

●沙勇忠 黄丽琚

Web挖掘的伦理问题研究*

摘要 Web挖掘中涉及到的伦理问题有:对用户隐私权的侵犯,对用户个性特征的忽视,知识产权问题,信息安全问题。解决这些问题应从多方面入手:制订伦理守则,运用技术手段践行伦理要求,用户对隐私权等信息权利认真追究,管理部门加强监管。参考文献16。

关键词 Web挖掘 网络伦理 信息伦理 知识产权

分类号 G251.6

ABSTRACT Ethical issues related to Web mining include the infringement of user's privacy, the neglect of user's personal characteristics, the intellectual property issue and the information security issue. To solve these problems, there are some approaches, such as ethic codes, technological measures, user's claim to privacy and strict management. 16 refs.

KEY WORDS Web mining. Network ethics. Information ethics. Intellectual property.

CLASS NUMBER G251.6

Web挖掘技术在给业界尤其是工商企业带来情报能力增进的同时,也带来了一系列伦理问题。本文对此进行分析和讨论,并提出可能的解决对策。

1 Web挖掘中涉及到的伦理问题

1.1 Web挖掘中对用户隐私权的侵犯

隐私权是公民依据习俗、道德或法律要求所具有的对个人领域、个人私事和个人信息的了解、拥有、公开的控制以及不受他人侵犯的权利。本文所探讨的隐私权集中在隐私权的信息方面,即用户享有对个人信息的了解、拥有、公开的控制以及不受他人侵犯的权利^[1]。

Web挖掘由于涉及到对个人数据的使用,可能会导致对一些重要的普遍价值的破坏,突出的是隐私权所受到的威胁。在Web使用挖掘中,用户只要登陆网站,网站的服务器日志就会记录下该用户的有关信息,如IP地址、进入和离开的日期与时间、点击的页面与链接、浏览的文件类型等,据此ISP可以对该用户的信息行为做出分析。虽然Web服务器识别的是Web用户而非个人,但当该用户再次出现在网上时,通过cookies他会被识别出是一个经常用户;尤为重要的是,ISP可以将其所分析的用户信息行为同其所掌握的用户个人数据结合起来,对用户的网络活动进行监控。这意味着如果ISP愿意,就可以在用户不知情的情况下,对用户的行为、偏好、习惯等进行透彻的了解,对用户进行“超级全景监狱”(panopticon)般的监视^[2],而这将限制用户的行为自由,并损害其人

格尊严,使用户对个人信息地了解、拥有、公开控制的权利丧失。据美国联邦贸易委员会的在线报告,被调查的1400家网站85%存在收集客户个人信息的行为,但只有14%的网站对其信息收集行为做了通告;而且1400家网站中只有2%的网站根据其全面的隐私政策对其信息收集行为提供了通告^[3]。多数网站对个人数据的收集和分析是在用户不知情且没有同意的情况下进行的,多数网站也没有从隐私保护的角度看待个人数据及其使用。实践中的另一个问题是,如果ISP将其对用户个人数据的分析结果出售给其他具有商业目的的厂商和机构,则可能会对用户的隐私权造成更大程度的侵犯。

Web内容挖掘和Web结构挖掘也可能侵犯用户的隐私权。与Web使用挖掘不同的是,Web内容挖掘和Web结构挖掘的数据是可以公开得到的,而不像Web使用挖掘其数据只有ISP或网站管理者才可得到。学术界关注的问题是,Web页面数据和Web结构数据一旦被挖掘,就会完全脱离原来的语境,并应用于截然不同的目的^[4]。是否值得对Web页面数据和Web结构数据进行保护?尼森鲍姆的回答是“值得”^[5],因为特定的Web页面数据和Web结构数据的聚合或与其他数据的结合,可能使不可确认的数据变成可以确认的(identifiable),产生被用户认为是有害的信息,从而对隐私权造成威胁。但究竟如何进行保护,显然不是一个能有效解决的问题。

