“十二五”科学技术专著丛书

# 网络隐私保护与信息安全

康海燕 著

NETWORK PRIVACY PROTECTION
AND INFORMATION SECURITY

北京邮电大学出版社
www.buptpress.com

“十二五”科学技术专著丛书

# 网络隐私保护与信息安全

康海燕　著

北京邮电大学出版社
www.buptpress.com

## 内 容 简 介

数据共享将带来巨大收益，然而，数据中的个人隐私泄露和信息安全将面临严峻挑战。本文围绕隐私保护数据发布(PPDP)的几个关键技术：隐私保护模型、匿名技术、差分隐私、基于隐私保护技术的应用、信息度量标准与算法、网络隐私保护的策略等问题展开讨论，主要涉及的技术和知识包括信息论、控制论、系统论、博弈论、管理学、数学、计算机安全、数据库系统、信息检索、数据挖掘、密码学、统计学、分布式处理和社会科学等，所以具有挑战性。本书内容的特点：理论和实践相结合。

**图书在版编目(CIP)数据**

网络隐私保护与信息安全 / 康海燕著. -- 北京：北京邮电大学出版社，2016.1

ISBN 978-7-5635-4621-3

Ⅰ. ①网… Ⅱ. ①康… Ⅲ. ①计算机网络—信息安全—安全技术 Ⅳ. ①TP393.08

中国版本图书馆 CIP 数据核字(2015)第 305963 号

---

**书　　名**：网络隐私保护与信息安全

**著作责任者**：康海燕　著

**责 任 编 辑**：徐振华　孙宏颖

**出 版 发 行**：北京邮电大学出版社

**社　　址**：北京市海淀区西土城路 10 号(邮编：100876)

**发 行 部**：电话：010-62282185　传真：010-62283578

**E-mail**：publish@bupt.edu.cn

**经　　销**：各地新华书店

**印　　刷**：北京九州迅驰传媒文化有限公司

**开　　本**：787 mm×1 092 mm　1/16

**印　　张**：11.75

**字　　数**：308 千字

**版　　次**：2016 年 1 月第 1 版　2016 年 1 月第 1 次印刷

---

ISBN 978-7-5635-4621-3　　定　价：29.80 元

# 前　言

21世纪是信息科学与技术极速发展的时代，信息成为一种重要的战略资源，信息的获取、存储、处理及安全保障能力成为一个国家的综合国力的重要组成部分。数据开放要走之路，让不同领域的数据真正流动、融合起来，才能释放大数据的价值。但是，原始的数据形式通常包含个人的敏感信息，发布这些数据会侵犯个人的隐私。现今各行各业收集数据的能力大大提升，随之为基于知识和信息的决策提供了广泛的机会。在利益或者规章的驱动下，不同的群体之间都有数据交换和共享的需求。然而，数据的收集、发布和分析（挖掘）要面对的一个重要问题是隐私泄露和信息安全，即在前网络时代，隐私在法律、政府、组织、个人的多重保护下，是相对安全的，而网络的出现，令现实社会中个人隐私权的有关问题延伸到了网络空间，由于网络社会的开放特征，使得个人隐私面临着严重的威胁。

数据发布的现行做法主要依赖于相关的政策法规和去标识符的简单处理。这种简易处理方法可能导致大量数据缺乏足够多的保护。为此，相关专家和科技工作者正在积极开展相关研究，开发数据发布隐私保护（PPDP）的有效方法，寻找保护数据隐私的同时提高数据的实用性的平衡。数据隐私保护已经成为一个新兴的、非常热门的研究领域，并且针对不同的数据发布场景提出了许多方法。

本书主要介绍新兴的数据隐私保护研究领域的产生背景、基础知识（当前隐私问题、隐私法律、隐私保护模型、数据匿名化、统计数据库、隐私保护数据分析、社交网络隐私等）、隐私保护技术、实现方法、商业应用、最新研究成果和进展。研究数据实际发布过程中遇到的挑战，明确PPDP和其他相关问题的不同以及区分这些差异所需的要求，并对今后的研究方向提出建议。

