的理论基础，也与我国的司法实践不相适应，且在我国刚完成《著作权法》第三次修订的大背景下讨论该问题也不具有现实意义。因此，下文主要讨论文本与数据挖掘如何在现有《著作权法》框架下寻求合法化解决路径。

4.2. 在《著作权法实施条例》中增设文本与数据挖掘例外

新修订的《著作权法》“合理使用”条款在原《著作权法》第22条的基础上明确了“三步检验标准”的适用，并增加了一项合理使用情形，即“法律、行政法规规定的其他情形。”该条款的设立虽然未改变我国封闭式的合理使用立法模式，但其为数字时代著作权法在新技术冲击下需要做出的调整留下了空间。

实践中，面对层出不穷的新技术，《著作权法》规定的12种合理使用情形已经不能满足当前网络环境下使用作品的现实需要。文本与数据挖掘技术大大拓宽了资源利用与科技创新的效率，具有巨大的科研潜力与经济价值[10]。而如果每次文本数据挖掘前都需要为全部涉及作品获取授权，那么研究人员需要花费巨大的人力、时间成本。且根据前文分析，文本与数据挖掘例外属于对原作品的“非表达性使用”，因此具有“转换性使用”目的的数据挖掘并不会不合理损害著作权人的权利，反而有利于《著作权法》促进科学文化繁荣的立法目的的实现[11]。综上，在《著作权法实施条例》中增设文本与数据挖掘例外即满足了实践中对文本与数据挖掘技术合法性认定的现实需求，又符合我国的司法现状，能够使《著作权法》的修订真正发挥实效。

4.3. 文本与数据挖掘合理使用的要件设置

文本与数据挖掘合理使用要件的设置应当在满足“三步检验标准”的同时，结合文本与数据挖掘自身的技术特点。

首先，在适用主体的设置上，一些国家将文本与数据挖掘合理使用的主体限定于非营利性的科研机构或公益机构中。但在目前大量研究机构与私营部门合作研究的背景下，将文本与数据挖掘合理使用的主体限定于科研机构或公益机构不利于该技术的应用，且科研机构在满足一定条件是本身也可以根据前述十二种合理使用情形进行豁免，此双重限制有架空文本与数据挖掘合理使用条款的风险。所以，对于要件的设置应当采取强化目的要件，而弱化主体要件的立法模式，不对挖掘主体进行限制，通过其他要件进行限制。

其次，目的要件作为文本与数据挖掘合理使用的关键要件应当符合“出于科学研究目的”而进行的挖掘。“出于科学研究目的”的数据挖掘在满足《著作权法》促进科学文化繁荣发展的立法目标的同时，避免了对文本数据挖掘技术的商业化使用，也满足了合理使用条款中的“三步检验标准”。另一方面，出于“科学研究目的”而进行的文本与数据挖掘，由于其最终实现的是社会公共利益，在此前提下让渡著作权人的部分利益，增进社会公共利益才是符合《著作权法》立法宗旨的[12]。

此外，合法获取作品是构成合理使用的前提条件，挖掘的对象要件应当设置为合法获取的作品，对于具体的作品类型可以不做限制。当然，在进行文本与数据挖掘时，研究人员也应当尽到合理的注意义务，避免对著作权人利益的不合理损害。文本数据挖掘技术在应用过程中应当严格控制对原作品“表达性使用”的数量及范围，仅在必要范围内，符合其使用目的的再现作品。此外，挖掘主体还应当采取合理措施防止复制件被非法利用以及二次传播，从而损害著作权人的利益。

5. 结语

随着数字技术的发展，面对各种新兴的作品利用方式，著作权法律制度也应当做出适当地调整以适应当下的现实需要。在文本与数据挖掘技术面临合法性困境的今天，将其认定为合理使用具有理论依据及现实意义。在现有著作权体系框架探索文本与数据挖掘的合法性路径，在《著作权法实施条例》中增设文本与数据挖掘合理使用情形。同时，明确该合理使用情形的具体适用要件，通过对文本数据来源的合法性以及科学研究目的的强调，来满足出版商的经济利益需要，并保护其正当利益不会不合理减损，平衡各权利主体之间的利益，对当下研究主体利用新技术进行科学研究具有重要意义。

参考文献

[1] 万勇. 人工智能时代著作权法合理使用制度的困境与出路[J]. 社会科学辑刊, 2021, 256(5): 93-102.

[2] 马治国, 赵龙. 文本与数据挖掘对著作权例外体系的冲击与应对[J]. 西北师大学报(社会科学版), 2021, 58(4): 107-115.

[3] 焦和平. 人工智能创作中数据获取与利用的著作权风险及化解路径[J]. 当代法学, 2022, 36(4): 128-140.

[4] Sag, M. (2019) The New Legal Landscape for Text Mining and Machine Learning. *Journal of the Copyright Society of the USA*, **61**, 346, 350.

[5] 王迁. 知识产权法教程[M]. 第 7 版. 北京: 中国人民大学出版社, 2021: 288-300.

[6] 华劼. 美国转换性使用规则研究及对我国的启示——以大规模数字化与数字图书馆建设为视角[J]. 同济大学学报(社会科学版), 2018, 29(3): 117-124.

[7] 万勇. 著作权法三步检验标准的误解澄清与本土重塑[J]. 上海政法学院学报(法治论丛), 2022, 37(4): 42-55.

[8] 袁锋, 徐琢. 新技术环境下图书馆限制与例外条款的问题与完善研究——兼论《信息网络传播权保护条例》的修订[J]. 图书馆杂志, 2022, 41(5): 31-38+55.

[9] 赵力. 文本与数据挖掘著作权合理使用的域外实践与借鉴[J]. 图书馆, 2022, 330(3): 63-69.

[10] 华劼. 合理使用制度运用于人工智能创作的两难及出路[J]. 电子知识产权, 2019(4): 29-39.

[11] 王文敏. 文本与数据挖掘的著作权困境及应对[J]. 图书馆理论与实践, 2020(3): 28-34.

[12] 董凡, 关永红. 论文本与数字挖掘技术应用的版权例外规则构建[J]. 河北法学, 2019, 37(9): 148-160.