1.2 Web挖掘中对用户个性特征的忽视

个性特征是人以一个独立而完整的个体存在的

* 本文系国家社会科学基金项目“网络伦理及其社会调控机制研究”(05CTQ006)研究成果之一。

最基本的特征之一,对个性特征的尊重是现代社会普遍认同的一种价值。

在 Web 挖掘中,个人的信息和数据不再被单独地看待和处理,而是成为大量数据中的一分子,用户个人信息的独立性遭到破坏,人的个体性在数据海洋中迷失。另一方面,与传统的信息检索技术不同,Web 挖掘技术是基于对大量数据的处理和挖掘得出潜在的有用模式和隐藏的信息,因而 Web 挖掘常常针对大量用户数据的某一方面的特征进行分析,其他的数据可能被舍弃,这也造成了用户个人信息的完整性不能得到保存。

Web 挖掘使用的数据常是匿名的。一是因为大量关于个人信息的细节性数据对于 Web 挖掘想要获得的潜在模式和隐藏信息并不重要,而被大量舍弃,挖掘者真正感兴趣的、有用的内容并不需要大量实名的信息来得到。二是 Web 挖掘所采用的数据是零散的,来自 Web 文档和 Web 活动的信息因为网络浏览的匿名性等特点,也具有了匿名的特征。

这样,在 Web 挖掘中,个人信息的独立性和完整性没有保障,用户的个性特征被淹没在数据的海洋中。Web 挖掘技术作为海德格尔所理解的一种“展现方式”,通过“限定”(仅从某种技术的需要给人和事物定位)和“强求”(尽可能加大技术力度,迫使事物进入非自然状态,使其无限度地为技术所用),迫使人进入技术系统的规制之中,以一种“非自然”的方式呈现,实现对人和信息活动的绝对控制^[6]。由此所产生的个性特征忽视问题值得在价值层面上予以关注。

1.3 Web 挖掘中涉及的知识产权问题

Web 挖掘所涉及的知识产权问题主要在 Web 内容挖掘方面。Web 内容挖掘对大量网页信息进行挖掘,从中提取有用的知识并加以运用。但 Web 网页作为一个整体是受版权保护的^[7]。对网页信息的挖掘是否需要得到网页拥有者的授权?挖掘出来的新知识是否要得到版权保护?如果需要,一定是 Web 挖掘者拥有吗?显然,要求 Web 挖掘都要得到网页拥有者的授权,Web 挖掘技术就无法得到应用,Web 挖掘的功能和价值就无法得以实现;而挖掘出来的新知识都得到版权保护,则不利于知识的传播,有悖知识产权制度鼓励知识创造的基本精神。因此,问题还是在于,如何在被挖掘信息和挖掘到的新信息的保护之间保持某种均衡。虽然人们现在还难以明确地认识和回答这一问题,但随着 Web 挖掘技术的进步及其更广泛应用,这一问题会上升为一个重要的主题并得到深入讨论。此外,Web 结构挖掘也会涉及知识产权问题,但程度较内容挖掘轻。

1.4 Web 挖掘中涉及的信息安全问题

信息安全主要有三方面要求:(1)保密性:控制那些存取信息的人;(2)完整性:确保信息更新和修改只能由经过授权的人进行;(3)可用性:确保授权用户总是能存取信息^[8]。Web 挖掘中这三者都受到不同程度挑战。Web 挖掘技术可以从大量数据中提取人们所需的信息,从中发现知识,使信息能更好地为人所用。但 Web 挖掘技术本身并不具有判断能力,进行挖掘时,软件工具本身并不知道它收集和获取的知识是否是被允许获取的,这就使信息的保密性受到挑战。Web 挖掘技术本身可能并不会威胁到信息的完整性和可用性,但因为 Web 挖掘技术的大量应用,导致信息的无授权使用和复制可能性的出现,从而威胁到信息的完整性和可用性。另外,保护专有和敏感信息是保证信息安全的基础,而诸如医疗记录、信用和财务记录等敏感信息一旦被存储在 Web 文档中,就有可能被挖掘出来,造成泄露,给相关的人造成伤害。