从构思到撰写的过程中，得到多位老师的鼓励和支持，感谢埃默里大学（Emory University）隐私保护研究小组的支持。感谢北京市优秀人才培养资助项目（2013E005007000001）、国家自然科学基金项目（61272513）、教育部人文社会科学青年基金项目（11YJC870011）的支持。

感谢我们研究生团队成员的共同努力，大家一起讨论的日子里，获益甚多。

感谢参与实验和撰写的研究生：孟祥、于东、刘建昆、苑晓姣、李清华、赵格。

感谢本文引用参考文献的所有作者，他们的工作和研究成果给了我极大帮助和启发，是他们刻苦钻研和辛勤工作的成果成就了隐私保护和信息安全这片天地。书中有些基本概念和基础知识已经比较成熟，为避免基础知识和基本概念的歧义，书中引述了许多著名学者和专家相关论文的内容。其中大部分已征得相关专家的认可，但由于各种原因仍有大量引述未能当面征求原著者意见，书中已尽量做出明确标注。在此对这些作者包括所有参考文献作者表示致意和衷心感谢，未尽事宜，敬请谅解！

由于作者水平有限，书中疏漏在所难免，诚请各界学者、专家和读者批评指正。

# 目　录

# 第1章 绪 论

据科学家预测，继实验科学、理论科学、计算机科学之后，数据密集型科学将成为人类科学研究的第四个范式(图灵奖获得者 Jim Gray 提出)。以大数据为代表的数据密集型科学将成为新技术变革的基石。大数据正在改变着世界，大数据已成为继云计算之后信息技术领域的另一个信息产业增长点。大数据“宝藏”将成为未来的新石油。但是大数据的共享和分析，增加了用户敏感数据泄露的风险，全国政协委员李晓明曾指出：“个人隐私保护是大数据时代的重要民生。”个人隐私保护正成为制约大数据发展的瓶颈，在网络时代，我们还没有找到一种方法可以完美守护我们的隐私，在这里，我们一起探讨。

## 1.1 研究背景

信息技术的迅速发展和互联网使用范围的扩大，更先进的信息采集、保存、共享和比较技术的出现，电子商务企业和政府部门对个人信息的大量收集和处理，为企业和国家带来了宝贵的知识与物质财富，与此同时，若不正确使用这些技术，将对个人隐私和数据安全构成威胁。隐私(Privacy)是一个逐渐为人们熟知和关注的话题。在一个特定的环境和时间点中，相对静态的隐私应该如何处理？

一个实例：Google 官司牵出公民隐私之忧。2005 年 8 月，美国司法部以打击网上黄色犯罪为由，要求美四大网络公司——美国在线、Microsoft、Yahoo、Google——提供有关网络搜索的数据信息(其中包括随机选择的网址和用户检索结果的数据)以协助调查，对于政府要求，除 Google 以外的 3 家很快满足，唯独 Google 坚决加以抵制，理由：①这样将侵犯用户隐私权，损害 Google 和用户建立的互信关系；②泄露公司搜索服务的商业秘密。Google 创始人塞尔吉·布林(Sergey Brin)表示保护隐私是 Google 的义务。2006 年，司法部将 Google 公司告上法庭。Google 与政府间的法律纠纷引发了关于因特网安全和公民网络隐私权的争议，网络隐私权的保护成为人们关注的焦点。因此，Google 与司法部的官司被认为是互联网时代美国网络公司与政府围绕隐私权问题爆发的“世纪大战”。“大战”结果：Google 抗争后占得上风。由于在舆论的压力面前，司法部只好作出重大让步，在庭审上，仅要求 Google 提交同用户搜索相关的 5 万个网址以及近 5 000 个搜索项，并承诺只对其中的 1 万个网址和 1 000 个搜索项进行研究。与最初的要求相比，司法部要求 Google 提供的信息量几乎缩小了 99.99%。而最终司法部连 5 000 个搜索项的要求也被拒绝了。Google 的代理辩护律师尼科·翁在公司网站上发表声明称：“裁决表明，无论是政府机构，还是其他任何人，在要求互联网公司提交数据时都没有特权。”一些分析人士认为，裁决对于 Google 公司以及隐私权保护者而言是一个巨大的胜利，Google 的维权行动将给美国的因特网管理规范带来新的启示。不过，围绕因特网安全和公民

