大数据时代如何保护隐私

数据挖掘无法避免 保护隐私必须多管齐下

在网络上，每个人都会或多或少，或主动或被动地泄露某些碎片信息。这些信息被大数据挖掘，就存在隐私泄露的风险，引发信息安全问题。面对汹涌而来的 5G 时代，大众对自己的隐私保护感到越来越迷茫，甚至有点不知所措。那么，你的隐私，大数据是怎么知道的呢？大家又该如何自我保护呢？

本版均资料图片

大数据啥都知道

大数据时代，每个人都有可能成为安徒生童话中那个“穿新衣”的皇帝。在大数据面前，你说过什么话，它知道；你做过什么事，它知道；你有什么爱好，它知道；你生过什么病，它知道；你家住哪里，它知道；你的亲朋好友都有谁，它也知道……总之，你自己知道的，它几乎都知道，或者说它都能够知道，至少可以说，它迟早会知道！

甚至，连你自己都不知道的事情，大数据也可能知道。例如，它能够发现你的许多潜意识习惯：集体照相时你喜欢站哪里呀，跨门槛时喜欢先迈左脚还是右脚呀，你喜欢与什么样的人打交道呀，你的性格特点都有什么呀，哪位朋友与你的观点不相同呀……

再进一步说，今后将要发生的事情，大数据还是有可能知道。例如，根据你“饮食多、运动少”等信息，它就能够推测出，你可能会“三高”。当你与许多人都在独立地购买感冒药时，大数据就知道：流感即将暴发了！其实，大数据已经成功地预测了包括世界杯比赛结果、股票的波动、物价趋势、用户行为、交通情况等。

当然，这里的“你”并非仅仅指“你个人”，包括但不限于，你的家庭，你的单位，你的民族，甚至你的国家等。至于这些你知道的、不知道的或今后才知道的隐私信息，将会把你塑造成什么，是英雄还是狗熊？这却难以预知。

数据挖掘就像“垃圾处理”

什么是大数据？形象地说，所谓大数据，就是由许多千奇百怪的数据，杂乱无章地堆积在一起。例如，你在网上说的话、发的微信、收发的电子邮件等，都是大数据的组成部分。在不知道的情况下被采集的众多信息，例如被马路摄像头获取的视频、手机定位系统留下的路线图、驾车的导航信号等被动信息，也都是大数据的组成部分。还有，各种传感器设备自动采集的有关温度、湿度、速度等万物信息，仍然是大数据的组成部分。总之，每个人、每种通信和控制类设备，无论它是软件还是硬件，其实都是大数据之源。

大数据利用了一种名叫“大数据挖掘”的技术，采用诸如神经网络、遗传算法、决策树、粗糙集、覆盖正例排斥反例、统计分析、模糊集等方法挖掘信息。大数据挖掘的过程，可以分为数据收集、数据集成、数据规约、数据清理、数据变换、挖掘分析、模式评估、知识表示等八大步骤。

不过，这些听起来高大上的大数据产业，几乎等同于垃圾处理和废品回收。

这并不是在开玩笑。废品收购和垃圾收集，可算作“数据收集”；将废品和垃圾送往集中处理场所，可算作“数据集成”；将废品和垃圾初步分类，可算作“数据规约”；将废品和垃圾适当清洁和整理，可算作“数据清理”；将破沙发拆成木、铁、布等原料，可算作“数据变换”；认真分析如何将这些原料卖个好价钱，可算作“数据分析”；不断总结经验，选择并固定上下游卖家和买家，可算作“模式评估”；最后，把这些技巧整理成口诀，可算作“知识表示”。

再看原料结构。大数据具有异构特性，就像垃圾一样千奇百怪。如果非要在垃圾和大数据之间找出本质差别的话，那就在于垃圾是有实体的，再利用的次数有限；而大数据是虚拟的，可以反复处理，反复利用。例如，大数据专家能将数据(废品)中挖掘出的旅客出行规律交给航空公司，将某群体的消费习惯卖给百货商店等。总之，大数据专家完全可以“一菜多吃”，反复利用，而且时间越久，价值越大。换句话说，大数据是很值钱的“垃圾”。

没有尽头的大数据挖掘

大数据挖掘，虽然能从正面创造价值，但是也有其负面影响，即存在泄露隐私的风险。隐私是如何被泄露的呢？这其实很简单，我们先来分解一下“人肉搜索”是如何侵犯隐私的吧！

一大群网友，出于某种目的，利用自己的一切资源渠道，尽可能多地收集当事人或物的所有信息；然后，将这些信息按照自己的目的提炼成新信息，反馈到网上与别人分享。这就完成了第一次“人肉迭代”。