2 Web 挖掘的可能辩护之伦理批判

为了解当前 Web 挖掘的实践以及 Web 挖掘专业人员关于伦理问题的态度,Lita van Wel 等对 20 位 Web 挖掘的专业人员进行了调查^[9]。结果发现,Web 挖掘者往往更乐于关注 Web 挖掘技术的优势而不是可能存在的道德风险,他们关于 Web 挖掘的辩护集中于 6 个方面,但这并不能证明无限度的使用 Web 挖掘技术是正确的。下面我们对 Lita van Wel 所调查的 Web 挖掘者的辩护观点从伦理的角度分别讨论。

2.1 Web 挖掘技术本身并没有带来新的伦理问题

这一观点源于这样的认识:计算机和信息技术所带来的新的问题都可以应用传统的伦理理论和伦理规范进行分析和解决。但问题是,Web 挖掘技术带来了哪些新的问题还需要重新认识。例如,“用户档案”在 Web 挖掘技术出现之前就存在,但 Web 挖掘技术基于对用户行为和特征的大量细节性数据的挖掘分析,极大地拓展了“用户档案”的范围和应用,为更高水平的决策制定提供了可能^[10]。虽然隐私权问题、个性特征忽视问题、知识产权问题、信息安全问题等,在 Web 挖掘技术尚未出现时它们已经存在,但 Web 挖掘技术“新的能力”究竟对这些道德价值和道德准则产生了怎样的影响?这需要在 Web 挖掘技术的新语境下重新认识和解释。正如约翰逊指出的,将传统的伦理理论和伦理规范简单地应用于新的环境是一种机械的做法,抹杀了计算机伦理学(网络伦理学;信息伦理学)的任务和使命^[11]。

2.2 有法律保护私人信息,并且有在线的隐私声明保证隐私,还担心什么呢?

这一观点认为已有现成的法律和隐私声明,我们还担心什么呢?然而事实是,现有的法律往往很不完备,或者完全滞后于技术发展。比如,现有的隐私保护法律往往只提出保护个人信息不受侵犯和不当使用。那么当个人信息以匿名的形式被大量采集并且使用的时候,是否也涉及了用户的隐私权呢?ISP的在线隐私声明可保证隐私,这一点更是受到质疑。用户访问一些Web站点,但他们并没有足够精力去了解自己的访问记录等是否被收集或利用,以及这些访问记录将被如何使用。而且大多隐私声明中使用的陈述太模糊,人们并不能据此做出正确判断。比如人们常会在网站的隐私声明中见到这样的句子:“不会与第三方共享这些信息”。这里的第三方究竟是指谁呢?这些信息不被用作商业用途,就能代表用户的信息权利没有受到侵犯么?

2.3 很多用户往往选择放弃对自己隐私权的控制,可以把它们作为公共信息来使用

这一观点认为,人们有权拒绝透露自己的个人信息,他们可以通过控制自己与私人以及公共组织的关系来控制自己的隐私权;如果用户放弃自己对隐私权的控制,这些信息被当作公共信息来收集是没有问题的。然而对大多数人而言,放弃对他们自己隐私权的控制往往是迫不得已的,或者是经过比较权衡后放弃的。因为人们只要使用网络,就会在网上留下“记录”(注册信息和使用信息),隐私信息就有被侵犯的危险。在理想的情况下,用户控制其隐私权的最好方式是远离网络,而这显然是一种代价高昂且不切实际的选择。多数情况下,网站会通过其隐私声明告知用户在特定情况下有权拒绝透露个人信息,但对用户来说,由于阅读隐私声明并做出比较判断需要付出不小的信息成本,并降低其信息活动的效率,往往会选择以隐私权的放弃或部分放弃来保证其信息活动的有效进行。因此,认为用户放弃隐私权控制的信息就是公共信息的观点是难以成立的。而根据前面的讨论,即使个人数据在网上可以公开得到,在没有当事人同意的情况下,为特定目的对当事人的公开数据进行挖掘和使用,在道德上也是有问题的。

2.4 许多收集到的信息并不涉及具体的个人,而且常常是匿名的

我们常会听到这样的一种说法:“如果我们不在乎你是谁,又怎么会侵犯到你的隐私权呢?”Web挖掘中个人信息的匿名性常被用作对于Web技术伦理问题的辩护。然而如前所述,这种匿名性本身就已造