隐私的争议并没有停止。

互联网的匿名性保护了用户的信息和网络使用安全，曾经网络上流行的一句话："On the Internet, nobody knows you're a dog"（互联网上没有人知道你是一条狗，如图 1-1 所示）。这是针对网络的虚拟性、匿名性所作的颇有几分夸张的描述。网络确实改变了过去那种社会交往与控制的模式，给人们创造了前所未有的信息空间。然而我们也常常能在各种媒体里面了解到发生在互联网上的侵犯隐私的恶性事件。当前对用户的隐私威胁最大的不是用于跟踪用户的 Cookie、间谍软件和用户浏览行为分析网站，而是我们日常使用的搜索引擎。大部分搜索引擎在用户使用其服务时，都会记录用户的 IP 地址、搜索的关键词、从搜索结果中跳转到哪个网站等信息，通过数据挖掘等技术，搜索服务商可以从这些信息中获得用户的身份、用户的爱好以及在网上的行为等隐私信息，并可能使用这些隐私信息进行商业活动。也就是说，今天在网络上不仅有人知道你是一条狗，而且还知道你是一条猎犬还是一条牧羊狗。例如，2008 年的央视 3·15 晚会揭露了一条重大消息：分众无线传媒技术有限公司（分众传媒子公司）掌握了中国 5 亿多手机用户中一半的手机用户信息。该公司对机主的信息进行详尽分类，精确到机主的性别、年龄、消费水平等，以"精确"发送垃圾短信，其中，仅郑州分众无线传媒技术有限公司的短信日发送量就达 2 亿条（仅仅一个企业，掌握了 2 亿多人的个人信息，如此令人骇然的现实印证了公众长期以来对个人信息保护的担忧）。无论是从 2008 年明星们的"艳照门"，2009 年的"艾滋女"闫德利事件，还是 2010 年汽车模特的"兽兽门"，一件件个人隐私信息被泄露后在网上掀起滔天巨浪的事件此起彼伏。而因这些信息泄露而遭到感情上的巨大创伤，更揭示出个人隐私被泄露传播已经成为一种专业化、规模化、商业化的运作，足以引起每个人的重视。

"On the Internet, nobody knows you're a dog."

图 1-1　互联网上没有人知道你是一条狗

例如，Netflix 是一种流行的在线电影租赁服务，为了提高电影推荐的准确性，最近公布了

一份包含 50 万用户电影爱好程度的数据集[2]，他开放了匿名的评论以及打分的信息，但是有人把它跟国际电影数据库 IMDB 匹配，结果把一个有同性恋倾向的人识别了出来，被告了。在美国马萨诸塞州，集体保险委员会（Group Insurance Commission，GIC）负责为州政府雇员购买健康保险。截止到 2002 年，GIC 已经收集到约 135 000 政府雇员及其家人的健康数据。由于这些数据被认为是匿名的，GIC 将这些数据的副本转给研究机构，并将另一份副本卖给商业公司。搜索引擎的查询日志可以进行用户行为分析，分析结果可以有效改进网络信息检索技术。但上述收集的数据包含大量用户敏感信息，如果数据发布者将这些原始数据直接进行发布，会泄露用户敏感信息及身份信息，如图 1-2 所示。


