

使用爬虫技术合法吗？

作者：杨迅 | 杨坚琪

网络爬虫，又称为“网页蜘蛛” (Spider)。它是一段自动执行的程序：向其提示统一资源定位(URL)之后，它从相应网页(也可以包括该网页上超链接指向的网页)自动下载信息¹。很多企业在数据收集阶段，或多或少需要借助“爬虫”技术。“爬虫技术”是不是合法，它是大数据行业的宠儿，还是乱局者？本文将从“爬虫技术”的应用场景出发，结合中外有关“爬虫技术”的司法实践，探讨“爬虫技术”的合法性边界。

一. 爬虫技术及其商业应用

从爬虫的技术历史上看，它几乎伴随着互联网行业共同发展。网络爬虫的信息检索功能极大提高了在浩如烟海的万维网检索、收集信息的效率，因而在产业界有着广泛的应用。

现实中，传统企业、互联网巨头和政府机关都收集和保有大量的数据，但是其对于数据的利用效率往往并不高，或者说对于数据的开发仍然没有达到最佳效率。知名的咨询公司 Forrester 在其报告中指出：“大多数公司估计他们只分析了已有数据的 12%，剩余 88%还没有被充分利用²。大量的数据孤岛和分析能力的缺乏是造成这种局面的主要原因。”对于“大数据”公司而言，使用爬虫技术和类似的自动搜索访问工具，可以打破“数据孤岛”，使得“数据”资源自由能够流通，从而充分利用该些没有充分发挥效能的数据，可以带来巨大的经济价值和社会效益。

.....
如您需要了解我们的出版物，
请联系：

Publication@llinkslaw.com

¹Wolfgang Glänzel 等：《Springer Handbook of Science and Technology Indicators》，《Springer Nature》2019 年 10 月 30 日，第 1103 页

² Forrester 数据研究报告：Big Data Hadoop Solutions, 2014 年第一季度, <https://go.forrester.com/press-newsroom/the-forrester-wave-big-data-hadoop-solutions-q1-2014/> (2020 年年 10 月 3 日访问)。

但是，爬虫技术的使用也饱受争议。对于已经占有大量数据的公司而言，纵然其可能没有充分利用这些数据，但其往往倾向于将这些数据视为其私产，进而拒绝第三方通过爬虫技术获取其收集的信息。他们指出了爬虫技术可能造成以下危害³：

- (1) 爬虫技术与 Ddos 攻击具有直接相关性，对网络安全形成巨大的威胁；
- (2) “被爬网站”采取的反爬措施降低了消费者的体验，进而降低了消费者的福利剩余；
- (3) 爬虫技术使得数据/信息的流通不受控制，使得“搭便车”行为更加便利，侵害了潜在权利人的正当利益；
- (4) 爬虫技术导致用户的个人信息和隐私受到极大威胁。

科技公司们还采取一系列以“反爬”为目的的技术手段，以保护其占有数据的安全性。常见的“反爬”措施其中包括公示 robots 协议，采取 IP 限制措施，采取 User Agent 鉴别机制，采取其他措施限制访问，如验证码、数据伪装、参数签名、隐藏验证和阻止调试等⁴。当然，就像无法阻止人类通过肉眼识别公开网站上的信息一样，只要信息是储存于置放在能够被公共访问的页面上的，就无法阻止爬虫在技术上绕过各种反制手段直接获取信息。

于是，除了技术手段之外，互联网巨头们也频频向政府和立法部门游说，以求在技术门槛之外寻求司法保护。从整体的司法历程来看，已有的司法判决主要关切两方面的问题：一方面是规范爬虫技术的使用规则，另一方面对“爬虫”获取对象的数据权益进行确权。对此，虽然在立法层面上尚无定论，但在具体案件中中美两国法院已经有了初步的回应。

二. 美国法律下的爬虫规制

美国法院很早就开始审理与爬虫获取数据有关案件。早期的法院判决，大体上围绕着普通法的“trespass”(非法侵入)以及《1986 计算机欺诈与滥用法》(Computer Fraud and Abuse Act of 1986, “CFAA”)的适用展开。尤其是，美国法院以普通法为原则，探究使用爬虫技术进入他人计算机系统获取数据，并在不影响他人正常使用情况下的合法性问题。