成了一定的伦理问题。在匿名的情况下,个人信息不再保持其完整性和独立性。而且Web挖掘对群体行为的揭示与特定的“个人数据”相结合,其不当应用也会对个人隐私造成侵犯。

2.5 Web挖掘技术往往可以提供更客观细节的信息,帮助决策

Web挖掘技术能为决策者提供更客观的基于大量原始数据的信息,从而为更加科学地制定决策提供可能。这些决策促使个性化服务的实现,往往使消费者享受更为便捷的服务。而对消费者而言,自己想要的信息不会出现在自己的视线范围内,不必为大量无用信息浪费时间。但是,这种说法只是考虑了Web挖掘这一技术的优势,并不代表Web挖掘技术带来的伦理问题不存在。

2.6 Web挖掘技术常用来定制个性化服务

识别用户的特定需求并对他们提供有针对性的个性化服务,是信息活动尤其是电子商务的重要追求,Web挖掘技术的应用使个性化定制服务实现的可能性大大增加。这也是对Web挖掘技术伦理问题辩护的一个理由。但如前所讨论的,我们不能以Web挖掘的技术优势去掩盖它潜藏的道德问题。

3 Web挖掘伦理问题的解决对策

3.1 制订伦理守则并予以明示

伦理守则(codes of ethics)是一种职业符合道德地运用专门化知识和实践职业技巧的一种承诺,体现了某一职业所积累的智慧。它在两个方面发挥作用:一方面,它告诉消费者和公众,其职业成员有确定的理想和价值并坚持特定的标准;另一方面,它告诉职业成员,什么是他们所期望的行为^[12]。对Web挖掘来说,由于ISP或厂商是Web挖掘行为的执行者,拥有和处理大量与用户相关的信息,他们作为行为人和占优势地位的一方,应制订伦理守则或专门的隐私政策,对其Web挖掘的行为标准和对用户可能产生的影响予以明示。具体来说,ISP或厂商在避免出现伦理问题以及伦理问题的解决上应做到:第一,在收集数据之前做出隐私声明,明示用户哪些个人信息将被收集以及这些信息将被用于何种用途。第二,对自己收集到的用户数据应妥善保管,恰当使用并及时向用户公示。第三,一旦发生伦理方面的纠纷,厂商应按照其伦理守则或隐私政策,采取积极措施予以解决。

3.2 运用技术手段予以保护

Web挖掘本身是一种技术系统,对Web挖掘伦理问题的解决自然应该诉诸技术的智慧。当技术被当做自律和防止信息失范行为的手段时,它同伦理原

则、伦理规范一样,成为信息伦理调控机制的一个组成部分。网络礼仪(Netiquette)技术、防杀病毒技术、加密技术、分级过滤技术、防火墙技术等与信息活动中的应用,即表明人们依靠技术手段践行伦理要求的一种努力。

国外现在已研制出 Privacy Enhancing Technique (PET) 软件^[13],如 Anonymous Remainers, Surf Anonymously, Cookie Busters 等。这些软件的功能大多在于帮助用户理解 ISP 的隐私政策,或以匿名的形式访问 Web 站点,以最大程度实现用户的个人信息不被 Web 挖掘工具挖掘到,保护用户个人信息。而在 Google 的数据采集礼貌问题中,为了避免对搜索的网站造成过大的负载,在 Crawler 中应采用轮询与延时的方法避免对单一服务器短时间内所带来的巨大冲击。从长远发展来说,制定 Web 挖掘软件或搜索引擎可以自动识别用户拒绝挖掘网页的拒绝挖掘标准(disallowing-mining standard),对用户和 Web 挖掘者双方来说,都是一种行之有效的技术解决方式。

3.3 用户对隐私权等信息权利的认真追究

例如,莲花公司曾宣布取消了一种叫做 Marketplace: Household 的涉及用户隐私的产品,即起因于始料不及而且相当大的公众舆论压力和集体追究^[14]。