图 1-2　数据泄露与隐私保护

在网络所带来的隐私权问题当中，一个关键的问题就是有关个人数据的权利问题。所谓个人数据，是指用来标识个人基本情况的一组数据资料，如姓名、年龄、性别、地址、社保号、信用卡号、驾照号、手机号、出生年月、收入、职业、个人爱好、银行资料以及受法律法规保护的数据等。具体而言，个人数据主要包括标识个人基本情况、标识个人生活和工作经历等情况、与网络有关的个人信息。主要包括以下 4 个方面的信息。①个人登录的身份、健康状况。网络用户在申请上网开户、个人主页、免费邮箱以及申请服务商提供的其他服务（购物、医疗、交友等）时，服务商往往要求用户登录姓名、年龄、住址、居民身份证编号、工作单位等身份和健康状况，服务商有义务和责任保守个人秘密，未经授权不得泄露。②个人的信用和财产状况，包括信用卡、电子消费卡、上网卡、上网账号和密码、交易账号和密码等。个人在网上消费、交易时，登录和使用的各种信用卡、账号均属个人隐私，不得泄露。③邮箱地址。邮箱地址同样是个人隐私，用户大多数不愿将之公开。掌握、搜集用户的邮箱并将之公开或提供给他人，致使用户收到大量的广告邮件、垃圾邮件或遭受攻击而不能正常使用，使用户受到干扰，显然也侵犯了用户的隐私权。④网络活动踪迹。个人在网上的活动踪迹，如 IP 地址、浏览踪迹、活动内容，均属个人的隐私。

思考：实名制下的隐私如何保证？实名制，尤其是网络实名制，在虚拟网络世界里以真实身份存在，规范了言行，注重了责任，也增大了个人信息泄露的可能性。如存款实名制、火车票实名制、手机实名制、快递业实名制、微博实名制，可以有效降低因身份虚拟造成的欺诈等现象的发生概率，对于加强监管、保障安全具有明显的作用。同时，个人信息保护问题也成为公众关注的焦点。

## 1.2　研究意义

随着信息技术的快速发展，社会正在收集大量数据（包括个人数据），这些数据无论对研究者、商业，还是对个人性化服务均有相当高的价值，如图 1-3 所示。因此这些数据的共享成为必然，但又给用户（个人、团体机构）的隐私泄露带来巨大的风险。我们的研究目标：最小化隐私泄露风险的同时，最大化数据的实用性。通过研究数据的隐私保护发布，具有以下几个方面的意义。此节主要侧重介绍面向大数据中个性化服务的隐私保护的研究意义。

(1) 解决大数据中个性化服务(检索)的隐私保护问题

个性化服务(检索)为提升搜索引擎查询结果的高质量服务(QoS) 提供了保障。以用户为中心的信息检索,需要收集和集成大量的用户资料/用户兴趣模式(User Profile,来自个人信息、个人兴趣和检索历史等),精确描述用户的个性特征和个性模型,研究用户的行为,理解他们的主要需求,根据这些需求改进和完善检索系统的组织和操作,向用户主动、及时、准确地提供所需信息,如 Google 和 Yahoo[3]。

然而,个性化服务(检索)面临着一个重要挑战:个性化数据隐私保护和信息安全。如何在成功进行高质量检索服务的同时保证隐私数据不被泄露是一个重要的课题。近几年隐私保护数据发布(Privacy-Preserving Data Publishing, PPDP)受到了广泛关注。隐私(Privacy)逐渐为人们熟知和关注。在前网络时代,隐私在法律、政府、组织、个人的多重保护下是相对安全的,而网络的出现,令现实社会中个人隐私权的有关问题延伸到了网络空间,由于网络社会的开放特征,使得个人隐私面临严重威胁。用户资料是网络公司最有价值的信息,因此,2010 年,国内有 360 和 QQ 之战,国外有 Facebook,Google,Wikileak 的隐私争论。如何保护用户个人信息的绝对安全成为最受关注的问题。隐私保护数据发布已经成为一个新兴的、热门的研究领域[2],并且针对不同的数据发布场景提出了许多方法。本研究将隐私保护技术应用于个性化 Web 检索中,具有重要的现实意义。