比如在 eBay Inc. vs. Bidder Edge Inc.⁵ 案件中，法院授予了 eBay Inc. 禁止 Bidder Edge Inc. 通过技术手段获取其数据的禁令，法院认为 Bidder Edge Inc. 获取 eBay Inc. 的数据虽然没有在现实中损害 eBay Inc. 的权利，但是如果允许它未经 eBay Inc. 许可获取 eBay Inc. 数据，必将引起其他公司效仿，因此这样的行为可能给 eBay Inc. 带来损害。在 Oyster Software Inc. vs. Forms Processing Inc.⁶ 案件中，法院甚至更进一步指出：无需证明损害，一旦第三方未经许可进入他人计算机空间，就属于侵权。但是，在随后的 TicketMaster vs. Ticket.com⁷ 一案，法院有了不同的看法。在该案中，纵然 TicketMaster 表明其收集、

³ 王超、程诗梦：《国内网络爬虫行业发展情况及监管思考》，载《江西通信科技》2019 年第 1 期，第 39-40 页。

⁴ “反爬虫和反反爬虫”，<https://cloud.tencent.com/developer/article/1163912>，以及“爬虫原理及反爬虫机制以及反爬虫方法的应对策略”，https://blog.csdn.net/Archer__Ye/article/details/89609280，最后访问于 2020 年 9 月 30 日。

⁵ eBay, Inc. v. Bidder's Edge, Inc., 100 F. Supp. 2d 1058 (N.D. Cal. 2000)

⁶ Oyster Software Inc. v. Forms Processing Inc., et al. N.D. Cal. Dec. 6, 2001, No. C-00-0724 JCS) 2001 WL 1736382

⁷ Ticketmaster Corp. v. Tickets.com, Inc., 2003 WL 21406289, (C.D. Cal. March 7, 2003)

整理购票信息花费了大量的精力，因此这些信息是有价值的，但法院仍然认为 Ticket.com 通过技术手段获取这些信息，没有损害 TicketMaster 的利益，因此不构成侵权。

2017 年的 *hiQ Labs, Inc. v. LinkedIn Corp.* 一案⁸则对爬取数据的边界作出了较为完整的论述。在本案中，加州北区法院在论述 hiQ 的抓取行为是否属于“未经授权”或“超出授权”行为时，将判决的核心重新调整到了“信息”的公开性上。法院认为，爬虫公开的信息并不构成 CFAA 意义上的“未经授权”或“超出授权”行为，因为公开信息不同于 CFAA 法条中阐明的“information from any protected computer”，其缺少相应的保护措施，所以不属于 CFAA 禁止的行为。

同时，加州北区法院依据(1)LinkedIn 的禁止爬取行为将其在职场社交领域的竞争优势“传导”到了职场分析数据领域，以及(2)网站上的“使用条款”给与了网站所有者过大的权力，这可能涉及违反加州当地宪法中言论自由相关的规定，因此支持了 hiQ 的临时禁令请求。在 2019 年 9 月，美国第九巡回上诉法院同样维持了加州北区法院的裁决。

这个具有里程碑意义的判决指出了爬虫技术使用的边界，即(1)如果爬虫抓取的是公开信息，为了保证特定数字服务市场的竞争性，最终是为了消费者福利的总体增长，以及(2)考虑到言论自由的问题，美国的司法机构们更愿意接受使用“爬虫”技术的合法性。相反，如果爬取的信息具有私有属性，则爬虫技术的使用可能存在合法性障碍。

三. 我国司法实践中的爬虫技术的合理使用和侵权风险

从政策角度观察：一方面，出于鼓励大数据产业发展和打破数据孤岛考虑，我国鼓励数据的流通，因此爬虫技术的使用存在积极价值；另一方面，纵容爬虫技术的滥用也确实更容易滋生不正当的“搭便车”现象。因此，中国法院在司法判决中往往以技术中立为原则，不否定爬虫技术的合法性，但就其具体使用，则严格限定边界和条件。