目前往往存在这种情况:用户的信息权利意识薄弱,对 Web 挖掘可能对自身信息权利所产生的影响不够关注且知之甚少,多数情况下默认 Web 挖掘者对其个人信息的随意使用。因此培养用户的信息权利意识,提高其信息素质(information literacy),对 Web 挖掘伦理问题的预防和解决具有重要意义。例如,目前 Web 挖掘中常用到的 cookies 技术,一旦采用会自动跟踪用户信息,对用户的隐私构成威胁;而现在浏览器都允许用户切断第三方的 cookies,只要在自己的浏览器上做一设置,就可以避免个人信息权利被侵犯。更为重要的是,信息权利觉醒的用户群体会形成一种制衡力量,他们对商家造成恶性后果的个人信息滥用进行集体追究或起诉,会对商家的 Web 挖掘行为形成一种制约,使二者在动态均衡的博弈中推进 Web 挖掘符合伦理地应用和发展。

3.4 管理部门加强监管

管理部门对 Web 挖掘中伦理问题的监管主要从 3 个方面进行:一是制定有关政策和行业行为标准(守则),促使 ISP 或厂商建立伦理管理制度,通过践行伦理守则提高道德建设水平。例如,近年来在国外企业中出现了首席隐私官(Chief Privacy Officer, CPO)这一新的职位,专门负责制定企业隐私政策,解决企业所面临的隐私问题^[15]。二是建立社会伦理监

督制度,接受举报并协调多种力量进行处理。例如,欧盟建立的热线制度就是一种重要的社会伦理监督制度,用户和服务商可向热线举报他们发现的有害信息,热线接到举报后展开调查并与 ISP 和警方合作,将有害信息迅速处理^[16]。这对 Web 挖掘的伦理监管也具借鉴价值。三是加强网络伦理和信息素质教育,在社会层面提高全体成员的信息道德水平和信息能力,促进 Web 挖掘等技术的合理利用和健康发展。

参考文献

- 1 董金华. 关于隐私权的理论探析. 理论探索, 2004(6)
- 2,6 沙勇忠. 信息伦理学. 北京:北京图书馆出版社, 2004
- 3 V. D. Bush, Beverly T. Venable, Alan J. Bush. Ethics and Marketing on the Internet: Practitioners' Perceptions of Societal, Industry and Company Concerns. Journal of Business Ethics, 2000, 23: 237 ~ 248
- 4,9 Lita van Wel, Lambert Royakkers. Ethical issues in Web data mining. Ethics and Information Technology, 2004, 6: 124 ~ 140
- 5 H. Nissenbaum. Toward an Approach to Privacy in Public: Challenges of Information Technology. Ethics & Behavior, 1997, 7(3)
- 7 薛虹. 网络时代的知识产权法. 北京:法律出版社, 2000
- 8 理查德·A·斯皮内洛著;刘钢译. 世纪道德——信息技术的伦理方面. 北京:中央编译出版社, 1999
- 10 B. A. Stead, J. Gilbert. Ethical Issues in Electronic Commerce. Journal of Business Ethics, 2001, 34: 75 ~ 85
- 11 J. F. Johnson. Immunity from the Illegitimate Focused Attention of Others: An Explanation of our Thinking and Talking about Privacy. In A. Vedder, editor, Ethics and the Internet, Intersentia, Antwerpen Groningen Oxford, 2001: 49 ~ 70
- 12 D. Johnson. Do IT Professionals Need a Code of Ethics. Business World, 1999, 8: 16 ~ 21
- 13 Herman T. Tavani. KDD, data mining, and the challenge for normative privacy. Ethics and Information Technology, 1999, 1: 265 ~ 273
- 14 R. O. Mason, etc. Ethics of information Management. Thousand Oaks, California: Sage publications, Inc. 1995
- 15 The Rise of the Chief Privacy Officer. http://www.business-week.com/careers/content/dec2000/ca20001214_253.htm
- 16 Internet Hotline providers in Europe Association. <http://europa.eu.int/ISPO/iap/projects/inhope.html>

沙勇忠 兰州大学管理学院副教授、博士、副院长。通信地址:兰州市。邮编 730000。

黄丽珺 兰州大学管理学院情报学专业硕士研究生。通信地址同上。

(来稿时间: 2006-08-21)