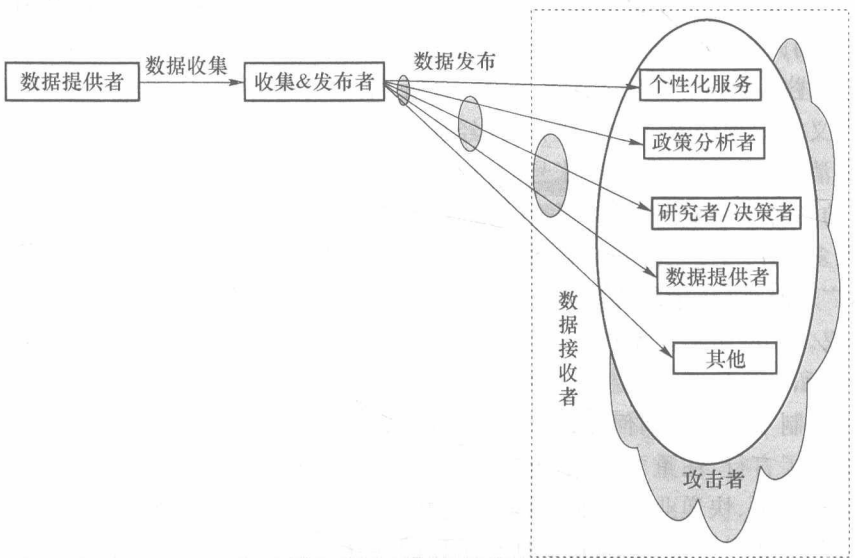


图 1-3 数据发布的意义

(2) 有助于新一代搜索引擎的健康发展(有效地解决了搜索引擎带来的隐私威胁)

网络确实改变了过去那种社会交往与控制的模式,给人们创造了前所未有的信息空间。然而,我们也常常能在各种媒体里面了解发生在互联网上的侵犯隐私的恶性事件。当前对用户的隐私威胁最大的不是用于跟踪用户的 Cookie、间谍软件和用户浏览行为分析网站,而是我们日常使用的搜索引擎。大部分搜索引擎在用户使用其服务时,都会记录用户的 IP 地址、搜索的关键词和跳转链接等信息。搜索服务商通过数据挖掘等技术可以从这些信息中获得用户的身份、爱好以及网上行为等隐私信息,并可能使用这些隐私信息进行商业活动。也就是说,今天在网络上不仅有人知道你是一条狗,而且还知道你是一条猎犬还是一条牧羊狗,如图

1-4 所示。

搜索引擎在我们的网络活动中正扮演着越来越重要的角色。那么如何保护用户在使用搜索引擎服务时的隐私？因此，结合隐私保护是成为个性化 Web 服务（检索）发展的一个必然趋势，其目的是研究新的算法或协议使得在不共享各方原始数据的情况下，进行高质量的信息检索。

图 1-4　互联网上总有人在监视

（3）避免隐私泄露，挽回经济损失

网络隐私问题已成为社会各界关注的焦点，并严重威胁着网络社会与网络经济的健康发展。2010 年 8 月 21 日，中国计算机学会青年计算机科技论坛“互联网上的个人隐私泄露与保护”报告会在北京举行。国家网络信息安全技术研究所所长、国家计算机网络应急技术处理协调中心副总工程师杜跃进在报告会上作了“互联网数据窃取威胁”的主题演讲。利益的驱动使数据窃取形成产业链，我国面临着严重的网络安全问题。由于受巨额利润的驱使，黑客产业链正在形成，利用木马程序、僵尸网络盗取银行账号密码、游戏装备，窃取个人隐私等网络犯罪行为让人防不胜防。互联网数据窃取给我国造成了巨大的经济损失，据 2006 年统计，我国网络攻击地下产业链的产值超过 2.38 亿元，每年给中国大陆造成的经济损失至少 76 亿元，目前可能会超过千亿元。

数据泄露代价高昂。据波士顿咨询公司（Boston Consulting Group，BCG）的一项调查显示：隐私比成本、易用性和安全性等更为用户所关心。据木星通信公司估计，2002 年这种对网络隐私的担心造成了高达 180 亿美元的经济损失。

波尼蒙研究所（Ponemon Institute）近期（2012 年）的一项调查显示，机构每丢失一条信息，估计将蒙受平均 200 美元的损失；每遭受一次全面的数据泄露，平均损失将高达 680 万美元。但这还仅仅只是金钱上的损失，如果把因数据泄露而导致的公司竞争力消失、收益下滑、诉讼缠身、声誉受损等问题也考虑在内，那么实际代价将更加难以估量。