我国法院倾向于认为爬虫技术本身并不违法，并且认定旨在规范爬虫规则的 Robots 协议也不具有强制性的约束力。在百度与奇虎 360 的“奇虎违反 robots 协议抓取数据”系列案件⁹中，北京一中院和北京高院均认为设置 robots 协议并不代表第三方就不可以利用“爬虫技术”抓取其数据，尤其是考虑到 robots 协议是在上世纪 90 年的网络资源缺乏的背景下制定的，而现在的技术条件与当时不可同日而语。因此，法院认为除非存在非常有限的例外情况，违反当事人设置的 robots 协议约定，利用爬虫技术抓取数据并不违法。

但是，在另一系列案件中，法院认为，如果使用爬虫技术获取数据的结果是：复制、传播和展示了他人受著作权保护的作品，或者构成不正当竞争，那么虽然爬虫技术本身并不违法，但是该种爬虫的使用方式则会构成侵权行为。大众点评诉爱帮网系列案件¹⁰最全面地阐释了这一观点。在该系列案件

⁸ *hiQLabs, Inc. v. LinkedIn Corp.*, 273 F. Supp.3d 1099 (N.D. Cal. 2017)

⁹ 参见(2013)一中民初字第 2668 号、(2013)高民初字第 3755 号、(2017)京民终 487 号判决书。

¹⁰ 参见(2009)一中民终字第 5031 号、(2010)海民初字第 4253 号以及(2011)一中民终字第 7512 号判决书。

中，大众点评曾分别以“著作权”和“不正当竞争”为由，起诉爱帮网未经授权抓取数据的行为，并获得法院认可。其中，尤为值得注意的是大众点评网以“反不正当竞争”为案由起诉的案件。在该案中，法院认为大众点评虽然不拥有点评和介绍信息的著作权，但是爱帮网获取该些信息用于公开展示，利用了大众点评网的劳动成果，对大众点评网构成不正当竞争。此案的判决具有里程碑意义，在之后的微博诉脉脉案¹¹、车来了诉酷米乐¹²、淘宝诉美景案¹³、腾讯诉头条案¹⁴中，数据拥有者均以“不正当竞争”为由起诉被告方违法使用或者抓取“大数据资源”的行为，并获得法院支持。

在这些案件中，法院虽然没有从正面肯定“大数据资源”的财产属性，但却以模棱两可的“无形财产权益”、“集合性民事权益”、“竞争优势”、“竞争性财产权益”等用语描述大数据在竞争法意义上的法律地位。这种裁判规则背后的规制逻辑是，商业主体利用爬虫技术获取第三方数据，且与原数据收集者形成竞争关系的，这样的爬虫技术使用行为可能构成不正当竞争。

四. 我国司法实践中的使用爬虫技术的刑事风险

法院还在在一系列刑事案件中表明：如果在爬虫抓取数据过程中存在其他的违法行为，或者造成法律禁止的严重后果的，那么该等爬虫使用行为非常可能产生刑事责任。目前，由于使用爬虫技术，导致刑事责任的，主要有以下几种情况：

- (一) 使用爬虫技术导致“被爬”网站长时间无法访问。在(2019)粤 0305 刑初 193 号案件中，被告人开发的爬虫软件在 2018 年 5 月 2 日 10 时至 5 月 2 日 12 时的两小时内，以每秒 183 次的频率访问“深圳市居住证系统”，导致“深圳市居住证系统”停止运行超过 2 小时，该等爬虫使用行为被认定违反了《刑法》第二百八十六条的规定，构成“破坏计算机信息系统罪”。
- (二) 绕开技术限制措施，进而使用爬虫技术下载网站后台数据。在(2019)鲁 0213 刑初 144 号案件中，被告人首先利用“SQL 注入漏洞”获取网站的后台管理权限，进而利用其编写的爬虫脚本程序侵入计算机信息系统，获取计算机系统内存储的大量数据，且该等数据并非在公开页面显示的数据，因此被告人的行为被认定违反了《刑法》第二百八十五条的规定，构成“非法获取计算机信息系统数据罪”。
- (三) 在授权目的之外，利用爬虫技术获取大量数据。(2019)浙 0602 刑初 636 号案件中，被告人北京瑞智华胜科技股份有限公司及其员工通过签署合作协议的方式获取了运营商服务器的登录权限，但是北京瑞智华胜科技股份有限公司通过部署恶意程序的方式保存服务器中的用户登录数据，并通过利用自行研发的爬虫程序调用数据库中的保存的数据和信息。经核实，该等信息可以用来登录淘宝、京东等网络平台。法院认为，在此情况下，超越授权范围使用爬虫非法采集、保存数据的行为，违反了《刑法》第二百八十五条第二款和第四款的规定，构成“非法获取计算机信息系统数据罪”。