接着，大家又在第一次人肉迭代的基础上，互相取经，再接再厉，交叉重复进行信息的收集、加工、整理等工作，于是，便诞生了第二次“人肉迭代”。如此循环往复，经过多次不懈迭代后，当事人或物的画像就跃然纸上了。如果构成“满意画像”的素材确实已经证实，至少主体是事实，“人肉搜索”就成功了。

几乎可以断定，只要参与“人肉搜索”的网友足够多，时间足够长，大家的毅力足够强，那么任何人都可能无处遁形。

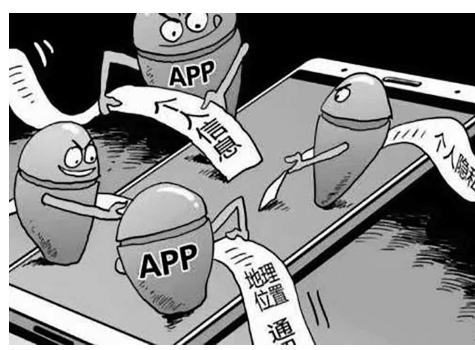
其实，所谓的大数据挖掘，在某种意义上说，就是由机器自动完成的特殊“人肉搜索”而已。只不过，这种搜索的目的，不再限于抹黑或颂扬某人，而是有更加广泛的目的，例如，为商品销售者寻找最佳买家、为某类数据寻找规律、为某些事物之间寻找关联等。总之，只要目的明确，那么，大数据挖掘就会有用武之地。

如果将“人肉搜索”与大数据挖掘相比，网友被电脑所替代；网友们收集的信息，被数据库中的海量异构数据所替代；网友寻找各种人物关联的技巧，被相应的智能算法替代；网友们相互借鉴、彼此启发的做法，被各种同步运算所替代。

各次迭代过程仍然照例进行，只不过机器的迭代次数更多，速度更快，每次迭代其实就是机器的一次“学习”过程。网友们的最终“满意画像”，被暂时的挖掘结果所替代。之所以说是暂时，那是因为对大数据挖掘来说，永远没有尽头，结果会越来越精准，智慧程度会越来越高，用户只需根据自己的标准，随时选择满意的结果就行了。

当然，除了相似性外，“人肉搜索”与“大数据挖掘”肯定也有许多重大的区别。例如，机器不会累，它们收集的数据会更多、更快，数据的渠道来源会更广泛。总之，网友的“人肉搜索”，最终将输给机器的“大数据挖掘”。

隐私到底该如何保护



必须承认，就当前的现实情况来说，大数据隐私挖掘的“杀伤力”，已经远远超过了大数据隐私保护的能力；换句话说，在大数据挖掘面前，当前人类有点不知所措。这确实是一种意外。自互联网诞生以后，在过去几十年，人们都不遗余力地将碎片信息永远留在网上。其中的每个碎片虽然都完全无害，可谁也不曾意识到，至少没有刻意去关注，当众多无害碎片融合起来，竟然后患无穷！

不过，大家也没必要过于担心。在人类历史上，类似的被动局面已经出现过不止一次了。从以往的经验来看，隐私保护与数据挖掘之间总是像“走马灯”一样轮换的——人类通过对隐私的“挖掘”，获得空前好处，产生了更多需要保护的“隐私”，于是，不得不再回过头来，认真研究如何保护这些隐私。当隐私积累得越来越多时，“挖掘”它们就会变得越来越有利可图，于是，新一轮的“挖掘”又开始了。历史地来看，人类在自身隐私保护方面，整体处于优势地位，在网络大数据挖掘之前，“隐私泄露”并不是一个突出的问题。

但是，现在人类需要面对一个棘手的问题——对过去遗留在网上的海量碎片信息，如何进行隐私保护呢？单靠技术，显然不行，甚至还会越“保护”，就越“泄露隐私”。

因此，必须多管齐下。例如从法律上，禁止以“人肉搜索”为目的的大数据挖掘行为；从管理角度，发现恶意的大数据搜索行为，对其进行必要的监督和管控。另外，在必要的时候，还需要重塑“隐私”概念，毕竟“隐私”本身就是一个与时间、地点、民族、文化等有关的约定俗成的概念。

对于个人的网络行为而言，在大数据时代，应该如何保护隐私呢？或者说，至少不要把过多包含个人隐私的碎片信息遗留在网上呢？答案只有两个字：匿名！只要做好匿名工作，就能在一定程度上，保护好隐私了。也就是说，在大数据技术出现之前，隐私就是把“私”藏起来，个人身份可公开，而大数据时代，隐私保护则是把“私”公开（实际上是没法不公开），而把个人身份隐藏起来，即匿名。

（来源：光明日报）