（4）有助于防止国家机密的泄露

除了对个人隐私构成潜在威胁，个性化服务中对海量信息的搜集、存储、共享和挖掘在一

定程度上还存在泄露国家机密、军事部署等核心安全信息的可能。攻击者利用个性化服务中所能搜集到的海量信息、个人用户信息和其他的一些数据信息，进行各种关联，可通过这些信息归纳出机密信息，例如，当前我国国防科研的目标和进展、大型国家上市公司的市场计划，甚至国家的军队部署等，从而对国家经济和国家机密造成极为严重的危害。因此，开展隐私保护技术研究对国家安全也具有重大意义。

(5) 有助于完善和推进信息安全技术发展

传统的信息安全技术(如身份认证、访问控制、加密、审计和入侵检测等)已经经过多年的发展，日趋成熟，对于数据挖掘过程中的隐私保护也是近些年学术界中研究的热点和焦点，在近年顶级的学术期刊和会议上均有不少相关工作发表，但针对个性化服务这一新兴而充满希望领域中的隐私保护技术研究还未真正展开，因此，研究隐私保护技术有助于完善和推进信息安全技术的发展。

(6) 为数据发布机构提供可靠的数据隐私保护技术

隐私是个人、团体或机构有权控制他们的信息何时、如何及何种程度被他人共享。检索可以被解释为代表一个人的意图和对信息的需求，并由此揭露出这个人的大量个人信息。例如，大多数用户对色情、谋杀等内容进行过搜索，这些搜索活动很容易被收集下来并有可能被公布，会引起搜索用户的疑虑，从而威胁到用户的隐私。

本项研究将完善隐私保护在数据中的应用，可以使数据机构更加快速、安全地发布数据，供社会团体、研究机构研究分析，由此增加数据利用价值，并且保证了用户的隐私不被泄露。解决了数据发布和分析过程中隐私泄露问题。通过对公开给外界的数据进行隐私保护，可以有效地将其中的敏感信息隐藏起来，最大限度地保护数据库中的个人隐私。发达国家越来越重视对社会网络数据中的个人信息的隐私保护，通过法规、标准手段加以保障，逐步形成了横跨立法、行政、司法的完整的信息安全管理体系。在美国，个人医疗记录属隐私，外人打听不到；学生的学习成绩也是隐私，老师不会将其公布。如果隐私被人侵犯，造成了精神或物质上的损害，美国人就会诉诸法律。这主要得益于美国与隐私保护相关的立法比较成熟。美国国会 1974 年通过了《隐私权法》，这是美国保障公民个人信息的最重要的基本法律。之后，又有《财务隐私权法》《联邦电子通信隐私权法》等不断补充进来。此外，美国各州还制定了一些保护本州公民隐私的细化法律。随着信息技术的发展，美国联邦政府以及各州也不断出台新法、“升级”老法，以保护公民隐私。

(7) 解决制约电子商务发展的最大障碍——网络隐私问题

2012 年，Check Point 软件技术有限公司和 YouGov 对 2 000 多位英国人的调查显示，因为过去 5 年里个人数据的不断被泄露和破坏，50%的受访者表示降低了对政府和公共部门的信任，44%的受访者表示降低了对私人企业的信任。Check Point 英国董事总经理 Terry Greer-King 指出：“公共和私人机构在过去 5 年中，数据泄露事件数量增长了十倍。因此，公众对机构处理数据安全的能力大失信任，而主动选择与未发生泄露的公司合作。这充分说明了机构保护敏感数据，防止数据落入他人之手的重要性”，如图 1-5 所示。

网络隐私问题和缺乏信任是网上交易增长的最大障碍。互联网产业是建立在企业和消费者的互信互利的基础上，隐私是信任最重要的组成部分，除非 Web 组织可以有效地解决隐私问题，否则他们极有可能将失去消费者的信任，从而使交易大打折扣。用户对隐私的极大关注和对互联网行业发展的关心，致使我们一定要注意保护数据的隐私。