¹¹ 参见(2016)京 73 民终 588 号判决书。

¹² 参见(2017)粤 03 民初 822 号判决书。

¹³ 参见(2018)浙 01 民终 7312 号判决书。

¹⁴ 参见(2019)津 0116 民初 2091 号判决书。

五. “爬虫技术”使用规则展望

我国在立法上并没有明确界定使用爬虫技术的合法边界，但是现有的司法判决力求达到保护数据原始收集人和使用人以及使用效率的利益平衡。这个平衡就在于：秉承“技术中立”的原则，不否定爬虫技术的合法性，允许其在不影响网络安全或者非不公平地侵犯商业秘密和个人信息，不与原始数据收集者直接竞争的前提下使用，从而增强数据的使用效率，增进社会福祉。换言之，我国司法以“帕累托”优化为主导思想，在不损害任何第三方现有利益的情况下，允许使用爬虫技术收集数据，进而挖掘数据的更大价值。

但是，我国法律对“爬虫技术”的规制，还有一些不明确之处。从立法上说，需要在规则层面上明确“数据”、“爬虫技术”以及“数据共享”的法律地位。在目前生效的法律渊源中，无论是国家立法或是司法解释，都没有对“数据抓取”或“爬虫技术”有所涉及，而前文判决大体上是在具体案件中法院的被动判决。对于类似我国这样的以成文法为唯一正式法律渊源的国家而言，缺少正式的成文规定，无疑给“爬虫技术”的使用蒙上不确定性。此外，一些地方政府(如深圳)基于当地的产业需求出台的地方性法规，有意无意创造了“数据权利”的概念，这样的立法活动本身也可能存在合法性缺失。在没有上位法支持的情况下，地方政府贸然创设“数据权利”也为“爬取”数据的合法性带来更大的不确定性。

从司法上而言，我国目前的司法裁判虽然已经形成了对使用爬虫技术合法性边界的初步规则，但是仍然有不少问题没有得到充分解析。比如，就“不正当竞争”而言，如何使用爬取的数据不认为构成竞争关系，这种竞争关系是否必须是直接的，抑或是功能上存在潜在的替代关系，或者压缩原始数据收集者发展空间也构成“竞争”？目前涉及爬虫技术的案件中，法院对于竞争关系的认定往往作广义解释，侧面反映了法院对规制爬虫技术使用的政策性思考。又比如，对于“侵入计算机系统”而言，如何界定侵入计算机系统，绕过口令、技术措施，还是违反网站上公开的网站使用声明都可以被认为是“侵入”行为？这些具体问题，关系到“爬虫技术”能不能被合法使用，仍然需要在今后的司法实践中进一步明确。

如您希望就相关问题进一步交流，请联系：



杨 迅
+86 21 3135 8799
xun.yang@llinkslaw.com

如您希望就其他问题进一步交流或有其他业务咨询需求，请随时与我们联系: master@llinkslaw.com

上海

上海市银城中路 68 号
时代金融中心 16/19 楼
T: +86 21 3135 8666
F: +86 21 3135 8600

北京

北京市建国门北大街 8 号
华润大厦 4 楼
T: +86 10 8519 2266
F: +86 10 8519 2929

深圳

深圳市南山区科苑南路 2666 号
中国华润大厦 18 楼
T: +86 755 3391 7666
F: +86 755 3391 7668

香港

香港中环皇后大道中 5 号
衡怡大厦 27 楼
T: +852 2592 1978
F: +852 2868 0883

伦敦

1/F, 3 More London Riverside
London SE1 2RE
T: +44 (0)20 3283 4337
D: +44 (0)20 3283 4323



www.llinkslaw.com



Wechat: Llinkslaw

本土化资源 国际化视野

免责声明:

本出版物仅供一般性参考，并无意提供任何法律或其他建议。我们明示不对任何依赖本出版物的任何内容而采取或不采取行动所导致的后果承担责任。我们保留所有对本出版物的权利。

© 通力律师事务所 2020