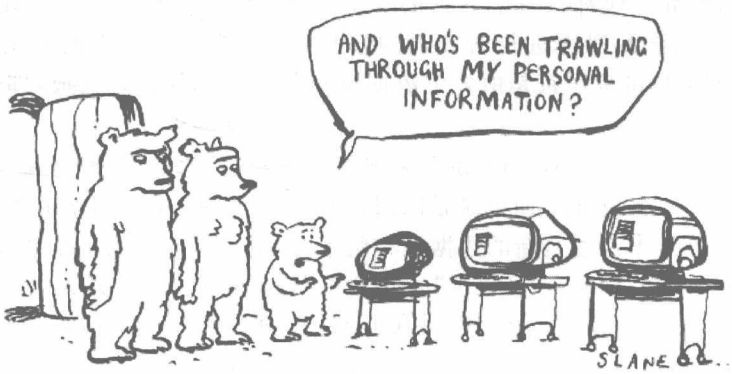


图 1-5 谁在查阅我的个人信息

## 1.3 国内外研究现状及发展动态分析

2001 年以来，隐私数据的发布得到重视和研究。2010 年，从 Facebook 到维基解密，再到基于 Google 街景的应用“I Can Stalk U”，使在线数据隐私问题成为全民关注的热点。在一个最新的调查中[2]，隐私保护数据发布绝大部分的工作都致力结构化或列表式数据。数据匿名的目标之一是设计一种隐私模型，如 $k$-anonymity、$l$-diversity、$t$-closeness、$m$-invariance 和 $\epsilon$-differential等模型，它们给出一个标准用于判断发布数据集(Published Dataset)是否提供了足够的隐私保护[4-15]。绝大多数实用模型都考虑攻击的具体类型(特定攻击)和假设攻击者有有限的背景知识。如 P. Samarati 和 L. Sweeney [4-6]提出了 $k$-匿名模型($k$-anonymity)，它要求发布表中的每个元组都至少与其他 $k-1$ 个元组在准标识属性上完全相同，能防止身份暴露(常导致属性暴露)。A. Machanavajjhala 等[7]进一步提出了 $l$-多样化模型($l$-diversity)，它要求每个 QI 分组中至少包含 $l$ 个不同的敏感属性取值，这一模型扩展了 $k$-匿名模型，对敏感属性的多样化提出了要求，并对敏感属性的多样化给出多种不同的解释。这一模型能防止直接的敏感属性泄露。Li Ninghui 等[8]提出的 $t$-接近模型($t$-closeness)，它要求每个等价类的敏感值的分布要接近于原始数据表中敏感属性的分布，这一模型能防止直接的敏感属性泄露。上述模型都是首先删除身份标识属性，然后对准标识属性进行概化。文献[10]证明最优的 $k$-匿名问题是 NP 难问题，由此引发人们设计出许多有效实现 $k$-匿名的启发式和近似算法[7,10,21]。同时，研究者们提出了若干关于 $k$-匿名的变异算法。G. Aggarwal 等利用聚类实现 $k$-匿名[9]。R. C. W. Wong 等[16]提出了$(\sigma,k)$-匿名模型，它在 $k$-匿名模型的基础上，要求每个 QI 分组中每个敏感属性的取值频率不能超过给定的值 $\sigma$。Xiao Xiaokui[17]提出了一种个性化的匿名模型，在分析隐私泄露的概率时，他们区分了两种情况：一种称为主键场景，其中单一个体对应的元组不能超过一个；另一种称为非主键场景，其中单一个体可以对应任意多个元组。

许多模型假设存在两种类型的属性：准标识属性和敏感属性。Wang Ke 等[18]提出了一种属性同时含有敏感值和标识值的框架。以上隐私模型是针对特定攻击进行保护和假设攻击者有有限的背景知识而构建的，大量的算法是为了满足以上某一种模型而改变数据集。如概化、抑制(去除)、排列和扰乱等[10, 20-24]。差分隐私(Differential Privacy)[19]是在任意知识背景下能保证隐私安全的观念下新兴起来的，但大部分工作仍停留在理论研究上。

然而，以上的研究工作都只考虑数据的一次发布。Wang Ke 和 B. C. M. Fung 最早对数据重发布可能存在的隐私泄露进行研究，并提出一种防止隐私泄露的方法[23]。但是，文献[23]假设数据表的全局准标识符是由多个数据发布版本的属性组合而成，即数据是“水平”更新的。

与[23]不同，文献[14，25，26]和本文则是考虑数据“垂直”更新环境下的重发布问题，即数据表的准标识符是固定的，而记录是随时间动态增加的。J. W. Byun 等提出了一种安全的匿名技术[25]，但该技术在匿名新的数据发布版本时，需要考虑所有的历史发布版本，因而效率不高；Xiao Xiaohui 和 Tao Yufei 提出“$m$-不变性”的概化原则[14]，其关键是引入伪概化技术来保证任何记录在不同数据发布版本中所在的 QI 组都具有相同的敏感属性值。最近，文献[26]提出一种单调递增匿名方式。然而，文献[14，26]的方法均不能避免出现隐私泄露。

最近，有不少工作基于数据集和事务处理数据而提出。特别是 G. Ghinita 等[27]定义了隐私等级 $p$，表示特定的敏感属性被关联的可能性为 $p$ 的倒数($1/p$)。Xu Yabo 等[28]提出了($h$，$k$，$p$)一致性隐私模型，$k$ 的倒数($1/k$)代表了将一个人关联的可能性，$h$ 代表与一个人关联的隐私项，$p$ 代表了攻击者的能力。ERASE[29]是一个为了净化文档(以关键词集作为模拟)而提出来的系统，它需要一个将关键词和被保护敏感实体联系起来的外部知识数据库。M. Terrovitis 等[30,31]提出 $k^m$-anonymity 模型，并不区分敏感项和非敏感项，其中，$k$ 的倒数代表与个人关联起来的可能性，$m$ 代表攻击者的能力。但是，他们没有防止在个人与潜在敏感项之间的链接攻击。匿名的概化实现算法上有全局重编码和局部重编码两类。在全局重编码[4,5,20,32,33]中，每个准标识属性的取值要求概化到概念层次树的同一层次。但是全局重编码通常会过度概化，从而损失了更多的信息。局部重编码[5,36]放弃了这一要求，准标识属性的取值可以概化到不同层次。童云海等[34]分析了单一个体对应多个记录的情况，提出一种保持身份标识属性的匿名方法，取得了一定的成果，保持隐私的同时进一步提高了信息有效性。

最近也有少数工作专注于查询日志匿名化研究[35-39]。R. Kumar 和 R. Jones 等[40,41]已经证明了基于哈希和简单绑定匿名方案的无效性或者隐私风险；A. Korolova 等[36]提供了一种基于差分隐私保护的匿名方法，但是匿名数据的效用有限。Hong Yuan 等[38]定义了 $k^\delta$-anonymity 模型，处理了稀疏查询关键词的问题，但是不能防止对个人和潜在敏感查询之间的连接攻击。P. kodeswaran 等人[39]将差分隐私和查询日志应用于交互的查询框架(PINQ)中，但是当在一定的查询上示范其有效性时，它并不清楚在个性化网络搜索中的互动机制是如何运作的。周水庚等[42]对部分隐私保护技术的基本原理、特点进行了阐述，重点指出了基于数据匿名化的隐私保护技术是当前该领域的研究热点，并指出了隐私保护技术的未来发展方向。Xu Yabo[43]将单一用户资料按照层次化结构组织，以达到隐私保护的目的，同时证明了检索性能的一个重大提升可以通过共享层次的用户资料信息实现，但是仅对单一用户资料进行处理，隐私保护程度是有限的。李太勇等[44]提出一种通过两次聚类实现 $k$-2 匿名的隐私保护方法。杨晓春等[45]提出了针对多敏感属性隐私数据发布的多维桶分组技术。

本研究可以解决隐私保护数据发布(PPDP)领域的区分敏感词与非敏感词、单一敏感词与多敏感词问题，同时，对隐私保护数据发布技术研究与个性化服务研究，具有方法论上的参考价值。

## 1.4 隐私保护研究目标

存储、处理和通信技术的快速发展正在改变着被私有企业和公共机构所采纳的信息系统